INTERNATIONAL DOCTORAL SCHOOL OF THE USC

Belén
Serrano Antón

PhD Thesis

Automated Segmentation and Quality Enhancement in Medical Imaging: Applications to Cardiology

Santiago de Compostela, 2025

Doctoral Programme in Materials Science

ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

TESE DE DOUTORAMENTO

AUTOMATED SEGMENTATION AND QUALITY ENHANCEMENT IN MEDICAL IMAGING: APPLICATIONS TO CARDIOLOGY

Autor

Belén Serrano Antón

Directores: Alberto Pérez Muñuzuri, Alberto Otero Cacho e José Ramón González Juanatey

Titor: Alberto Pérez Muñuzuri

PROGRAMA DE DOUTORAMENTO EN CIENCIA DE MATERIAIS

SANTIAGO DE COMPOSTELA

Contents

El que es fiel en lo mínimo, lo es también en lo mucho; y el que es injusto en lo mínimo, también lo es en lo mucho.

– Lc, 16, 10

Agradecimientos

Queda en este documento recogida una ínfima parte de toda la investigación y aprendizaje que he llevado a cabo durante estos 4 años. Como se pueden imaginar, ha sido un camino lleno de piedras y dudas. Con días grises y lluviosos. Pero ¡ay...! Con tantos días de sol, ¡que casi no parecía Galicia! Un camino en el que he aprendido mucho, etapa a etapa, en lo académico, lo profesional, pero, sobre todo, en lo personal.

Como en casi todos los líos en los que me meto en la vida, no he estado sola. Este caso no es una excepción. A mi lado han pasado personas maravillosas; algunas ya estaban antes de todo esto y otras llegaron para quedarse. ¡Tengo tanto que agradecer!

En primer lugar, a mis directores de tesis. A los Albertos, por su guía, acompañamiento y su templanza. La misma templanza que, en ocasiones, tanto me costó entender y de la que he acabado contagiándome poco a poco (muuuy poco a poco...).

A José Ramón, por acercarme a los aspectos más clínicos de la investigación. Sin duda, un aspecto fundamental de este trabajo. No puedo olvidarme del equipo de cardiología del CHUS, de María, Diego y Brais. Gracias por enseñarme el día a día y la verdadera aplicabilidad de todo este trabajo. Por despertar la ilusión de que el futuro de la medicina también está en proyectos como este.

Como si de un agente secreto se tratase, durante estos 4 años he vivido una doble vida. A veces en la empresa, a veces en la universidad. Y es que un doctorado industrial es así, con lo mejor y lo peor de los dos mundos. Tengo que advertirles que llevar dos vidas no es tarea fácil: conlleva sacrificio y, sobre todo, mucha discreción. Aunque, después del doble grado, ya llevaba algo de entrenamiento je,je. La ventaja es poder compartir el camino por partida doble.

Gracias a mis compañeros de Flow. A Manuel, quien ha estado conmigo desde el primer día. Con el que tantos momentos y confidencias he compartido. Recuerdo con tanta alegría los paseos al sol, las divagaciones de última hora y ese lenguaje no verbal que creamos a nuestra medida.

A Gemma, por su alegría, por su buen humor y por entenderme tan bien. Pero... sobre todo, sobre todo... ¡porque sin ti jamás hubiera conseguido las entradas del Eras Tour! Nunca lo olvidaré. *It's fearless*. Recuerda: te avisé.

A Andrea, por su mirada soñadora, su sonrisa inagotable y su mente elocuente. Gracias por siempre estar ahí, por alegrarnos todos los días desde primera hora.

A mi Team Imagen, quienes me han aguantado en esta última etapa. Por darme un lugar donde ser yo misma, por dejar un hueco en la pizarra para dibujar mis personajes y por no desesperar (demasiado) con mis manías.

A Santi, por siempre darme el perfil bueno y estar siempre a mi lado y dispuesto a echarme un cable. Por esa conexión inalámbrica y genuina que hace que, al mirarnos de reojo, todo entre en sintonía.

A Irene, por su comprensión y apoyo, su dulzura y ser mi cómplice a la hora de trabajar en

penumbras en el despacho.

A Óscar, por ponerme el reto diario de no enloquecer con sus comentarios perspicaces y torpes, su risa y ser mi toma de tierra cuando hay que liberar tensión.

A Agustín, por sus pocas, pero acertadas palabras.

A Juan, por su bondad, su constante amabilidad y el soportar mis preguntas: Juan, ¿qué preferirías...?

Al otro Santi, al que tiene pel... bueno, a Paramés, por sus palabras de ánimo, por esa inquietud rebosante y esas ganas de aprender que contagian a cualquiera.

Al otro lado del ring, a mis compañeros del GFNL. Aquellos que estabais y a los que están. A Alejandro, por su amistad, por ser mi confidente, el hombro en el que apoyarme y mi terapeuta a tiempo parcial. Gracias por compartir tantos momentos, viajes y memes. Aunque no volveré a confiar en ti para una ruta "facilita" de montaña.

A Sara, por ser una de las mejores y más graciosas personas que conozco. El mundo, señores, se está perdiendo un gran talento.

A Alba, quien no solo se infiltró en el GFNL, sino también en mi vida. Gracias por tu sensibilidad y por compartir y hacer nuestros los paseos, las meriendas, los conciertos y las foliadas.

Seguramente ya hayan advertido mi poco acento gallego. Y es que yo no soy de aquí, soy de otro lado. Me aventuré a hacer el doctorado en una ciudad nueva, de la que apenas conocía nada. Llegué a Santiago sin entender gallego, ni la retranca, nin sabía o que era unha foliada, nin sequera entendía aos galegos. E mirade, mirade, tanto aprendín que acabei por pintar una raia azul no meu corazón para sempre.

O maior tesouro desta terra son dous galegos, máis galegos que os pementos de Padrón ou a tortilla de Betanzos. Os meus compañeiros de doutoramento, de piso e de vida. Os meus "todo está ben".

A Javi, que apareceu de xeito casual, case por sorpresa. Chegou á miña vida e, do mesmo xeito que fixo coa miña alacena, organizou e limpou todo o que estaba ciscado. Por moito que chova, sempre sae o sol na nosa casa. Só fai falta que esteas ti para enchelo todo (ás veces literalmente), para nos contaxiares coa túa alegría e as túas inesgotábeis ocorrencias disparatadas. Non teño folios suficientes e creo que habería que inventar unha nova linguaxe para expresar todo o amor que sinto por ti. De non te ter atopado, buscaríate toda a vida.

A Alfredo, o meu refuxio na neve, a aperta cálida á cal recorrer en caso de emerxencia emocional, o cociñeiro particular dos mércores e o mellor guía de sendeiros que se pode atopar. Grazas pola túa proximidade e a túa xenerosidade sen límites. Grazas tamén por lle devolveres a cordura ao piso, sobre todo cando Javi leva unha cullerada de azucre de máis na macela.

Outra das cousas que Galiza fixo por min foi redescubrirme o amor polo folclore. Grazas a Álex, o meu profe de baile tradicional galego, non só por ensinarme a bailar, senón por mostrarme o marabilloso patrimonio inmaterial que temos a obriga de coidar e respectar. Grazas a ti e a todos os meus compañeiros de baile, porque o que comezou como unha afección acabou converténdose na miña forma favorita de desconexión.

A aqueles que chegaron de rebote e ficaron. A Gino e Pablo, as persoas máis singulares que coñecín. Grazas por terdes sempre a porta da vosa casa aberta para min, por contardes comigo para me levardes de aventuras e por me facerdes partícipe das vosas picarescas peripecias.

A Juli, pola confianza e o agarimo. Por ser sempre o "sí" a calquera plan proposto e por percorrer comigo as beiras do Sar un número non numerábel de veces.

A Alba Lalín, por ser partícipe das nosas andanzas, pola confianza e... polo cocido.

Á Conxi, por esperarme todos os domingos no mesmo lugar e á mesma hora. Oxalá soubeses canto me gusta ir verte e o ben que me fas.

Aunque esta tierra me ha traído muchas cosas buenas, nunca olvido la mía, mi querida Segovia. En la que tantas cosas he vivido y en la que tanto cariño he recibido.

A Ainoa, mi hermana, por estar a mi lado desde los tres años, por nuestras conversaciones infinitas al teléfono fijo, las numerosas rutas por la montaña en las que solo yo me sabía la ruta, por enseñarme a tener paciencia y esperar(te), por seguir teniendo esa inocencia y no perder la niña que llevas dentro.

Al resto del clan, a Paloma, la Martínez y Sonia, porque junto a vosotras he recuperado una parte de mí que había dejado atrás sin darme cuenta. Por abrirme los chacras y resetearme en cada viaje. En definitiva, gracias, porque sé que siempre habéis estado ahí.

A los pilares de mi vida, a mis padres, por su apoyo siempre, su ayuda y darme todo lo que he necesitado para llegar tan lejos como he llegado. Creo que no os lo digo lo suficiente, pero gracias. Gracias por ser vosotros quienes me abristeis las primeras puertas y me disteis la oportunidad para que yo pudiera ser lo que siempre quise, lo que soy hoy. Gracias por confiar en mí, por ponerme siempre como una prioridad. Os quiero mucho.

Gracias a mi familia, quienes mantenéis a flote el barco y seguís de cerca todos mis pasos.

A Coffee, por ser el gato más bueno del mundo y recibirme siempre con el mejor de sus maullidos.

El Erasmus..., !qué experiencia!. To Mr. Bean, for inspiring me with that elegant and determined personality, for being by my side every day and on every journey.

A Mai y Vero, por todos los viajes y risas que pasamos y pasaremos, porque, aunque cada una esté en una punta del mundo, seguís siendo mi punto de apoyo si me desestabilizo.

A quienes compartieron conmigo los inicios de mi vida académica, mis compañeros de carrera y amigos.

A David, por toda su ayuda, su paciencia, su gran corazón y llevar la palabra amistad al máximo nivel. Por estar ahí, a pesar de la distancia, a pesar de los altibajos. Por cada conversación, cada viaje y aguantar ese bullying cariñoso.

A Clara, por su espontaneidad, generosidad y su buen hacer. Me encantaría decir que eres la mejor compañera de conciertos, pero para eso tendrías que venir a alguno... Te lo perdono solo por ir conmigo los 5 de enero al Parque de Atracciones a pesar de los vértigos y llevarme a las Fragas do Eume sin perdernos mientras mi actuación como copiloto brillaba por su ausencia. Eres una persona excepcional, a la que metería en la maleta para que fuera conmigo a todos lados (si no midieras 1.84, claro, seguro, ...).

Abstract

This thesis is part of an industrial PhD program, conducted in collaboration with a company specializing in the development of innovative solutions in the field of medical imaging, with a particular emphasis on non-invasive cardiology. In this context, one of the main challenges in diagnosing coronary artery disease is the accurate evaluation of the functional significance of a stenosis. Currently, the reference standard for this analysis is Fractional Flow Reserve (FFR), an invasive technique that requires the insertion of a pressure catheter into the coronary artery during angiography to measure the pressure drop and determine the hemodynamic relevance of the stenosis. As a non-invasive alternative, the FFR_{CT} technique relies on obtaining computed tomography (CT) images and using advanced computational models to estimate fractional flow reserve without the need for an invasive procedure. Accurate coronary artery segmentation in CT images is essential for its proper implementation.

The development of this technology requires close collaboration between academic research, industry, and the clinical field. The combination of these three pillars enables addressing the problem from a multidisciplinary perspective, integrating knowledge in medical imaging, artificial intelligence, and cardiology. The availability of medical data and the expertise of healthcare professionals are essential for guiding the development process and ensuring the applicability of the solutions in real clinical environments.

Given this context, the present work focuses on medical image processing from a comprehensive perspective, encompassing acquisition, analysis, quality enhancement, segmentation, and the automation of these processes. Automation plays a key role both at the enterprise and clinical levels, as it allows the integration of efficient solutions into daily medical practice, providing fast and reliable results that optimize diagnosis and decision-making.

The structure of this thesis is as follows:

Chapter 1 sets the context of the thesis, highlighting the relevance of cardiovascular diseases and the importance of accurate evaluation of the coronary arteries and aorta. The anatomical and functional foundations of the heart are presented, along with an introduction to computed tomography (CT), its applications in cardiology, and its limitations, especially regarding the capture of moving structures and the presence of metal implants such as stents.

Additionally, the main clinical procedures for diagnosing coronary artery disease (CAD) and aortic valve stenosis (AVS) are reviewed, analyzing the need for more efficient and automated methods. Finally, the growing impact of artificial intelligence (AI) in medical image analysis is examined, and the thesis objectives are presented, focusing on the development of automatic AI-based tools for the segmentation and analysis of cardiovascular structures in CT images.

Chapter 2 describes the general methodology employed in this work. First, the image acquisition methods are presented, focusing on computed tomography (CT) and computed tomography angiography (CTA), which uses contrast media to visualize the lumen of blood

vessels. The process of dataset generation used in this study is also detailed, which was obtained through manual segmentation, a laborious but essential procedure to ensure the quality and precision of the automated models.

In the second part of the chapter, the architecture of the main neural networks implemented in this work is analyzed, categorized into three main types: Convolutional Neural Networks (CNNs), Transformer-based models, and the Mamba architecture. While CNNs have been widely used in medical image segmentation due to their ability to capture spatial features, Transformers have shown effectiveness in medical image segmentation by capturing complex spatial relationships, although their high computational cost remains a challenge. Mamba, on the other hand, is a more recent approach that aims to improve efficiency in processing long sequences, optimizing resource usage without sacrificing global modeling capability.

Finally, the workflow followed in subsequent chapters is presented, outlining the methodological structure guiding the development of the study.

Chapter 3 presents the first set of results of this thesis, focusing on reducing artifacts caused by metal implants in CT images for the diagnosis and treatment of head and neck cancer. This work is developed in collaboration with the University of Udine and the Centro di Riferimento Oncologico di Aviano IRCCS in Italy, as part of a predoctoral stay.

In this type of imaging, artifacts appear when the patient has metal implants in the oral region, affecting image quality and complicating its interpretation. The chapter describes in detail the methods used to generate the dataset, as well as the AI tools developed to mitigate these effects. The results obtained are presented, and the different strategies implemented are compared.

This research not only aims to improve image quality in oncology but also lays the foundation for the transfer of knowledge to the cardiology field.

Chapter 4 focuses on the automatic segmentation of the coronary tree from contrast-enhanced CT images and is divided into two distinct parts. The first part addresses coronary artery segmentation in a conventional manner, using the original images and applying various methodologies based on convolutional neural networks (CNNs). Additionally, an analysis is conducted on the impact of the number of data on the results obtained, given that medical image datasets are often limited and lack annotations, posing a significant challenge in developing precise models.

The second part of the chapter presents an innovative approach through an unsupervised segmentation methodology that eliminates the need for previously annotated datasets. This new approach arises in response to the limitations identified in the previous methodology, where the image resolution was insufficient for accurate segmentation. The need for higher computational power and a more robust dataset was also identified in order to overcome the challenges related to quality and precision.

In this context, a preprocessing process of the original images is introduced with the goal of improving their resolution. This enhancement in image quality facilitates more precise and efficient segmentation.

Chapter 5 focuses on the segmentation of calcium in the thoracic aorta, a crucial aspect in the context of aortic valve stenosis (AVS). In some AVS cases, the TAVI (Transcatheter Aortic Valve Implantation) procedure is performed, in which a new aortic valve is implanted through a catheter. However, during this procedure, there is a risk that fragments of calcified plaque could be displaced by the passage of the catheter, leading to severe cerebrovascular events.

To mitigate this risk, it is necessary to assess the amount of calcified plaque in the aorta and

in specific regions. This helps determine the need to use carotid protection devices to prevent plaque from traveling to the brain arteries. In this chapter, the methodology for segmenting calcium in contrast-enhanced CT images, routinely used in clinical practice to assess these pathologies, is presented.

The proposed methodology seeks to establish a standard for aortic calcium segmentation. To do this, it is validated with real clinical data and automated using artificial intelligence techniques, leveraging the previously generated dataset.

A key aspect addressed throughout the development and analysis of all the algorithms has been process automation, both in generating results and in their visualization. This approach has streamlined the result acquisition process and optimized its presentation, facilitating integration into the daily workflows of the company. This approach has been critical for both the company and the clinical field, where the speed and accuracy of result delivery are crucial.

The thesis concludes with the general **Conclusions** of this work. In this final chapter, the main advances achieved during the research are summarized, highlighting the validation and improvement of methodologies for the segmentation of cardiovascular structures and the extraction of relevant clinical parameters. Automated segmentation, along with the development of models based on artificial intelligence techniques, has contributed to greater efficiency in the diagnosis and analysis of cardiovascular diseases. Despite limitations in the amount of data and clinical resources, the work conducted demonstrates significant potential for future applications, particularly in cardiovascular risk prediction and treatment planning. The project remains in development, with the goal of enhancing its capabilities and extending its application to new clinical areas.

Resumen

Esta tesis se enmarca dentro de un programa de doctorado industrial, llevado a cabo en colaboración con una empresa especializada en el desarrollo de soluciones innovadoras en el ámbito de la imagen médica, con especial énfasis en la cardiología no invasiva. En este contexto, uno de los principales retos en el diagnóstico de la enfermedad arterial coronaria es la evaluación precisa de la significancia funcional de una estenosis. Actualmente, el estándar de referencia para este análisis es la reserva fraccional de flujo (FFR), una técnica invasiva que requiere la introducción de una guía de presión en la arteria coronaria durante una angiografía para medir la caída de presión y determinar la relevancia hemodinámica de la estenosis. Como alternativa no invasiva, la técnica FFR_{CT} se basa en la obtención de imágenes de tomografía computarizada (CT) y el uso de modelos computacionales avanzados para estimar la reserva fraccional de flujo sin necesidad de un procedimiento invasivo. El flujo de trabajo de esta técnica consta de varias etapas: primero, se adquieren imágenes de CT con contraste para visualizar el lumen de las arterias coronarias. A continuación, se realiza la segmentación del árbol coronario para extraer su geometría tridimensional. Finalmente, esta geometría se utiliza en una simulación de dinámica de fluidos computacional (*Computational Fluid Dynamics*, CFD), la cual permite calcular parámetros clínicos de interés, como la caída de presión a lo largo de las arterias y la reserva fraccional de flujo.

Para su correcta implementación, es esencial la segmentación precisa de las arterias coronarias en las imágenes de CT, ya que cualquier error en esta etapa inicial puede afectar significativamente la calidad de los resultados obtenidos en la simulación y, por ende, la fiabilidad del diagnóstico.

El desarrollo de esta tecnología requiere una estrecha colaboración entre la investigación, la industria y el ámbito clínico. La combinación de estos tres pilares permite abordar el problema desde una perspectiva multidisciplinar, integrando conocimientos en tratamiento de imagen médica, programación e inteligencia artificial aplicada a la cardiología. La disponibilidad de datos médicos y la experiencia de los profesionales sanitarios son esenciales para guiar el proceso de desarrollo y asegurar la aplicabilidad de las soluciones en entornos clínicos reales.

Dado este contexto, el presente trabajo se centra en el procesamiento de imágenes médicas desde una perspectiva integral, abarcando su adquisición, análisis, mejora de calidad y segmentación, así como la automatización de estos procesos. La automatización juega un papel clave tanto a nivel empresarial como clínico, ya que permite la integración de soluciones eficientes en la práctica médica diaria, proporcionando resultados rápidos y fiables que optimizan el diagnóstico y la toma de decisiones. Además, las soluciones implementadas han sido de carácter general, lo que ha permitido que los avances logrados en esta investigación hayan servido de base para abordar nuevos retos en otras áreas. Esto ha asegurado una continuidad en la exploración y aplicación de estas metodologías en distintos contextos, ampliando su impacto más allá del diagnóstico de la enfermedad arterial coronaria.

El contenido de esta tesis se estructura de la siguiente manera:

El **Capítulo 1** establece el marco clínico de este estudio, proporcionando una base sólida para comprender el contexto en el que se desarrolla el trabajo. En él, se presentan los fundamentos anatómicos y funcionales del corazón, con especial énfasis en las arterias coronarias y la aorta, que son el foco principal de esta investigación. Además, se introduce la tomografía computarizada (CT), describiendo su papel en cardiología, sus aplicaciones más relevantes y las limitaciones que presenta, en particular, las dificultades asociadas a la obtención de imágenes de estructuras en movimiento, como el corazón, y la interferencia generada por la presencia de implantes metálicos, como los stents.

Asimismo, se presentan los fundamentos de los principales procedimientos clínicos empleados en el diagnóstico de la enfermedad arterial coronaria (*Coronary Artery Disease*, CAD) y la estenosis aórtica (*Aortic Valve Stenosis*, AVS), destacando la necesidad de desarrollar métodos no invasivos, eficientes y automatizados que mejoren la precisión y seguridad del diagnóstico. Por último, se analiza el impacto creciente de la inteligencia artificial (IA) en el procesamiento de imágenes médicas y su papel en la medicina personalizada, antes de exponer los objetivos específicos de esta tesis.

El **Capítulo 2** describe la metodología general empleada en este trabajo. En primer lugar, se presentan los métodos de adquisición de imágenes, centrándose en la tomografía computarizada (CT) y la angiografía por tomografía computarizada (CTA), que utiliza medio de contraste para visualizar el lumen de los vasos. Asimismo, se detalla el proceso de generación de los conjuntos de datos utilizados en este estudio, los cuales han sido obtenidos mediante segmentación manual, un procedimiento laborioso pero esencial para garantizar la calidad y precisión de los modelos automatizados.

En la segunda parte del capítulo, se analiza la arquitectura de las principales redes neuronales implementadas en este trabajo, clasificadas en tres categorías principales: redes convolucionales (*Convolutional Neural Networks*, CNNs), modelos basados en *Transformers* y la arquitectura *Mamba*. Mientras que las CNNs han sido ampliamente utilizadas en la segmentación de imágenes médicas debido a su capacidad para captar características espaciales, los *Transformers* han demostrado ser efectivos en segmentación al capturar relaciones espaciales complejas, aunque su alto costo computacional sigue siendo un desafío. *Mamba*, por su parte, es un enfoque más reciente que busca mejorar la eficiencia en el procesamiento de secuencias largas, optimizando el uso de recursos sin perder capacidad de modelado global. Dentro de este contexto, la arquitectura por excelencia en la que se basan el resto de implementaciones de este estudio es la U-Net, debido a su capacidad para combinar información de bajo y alto nivel mediante conexiones *skip*, lo que la hace especialmente efectiva para tareas de segmentación en imágenes médicas.

Finalmente, se presenta el flujo de trabajo seguido en los capítulos posteriores, estableciendo la estructura metodológica que guía el desarrollo del estudio.

El **Capítulo 3** presenta el primer conjunto de resultados de esta tesis, enfocado en la reducción de artefactos generados por implantes metálicos en imágenes de tomografía computarizada (CT) para el diagnóstico y tratamiento del cáncer de cabeza y cuello. Este trabajo se desarrolla en colaboración con la Universidad de Údine y el Centro di Riferimento Oncologico di Aviano IRCCS en Italia, en el marco de una estancia predoctoral.

En este tipo de imágenes, los artefactos aparecen cuando el paciente presenta implantes metálicos en la región bucal, afectando la calidad de la imagen y dificultando su interpretación. En el capítulo se describen en detalle los métodos utilizados para la generación del *dataset*,

así como las herramientas de inteligencia artificial desarrolladas para mitigar estos efectos. Se presentan los resultados obtenidos y se comparan las distintas estrategias implementadas.

Esta investigación no solo busca mejorar la calidad de imagen en oncología, sino que sienta las bases para la transferencia de conocimiento al ámbito de la cardiología.

El **Capítulo 4** se centra en la segmentación automática del árbol coronario a partir de imágenes de tomografía computarizada (CT) con contraste, y se estructura en dos partes diferenciadas. En la primera, se aborda la segmentación de las arterias coronarias mediante un enfoque inicial más simple, utilizando las imágenes originales y aplicando diversas metodologías basadas en redes neuronales convolucionales (CNNs). Además, se realiza un análisis sobre el impacto del número de datos en los resultados obtenidos, dado que los conjuntos de datos en el ámbito de la imagen médica son frecuentemente limitados y carecen de anotaciones, lo que plantea un desafío importante en el desarrollo de modelos precisos.

La segunda parte del capítulo presenta un enfoque innovador mediante una metodología de segmentación no supervisada, que prescinde de la necesidad de un conjunto de datos anotados previamente. Este nuevo enfoque surge como respuesta a las limitaciones identificadas en la metodología anterior, en la que la resolución de las imágenes resultaba insuficiente para lograr una segmentación precisa. Asimismo, se identificó la necesidad de contar con mayor poder computacional y un *dataset* más robusto, con el fin de superar los desafíos relacionados con la calidad y precisión de los resultados obtenidos.

El **Capítulo 5** se enfoca en la segmentación del calcio en la aorta torácica, un aspecto crucial en el contexto de la estenosis aórtica (AVS). En algunos casos de AVS, se realiza el procedimiento de TAVI (*Transcatheter Aortic Valve Implantation*), en el cual se implanta una nueva válvula aórtica a través de un catéter. Sin embargo, durante este procedimiento, existe el riesgo de que fragmentos de placa calcificada se desplacen debido al paso del catéter, lo que podría dar lugar a eventos cerebrovasculares graves.

Para mitigar este riesgo, se evalúa la cantidad de placa calcificada en la aorta por regiones específicas. Esto tiene el objetivo de determinar la necesidad de utilizar dispositivos protectores en las carótidas, previniendo así el desplazamiento de la placa hacia las arterias cerebrales. En este capítulo, se presenta la metodología para segmentar el calcio en imágenes de tomografía computarizada (CT) con contraste, que se utilizan en la práctica clínica habitual para la preparación del procedimiento TAVI.

La metodología propuesta busca establecer un estándar de segmentación del calcio aórtico, ya que la segmentación de calcio suele hacerse en imágenes CT sin contraste. Para ello, se valida con datos clínicos reales y se automatiza mediante técnicas de inteligencia artificial, utilizando el *dataset* previamente generado.

Un aspecto clave abordado a lo largo del desarrollo y análisis de todos los algoritmos ha sido la automatización de procesos, tanto en la generación de resultados como en su visualización. Este enfoque ha permitido agilizar la obtención de resultados y optimizar su presentación, facilitando así su integración en los flujos de trabajo diarios de la empresa. Este enfoque ha sido fundamental tanto para la empresa, como para el ámbito clínico, donde la rapidez y precisión en la entrega de resultados son cruciales.

La tesis finaliza con las **Conclusiones** generales de este trabajo. En este último capítulo, se resumen los principales avances alcanzados durante la investigación, destacando la validación y mejora de metodologías para la segmentación de estructuras cardiovasculares y la obtención de parámetros clínicos relevantes. La segmentación automatizada, junto con el desarrollo de modelos basados en técnicas de inteligencia artificial, ha permitido avanzar hacia una

mayor eficiencia en el diagnóstico y análisis de enfermedades cardiovasculares. A pesar de las limitaciones en la cantidad de datos y recursos clínicos, el trabajo realizado muestra un importante potencial para futuras aplicaciones, especialmente en la predicción del riesgo cardiovascular y la planificación de tratamientos. El proyecto sigue en desarrollo, con el objetivo de mejorar su capacidad y extender su uso a nuevos ámbitos clínicos.

Resumo

Esta tese enmárcase dentro dun programa de doutoramento industrial, levado a cabo en colaboración cunha empresa especializada no deseño de solucións innovadoras no ámbito da imaxe médica, con especial énfase na cardioloxía non invasiva. Neste contexto, un dos principais retos na diágnose da enfermidade arterial coronaria é a avaliación precisa da relevancia funcional dunha estenose.

Na actualidade, o estándar de referencia para esta análise é a reserva fraccional de fluxo (FFR), unha técnica invasiva que require a introdución dunha guía de presión na arteria coronaria durante unha anxiografía para medir a caída de presión e determinar a importancia hemodinámica da estenose. Como alternativa non invasiva, a técnica $\mathrm{FFR_{CT}}$ baséase na adquisición de imaxes de tomografía computarizada (CT) e no uso de modelos computacionais avanzados para estimar a reserva fraccional de fluxo sen necesidade dun procedemento invasivo.

O fluxo de traballo desta técnica consta de varias etapas: en primeiro lugar, adquírense imaxes de CT con contraste para visualizar o lumen das arterias coronarias. A continuación, realízase a segmentación da árbore coronaria para extraer a súa xeometría tridimensional. Finalmente, esta xeometría emprégase nunha simulación de dinámica de fluídos computacional (*Computational Fluid Dynamics*, CFD), que permite calcular parámetros clínicos de interese, como a caída de presión ao longo das arterias e a reserva fraccional de fluxo.

A segmentación é o proceso de identificar e seleccionar os píxeles correspondentes a unha estrutura específica dentro dunha imaxe, como as arterias coronarias nas imaxes de tomografía computarizada (CT). Neste caso, trátase de seleccionar, en cada corte ou *slice* da imaxe, os píxeles que pertencen ás coronarias e separalos do resto dos tecidos e estruturas.

Para obter xeometrías 3D que representen fielmente a árbore coronaria de cada paciente, é fundamental unha segmentación precisa das arterias coronarias nas imaxes de CT, xa que calquera erro nesta etapa inicial pode afectar de maneira significativa á calidade dos resultados da simulación e, en consecuencia, á fiabilidade do diagnóstico. A segmentación manual da árbore coronaria en imaxes de CT é unha tarefa laboriosa que require coñecemento experto en imaxe médica e anatomía do corazón, ademais de ser un proceso dependente do usuario, o que pode introducir variabilidade nos resultados. Por este motivo, o obxectivo é automatizar este proceso para garantir unha segmentación máis eficiente, reproducible e precisa.

A automatización deste proceso lévase a cabo principalmente mediante o uso de redes neuronais nun enfoque de aprendizaxe supervisada, o cal require dispor dun conxunto de datos anotados con precisión. Porén, no ámbito da imaxe médica, a obtención destes datos é especialmente complexa, xa que existen poucos conxuntos de datos públicos que conten coa resolución e calidade necesarias para este tipo de segmentación. Por este motivo, no presente traballo desenvolvéronse conxuntos de datos propios, garantindo que os modelos poidan adestrarse con datos de alta calidade e adaptados aos requirimentos específicos do problema.

O cumprimento deste obxectivo require un enfoque multidisciplinar, combinando coñecementos en imaxe médica, ciencias da computación e medicina. A interacción entre estas áreas foi fundamental para desenvolver solucións eficaces e aplicables nun entorno clínico real. Este carácter integral non só permitiu avances na segmentación automática da árbore coronaria, senón que tamén deu lugar ao desenvolvemento de dous subproxectos adicionais, que se recollen neste documento. Estes subproxectos, aínda que distintos na súa aplicación, comparten a mesma base metodolóxica e reflicten a versatilidade das técnicas desenvolvidas, demostrando a súa aplicabilidade en distintos contextos dentro do ámbito da imaxe médica.

O primeiro destes subproxectos céntrase na redución de artefactos metálicos en imaxes de tomografía computarizada (CT), un problema común que afecta a calidade das imaxes e pode dificultar a súa interpretación clínica. Estes artefactos prodúcense debido á alta densidade dos materiais metálicos, como implantes dentais ou próteses, que xeran distorsións na imaxe ao interactuar cos raios X, provocando rexións con perda de información e sombras non desexadas. Para abordar esta problemática, desenvolveuse unha solución baseada en redes neuronais, especificamente deseñada para mellorar a calidade das imaxes utilizadas no diagnóstico e tratamento do cancro de cabeza e pescozo. A implementación deste enfoque permite obter imaxes máis nítidas e fiables, optimizando así o proceso de diagnóstico e planificación do tratamento. Este traballo detállase no Capítulo 3.

Ademais, estes artefactos tamén afectan ás imaxes cardíacas, como no caso dos *stents* coronarios, onde poden comprometer a avaliación precisa da anatomía e funcionalidade arterial. Por este motivo, o traballo desenvolvido na redución de artefactos metálicos en oncoloxía serve de base para a súa posible aplicación en cardioloxía, ampliando o alcance e a utilidade das solucións implementadas.

O segundo proxecto derivado céntrase na segmentación do calcio na aorta torácica, un aspecto clave na avaliación da estenose aórtica (*Aortic Valve Stenosis*, AVS). Nalgúns casos, o tratamento mediante TAVI (*Transcatheter Aortic Valve Implantation*) implica o risco de que fragmentos de placa calcificada se desprendan durante a intervención, aumentando a probabilidade de eventos cerebrovasculares. Para mitigar este risco, cuantifícase e localízase a placa calcificada na aorta por rexións específicas, o que permite determinar a necesidade de dispositivos protectores nas carótidas. Este traballo, detallado no Capítulo 5, contribúe a mellorar a seguridade e planificación do procedemento.

O desenvolvemento destas tecnoloxías require unha estreita colaboración entre a investigación, a industria e o ámbito clínico. A combinación destes tres piares permite abordar o problema desde unha perspectiva multidisciplinar, integrando coñecementos en tratamento de imaxe médica, programación e intelixencia artificial aplicada á cardioloxía. A dispoñibilidade de datos médicos e a experiencia dos profesionais sanitarios son esenciais para guiar o proceso de desenvolvemento e asegurar a aplicabilidade das solucións en contornas clínicas reais.

Dado este contexto, o presente traballo céntrase no procesamento de imaxes médicas desde unha perspectiva integral, abarcando a súa adquisición, análise, mellora de calidade e segmentación, así como a automatización destes procesos. A automatización desempeña un papel clave tanto a nivel empresarial como clínico, xa que permite a integración de solucións eficientes na práctica médica diaria, proporcionando resultados rápidos e fiables que optimizan o diagnóstico e a toma de decisións.

O contido desta tese estrutúrase do seguinte xeito:

O **Capítulo 1** establece o marco clínico deste estudo, proporcionando unha base sólida para comprender o contexto no que se desenvolve o traballo. Nel preséntanse os fundamentos

anatómicos e funcionais do corazón, con especial énfase nas arterias coronarias e na aorta, que constitúen o foco principal desta investigación. Ademais, introdúcese a tomografía computarizada (CT), describindo o seu papel na cardioloxía, os tipos de imaxes empregadas nesta tese, as súas aplicacións máis relevantes e as limitacións que presenta. En particular, analízanse as dificultades asociadas á obtención de imaxes de estruturas en movemento, como o corazón, así como a interferencia xerada pola presenza de implantes metálicos, como os stents.

Tamén se presentan os fundamentos dos principais procedementos clínicos utilizados no diagnóstico da enfermidade arterial coronaria (*Coronary Artery Disease*, CAD) e da estenose aórtica (*Aortic Valve Stenosis*, AVS), destacando a necesidade de desenvolver métodos non invasivos, eficientes e automatizados que melloren a precisión e a seguridade do diagnóstico. Por último, analízase o impacto crecente da intelixencia artificial (IA) no procesamento de imaxes médicas e o seu papel na medicina personalizada, antes de expoñer os obxectivos específicos desta tese.

Este capítulo tamén pretende achegar o problema clínico a aqueles lectores que proceden de disciplinas alleas á medicina. Estruturouse de maneira que os conceptos anatómicos, fisiolóxicos e tecnolóxicos se presenten de forma clara e accesible, permitindo unha mellor comprensión do contexto clínico no que se enmarca esta investigación. Deste xeito, facilítase que profesionais doutros ámbitos, como a enxeñaría ou a computación, poidan entender a relevancia e os desafíos do problema abordado, promovendo así un enfoque verdadeiramente multidisciplinar.

O **Capítulo 2** describe a metodoloxía xeral empregada neste traballo, estruturada en dous grandes bloques. En primeiro lugar, preséntanse os métodos de adquisición de imaxes, centrándose na tomografía computarizada (CT) e na anxiografía por tomografía computarizada (CTA), que emprega medio de contraste para visualizar o lumen dos vasos. Ademais, detállanse os parámetros de adquisición, que definen o tamaño e a resolución das imaxes, aspectos clave para obter datos precisos e de alta calidade. O CT emprégase para a segmentación do calcio nas arterias, xa que, aínda que non permite visualizar o interior dos vasos, proporciona unha imaxe detallada das estruturas calcificadas. Pola súa banda, o CTA, grazas ao uso de contraste, permite observar tanto o lumen dos vasos como as áreas calcificadas, proporcionando unha visión máis completa e precisa para o diagnóstico e o tratamento.

Ademais, abórdanse as consideracións éticas relacionadas co tratamento das imaxes médicas. É fundamental garantir a seguridade e a legalidade no manexo dos datos, especialmente no contexto das imaxes clínicas, debido á súa natureza confidencial. Estas consideracións son esenciais para asegurar o cumprimento das normativas de privacidade e protección de datos, así como para garantir que as imaxes se utilicen de maneira responsable e ética na investigación.

Tamén se detalla a metodoloxía de segmentación manual das coronarias en imaxes de CTA. Este proceso consiste na identificación e selección dos píxeles que corresponden ao lumen das arterias coronarias en cada *slice* da imaxe, un procedemento laborioso pero esencial para garantir a calidade e precisión dos modelos automatizados.

Nesta segunda parte do capítulo, cámbiase o enfoque para abordar o aspecto máis computacional da metodoloxía empregada neste estudo. A continuación, explóranse as redes neuronais e as arquitecturas utilizadas neste traballo, comezando cunha descrición das principais arquitecturas implementadas. Primeiro abórdanse as redes convolucionais (*Convolutional Neural Networks*, CNNs), os modelos baseados en *Transformers* e a arquitectura *Mamba*. Mentres que as CNNs foron amplamente empregadas na segmentación de imaxes

médicas debido á súa capacidade para capturar características espaciais, os *Transformers* demostraron ser efectivos ao captar relacións espaciais complexas, aínda que o seu alto custo computacional segue sendo un desafío. Pola súa parte, *Mamba* é un enfoque máis recente que busca mellorar a eficiencia no procesamento de secuencias longas, optimizando o uso de recursos sen perder capacidade de modelado global.

Dentro deste contexto, a arquitectura de referencia sobre a que se basean as demais implementacións deste estudo é a U-Net, debido á súa capacidade para combinar información de baixo e alto nivel mediante conexións *skip*, o que a fai especialmente efectiva para tarefas de segmentación en imaxes médicas.

Posteriormente, o capítulo afonda nas variantes 2.5D e 3D dos modelos, onde se comparan as vantaxes e desvantaxes de traballar con varias imaxes en 2D en diferentes planos e 3D, discutindo como estas opcións impactan na precisión da segmentación e na eficiencia computacional. A continuación, trátase o concepto de transferencia de coñecemento (*Transfer Learning*), unha técnica que permite aproveitar modelos preentrenados en grandes conxuntos de datos, mellorando o rendemento con menos datos específicos do dominio.

Tamén se analizan as funcións de perda e as métricas de avaliación, destacando como estas varían segundo o tipo de problema a resolver. No caso da xeración de imaxes, como na redución de artefactos, as funcións de perda e métricas oriéntanse á minimización das diferenzas entre a imaxe xerada e a orixinal. En contraste, na segmentación de estruturas como a árbore coronaria, a aorta ou o calcio, as funcións de perda e métricas céntranse máis na precisión dos límites segmentados e na correcta identificación das rexións de interese.

Finalmente, preséntase o fluxo de traballo metodolóxico, que proporciona a estrutura que guía o desenvolvemento do estudo nos capítulos posteriores.

O **Capítulo 3** presenta o primeiro conxunto de resultados desta tese, centrado na redución de artefactos xerados por implantes metálicos en imaxes de tomografía computarizada (CT) para o diagnóstico e tratamento do cancro de cabeza e pescozo. Este traballo desenvolvese en colaboración coa Universidade de Údine e o Centro di Riferimento Oncologico di Aviano IRCCS en Italia, no marco dunha estadía predoutoral.

Neste tipo de imaxes, os artefactos aparecen cando o paciente presenta implantes metálicos na rexión bucal, afectando á calidade da imaxe e dificultando a súa interpretación. No capítulo descríbense en detalle os métodos utilizados para a xeración do *dataset*, o cal está composto por imaxes que presentan o artefacto debido á tecnoloxía utilizada, así como imaxes sen dito artefacto, que guían ao modelo para mitigar o efecto. O procedemento para xerar este *dataset* foi completamente automatizado, o que permitiu unha maior eficiencia na creación do conxunto de datos e facilitou a implementación das ferramentas de intelixencia artificial desenvolvidas para mitigar estes efectos. As diferentes propostas baséanse en distintas arquitecturas e funcións de perda específicas para este problema. Présentanse os resultados obtidos e comparanse as distintas estratexias implementadas para avaliar a efectividade das solucións propostas.

Esta investigación non só busca mellorar a calidade de imaxe en oncoloxía, senón que establece as bases para a transferencia de coñecemento ao ámbito da cardioloxía.

O **Capítulo 4** céntrase na segmentación automática da árbore coronaria a partir de imaxes de tomografía computarizada (CT) con contraste, e estrúturase en varias partes para explorar diferentes enfoques. En primeiro lugar, analízase como as redes neuronais convolucionais (CNNs) poden ser utilizadas para a segmentación da árbore coronaria, explorando distintos enfoques en 2D e 3D para avaliar as súas capacidades e determinar ata onde poden chegar estas redes para realizar unha segmentación precisa. Ademais, experimentase coa segmentación de

dúas estruturas diferentes: as coronarias, que son relativamente pequenas e ocupan moi pouco espazo respecto ao fondo da imaxe, e as coronarias xunto coa aorta. Estúdase o impacto destas diferenzas nos resultados obtidos, considerando como a segmentación dunha estrutura máis grande, como a aorta, pode influír na precisión e eficacia da segmentación das coronarias.

Ademais, investíganse os efectos do número de datos dispoñibles no adestramento das redes neuronais, xa que os conxuntos de datos no ámbito da imaxe médica adoitan ser limitados e carecen de anotacións completas, o que representa un reto á hora de adestrar modelos robustos. Tamén se analiza o impacto do *transfer learning*, unha técnica que permite aproveitar modelos previamente adestrados con grandes volumes de datos, o que pode mellorar os resultados cando se adestra con conxuntos de datos pequenos. Estes factores explóranse para comprender mellor as limitacións e as posibilidades das redes neuronais na segmentación automática da árbore coronaria.

A análise preliminar identificou varias limitacións significativas, sendo a principal a resolución insuficiente das imaxes orixinais. Esta deficiencia impide a xeración de xeometrías precisas necesarias para a simulación ou a segmentación de lesións prominentes, afectando á calidade dos resultados obtidos. Diante desta problemática, a segunda parte do capítulo adopta unha estratexia diferente. En primeiro lugar, aplícase un proceso de preprocesado para mellorar a resolución das imaxes, o que permite obter detalles máis finos e precisos. A continuación, implémentase unha nova estratexia de segmentación que permite extraer imaxes pequenas e centradas no vaso, mellorando a precisión na segmentación das estruturas de interese.

Ademais, introdúcese unha solución innovadora baseada nunha metodoloxía de segmentación non supervisada, eliminando a necesidade de crear un conxunto de datos anotados de novo. Este algoritmo aplícase mediante dous enfoques complementarios: un en 2.5D e outro cunha visión perpendicular ao vaso, o que mellora a precisión e a eficiencia do proceso de segmentación, superando as limitacións inherentes ás aproximacións previas.

Tamén, compárase o enfoque non supervisado con redes neuronais do estado da arte, avaliando o seu rendemento na segmentación de imaxes médicas. Esta comparación permite identificar as vantaxes e limitacións da nova solución fronte aos métodos tradicionais baseados en redes neuronais supervisadas, proporcionando unha valoración crítica da súa efectividade en diferentes escenarios.

O **Capítulo 5** aborda a segmentación do calcio na aorta torácica, un factor crucial para avaliar o risco durante o procedemento de TAVI (Implantación de Válvula Aórtica Transcateter) en pacientes con estenose aórtica (AVS). Este proceso pode implicar a mobilidade de fragmentos de placa calcificada debido ao paso do catéter, o que aumenta o risco de eventos cerebrovasculares graves.

Para reducir este risco, avalíase a distribución do calcio na aorta para identificar áreas de maior acumulación e determinar a necesidade de dispositivos protectores nas carótidas, co fin de previr o desprazamento da placa cara as arterias cerebrais. Neste capítulo, descríbese a metodoloxía de segmentación do calcio en imaxes de tomografía computarizada (CT) con contraste, amplamente utilizadas na práctica clínica para a planificación e avaliación do procedemento TAVI.

A metodoloxía proposta céntrase en establecer un estándar robusto para a segmentación do calcio aórtico, un proceso tradicionalmente realizado en imaxes CT sen contraste. Con todo, ao utilizar imaxes de anxiografía por tomografía computarizada (CTA) con contraste, os brillos e as características do paciente poden variar considerablemente, o que require unha adaptación específica ás condicións individuais de cada imaxe.

Inicialmente, a segmentación do calcio aórtico realízase de forma manual para garantir unha segmentación precisa das diferentes rexións do calcio aórtico, dado que estas imaxes presentan variacións debido ás características dos pacientes e ao uso do contraste. Posteriormente, este proceso automatízase mediante o uso de técnicas avanzadas de intelixencia artificial, especificamente cun enfoque de segmentación multiclase, o que significa que o modelo é capaz de identificar e segmentar diferentes rexións do calcio na aorta, clasificándoas segundo a súa localización. Isto non só mellora a precisión na identificación das rexións calcificadas, senón que tamén optimiza o proceso, eliminando a dependencia da intervención manual e aumentando a eficiencia da análise.

Un aspecto clave abordado ao longo do desenvolvemento e análise de todos os algoritmos foi a automatización de procesos, tanto na xeración de resultados como na súa visualización. Este enfoque permitiu axilizar a obtención de resultados e optimizar a súa presentación, facilitando así a súa integración nos fluxos de traballo diarios da empresa. Este enfoque foi fundamental tanto para a empresa como para o ámbito clínico, onde a rapidez e precisión na entrega de resultados son cruciais. A metodoloxía é validada utilizando datos clínicos reais, o que asegura a súa aplicabilidade e fiabilidade en contextos clínicos reais.

A tese finaliza coas **Conclusións** xerais deste traballo. Neste capítulo de conclusións, resúmense os principais avances acadados durante a investigación, destacando a validación e mellora de metodoloxías para a segmentación de estruturas cardiovasculares e a obtención de parámetros clínicos relevantes. A segmentación automatizada, xunto co desenvolvemento de modelos baseados en técnicas de intelixencia artificial, permitiu avanzar cara a unha mellor eficiencia no diagnóstico e análise de enfermidades cardiovasculares. A pesar das limitacións no número de datos e recursos clínicos, o traballo realizado amosa un importante potencial para futuras aplicacións, especialmente na predición de risco cardiovascular e na planificación de tratamentos. O proxecto continúa en desenvolvemento, co obxectivo de mellorar a súa capacidade e estender o seu uso a novos ámbitos clínicos.

Chapter 1

Introduction

Cardiovascular diseases (CVDs) remain a leading global health challenge, with their prevalence and impact steadily increasing over time. According to the 2019 Global Burden of Disease study, the number of CVD cases nearly doubled over three decades, rising from 271 million in 1990 to 523 million in 2019. This increase has been accompanied by a significant rise in mortality, with deaths growing from 12.1 million in 1990 to 18.6 million in 2019 [1].

In Europe, CVDs continue to impose a considerable public health burden. Data from the European Society of Cardiology (ESC) highlight that these conditions are the leading cause of death among ESC member countries, responsible for over 3 million deaths annually [2]. Furthermore, the 2023 ESC report underscores that CVDs account for 11% of total healthcare expenditure across the European Union. Disparities in mortality are evident within the region, with medium-income countries bearing a heavier burden. In such nations, CVDs contribute to 53% of female and 46% of male deaths, compared to 34% and 30% in high-income countries, respectively [3].

In Spain, cardiovascular diseases are the foremost cause of mortality. Risk factors such as hypertension, smoking, type 2 diabetes, obesity, and hypercholesterolemia are highly prevalent. Specific to atherosclerosis [4], the EVA (Evaluación de la Aterosclerosis) study revealed a 40% prevalence of coronary atherosclerosis in asymptomatic populations, with men (51%) being more affected than women (25%) [5]. This growing prevalence and associated mortality underscore the critical need for innovative strategies to address these diseases at both global and regional levels.

To address the growing burden of cardiovascular diseases (CVDs), medical imaging has become indispensable in both clinical management and research, as it provides critical, non-invasive insights into coronary structure and function, enabling more precise diagnoses and tailored treatment strategies.

Currently, a variety of imaging modalities and protocols are employed, including computed tomography (CT), magnetic resonance imaging (MRI), transthoracic and transesophageal echocardiography (TTE and TEE), and abdominal aortic ultrasound. These modalities are selected based on patient-specific factors such as hemodynamic stability, renal function, and tolerance to procedures [6]. Among these, Coronary Computed Tomography Angiography (CCTA) stands out as the gold standard for coronary artery imaging due to its high resolution and non-invasive nature [7]. By employing intravenous radioopaque contrast agents, CCTA generates sharp, detailed images that distinguish blood-filled regions from surrounding tissues. This capability makes CCTA invaluable for diagnosing and assessing coronary

conditions, despite its requirement for radiation exposure. Additionally, Coronary Computed Tomography Angiography (CCTA) enables detailed anatomical and functional assessments, such as measuring the caliber of blood vessels, segmenting structures, and reconstructing 3D geometries of coronary arteries or other cardiovascular structures. It also supports advanced hemodynamic evaluations, such as Fractional Flow Reserve derived from CT (FFR_{CT}), further enhancing its utility in both clinical and research contexts [8, 9].

Coronary Computed Tomography Angiography (CCTA), while highly valuable, has limitations, particularly in terms of resolution and susceptibility to artifacts, which can compromise image clarity and diagnostic accuracy. These artifacts may arise from factors like patient movement, metal implants, or calcifications. Moreover, in cardiology, the small size and dynamic motion of coronary arteries present additional challenges, making precise imaging and analysis crucial [8, 9]. These constraints highlight the need for continuous advancements in imaging techniques and computational analysis to enhance CCTA's capabilities.

These challenges in cardiac imaging have driven the integration of artificial intelligence (AI) into the field, where it plays a pivotal role in advancing precision medicine. AI enhances the accuracy and efficiency of cardiac and aortic imaging techniques while automating complex processes, leading to faster and more reliable outcomes [10]. Neural networks, particularly U-Net-based architectures, are increasingly employed for tasks like coronary artery segmentation, aortic structure delineation, pathology detection, and risk stratification [11]. The research community is actively pursuing high-fidelity, efficient segmentation methods that minimize the need for user corrections, addressing critical clinical and research demands [7].

This Introduction section begins with the anatomy of the coronary system, including the aorta and coronary arteries (Section 1.1), and introduces the basics of computed tomography (CT) medical imaging (Section 1.2). It then discusses key clinical procedures, including the diagnosis and treatment of coronary artery disease (CAD) and aortic valve stenosis (AVS) (Section 1.3). Finally, the integration of artificial intelligence in medical image processing is explored, particularly its impact on imaging segmentation accuracy, efficiency and personalized medicine (Section 1.4).

1.1 Coronary Anatomy and Physiology

Anatomy is not merely a descriptive discipline; it forms the fundamental basis for understanding physiological processes and recognizing pathological changes, which are essential for both medical education and clinical practice [12]. The heart, as a central organ in the circulatory system, requires a continuous supply of blood to sustain its function and support life. It works as a powerful pump, sending oxygen-rich blood throughout the body and returning oxygen-poor blood to the lungs for reoxygenation. This complex system of circulation is vital for delivering oxygen and nutrients to all tissues, from major organs like the brain to the smallest peripheral tissues [13].

Figure 1.1 presents a detailed diagram of the human heart, illustrating the circulation of blood through its chambers and major vessels. White arrows indicate the direction of blood flow, distinguishing between deoxygenated (blue) and oxygenated blood (red). The blood flow functions as follows: blood enters the right atrium from the body, flows into the right ventricle, and is then pushed into the lungs through the pulmonary arteries. There, the blood is oxygenated and it releases carbon dioxide. Then, it returns to the heart via the pulmonary veins, entering the left atrium. Right after, it moves into the left ventricle. Finally, the left ventricle pumps the

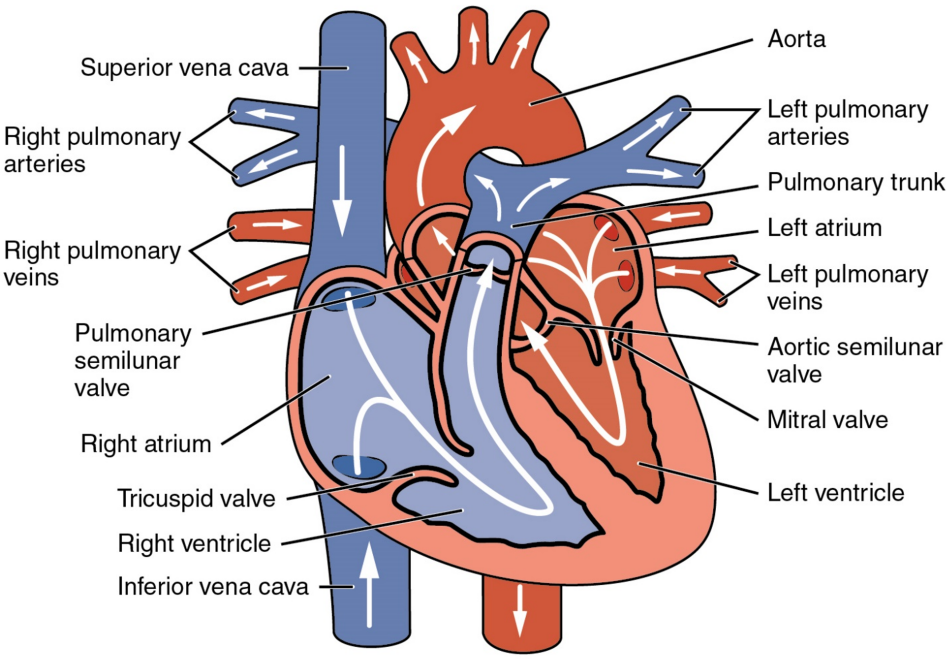


Figure 1.1: Heart and blood flow dynamics. The figure highlights the major blood vessels and the direction of blood circulation, including both systemic and pulmonary circuits, to demonstrate the flow through the heart's chambers. Image from Wikimedia Commons.

oxygenated blood through the aorta, which distributes it throughout the body [14].

In the following sections, we dive into the anatomy and physiology of the aorta (Section 1.1.1) and the coronary arteries (Section 1.1.2), both of which are crucial to this study. We also explore calcium deposits (Section 1.1.3), as these deposits can form in both the aorta and coronary arteries, leading to stenosis and blockages, reducing arterial flexibility, and ultimately compromising proper function.

1.1.1 Aorta

The aorta, originating from the left ventricle, is the largest elastic artery in the human body. It carries oxygen-rich blood from the heart and branches out into a network of arteries, ensuring that every organ and tissue receives adequate oxygen and nutrients. It plays a key role in systemic circulation, maintaining the supply needed for the cells' proper functioning.

The aorta is divided into two main segments: the thoracic aorta (TA) and the abdominal aorta (AA), which are separated by the diaphragm [15]. This section will specifically focus on the thoracic aorta, which is further divided into four distinct regions: the aortic root, the ascending aorta, the arch, and the descending aorta, whose location and major components are illustrated in Figure 1.2.

The thoracic aorta is highly elastic, particularly in its proximal segments, which enables it to maintain diastolic pressure and succpport blood flow to the peripheral circulation. This compliance is greater in the thoracic aorta compared to the abdominal aorta, reflecting its functional significance [15].

The ascending aorta measures approximately 5 to 7 cm in length, with a diameter ranging from 2.5 to 3.0 cm. It is divided into two parts: the aortic root and the tubular ascending aorta

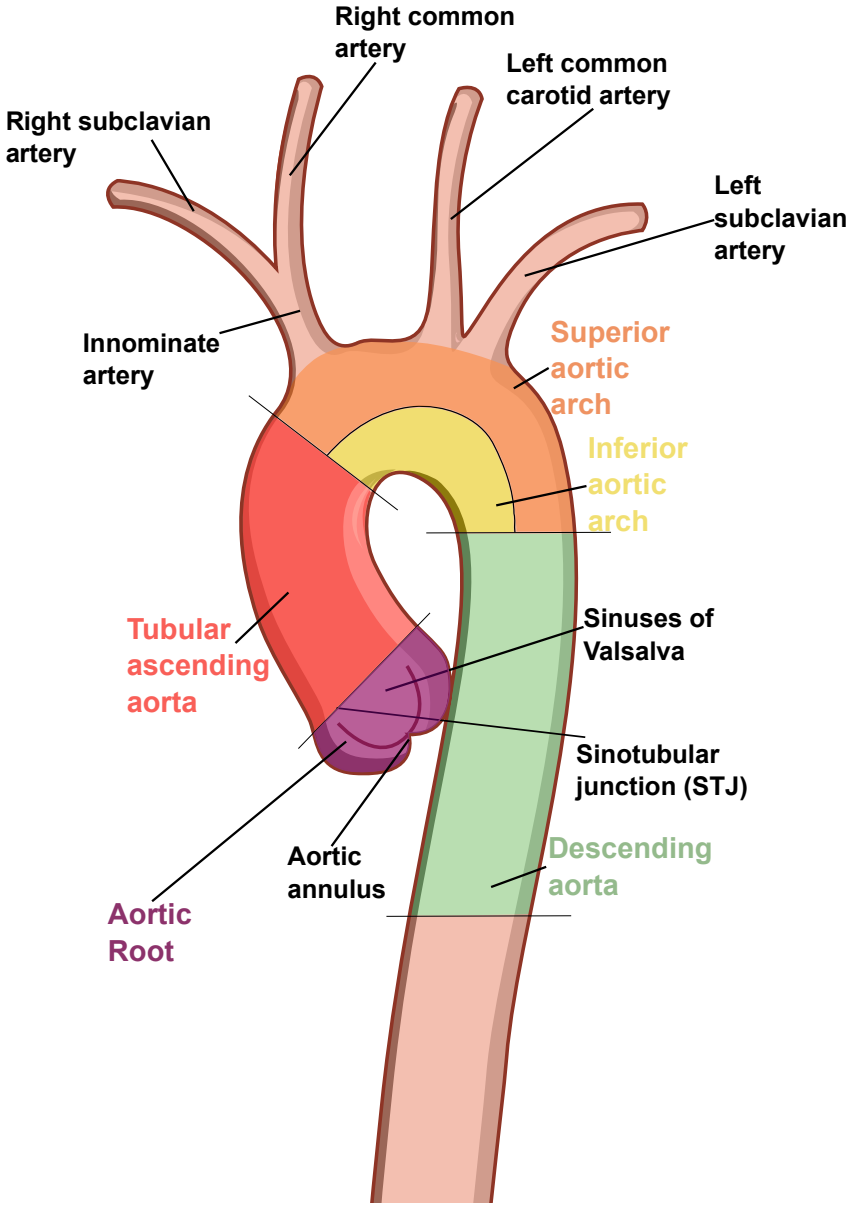


Figure 1.2: Diagram of the thoracic aorta segmented by regions. The labeled sections include the aortic root (purple), tubular ascending aorta (red), aortic arch with inferior and superior portions (yellow and orange, respectively), and descending aorta (green). Image modified from Servier Medical Art.

[16], represented by purple and red colors in Figure 1.2, respectively. The aortic root itself begins at the aortic annulus and extends to the sinotubular junction (STJ), where it transitions into the tubular segment of the ascending aorta. It consists of several key components, including the aortic annulus, the aortic valve, the sinuses of Valsalva, and the origins of the coronary arteries [16, 17]. The aortic annulus is a virtual circular line located at the base of the aortic cusps. It is positioned distally to the anatomical ventricular-aortic junction, where cardiac muscle fibers give way to smooth muscle cells of the aortic wall, as well as to the hemodynamic junction, where the cusps attach to the sinus walls in a crown-like pattern [15]. At the sinotubular junction, a similar virtual line runs through the tips of the valve commissures. It is important to note that the wall of the sinuses of Valsalva is thinner (approximately 2 mm) compared to the thicker (4 mm) wall of the aorta [15].

The ascending aorta, from which no branches emerge, then continues into the aortic arch [16]. The arch itself ascends to the left of the trachea and arches diagonally, descending beside the fourth thoracic vertebra to transition into the descending thoracic aorta. This portion of the aorta is notable for its complex relationships with adjacent anatomical structures, and variations in its anatomy are relatively common [16]. The superior convexity of the arch gives rise to three major arterial branches in an anterior-posterior orientation: the brachiocephalic (innominate) artery, the left common carotid artery, and the left subclavian artery [16], labeled in Figure 1.2.

1.1.2 Coronary Arteries

The coronary arteries supply the heart muscle with oxygenated blood, ensuring its continuous activity. If the flow through these arteries is compromised, as seen in cases of blockage or stenosis, it can lead to severe conditions such as myocardial infarction, where a portion of the heart muscle may be damaged or die [18]. This highlights the critical importance of the coronary arteries in maintaining cardiac health and function.

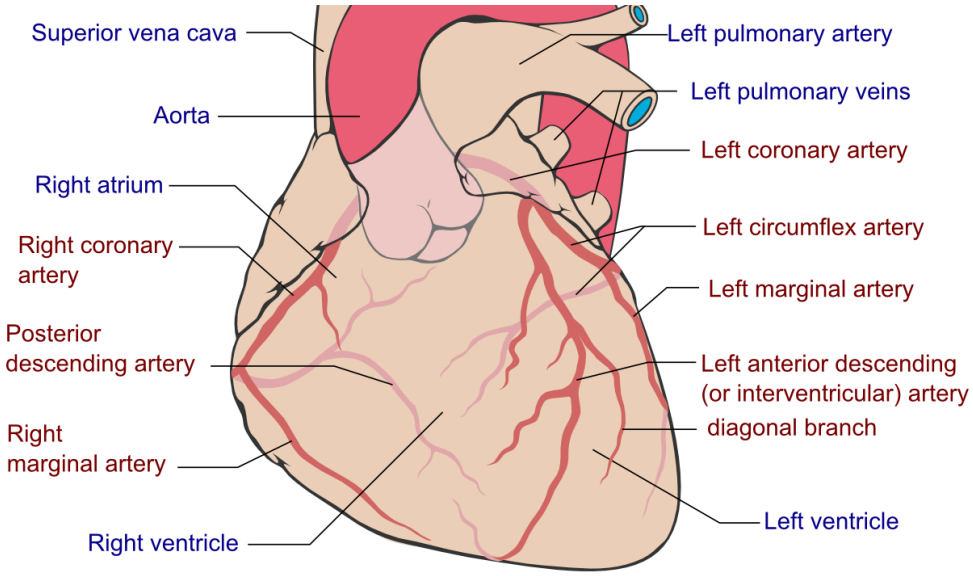
The coronary arterial system begins at the aortic root, specifically within the sinuses of Valsalva [19] (see Section 1.1.1). As shown in Figure 1.3a, the coronary arteries (depicted in red) surround the heart and are divided into two main branches depending on whether they supply the left or right side of the heart.

The right coronary artery (RCA) emerges from the right sinus of Valsalva and travels through the atrioventricular groove, descending towards the right border of the heart. Its largest branch, the right marginal artery, usually originates proximally before reaching the apex. It primarily supplies blood to the right atrium, right ventricle, and portions of the posterior heart [20].

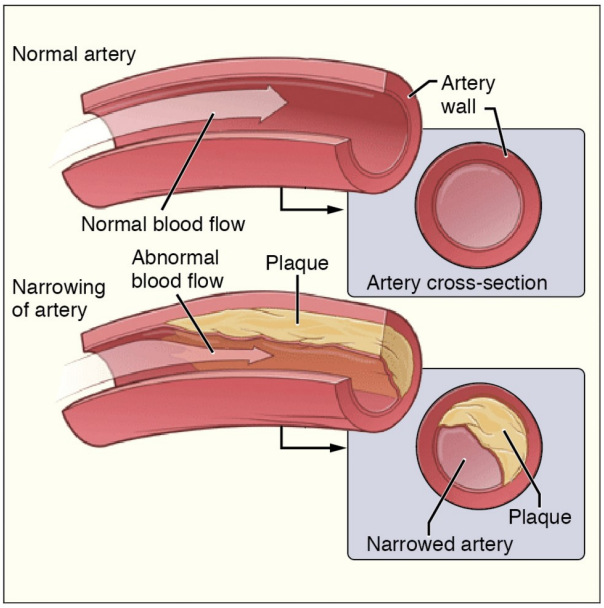
The left coronary artery (LCA), originating from the left sinus of Valsalva, runs between the left atrial appendage and the pulmonary trunk, and divides into the circumflex and anterior interventricular arteries, supplying significant portions of the heart, including the left ventricle, ventricular septum, and left atrium [20].

Although the anatomy of these vessels is well-documented, there remains a broad spectrum of anatomical variations and anomalies that are still debated within the scientific community [19].

Also the vessel caliber varies along its length, with larger diameters at the beginning and smaller at the distal end. Dodge et al. [21] reveals the following vessel calibers, although they can vary depending on the individual and the presence of lesions. The left main artery measures 4.5 ± 0.5 mm, the proximal left anterior descending coronary artery (LAD) is 3.7 ± 0.4 mm,



(a)



(b)

Figure 1.3: Coronary artery anatomy and plaque accumulation. (a) Illustration of the heart (brown) with the coronary arteries and aorta in red. The main vessels and heart chambers are labeled for reference. (b) Cross-section of a normal coronary artery compared to one with plaque buildup, leading to vessel narrowing and potential blood flow restriction. Images from Wikimedia Commons.

and the distal LAD narrows to 1.9 ± 0.4 mm. The proximal diameter of the right coronary artery (RCA) ranges from 3.9 ± 0.6 mm to 2.8 ± 0.5 mm ($p < 0.01$), and for the left circumflex artery, it varies between 3.4 ± 0.5 mm and 4.2 ± 0.6 mm ($p < 0.01$). Additionally, women generally have smaller epicardial arterial diameters than men, with an average difference of -9% ($p < 0.001$).

1.1.3 Calcium Deposits

Calcifications are a common component found in atherosclerotic plaques, which are focal thickenings of the intimal layer of arteries. These plaques form due to the accumulation of various substances, including foamy macrophages, blood products, smooth muscle cells, lipids, collagen, necrotic debris, and calcium [22, 23, 24]. The presence of calcifications typically indicates a more advanced stage of plaque development and is often associated with stable plaques [23]. However, microcalcifications, which are small, speckled calcifications, can be present in high-risk plaques prone to rupture [22].

The precise mechanisms behind calcification in atherosclerotic plaques remain under investigation. However, several theories and contributing factors have been proposed [22, 24]:

- Inflammation: Inflammatory cells within the plaque may release factors that promote calcification.
- Cell death: The death of cells within the plaque releases substances that can stimulate calcification. Necrotic debris from these cells may act as a nucleus for calcium deposits.
- Mechanical stress: The mechanical forces exerted on the artery wall, such as pressure and stretching, may also play a role in promoting calcification.

As plaques heal, they may undergo extensive calcification and fibrosis, potentially leading to moderate or severe stenosis. Figure 1.3b illustrates an example of plaque accumulation in a previously healthy blood vessel. The upper part of the subfigure represents a normal vessel, while the lower part shows the same vessel affected by plaque buildup. On the right side, the cross-sectional view highlights how the diameter of the lumen has been significantly reduced due to the accumulation, leading to stenosis and potentially restricting blood flow.

Positive remodeling can also occur, where the arterial wall expands outward. This is often a sign of vulnerable plaques with a larger necrotic core [22].

When a plaque ruptures, it can trigger the formation of a thrombus (blood clot), obstructing the artery and potentially leading to severe cardiovascular events such as myocardial infarction (heart attack) or stroke [22, 23, 24].

1.2 Medical Imaging

Medical imaging is a fundamental component in modern healthcare, providing visual representations of the human body's internal structures and allowing the examination of tissues and organs beneath the skin and bones, which is essential for early detection and diagnosing abnormalities, guiding disease treatment, and making precise interventions [25, 26]

Despite some disadvantages, such as the exposure to radiation in techniques like CT (Computed Tomography), non-invasive imaging modalities, have greatly expanded diagnostic

capabilities. For instance, CT is highly effective for examining the heart, abdomen, bones, and spinal cord due to its high-resolution imaging, while MRI (Magnetic Resonance Imaging) offers valuable insights for soft tissue and neurological conditions without exposing patients to ionizing radiation, making it a safer alternative in some cases [25, 27, 28].

The fundamentals of computed tomography (CT) are addressed in Section 1.2.1. Image artifacts, which are discrepancies between the attenuation values in the reconstructed image and the scanned object, are discussed in Section 1.2.2. Finally, we focus on cardiac CT, exploring its specific features and limitations in Section 1.2.3.

1.2.1 Computed Tomography

Computed tomography (CT) is an advanced imaging technology that extends the principles of X-ray imaging to generate detailed cross-sectional images of the human body. In fact, according to Webster's dictionary, the term "tomography" originates from the Greek word "tomos", meaning "a technique in X-ray imaging where a single plane is captured, and the structures in other planes are excluded" [29].

X-rays, a form of electromagnetic radiation with shorter wavelengths and higher energy compared to visible light, are utilized in medical imaging due to their ability to penetrate the body and interact with various tissues differently. As X-rays pass through the body, they are attenuated—or weakened—to different degrees based on the density and atomic number of the materials they encounter. Denser materials, like bone, attenuate X-rays more significantly than softer tissues, creating the contrast observed in X-ray images and enabling visualization of internal structures. The interactions between X-rays and matter occur primarily through the photoelectric effect, Compton scattering, and coherent scattering, with each playing a role in forming the image depending on the material's atomic structure [29].

CT imaging builds upon these principles but employs more sophisticated techniques. The fundamental concept of CT involves acquiring multiple X-ray projections from various angles around the patient. This is achieved through a synchronously rotating X-ray source and detector system that collects attenuation data from different paths as the X-rays pass through the body. These measurements are then processed using advanced algorithms to reconstruct cross-sectional images of the scanned region. The reconstructed images display the spatial distribution of X-ray attenuation coefficients within the body, providing precise anatomical details of internal structures [29].

Attenuation values in CT imaging play a critical role in distinguishing different types of tissues. Water and soft tissue, for instance, have nearly identical attenuation coefficients because a large proportion of soft tissue is composed of water. This similarity makes water phantoms ideal for calibrating CT systems to ensure accuracy in representing soft tissues [29].

To quantify these attenuation differences, CT imaging employs a standardized scale called Hounsfield units (HU), named after the inventor of CT, Sir Godfrey Hounsfield. This scale is based on the linear attenuation coefficient of water, which is assigned a value of zero HU. Air, with an attenuation coefficient close to zero, has a value of -1000 HU, while soft tissues such as fat, muscle, and organs range from -100 HU to 60 HU. More attenuating structures like cortical bones range from 250 HU to over 1000 HU, allowing for a comprehensive differentiation between tissues of varying densities and compositions [29]. In general, the intensity scale used in the reconstructed image, known as the CT number (HU scale), is defined by Equation 1.1.

$$\text{CT number} = \frac{(\mu - \mu_{\text{water}})}{\mu_{\text{water}}} \times 1000 \tag{1.1}$$

where μ_{water} represents the linear attenuation coefficient of water [29].

Since the composition of bones and tissues is generally consistent, aside from minor variations, this allows for the generic identification of different structures in CT scans. However, iodine, which has a higher attenuation coefficient, is often used as a contrast agent in CT imaging, particularly in angiography, to enhance the visibility of blood vessels [29]. This creates a challenge in identifying the vessel lumen, since the brightness varies in each scan. Therefore, dynamic algorithms based on the local brightness of the patient's lumen are required for accurate detection [30, 31, 32].

One of the key advantages of CT imaging over traditional radiography is its ability to enhance contrast resolution, providing detailed differentiation between tissues with similar densities. This is achieved by eliminating superimposed structures, which is a limitation in conventional X-rays, and enabling a clearer, more precise visualization of individual organs and tissues. Moreover, CT technology allows for the creation of three-dimensional representations of the scanned area, aiding in detailed anatomical assessments and surgical planning. This ability to visualize the body in such detail has transformed diagnostic imaging, making CT an indispensable tool in modern medicine [26, 29].

However, despite its advantages, CT imaging does come with the concern of radiation exposure. While the benefits of CT imaging in diagnosing and managing diseases often outweigh the risks, careful management of radiation dose is crucial. Techniques such as optimizing scan parameters, employing tube current modulation based on patient size, and using iterative reconstruction algorithms can minimize exposure without compromising diagnostic quality. These advancements not only enhance patient safety but also uphold the diagnostic integrity that makes CT a powerful tool in clinical practice [26].

1.2.2 Image Artifacts

Image artifacts in computed tomography (CT) represent discrepancies between the reconstructed image values and the actual attenuation coefficients of the scanned object. While nearly every CT image contains some form of artifact, particular attention is given to those that significantly impact a radiologist's performance. Since artifacts can significantly affect image quality and potentially influence clinical diagnoses [29].

CT image artifacts can be classified into four main types: streaking, shading, rings and bands, and miscellaneous artifacts [29].

- Streaking artifacts appear as bright or dark lines across the image, often resulting from inconsistencies in projection data that become amplified during the reconstruction process. While they are generally not a major diagnostic concern, a high number of pronounced streaks can severely degrade image quality.
- Shading artifacts manifest as areas of increased or decreased brightness near high-contrast objects, such as bones or air, and can lead to misdiagnosis by mimicking actual pathological conditions.
- Rings and bands appear as circular patterns or bands superimposed on the image; full rings are less problematic than partial rings, which may resemble specific pathologies.

- Miscellaneous category includes less common artifacts, such as basket weave patterns that can arise during image resizing.

Throughout this project, it has been necessary to address various types of artifacts, particularly those affecting CCTA image quality and segmentation accuracy. Figure 1.4 illustrates examples of these artifacts, showing how motion artifacts can alter a coronary artery vessel morphology and how stents introduce high-density artifacts that obscure details. Addressing these challenges is crucial for improving automated analysis and ensuring reliable clinical assessments.

- Motion Artifact: This occurs when there is a brief movement of a patient or an object during a CT scan, resulting in inconsistent measurements. This movement leads to distortions in the reconstructed image, typically appearing as streaks [26, 29]. Two examples of motion artifacts in CCTA (Coronary Computed Tomography Angiography) can be seen in Figure 1.4a to 1.4d. The figure displays four axial slices along the distal direction of the vessel. In (a), the vessel appears normal; however, as we progress through the slices, the vessel begins to blur, eventually forming a ring-like structure. This artifact does not correspond to the actual anatomy, as there is no presence of calcium or metallic material in the vessel. In Figure 1.4e, we see the ostium of the left coronary artery and its proximal segment. In this region, the middle part appears brighter than the adjacent areas due to a motion artifact.

- Blooming Artifact: This occurs when high-density objects, such as calcifications or stents, appear larger than their actual size in the reconstructed image. The intensity from these dense objects tends to spill over into surrounding areas, making it difficult to delineate their boundaries. This issue is particularly critical when assessing the narrowing of a vessel's lumen. If blooming from a stent extends into the luminal space, it becomes nearly impossible to accurately evaluate vessel integrity [29, 33]. In cases where contrast agents are used in imaging, calcified plaques appear even brighter, making boundary delineation more challenging [34, 35, 36]. In Figure 1.4f, a stent placed in a coronary artery is observed. The section inside the stent exhibits also blooming artifacts, which are not present in the calcium deposit at the ostium in Figure 1.4e (which appears as a solid white mass). Another example can be seen in Figure 1.6, where calcium deposits in the aortic valve appear larger than their actual size due to blooming artifacts, which result from high-density materials causing an exaggerated signal in CT imaging.

- Metal Artifact: This is caused by dense metallic objects, such as implants or prostheses, within the scan area. Metal artifacts occur due to a combination of factors, including beam hardening, scatter, and nonlinear partial volume effects. They often appear as streaks or distortions radiating from the metal, obscuring surrounding anatomical structures. This can complicate diagnosis and reduce the clarity of important regions near the metallic object [33].

- Partial Volume Effect: This artifact arises when a single voxel in the CT scan contains multiple tissue types, resulting in an averaged attenuation value that does not accurately reflect any individual tissue. The effect becomes more pronounced as slice thickness increases, since thicker slices may encompass different structures with varying attenuation coefficients. For instance, as slice thickness increases from 1 mm to 10 mm,

the likelihood of partial volume artifacts also increases, leading to a loss of detail and accuracy in the reconstructed image [29, 33].

The causes of CT artifacts are varied and can stem from the physics of X-ray interactions, system design limitations, patient characteristics, and operator errors. To address these artifacts, methods can be divided into two main categories: artifact correction and artifact avoidance. Artifact correction involves algorithms and techniques designed to reduce or eliminate artifacts in the reconstructed image. Many research efforts focus on these strategies, although some methods remain proprietary. On the other hand, Artifact avoidance relies on optimal system design and proper scanner operation. Manufacturers provide best practices and protocols to minimize artifacts, while effective training for operators is crucial for successful implementation of these recommendations [29].

In summary, CT image artifacts pose significant challenges, originating from various sources, and can greatly compromise diagnostic quality. Recognizing these artifacts, understanding their limitations, and knowing how to address them are essential for ensuring accurate solutions and improving the reliability of diagnostic outcomes. A key objective of this thesis is to reduce these artifacts, particularly those caused by metallic implants, in order to enhance image clarity and diagnostic precision in clinical applications.

1.2.3 Cardiac Imaging

The heart is a constantly moving organ, making static imaging inherently challenging. During the cardiac cycle, two important phases define its movement: end-systole, when the left ventricle (LV) is in full contraction, and end-diastole, when the LV is most expanded. This dynamic process necessitates specific imaging strategies to effectively "freeze" the heart's motion, which is typically achieved through Computed Tomography Angiography (CTA). However, cardiac imaging may also involve dynamic imaging, such as angiography, where interpreting real-time movement of coronary structures presents a separate set of challenges.

Due to this anatomic and physiologic nature, accurately imaging coronary vessels demands high spatial resolution, optimized temporal resolution, and careful management of radiation exposure, with each aspect posing its own set of challenges:

- Temporal Resolution: precise temporal resolution is essential to minimize motion artifacts from the heart's movement. Electrocardiogram (EKG) gating techniques allow image capture to align with the cardiac cycle's low-motion phases, end-systole and end-diastole, reducing blurring.

- Spatial Resolution: The small dimensions of coronary vessels necessitate high spatial resolution to clearly visualize vessel structures and identify any stenosis (remember that coronary vessels have a diameter around 4 mm). Achieving this level of detail without increasing noise levels requires adjustments in X-ray flux, adding another layer of technical complexity.

- Radiation Dose: In earlier cardiac CT imaging techniques, high radiation doses were necessary due to the need for continuous coverage as the scanner moved slowly along the heart. This slower movement ensured that no area of the heart was missed, but it also meant that the same regions received repeated radiation exposure. With advancements in

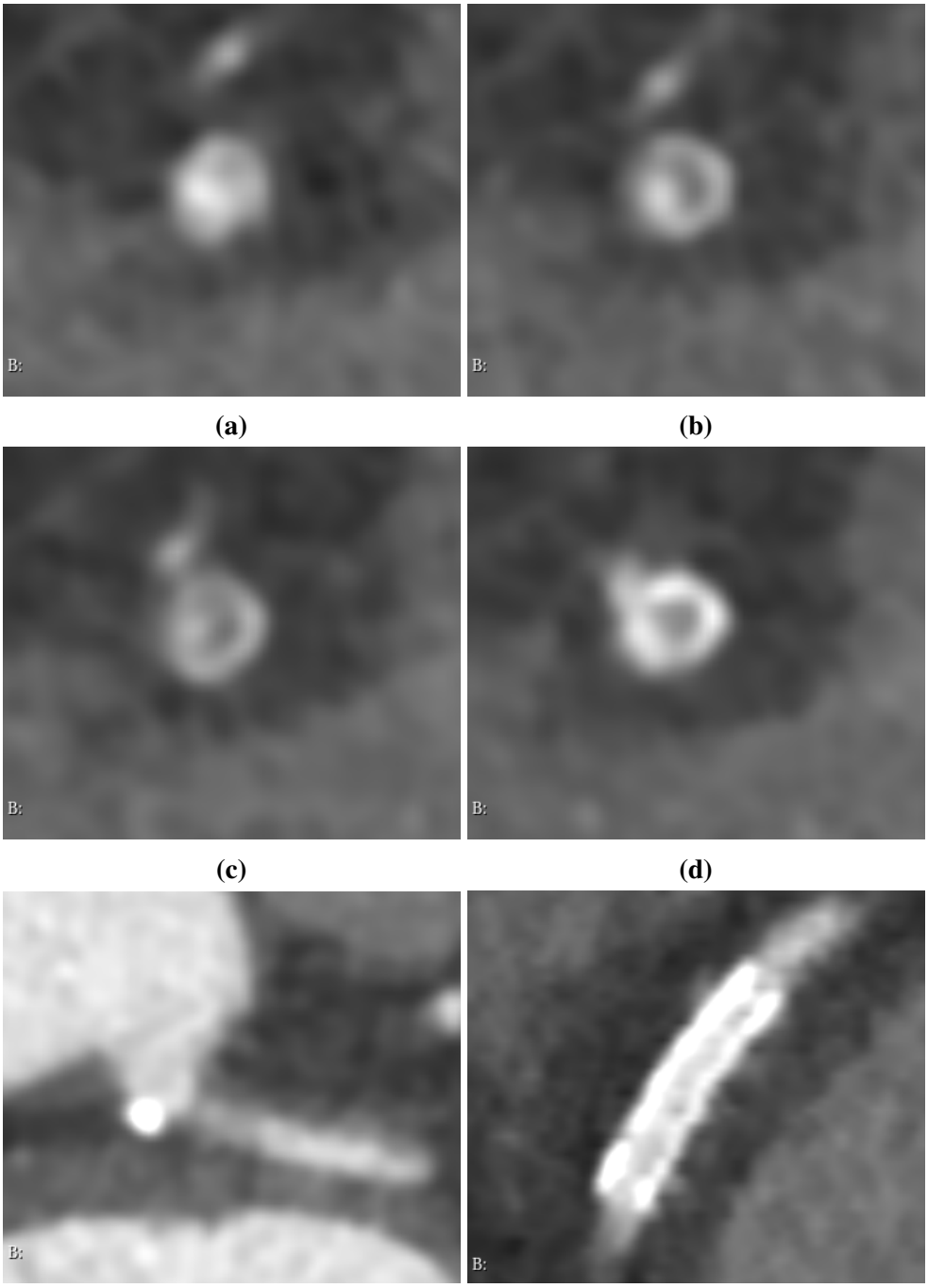


Figure 1.4: Artifacts in CCTA affecting coronary artery visualization and segmentation. (a–d) Motion artifact in the right coronary artery (RCA), showing three axial slices in the distal direction where the vessel appears ring-shaped. (e) Motion artifact in the left coronary artery (LCA), where lumen brightness variations are visible. (f) Stent placed in a vessel. showed in coronal view. Blooming artifact and metal artifact obscure the lumen, making accurate segmentation challenging. Voxel size is 0.25mm^3. Images provided by FlowReserve Labs S. L.

CT technology, modern scanners can now capture the entire heart in a single, rapid scan or with fewer passes. This approach significantly reduces the total radiation dose, as each area is exposed fewer times while maintaining high image quality. Consequently, newer CT systems enable effective cardiac imaging with much lower radiation exposure than was previously possible.

- Coverage: Comprehensive cardiac imaging requires capturing the entire heart volume, typically $10 - 16$ cm in the z-axis, within a single breath-hold. Early scanners lacked sufficient z-coverage, leading to multi-segment scans that increased radiation exposure and created the potential for motion artifacts due to vessel displacement between cycles. Newer CT scanners now offer larger z-coverage, enabling single-capture scans and reducing these issues.

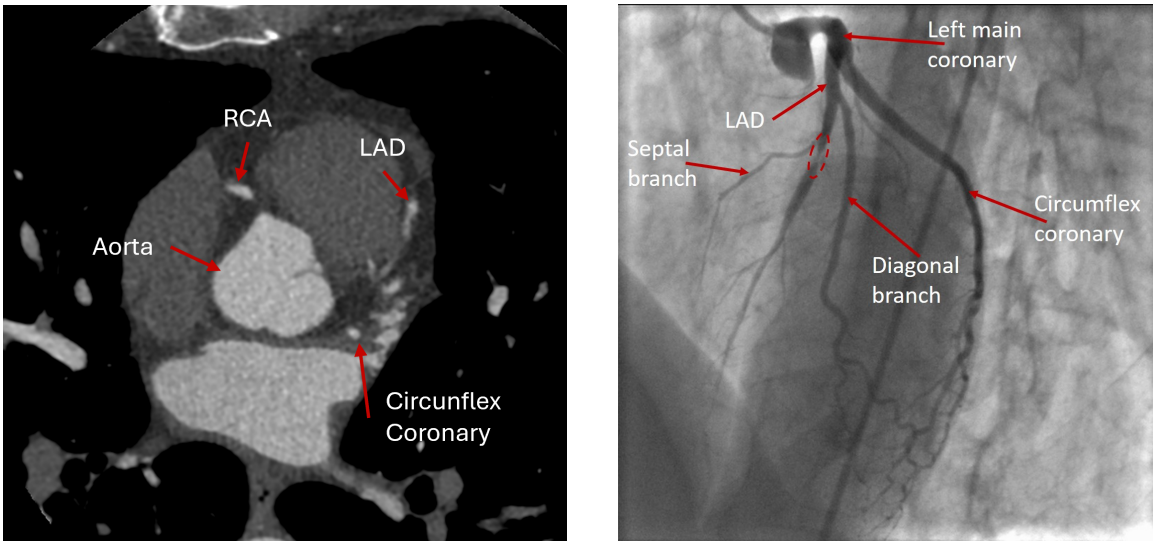
Cardiac imaging techniques generally fall into two categories: Coronary Artery Imaging (CAI) and Coronary Artery Calcification (CAC).

Coronary Artery Imaging (CAI) Coronary Artery Imaging (CAI) visualizes the heart's vascular structure, primarily to detect stenosis and plaque deposits within coronary arteries. CAI can also be used to assess the motion of heart muscles, which adds a functional perspective to structural assessment. However, CAI requires high scanner performance because it needs to accurately portray vessel size while also "freezing" cardiac motion, thus necessitating high temporal resolution. To optimize this, electrocardiogram (EKG) gating synchronizes data acquisition with the heart's cycle, capturing images during the end-systolic and end-diastolic phases, when motion is minimal. Even slight variations in the cardiac cycle can introduce motion artifacts, which can significantly compromise image quality if not addressed during image acquisition [29].

In this thesis, we will employ Computed Tomography Angiography (CTA) images for coronary artery imaging (CAI), particularly for diagnosing coronary artery disease (CAD). CTA captures images by introducing contrast agents to visualize the lumen of coronary arteries, allowing the imaging system to distinguish soft tissue from blood flow within the vessels. Typically, CTA is performed during the diastolic phase, when the heart is most relaxed and coronary arteries are fully dilated, minimizing motion artifacts and providing a clear view of the lumen. Although CTA offers high spatial resolution, the procedure requires high temporal resolution to maintain accuracy in imaging a constantly moving organ.

In addition to CTA, coronary angiography is an established method of imaging coronary arteries, often used for its finer spatial resolution (100-200 μm), which allows detailed visualization of the artery's inner lumen. However, angiography is an invasive procedure that involves threading a catheter into the coronary arteries and introducing contrast dye to highlight blood flow on X-ray images. While this approach provides exceptional resolution for assessing narrowings and blockages, it exposes both the patient and the medical team to higher levels of radiation. Additionally, because angiography visualizes the lumen through pre- and post-contrast image subtraction, it does not reveal calcified deposits or details about plaque composition. Thus, while CTA and coronary angiography each have distinct strengths, CTA remains a more suitable choice for the non-invasive evaluation of coronary artery disease (CAD) in this study.

An example of both modalities can be seen in Figure 1.5. In Figure 1.5a, a CTA scan provides a detailed view of the aorta and major coronary arteries, including the Right Coronary



(a) **(b)**

Figure 1.5: Coronary artery imaging modalities. (a) CTA image highlighting the aorta and major coronary arteries, including the Right Coronary Artery (RCA), Circumflex Artery (Cx), and Left Descending Artery (LDA). Image provided by FlowReserve Labs S. L. (b) Angiographic view of the Left Coronary Artery (LCA), where a lesion is outlined by a dashed circle. The LCA, Circumflex (Cx), Diagonal, and Septal branches are also labeled for reference. Case courtesy of Stefan Tigges, Radiopaedia.org, rID: 95338. For both images, voxel size is 0.25mm^3.

Artery (RCA), Circumflex Artery (Cx), and Left Descending Artery (LDA), allowing for anatomical evaluation. Figure 1.5b shows an angiographic image of the Left Coronary Artery (LCA), where a lesion is highlighted with a dashed circle. This technique enables real-time visualization of blood flow and vascular abnormalities. One notable aspect is that in CTA, areas containing contrast, such as blood vessels, appear brighter, typically in white. In contrast, in angiography, structures filled with contrast appear darker, almost black. This difference is due to the imaging principles of each modality: CTA relies on X-ray attenuation, where contrast-enhanced blood absorbs more radiation and appears brighter, while in angiography, contrast media block X-ray passage, creating negative images where contrasted areas appear darker.

Coronary Artery Calcification (CAC) Coronary Artery Calcification (CAC) imaging is a technique focused on detecting and quantifying calcium deposits within the coronary arteries and the aorta. This imaging method serves primarily as a screening tool to assess the risk of cardiac events, particularly in asymptomatic patients, by providing a measure of the calcium burden in coronary vessels and the aortic valve. The presence of coronary and aortic calcium is a well-established indicator of atherosclerosis (coronary artery disease, CAD) and calcific Aortic Valve Stenosis (AVS), respectively, making CAC imaging crucial for evaluating cardiovascular risk and valve health.

CAC imaging is typically performed with low-dose CT scans, acquired during the diastolic phase of the cardiac cycle to minimize motion artifacts that could interfere with image clarity. However, due to the lower resolution between slices in these low-dose images, CAC imaging

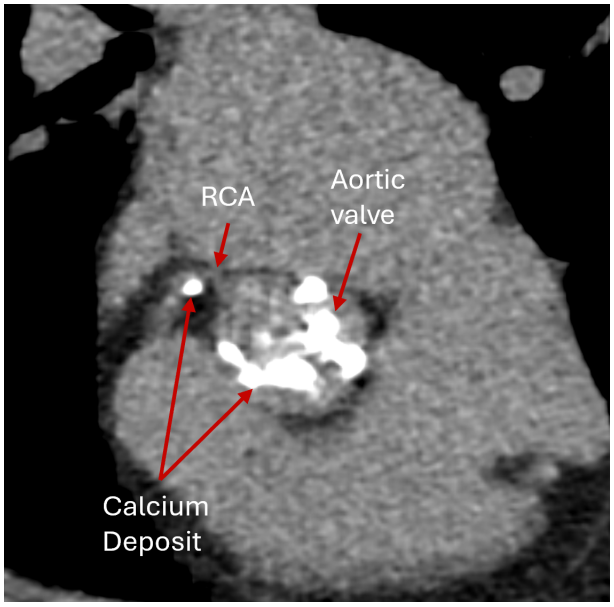


Figure 1.6: Non-contrast-enhanced CT image showing the ostium of the Right Coronary Artery (RCA) and the aortic valve. High-density regions corresponding to calcium deposits appear brighter, highlighting areas of calcification within the valve and arterial walls. Voxel size is $0.48 \times 0.48 \times 2.5\,\mathrm{mm}^3$. Image provided by FlowReserve Labs S. L.

is more susceptible to specific artifacts. One common artifact is the partial volume effect.This blending effect can lead to an overestimation of the extent or size of calcification.

Additionally, the high brightness of calcium on CT images can cause blurring artifacts, where the intense signal from calcium spills over into adjacent areas, obscuring clear boundaries and affecting the accurate delineation of plaque structures. These artifacts not only complicate the assessment of calcified regions but also impact the consistency of calcium scoring across different scans, which is crucial for monitoring disease progression over time.

Figure 1.6 presents a non-contrast-enhanced CT scan, capturing the ostium of the RCA along with the aortic valve. The image reveals bright, high-density areas corresponding to calcium deposits, which are key indicators of vascular disease. A blooming artifact is also visible, causing the calcifications to appear larger than their actual size.

In this thesis, CAC imaging will be used alongside other modalities to capture and analyze calcifications in both coronary arteries and the aortic valve, aiding in the assessment and diagnosis of aortic valve stenosis and coronary artery disease.

1.3 Clinical Procedures

The next section will explain the key clinical procedures relevant to this work.

In Section 1.3.1, we discuss coronary artery disease, covering its diagnosis through invasive techniques, treatment options, and non-invasive diagnostic alternatives. Section 1.3.2 focuses on the diagnostic and treatment methods for aortic valve stenosis, exploring the different imaging techniques used for its assessment and the available therapeutic approaches. Understanding these procedures is essential for evaluating the role of advanced imaging and computational methods in improving diagnosis and treatment outcomes.

1.3.1 Diagnosis and Treatment of Coronary Artery Disease (CAD)

As stated at the beginning of the Introduction (Chapter 1), coronary artery disease (CAD) is a leading cause of death worldwide. This condition affects the primary blood vessels that supply oxygen-rich blood to the heart, known as the coronary arteries (see Section 1.1.2). In CAD, blood flow to the heart muscle is reduced, often due to the buildup of fats, cholesterol, and other substances on the arterial walls, a condition known as atherosclerosis. This accumulation, referred to as plaque, narrows the arteries over time (see Section 1.1.3). CAD can develop gradually over many years, and when blood flow becomes significantly restricted, the heart experiences a lack of oxygen, leading to symptoms like chest pain (angina), shortness of breath, or fatigue.

1.3.1.1 Invasive Coronary Angiography (ICA)

Early diagnosis of CAD is crucial, as symptoms and consequences can worsen considerably if left untreated. Traditionally, the gold standard for diagnosing CAD was through Invasive Coronary Angiography (ICA), an imaging procedure that provides a detailed view of the coronary arteries. In ICA, a radio-opaque dye is injected into the bloodstream, and X-rays are used to create a real-time, moving image of the coronary tree (see Section 1.2.3). This allows for the identification of arterial narrowings, or stenoses, which can reduce blood flow and lead to myocardial ischemia [37, 38, 39]. The procedure of an Invasive Coronary Angiography (ICA) involves several key steps. First, a catheter is inserted through an access point, typically in the groin or wrist, and carefully guided through the vascular system toward the coronary arteries. Once the catheter reaches the target artery, a contrast agent is injected, enabling the visualization of the coronary vasculature under X-ray imaging. This allows clinicians to assess the presence of any abnormalities, such as stenotic regions where blood flow is restricted, as indicated by the arrow in Figure 1.7, that illustrates the procedure.

A stenosis of 50% or greater is typically considered clinically significant and used as a threshold for guiding treatment decisions, whether through medical management or revascularization. However, visual assessment of stenosis is often imprecise and not highly reproducible, particularly when compared to functional assessments like Fractional Flow Reserve (FFR) [39, 37].

1.3.1.2 Fractional Flow Reserve (FFR)

Fractional flow reserve (FFR) is a diagnostic procedure performed during invasive coronary angiography (ICA) to evaluate the severity of a stenosis, or a narrowing, in a coronary artery [39, 37, 40]. By directly measuring the pressure difference across a stenotic segment, FFR helps determine whether the narrowing significantly impedes blood flow to the heart muscle, a condition that can lead to myocardial ischemia [37, 40, 41].

During the FFR procedure, a specialized pressure-sensing guidewire is inserted and advanced distal to the coronary artery stenosis, guided by ICA images. To ensure an accurate pressure measurement, a vasodilator, usually adenosine, is administered to induce maximal hyperemia (increased blood flow) in the artery [37, 42, 43]. Under these conditions, pressure is measured in both the aorta (P_a) and distal to the stenosis (P_d), allowing FFR to be calculated as the ratio P_d/P_a [41]. Following the guidelines, An FFR value of 0.80 or below signifies

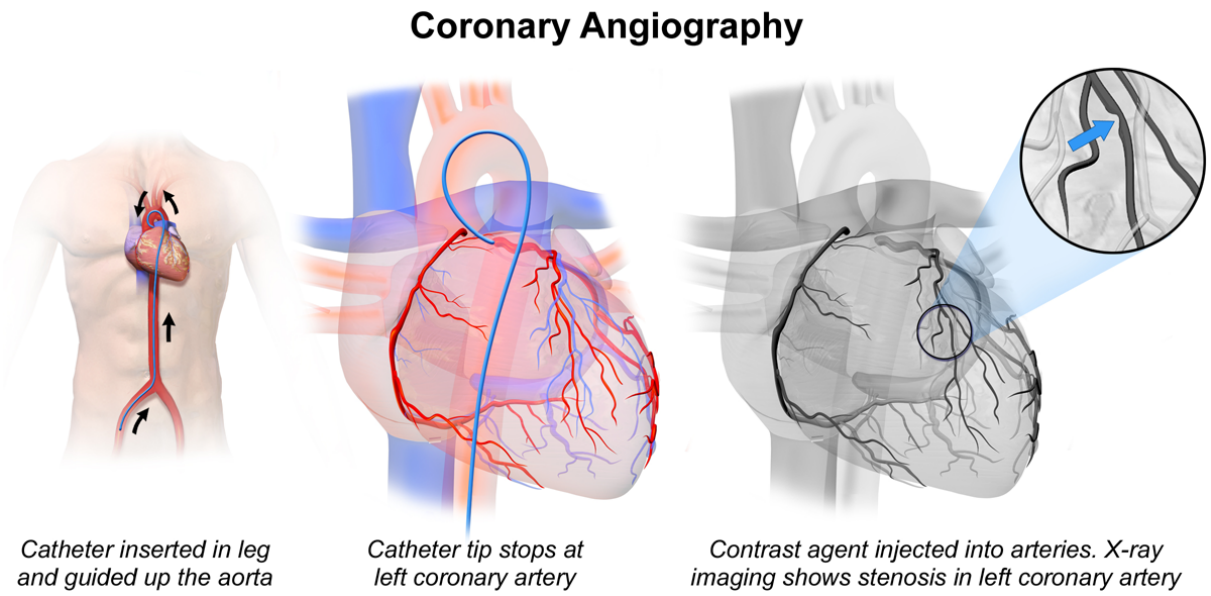


Figure 1.7: Diagram illustrating the process of coronary angiography. Left: A catheter is inserted through the groin and guided toward the coronary arteries. Center: The catheter reaches the desired coronary artery. Right: Contrast agent is injected, allowing visualization of the coronary vasculature under X-ray imaging. The arrow indicates a stenotic region, where blood flow is restricted. Image from Medical gallery of Blausen Medical 2014, by Bruce Blaus, published in WikiJournal of Medicine 1 (2). DOI: 10.15347/wjm/2014.010.

that the stenosis significantly restricts blood flow, enough to potentially cause ischemia in the downstream myocardium [44], leading to revascularization.

FFR is widely recognized as the gold standard for assessing the functional significance of intermediate coronary lesions, defined as stenoses that appear to obstruct $50-90\%$ of the coronary artery's diameter based on angiographic images [37, 45, 39, 42]. It is considered more reliable than alternative noninvasive tests like exercise electrocardiography or stress echocardiography for identifying hemodynamically significant stenoses [37]. Importantly, FFR-guided revascularization has demonstrated better clinical outcomes than angiography-guided procedures, with studies showing that it reduces rates of mortality, non-fatal myocardial infarction (MI), and the need for repeat revascularizations. Using FFR to guide percutaneous coronary intervention (PCI) or coronary artery bypass grafting (CABG) decisions can prevent unnecessary interventions and improve patient outcomes [37].

Despite its clinical benefits, the adoption of FFR in routine practice remains limited due to technical, economic, and procedural factors. In fact, it carries inherent risks associated with coronary angiography, such as bleeding at the catheter insertion site, allergic reactions to contrast dye, potential damage to the coronary artery, and arrhythmias. Furthermore, the administration of adenosine, used to induce hyperemia during the procedure, may also lead to hypotension, bradycardia, bronchospasm, and other transient side effects [37]. Moreover, there are risks of misinterpretation of FFR results due to the impact of the pressure wire on measurements [37, 45], as well as potential inaccuracies in specific clinical situations, including acute coronary syndromes and the presence of coronary stents [37].

Emerging technologies such as CT-derived FFR (FFR_{CT}) are being developed to provide

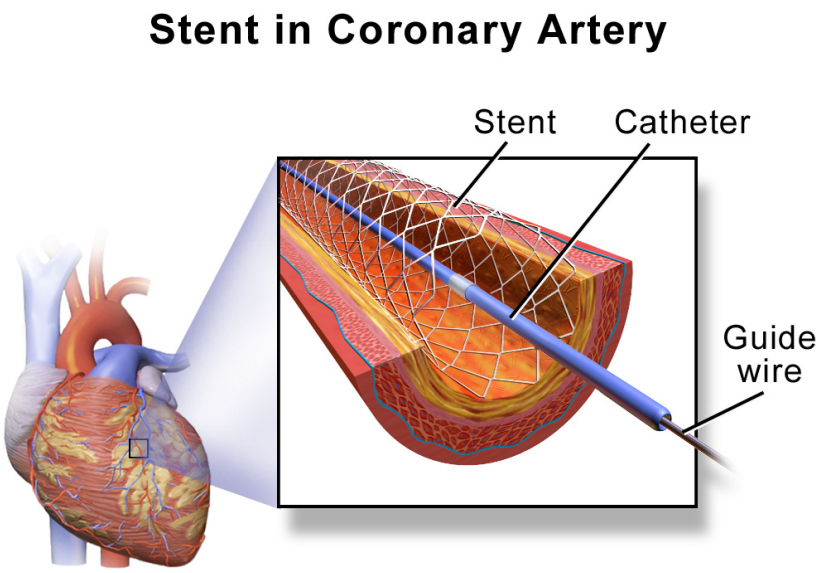


Figure 1.8: Diagram illustrating the placement of a stent in a coronary artery. The diagram shows the heart and a magnified view of the catheter and guidewire navigating through the artery. The stent is expanded at the site of the narrowing, helping to restore normal blood flow. Image from Blausen.com staff (29 August 2014). Image from Medical gallery of Blausen Medical 2014. WikiJournal of Medicine 1 (2). doi: 10.15347/WJM/2014.010. Wikidata Q44276831. ISSN 2002-4436.

a noninvasive alternative to traditional FFR. These methods use CT angiography images combined with computational fluid dynamics to estimate FFR values without the need for invasive procedures.

1.3.1.3 Percutaneous Coronay Intervention (PCI)

When diagnostic procedure, such as an FFR measurement of a value below 0.80, revascularization via percutaneous coronary intervention (PCI) becomes essential to restore sufficient blood flow to the heart. After identifying the lesion through invasive coronary angiography (ICA), a catheter is introduced, along with an injected contrast agent, similar to the method used for placing the pressure sensor for FFR measurement [46].

The next step involves a procedure known as balloon angioplasty. A second catheter, equipped with a balloon at its tip, is advanced over the guidewire to the site of the blockage. The cardiologist inflates the balloon once it is positioned correctly, which effectively compresses the plaque against the arterial wall, widening the artery and restoring blood flow [46, 47].

In many cases, the cardiologist will also place a stent—a small mesh tube—within the newly widened section of the artery (see procedure diagram in Figure 1.8). This stent remains in the artery permanently, helping to keep it open and reducing the likelihood of future blockages, a condition known as restenosis. Recent research advocates for the use of drug-eluting stents, which are coated with medication that gradually releases to reduce the risk of artery re-narrowing. However, further studies are still ongoing to explore their long-term efficacy and potential improvements [47].

1.3.1.4 Fractional Flow Reserve – Computed Tomography (FFR_{CT})

Both ICA and FFR are invasive diagnostic procedures that expose patients to radiation and carry inherent risks, such as arterial punctures and plaque dislodgement from catheter insertion. In clinical practice, between half and two-thirds of ICAs are performed without any subsequent intervention (PCI), which means that many patients undergo invasive diagnostic testing unnecessarily, as the procedures reveal no need for further invasive treatment [48, 42]. This highlights the need for more accurate, non-invasive diagnostic strategies for stable coronary artery disease (CAD), to avoid exposing patients to these risks.

In recent years, FFR_{CT} (Fractional Flow Reserve derived from Computed Tomography) has emerged as a non-invasive alternative to traditional FFR. FFR_{CT} uses coronary CT angiography (CCTA) combined with computational fluid dynamics (CFD) to assess the functional impact of coronary stenosis, eliminating the need for catheterization [41, 49]. CCTA has been widely accepted as a non-invasive method to rule out significant coronary artery disease and reduce the need for further invasive testing.

FFR_{CT} leverages computational fluid dynamics principles, treating blood as an incompressible Newtonian fluid, allowing the application of 3D Navier-Stokes equations to compute coronary flow, velocity, and pressure both at rest and during hyperemia [40, 43, 42]. Recent advancements have also led to the development of faster, reduced-order models that can be performed on-site without the need for supercomputers, though these models are limited when applied to small vessels, side branches, bifurcations, or eccentric lesions [43, 45].

The calculation of FFR from computational fluid dynamics requires precise geometry of the coronary arteries, which is obtained through the segmentation of the vessel lumen from the CCTA. This segmentation process, often performed manually, is tedious and prone to errors. As a result, there is a growing need for the automation of this procedure to ensure it is not only fast but also less dependent on user variability. Achieving this automation is one of the key objectives of this thesis.

1.3.2 Diagnosis and Treatment of Aortic Valve Stenosis (AVS)

Aortic valve stenosis (AVS) is a progressive cardiac disorder characterized by narrowing of the aortic valve, obstructing blood flow from the left ventricle to the aorta , which results in increased afterload [50, 51, 52]. Consequently, left ventricular hypertrophy (LVH) develops as the heart works harder to pump blood through the narrowed valve. As AVS progresses, LVH can lead to diastolic dysfunction, systolic dysfunction, myocardial ischemia, and eventually heart failure [51, 52].

This disease is one of the most common valvular heart diseases. It affects approximately 3% of adults in developed countries and its prevalence increases with age [53]. The main causes of AVS are as follows:

- Degenerative Calcified AVS: The most common cause of AVS in developed countries [50, 53], where aging of the aortic valve leads to calcium deposits on the valve leaflets, causing stiffness and narrowing [50, 51].
- Congenital AVS: Often due to a bicuspid aortic valve (BAV). Is present in 40% of patients over 70 and 60% of patients under 70 years of age [53].

- Rheumatic AVS: This form of AVS is more prevalent in developing countries and results from rheumatic fever, which can damage the valve leaflets and lead to stenosis [51].

The diagnosis of AVS involves a structured, multi-modality approach as outlined by the European Association of Cardiovascular Imaging (EACVI) [54].

1.3.2.1 Transthoracic echocardiography (TTE)

This modality is the primary imaging tool for assessing aortic valve stenosis (AVS), providing critical information on valve anatomy, function, and cardiac remodeling. Transthoracic echocardiography (TTE), a non-invasive ultrasound technique, produces real-time images of the heart's chambers, valves, and blood flow dynamics by positioning a transducer on the chest. Additionally, Doppler echocardiography is integrated into TTE, enabling precise measurement of blood flow velocities and gradients across the aortic valve. This combination offers key parameters essential for evaluating aortic valve stenosis severity, such as maximum aortic velocity, mean pressure gradient, and aortic valve area.

- Maximum aortic velocity refers to the highest blood flow velocity across the stenotic valve, measured in meters per second (m/s). Higher velocities indicate more severe narrowing.
- Mean pressure gradient quantifies the average pressure difference between the left ventricle and the aorta during systole. Values ≥ 40 mmHg typically denote severe AVS.
- Aortic valve area (AVA) represents the effective opening area of the valve, calculated using the continuity equation. An AVA ≤ 1 cm^2 confirms severe stenosis.

1.3.2.2 Electrocardiogram-gated Computed Tomography (ECG-CT) and Calcium Scoring

When echocardiographic findings for aortic valve stenosis (AVS) are discordant, other imaging modalities are essential for accurate assessment. Stress echocardiography (SE) can help differentiate between true and pseudo-severe AVS, particularly in low-flow, low-gradient cases. In low-flow, low-gradient, the flow across the valve is reduced due to poor left ventricular function, despite the presence of severe stenosis. This condition is characterized by reduced stroke volume and a mean pressure gradient below 40 mmHg, despite a small aortic valve area (AVA) ($\leq 1.0, \text{cm}^2$). This paradox makes it challenging to assess severity using conventional echocardiography. Stress Echocardiography (SE) helps by evaluating the heart's contractile reserve and guiding the need for intervention. Additionally, computed tomography (CT) provides calcium scoring, while cardiac magnetic resonance imaging (CMR) offers further insights into left ventricular function and myocardial fibrosis, aiding in accurate diagnosis and risk stratification [54, 53].

Calcium scoring is a key tool for assessing aortic valve stenosis (AVS) and is endorsed by the European Society of Cardiology as a reliable predictor of severe AVS. The Agatston method is commonly used for scoring, focusing on the non-contrast, ECG-synchronized CT images of the aortic valve region. Severe AVS is indicated by thresholds of 1300 Agatston units for women and 2000 for men. In addition to the Agatston method, alternative techniques like volume scoring and mass scoring are being explored for potentially more precise assessments of valve calcification [51, 55, 56].

1.3.2.3 Transcatheter Aortic Valve Implantation (TAVI)

Aortic valve replacement (AVR) is the only effective treatment for severe aortic valve stenosis (AVS), with the timing of intervention based on both disease severity and symptom presence [57]. There are two primary AVR approaches: Surgical Aortic Valve Replacement (SAVR) and Transcatheter Aortic Valve Implantation (TAVI). While SAVR has historically been the standard treatment for severe AVS, TAVI has emerged as a less invasive alternative, particularly suited for patients at high surgical risk or those deemed inoperable due to other health factors [51, 58].

TAVI replaces the diseased aortic valve with a prosthetic valve delivered via a catheter, thereby avoiding the need for open-heart surgery [59]. The transfemoral approach is the most commonly used access route due to its minimally invasive nature, though alternative pathways—such as transapical, subclavian, transaortic, and transcarotid—may be selected based on patient-specific anatomy and comorbidities.

Prior to performing TAVI, patients undergo a comprehensive preoperative assessment to ensure they are suitable candidates and to guide procedural planning. This evaluation typically includes [53, 60]:

- Echocardiography: Used to confirm AVS severity, evaluate ventricular function, and detect any additional valvular abnormalities.
- Computed Tomography (CT): Provides detailed imaging of the aortic valve, aortic root, coronary arteries, and peripheral vessels, which aids in selecting the appropriate prosthesis size and refining procedural planning.
- Computed Tomography Angiography (CTA): Assesses the anatomy of the aorta, coronary arteries, and peripheral vessels to determine the optimal catheter access route.

Although TAVI is less invasive than surgical aortic valve replacement (SAVR), it still carries risks and potential complications, such as stroke, bleeding, paravalvular leakage, vascular injury, and heart block that may require pacemaker implantation. TAVI continues to evolve with the development of new devices, techniques, and applications, with future research focusing on reducing stroke incidence, enhancing valve durability, and expanding TAVI indications to include low-risk AVS patients [60].

Additionally, while the calcium score and its measurement through ECG-gated CT provide valuable information, they do not capture the entire aorta, excluding views of the aortic arch, a high-risk region due to its curved morphology, which promotes plaque accumulation, and descending aorta [61, 62]. To address this limitation in decision-making and procedural planning, complete aortic imaging is achieved through CT angiography (CTA), which covers the full thoracic aorta. An important objective of this thesis is to explore the feasibility of obtaining regional calcium scores directly from these CTA scans, thus enabling targeted analysis of each section of the thoracic aorta without additional imaging tests for the patient.

1.4 Artificial Intelligence in Medical Image Processing

Artificial intelligence (AI) has revolutionized medical imaging, particularly with the advent of deep learning and neural networks. Over the past decade, AI has been widely adopted for

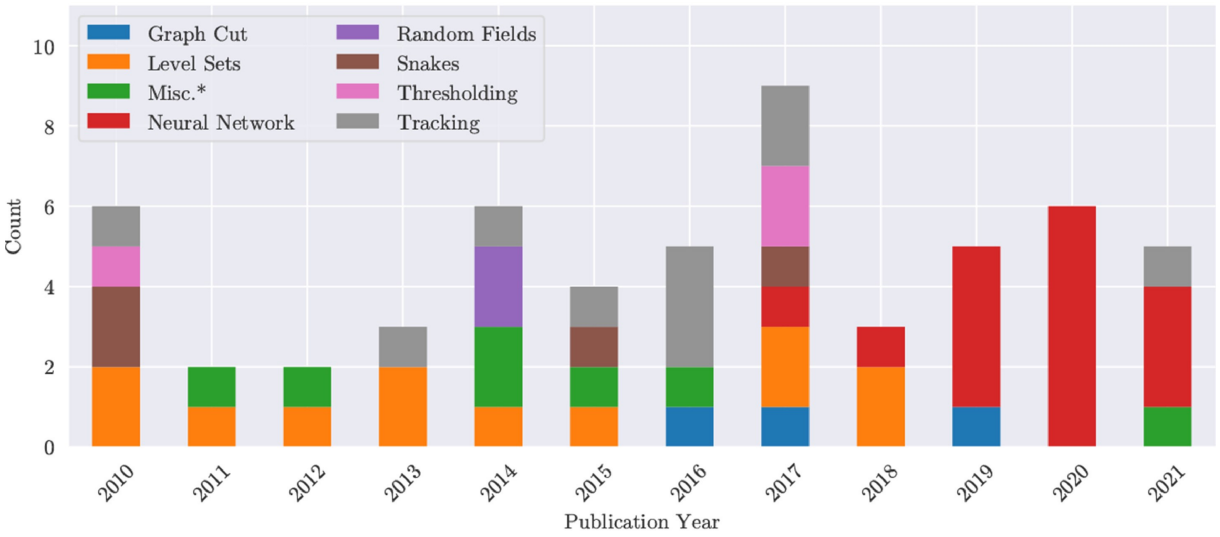


Figure 1.9: Number of published papers on artificial intelligence methods for coronary artery segmentation from 2014 to 2021. The y-axis represents the number of publications per year, while the x-axis denotes the year of publication. The figure highlights the increasing research interest in neural network segmentation techniques in the last years. Figure from [7].

tasks such as image segmentation, classification, and artifact reduction, significantly enhancing the precision and efficiency of image interpretation [11].

The role of AI is particularly pronounced in the context of personalized medicine, where tailored approaches to diagnosis and therapy rely on detailed, patient-specific insights derived from imaging data. AI-driven solutions provide unparalleled accuracy, enabling clinicians to make decisions that account for the unique anatomical and physiological characteristics of each patient.

AI models for image processing fall into two main categories: unsupervised and supervised learning, each with unique advantages and challenges depending on the dataset and objectives.

In unsupervised learning, models learn patterns and structures within the data without relying on pre-existing labels. Since no annotated dataset is required, unsupervised methods can be applied even when labeled medical data is unavailable. In this project, clustering algorithms—a type of unsupervised learning—are used for coronary artery segmentation, capitalizing on the natural brightness pattern of vessels. In CTA images, coronary arteries tend to appear brighter at the center, gradually fading toward the edges. This gradient in intensity enables clustering algorithms to distinguish the vessel boundaries effectively without requiring predefined labels.

Supervised learning, on the other hand, relies on labeled datasets where each image or structure is annotated. In medical imaging, acquiring high-quality annotated datasets is challenging, as they are often limited due to privacy concerns or require significant expert involvement. Additionally, if an annotated dataset is found, it may not have the desired resolution, and in such cases, the segmentation labels may no longer be valid. This misalignment between resolutions can render the dataset incompatible and necessitate re-annotation.

Neural networks, particularly convolutional neural networks (CNNs), are a cornerstone

of supervised learning, especially for image analysis tasks. Inspired by the structure of biological neurons, neural networks consist of interconnected nodes, or "neurons", that process information collectively. CNNs take advantage of this architecture to capture intricate spatial patterns, making them especially effective for precise image analysis.

In cardiac imaging research, the use of CNNs like U-Net, has expanded rapidly. Figure 1.9 illustrates the number of published articles (y-axis) per year (x-axis) for each topic indicated in the legend. The trend shows a significant increase in research activity over time, with a notable shift in methodology. Since 2019, the majority of studies in this field have adopted deep learning techniques [7], reflecting the growing impact of artificial intelligence, and, in particular, in medical imaging and analysis.

This growth is fueled by advancements in computational resources and the remarkable accuracy of these networks post-training. However, as imaging technologies such as CT scanners evolve to offer higher resolution, the corresponding demands for computational memory and processing power have also risen, presenting new challenges alongside opportunities [7].

We implemented various types of convolutional neural networks (CNNs) and other advanced architectures to address key challenges in medical imaging, including image segmentation and metal artifact reduction. However, these techniques have notable limitations, such as the dependency on high-quality datasets and precise annotations, which are often unavailable or lack sufficient accuracy for clinical use. Manual annotation is laborious and error-prone. Additionally, the significant computational requirements of training complex models necessitate careful trade-offs to balance accuracy, resource efficiency, and clinical practicality.

We navigated these challenges by prioritizing a balanced approach that integrates accuracy, computational efficiency, and clinical applicability. Through careful consideration of resource constraints and the practical demands of real-world workflows, we optimized our methods to deliver reliable and efficient solutions tailored to the complexities of medical imaging.

1.5 Objectives

After establishing the theoretical and clinical framework of this work, the importance and necessity of fast, automatic, and accurate medical image interpretation have become evident. This need is even more critical when non-invasive methods can replace high-risk invasive procedures, improving patient safety and clinical decision-making.

Furthermore, this project is part of an industrial initiative within a recently founded company with a well-defined primary objective: the development of an automated FFR_{CT} computation system, which requires automatic segmentation of coronary arteries. To achieve this goal, the project is divided into the following sub-objectives:

- Image Quality Enhancement: The accuracy of coronary artery segmentation in CT images is highly dependent on image quality and the presence of artifacts. Therefore, the first objective is to study the characteristics of CT images and develop methods to enhance their quality. These improvements are not only crucial for better segmentation but also for providing clinicians with clearer, more detailed images, facilitating better visualization and more accurate diagnoses.

- Development of a High-Quality Annotated Datasets: Creating a reliable and well-annotated dataset is fundamental for training and evaluating different automatic segmentation methodologies. Given that annotated medical datasets are scarce and challenging to obtain, this work aims to build a robust dataset to serve as a reference for future research in the field.
- Influence of Data on AI-Based Segmentation Performance: Investigating the impact of data quantity and quality on the performance of deep learning-based segmentation models. This analysis helps determine the minimum amount of data required for generalization and assesses the trade-offs between dataset size, model complexity, and performance.
- Software Integration for Full Automation: Developing an integrated software solution that combines the different segmentation methodologies, allowing not only segmentation but also visualization, editing, and seamless integration with CFD (Computational Fluid Dynamics) simulation software. This will enable a fully automated pipeline for extracting clinically relevant parameters, such as FFR_{CT}, directly from CT images, reducing manual workload and increasing efficiency.
- Expansion to Other Medical Imaging Applications: Extending the methodologies developed in this work to other types of medical imaging and clinical problems, paving the way for new research directions and potential applications beyond coronary artery analysis. This includes calcium segmentation, aortic analysis, and other cardiovascular applications, ultimately contributing to better diagnosis and patient management.

Chapter 2

Methodology

Part of the text and figures for this chapter are reproduced from the following publications

(I) Serrano-Antón, B.[1,2,3], Otero-Cacho, A.[1,2,3], López-Otero, D.[4,5], Díaz-Fernández, B.[4,5], Bastos-Fernández, M.[4,5], Pérez-Muñuzuri, V.[3,6], González-Juanatey, J.R.[4,5], P. Muñuzuri, A.[2,3]. Coronary Artery Segmentation Based on Transfer Learning and UNet Architecture on Computed Tomography Coronary Angiography Images. In IEEE Access, vol. 11, pp. 75484-75496, 2023. IEEE. DOI: www.doi.org/10.1109/ACCESS.2023.3293090.

(II) Serrano-Antón, B.[1,2,3], Insúa Villa, M.[1], Pendón Minguillón, S.[1], Paramés-Estévez, S.[1,2,3], Otero-Cacho, A.[1,2,3], López-Otero, D.[7,5], Díaz-Fernández, B.[4,5], Bastos-Fernández, M.[4,5], González-Juanatey, J.R.[4,5,8], P. Muñuzuri, A.[2,3]. Unsupervised clustering based coronary artery segmentation. In BioData Mining, vol. 18, no. 1, pp. 1–23, 2025. BioMed Central. DOI: https://doi.org/10.1186/s13040-025-00435-y.

[1] FlowReserve Labs S. L., Santiago de Compostela, 15782, Galicia, Spain.
[2] Galician Center for Mathematical Research and Technology (CITMAga), Santiago de Compostela, 15782, Galicia, Spain.
[3] Group of Nonlinear Physics, University of Santiago de Compostela, Santiago de Compostela, 15782, Galicia, Spain.
[4] Cardiology and Intensive Cardiac Care Department, University Hospital of Santiago de Compostela, Santiago de Compostela, 15706, Galicia, Spain.
[5] Centro de Investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), Madrid, 28029, Madrid, Spain.
[6] Institute CRETUS, Group of Nonlinear Physics, University of Santiago de Compostela, Santiago de Compostela, 15705, Galicia, Spain
[7] Cardiology and Intensive Care Department, University Hospital of Pontevedra, Pontevedra, Galicia 36161, Spain
[8] Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Santiago de Compostela, 15706, Galicia, Spain.

In this chapter, we present the key medical imaging modalities utilized in this study and the foundations of the deep learning architectures implemented for image segmentation and artifact reduction. We begin by exposing the ethical statement we followed in the study, in Section 2.1, as ethical considerations are fundamental when working with medical images. Ensuring patient privacy, data security, and compliance with legal frameworks is critical to maintaining the integrity and reliability of the research while protecting sensitive medical information.

Then, we detail the different CT imaging techniques employed, emphasizing their clinical relevance and role in generating high-quality datasets for training machine learning models (Section 2.2).

Following this, we introduce the deep learning architectures used, ranging from convolutional neural networks (CNNs) and U-Net to more recent and advanced models like Transformers and Mamba (Section 2.3). Each of these architectures offers unique advantages in handling medical imaging data, particularly for segmentation tasks, and their implementation is tailored to address the challenges inherent to medical image analysis. In addition, we discuss the primary loss functions and evaluation metrics used to train and validate these models, ensuring reliable performance assessment.

Finally, we outline the main workflow pipeline (Section 2.4), describing the steps from data preprocessing to model training, optimization, and inference. This structured approach provides a comprehensive methodology for tackling the complexities of medical image analysis.

2.1 Ethics Statement

In this study, we are committed to complying with all relevant ethical and legal guidelines concerning the use of medical imaging data. The handling of sensitive patient information carries significant ethical responsibilities, as its misuse could lead to privacy violations and ethical concerns. To mitigate these risks, all data used in this research have been fully anonymized, and informed consent has been obtained from all patients, regardless of the imaging modality.

The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki (1964) and its subsequent ratifications (Tokyo 1975, Venice 1983, Hong Kong 1989, Somerset West 1996, Scotland 2000, Seoul 2008, and Fortaleza 2013). It also adheres to Royal Decree 1090/2015 (December 24) on clinical trials, particularly Article 38 on good clinical practices, and the Convention on Human Rights and Biomedicine (Oviedo, April 4, 1997, and its subsequent updates).

To ensure patient anonymity, all collected clinical data were separated from personally identifiable information in compliance with the Personal Data Protection Law (Organic Law 15/1999, December 13) and its regulatory developments (RD 1720/2007, December 21). Additionally, we followed the principles outlined in Law 41/2002 (November 14), which regulates patient autonomy, rights, and obligations concerning medical information and documentation, as well as Law 3/2001 (May 28) and Law 3/2005 (March 7), which regulate informed consent and patient medical records. Furthermore, Decree 29/2009 (February 5) governing access to electronic medical records was strictly followed.

Patient data were collected in a Case Report Form (CRF), ensuring encryption and anonymization. Only the research team and authorized health authorities, bound by confidentiality obligations, had access to these data. Any data shared with third parties were fully anonymized to prevent identification. Once the study was completed, all patient data were securely destroyed.

Data processing, communication, and transfer adhered to the General Data Protection Regulation (GDPR), Regulation (EU) 2016/679 of the European Parliament and the Council (April 27, 2016). The data collected were used exclusively for research purposes outlined in this study and retained only for the necessary duration, in compliance with applicable legal requirements.

As this is a retrospective study based on archived medical records and imaging data, which does not alter standard clinical practice, the Ethics Committee determined that patient informed consent and full anonymization were sufficient to conduct this research.

2.2 CT Imaging Modalities

This section describes the various CT imaging modalities utilized in this study, each selected based on specific clinical and research objectives. These techniques differ in resolution, contrast usage, and target structures, primarily focusing on cardiovascular visualization, including the aorta, coronary arteries, and atherosclerotic plaque assessment.

Section 2.2.1 details the use of coronary computed tomography angiography (CCTA) for aorta and coronary artery segmentation, along with image quality enhancement techniques and segmentation processes. These methods facilitate the creation of annotated datasets for supervised and unsupervised learning algorithms discussed in Chapter 4.

Additionally, Section 2.2.2 addresses non-contrast CT for calcium scoring and CTA used for pre-TAVI planning. In both cases, the primary goal is calcium segmentation and the generation of annotated datasets for extracting key cardiological parameters in the thoracic aorta, ultimately enabling automated analysis (further elaborated in Chapter 5). Representative images accompany each modality to illustrate their applications.

2.2.1 Coronary CT Angiography (CCTA)

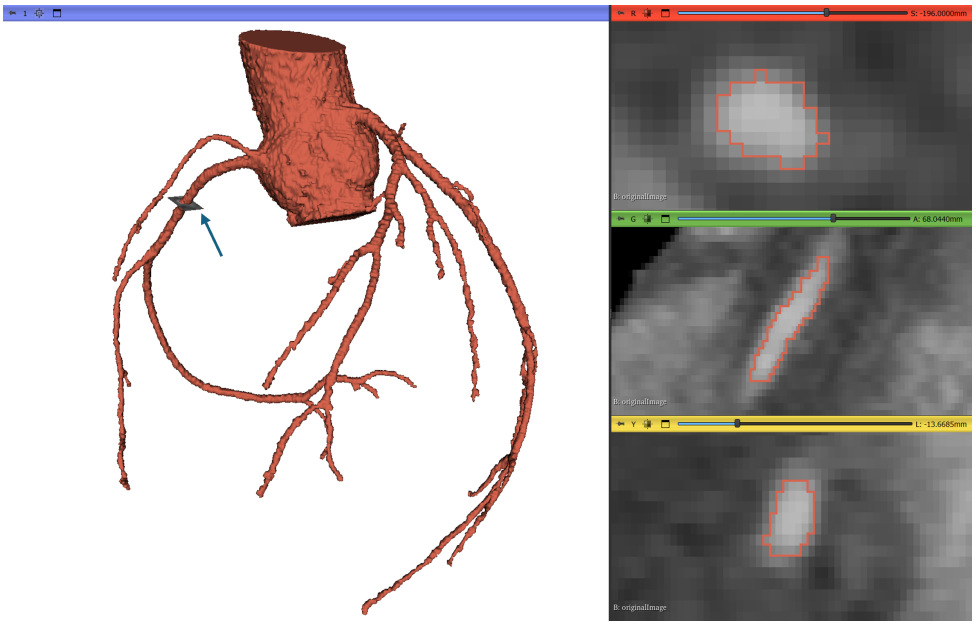
Coronary CT Angiography (CCTA), introduced in Section 1.2.3 and Section 1.3.1, has emerged as a front-line tool for detecting coronary artery disease (CAD) in recent years. This non-invasive imaging technique uses X-rays and contrast material to produce detailed images of the coronary arteries [63]. Since its clinical adoption in the early 2000s, CCTA has undergone significant advancements in both hardware and software technologies [48].

CCTA images are analyzed to assess stenosis, atherosclerotic plaque composition, vascular remodeling, and other relevant characteristics. These analyses can be qualitative or quantitative, enabling measurements such as the percentage of stenosis or calcium volume [48].

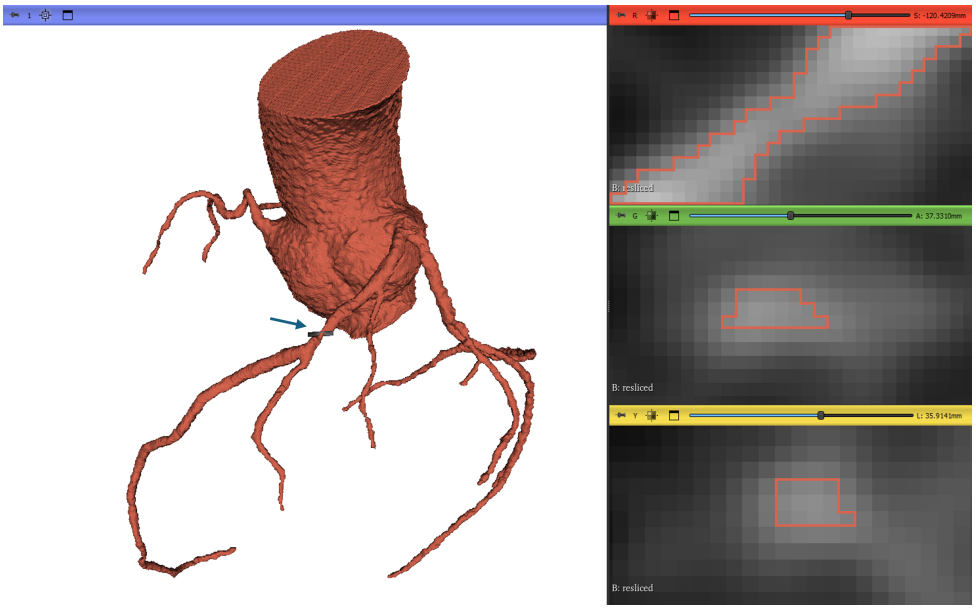
Recent advancements include improvements in spatial and temporal resolution. The latest generation CT scanners produce sharper images and reduce motion artifacts, enhancing diagnostic accuracy [63].

In this study, the resolution of the CCTA images is approximately $0.38 \times 0.38 \times 0.625\,\text{mm}^3$, where the x and y dimension can vary up to $0.293 - 0.508$ mm, with a standard deviation of 0.0389. Each image consists of $512 \times 512 \times 256$ pixels, providing detailed insights into vascular structures and abnormalities. These images were acquired using the Revolution CT by GE Healthcare, introduced in 2017 at the University Hospital in Santiago de Compostela. This scanner utilizes a 256-slice detector array with a 160 mm axial coverage per rotation, enabling a complete heart image without interruptions. This capability is crucial for reducing motion artifacts and improving diagnostic accuracy.

From these images, two datasets were generated for further analysis.



(a)



(b)

Figure 2.1: Visualization of 3D geometries in 3D Slicer [64]. The blue arrow indicates the zoomed region shown on the right, corresponding to axial, coronal, and sagittal planes (red, green, and yellow planes). (*a*) Geometry from the dataset with original image spacing ($0.45 \times 0.45 \times 0.625$ mm). The arrow points to the axial plane (red) on the geometry. (b) Geometry from the resampled dataset (isotropic voxel size 0.25 mm). The arrow highlights the location of a lesion, with the sagittal plane (yellow) intersecting the geometry. Medical images provided by FlowReserve Labs S. L.

2.2.1.1 CCTA Dataset Generation. Manual Segmentation

In this study, CCTA images were processed to segment the aorta and coronary arteries through a systematic approach conducted manually using 3D Slicer v.5.6.2 [64]. The workflow consist of the following steps [65]:

Step 1: Pre-Segmentation Calibration The segmentation process began by calibrating the brightness window in the analysis software. This involved measuring the average brightness of blood in the region between the aorta and the ostium, where the coronary arteries originate, in Hounsfield Units (HU) [29] (see Section 1.2.1). This value was then used to optimize the visibility of coronary arteries by adjusting the density range highlighted in the images. The brightness window represents the range of HU values displayed in the image, and proper calibration ensures that key anatomical details are accurately visualized.

Step 2: Vessel Segmentation After calibration, the expert manually traced the coronary arteries to segment them:

- The minimum brightness at vessel edges was used to accurately define the vessel's boundaries.
- The maximum brightness was set using the proximal aorta as a reference for blood contrast enhancement.

Calcified atherosclerotic plaques, with HU values typically ranging from 600 to 1200 HU, were carefully distinguished from blood (200–600 HU) [66, 67, 68, 69]. This distinction ensured accurate delineation, especially in vessels with calcifications or irregular structures.

Step 3: Adjustments for Patient-Specific Variability Brightness and density thresholds were tailored for each patient to account for individual differences in tissue attenuation and blood density. This patient-specific approach ensured that the segmentation was aligned with the unique anatomical and pathological features of each case.

Patients with stents were excluded from the dataset due to the significant imaging limitations caused by metal artifacts. The presence of metallic stents generates beam hardening and blooming artifacts in CT scans, obscuring the vessel lumen and making it difficult to accurately assess vascular structures (see Section 1.2.2). Since the primary goal of this study involves precise segmentation and analysis of vascular regions, including calcium quantification and lumen visualization, the exclusion of these cases ensures higher data quality and prevents inaccuracies in model training and evaluation.

Once the segmentation technique was established, two datasets were created, consisting of the CCTA images of patients and their respective segmentation masks of the aorta and/or coronary arteries. The segmentation masks are binary representations of the original images, where the background is labeled as 0 and the vessel as 1.

1. The first dataset consists of 88 patients, with segmentation performed while preserving the original resolution. These patients were selected based on image clarity and the absence of calcium-related or stenosis lesions, diagnosed by clinicians. In rare cases (fewer than 5% of patients) where minimal calcification was present, the calcium was removed (not segmented as lumen) from the vessel while maintaining the blood flow region.
2. The original mean resolution of 0.38 mm is insufficient for creating geometries smooth enough for fluid simulations or for accurately segmenting stenoses with a minimum

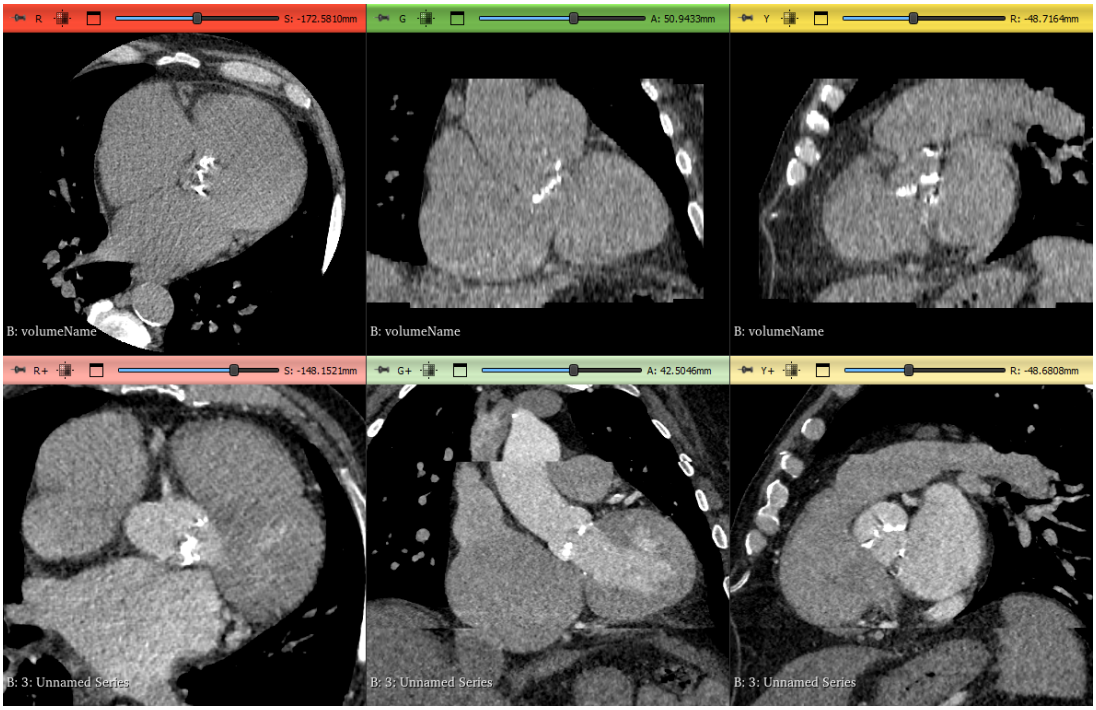


Figure 2.2: Multiplanar visualization of CT scans for calcium scoring and TAVI planning. Axial (red), coronal (green), and sagittal (yellow) planes are shown for two types of scans. The first row depicts CT-AVC scans highlighting aortic valve calcifications where voxel size is $0.48 \times 0.48 \times 2.5\,\text{mm}^3$. The second row presents pre-TAVI CTA scans, which exhibit artifacts such as displacement and contrast variations caused by the need for multiple acquisitions, most noticeable in the coronal plane (green). Voxel size is $0.714 \times 0.714 \times 0.625\,\text{mm}^3$ Images provided by FlowReserve Labs S. L.

caliber of 0.5 mm. To address this, a resampling process was applied to create a second dataset. Using the *Resample Scalar Volume* module in 3D Slicer [70], all original images were resampled to isotropic 0.25 mm spacing with bSpline interpolation. This dataset consists of 32 patients, including 10 from the first dataset and an additional 22 with clinically diagnosed lesions.

Figure 2.1 illustrates an example of the 3D geometries generated for both datasets using 3D Slicer. In Figure 2.1a, a segmentation from the first dataset is shown, capturing various bifurcations and distal vessels. However, due to the original pixel size, the geometry appears irregular with sharp edges. In contrast, Figure 2.1b presents a refined and anatomically accurate reconstruction from the second dataset, where vessels appear smoother due to segmentation from an isotropic 0.25 mm volume. The blue arrow highlights a lesion, visible in the axial plane (red), indicating a significant stenosis.

2.2.2 Computed Tomography Calcium Scoring (CT-AVC)

Aortic valve stenosis (AVS) is a progressive condition characterized by calcification and fibrocalcific remodeling of the aortic valve leaflets, leading to restricted blood flow (see Section 1.3.2). Severe cases require timely intervention, usually aortic valve replacement (AVR),

to prevent complications. While standard echocardiographic measures such as Vmax, mean gradient, and aortic valve area (AVA) are commonly used for diagnosis, discrepancies arise in approximately 25% of cases, complicating clinical decision-making [50]. To improve diagnostic accuracy, CT aortic valve calcium scoring (CT-AVC) has been introduced as a complementary tool, utilizing calcification quantification to assess AVS severity [71].

CT-AVC analysis typically employs imaging protocols with a slice thickness of 2.5 or 3 mm, 120 kVp, and variable mAs tailored to the patient's body. In our study, the images were acquired at a resolution of $0.48 \times 0.48 \times 2.5\,\text{mm}^3$, ensuring sufficient detail for calcium scoring while aligning with standard practices for reproducibility. This resolution facilitates visualization and quantification of calcium deposits in the valve leaflets, using similar methodologies to coronary artery calcium scoring (CAC).

An example of CT-AVC imaging is shown in the first row of Figure 2.2, illustrating the axial, coronal, and sagittal planes (red, green, and yellow), where the calcification in the aortic valve leaflets is clearly visible.

2.2.3 Pre-TAVI CTA

Computed Tomography Angiography (CTA) is critical for planning transcatheter aortic valve implantation (TAVI) (more information can be found in Section 1.3.2). The primary goal of these scans is to assess vascular access routes, determining the feasibility of a transfemoral approach or the necessity of alternative access.

Additionally, CTA provides a complete view of the aorta, enabling the identification and analysis of calcifications across different regions, which is central to our study.

Acquired at $0.714 \times 0.714 \times 0.625\,\text{mm}^3$ resolution, where the x and y dimensions can range from 0.594 mm to 0.914 mm, with a standard deviation of 0.0938 mm. These images offer high detail but pose challenges such as motion artifacts and contrast variations. These artifacts arise because the extensive anatomical area covered in the scan requires multiple acquisitions, leading to potential misalignments and inconsistencies between sections. For example, in the second row of Figure 2.2 illustrates a coronal plane (green), showing displacement of the aorta and brightness fluctuations between frames. These challenges significantly increase the complexity of accurate analysis and segmentation in our study (detailed in Chapter 5).

2.3 AI Methodologies

In this section, we introduce the fundamental neural network architectures used in this study, all focused on medical image analysis. These models follow a structured evolution, starting with Convolutional Neural Networks (CNNs) and the U-Net (Section 2.3.1), a widely used network for biomedical image segmentation. Variants such as VGG, ResNet, MobileNet, and U-Net++ build on this U-shaped structure, enhancing feature extraction and efficiency.

Beyond CNNs, Transformers introduced the self-attention mechanism, allowing for a global receptive field, which significantly improved segmentation performance (Section 2.3.2). However, attention mechanisms are computationally expensive, making them challenging to scale. To address this, Vision Transformers (ViTs) and Swin Transformers were introduced, offering more efficient hierarchical and window-based attention mechanisms, making them more suitable for high-resolution medical images.

More recently, Mamba has emerged as a novel alternative (Section 2.3.3), replacing both convolutions and attention mechanisms with state-space models (SSMs), which improve computational efficiency while maintaining high segmentation accuracy. The transition from CNNs to Transformers and Mamba reflects the ongoing pursuit of scalable, precise, and efficient deep-learning solutions in medical imaging.

When working with medical imaging, our initial approach focuses on 2D architectures, as they provide a fundamental and computationally efficient way to process and analyze images. However, given that medical imaging data is inherently volumetric, we also explore 2.5D and 3D approaches. These methods allow for a more comprehensive understanding of spatial relationships by incorporating depth information, which is particularly useful in segmenting complex anatomical structures. Section 2.3.4 discusses the transition from 2D to 3D models, highlighting their advantages and challenges, such as increased computational demands and data requirements. Additionally, we address transfer learning, a technique that allows the use of pre-trained models to improve performance specially when data is scarce.

A crucial component of neural network training is the loss function (Section 2.3.5), which guides the learning process by quantifying the error between the predicted segmentation and the ground truth. By iteratively adjusting network weights to minimize this error, the model improves its performance over time. In this work, we employ several loss functions tailored for medical image segmentation, such as Dice Similarity Coefficient (DSC) loss and cross-entropy loss, both of which address challenges like class imbalance and fine boundary delineation.

To assess the effectiveness of our segmentation models, we use evaluation metrics that objectively measure performance (Section 2.3.6). These metrics compare the predicted segmentation with the ground truth, providing insight into the accuracy and reliability of the model. Commonly used metrics in this study include the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU), which evaluate spatial overlap, as well as precision and recall to assess segmentation quality in different clinical scenarios.

2.3.1 Convolutional Neural Networks (CNNs) Architectures

The development of neural networks, particularly Convolutional Neural Networks (CNNs), has significantly advanced AI's capabilities in image processing. Before CNNs, the standard approach to training neural networks for image processing involved flattening images into lists of pixels and passing them through a feedforward neural network. However, this method discarded essential spatial information, severely limiting the network's ability to capture spatial relationships within the image.

CNNs, first introduced around 1989 by Yann LeCun et al. [72], revolutionized this process by preserving the two-dimensional structure of images. Unlike traditional feedforward networks, which treat each input as independent data points, CNNs process spatial information, capturing hierarchical features across the entire image while maintaining computational efficiency. This efficiency arises from the network's architecture, which utilizes shared weights and biases across spatial regions, thereby reducing the number of parameters and minimizing computational costs [73]. Consequently, CNNs have become a cornerstone in medical imaging applications, enabling precise analysis of complex image data while preserving the critical spatial context necessary for accurate diagnosis.

CNNs consist of several key operations that allow them to extract features from images efficiently:

- Convolution: The convolution operation is the foundational layer of a CNN. It involves a small matrix called a filter (or kernel), typically of size 3×3, 5×5, which slides (or convolves) across the input image. The stride, or the step size of the filter, determines how much the filter shifts at each step (e.g., a stride of 1 moves the filter by one pixel, whereas a stride of 2 moves it by two pixels, reducing output size). Each position of the filter multiplies and sums the input pixel values by the corresponding filter values, creating an output called a feature map.

  - Padding is an additional parameter applied to control output size. By adding extra pixels around the image border (often with zeros), *same padding* maintains the input size, while *valid padding* does not add borders, reducing the output size as the filter moves within the original boundaries.
  - After the convolution operation, an activation function—most commonly ReLU (Rectified Linear Unit)—is applied to introduce non-linearity, which enables the model to capture complex features. ReLU converts negative values to zero, increasing the network's ability to learn non-linear patterns while improving computational efficiency.

Following the example in Figure 2.3, where we have an input image (I) of size 4×3 (width $\times$ height) and a kernel (K) of size $m = 2 \times n = 2$, we calculate the output feature map size using Equation 2.1 and Equation 2.2:

$$\text{Output Width} = \frac{\text{Input Width} - \text{Kernel Width} + 2 * \text{Padding Width}}{\text{Stride}} + 1 \tag{2.1}$$

$$\text{Output Height} = \frac{\text{Input Height} - \text{Kernel Height} + 2 * \text{Padding Height}}{\text{Stride}} + 1 \tag{2.2}$$

Thus, the resulting feature map (S) has dimensions 3×2, since we assume stride $= 1$ and no padding. Each element of the feature map, S_{ij}, is calculated by placing the kernel (K) at position (i, j) in the input image (I) and taking the weighted sum of the overlapping elements, as described by Equation 2.3.

$$S_{i,j} = \sum_{a=0}^{m-1} \sum_{b=0}^{n-1} I_{i+a,j+b} K_{a,b}, \tag{2.3}$$

where $i \in [0,3]$ and $j \in [0,2]$, covering the full extent of the input image to compute each element in the 3×2 feature map.

However, if we want to preserve the same input size after the convolution, one way to achieve this is by changing the kernel size to 3×3 and adding padding. With a 3×3 kernel and padding of 1 (see the padded image, P_z, in Figure 2.3), the dimensions of the output feature map will match the input dimensions. This can be proven by using the general formula for output dimensions, Equation 2.1 and Equation 2.2:

$$\text{Output Width} = \frac{3 - 3 + 2 \cdot 1}{1} + 1 = 3$$

$$\text{Output Height} = \frac{4-3+2\cdot 1}{1} + 1 = 4$$

- Pooling: Pooling layers reduce the spatial dimensions of the feature maps while preserving the most essential information. This downsampling helps reduce the computational load, controls overfitting, and provides some spatial invariance. The two common types of pooling are:

  – Max Pooling: This operation selects the maximum value within a defined region, such as 2×2 or 3×3 pixels, with a stride of 2, which reduces the feature map size by half.

  – Average Pooling: Instead of selecting the maximum, average pooling takes the average of the pixel values within the filter region, providing a smoother, less detailed downsampled feature map.

  Each pooling layer reduces the image dimensions, which decreases memory usage and computational demand, allowing the network to focus on the most prominent features across the image.

  For example, applying a Max Pooling operation to the input image (I) in Figure 2.3, the output dimensions are again computed with Equation 2.1 and Equation 2.2 (same as with the convolution operation). The Max Pooling output (P) is given by Equation 2.4.

$$P_{i,j} = \max_{0 \le a < m, 0 \le b < n} I_{i+a,j+b}, \tag{2.4}$$

  where $i \in [0,3]$ and $j \in [0,2]$, assuming stride $= 1$.

- Fully Connected Layers: Fully connected (FC) layers are generally used at the end of a CNN for tasks like classification or regression. In an FC layer, each neuron connects to every neuron in the previous layer, consolidating all features extracted by previous convolutional and pooling layers into a single output, such as class probabilities. However, in tasks like segmentation, where the spatial arrangement of pixels is crucial, FC layers are often omitted to preserve the image's spatial context, enabling accurate per-pixel predictions rather than global classifications.

- Activation functions play a crucial role in neural networks by filtering relevant information while suppressing less important signals. Instead of transmitting all input data indiscriminately, these functions ensure that only meaningful features are passed forward.

  The primary goal of an activation function is to transform the weighted sum of inputs into an output value that either serves as input for the next layer or as the final output of the network. This transformation must be **non-linear**, as non-linearity is what enables neural networks to model complex relationships beyond simple linear mappings.

  When the activation function is applied at the output layer, its choice depends on the specific problem being solved. Below are some commonly used activation functions [74]:

– ReLU (Rectified Linear Unit). ReLU introduces non-linearity while mitigating vanishing gradient issues. It is defined as:

$$f(x) = \max(0, x) \tag{2.5}$$

However, ReLU can suffer from the *dying ReLU* problem, where neurons output zero for all negative values, hindering learning.

– Sigmoid. The sigmoid function maps inputs to a range between 0 and 1, making it useful for binary classification. It is given by:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.6}$$

Despite its usefulness, sigmoid has limitations such as vanishing gradients and outputs that are not zero-centered, which can slow down training.

– Softmax. Softmax generalizes sigmoid for multi-class classification by converting logits into probabilities. Each output value lies between 0 and 1, and the sum of all outputs is always 1. It is defined as:

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{2.7}$$

While useful, it can be sensitive to large input values, leading to numerical instability.

During backpropagation, gradients are used to update the weights of a neural network. However, in deep networks, gradients can become extremely small as they propagate backward through multiple layers. This issue, known as the *vanishing gradient problem*, occurs when activation functions squash input values into a narrow range, making gradient updates negligible. Sigmoid and tanh activations are particularly prone to this problem since their derivatives approach zero for large positive or negative inputs. As a result, earlier layers learn very slowly or stop updating altogether. This problem can hinder deep network training, which is why ReLU and its variants are often preferred [74].

Choosing the right activation function depends on the network architecture and the specific task, balancing computational efficiency and gradient flow for optimal learning.

- Transposed Convolutions: In contrast to the convolutional and pooling layers, which typically reduce or maintain the spatial dimensions of the input, transposed convolutions are designed to upsample the spatial resolution of feature maps. This is particularly useful in tasks like semantic segmentation, where the output must have the same spatial dimensions as the input to provide pixel-level classification.

  Transposed convolutions, also referred to as fractional-stride convolutions [75], reverse the downsampling effect of standard convolutions. By sliding the kernel over the input with a specified stride and padding, transposed convolutions effectively reconstruct higher-resolution outputs. Consider an input tensor X of size $H \times W$ and a convolution

Input (I)

a	b	c	d
e	f	g	h
i	j	k	l

Kernel (K)

w	x
y	z

Input with Zero Padding (P_z)

0	0	0	0	0	0
0	a	b	c	d	0
0	e	f	g	h	0
0	i	j	k	l	0
0	0	0	0	0	0

Figure 2.3: Illustration of an input image, a zero-padded input image and a kernel for convolution operation.

kernel K of size $k \times k$ with a stride s and padding p. The spatial dimensions of the output Y are given by:

$$H_{\text{out}} = s \cdot (H-1) - 2p + k, \quad W_{\text{out}} = s \cdot (W-1) - 2p + k \tag{2.8}$$

Each element of the input is multiplied by the kernel, and the resulting patches are summed to produce the final output. Unlike regular convolutions, transposed convolutions ensure that each spatial position in the input contributes to multiple positions in the output, enabling effective upsampling.

Figure 2.4 provides a visual example of how a convolutional neural network (CNN) processes an image through hierarchical feature extraction. In Figure 2.4a, the original image highlights a calcified region in the aortic arch. As the image moves through the layers of a U-Net encoder, multiple convolutional filters extract different features, resulting in several feature maps at each level. Each feature map captures distinct information—Figure 2.4b, a high-resolution feature map, shows strong activations in the calcium region, indicating that the network detects it as a significant feature. In Figure 2.4c, the focus shifts towards edges and background structures, reflecting how different filters specialize in capturing textures and boundaries. As the network progresses to deeper layers, the spatial resolution decreases, but the representations become more abstract. In Figure 2.4d, a lower-resolution feature map highlights both the calcium deposits and the edges, suggesting a mid-level feature representation.

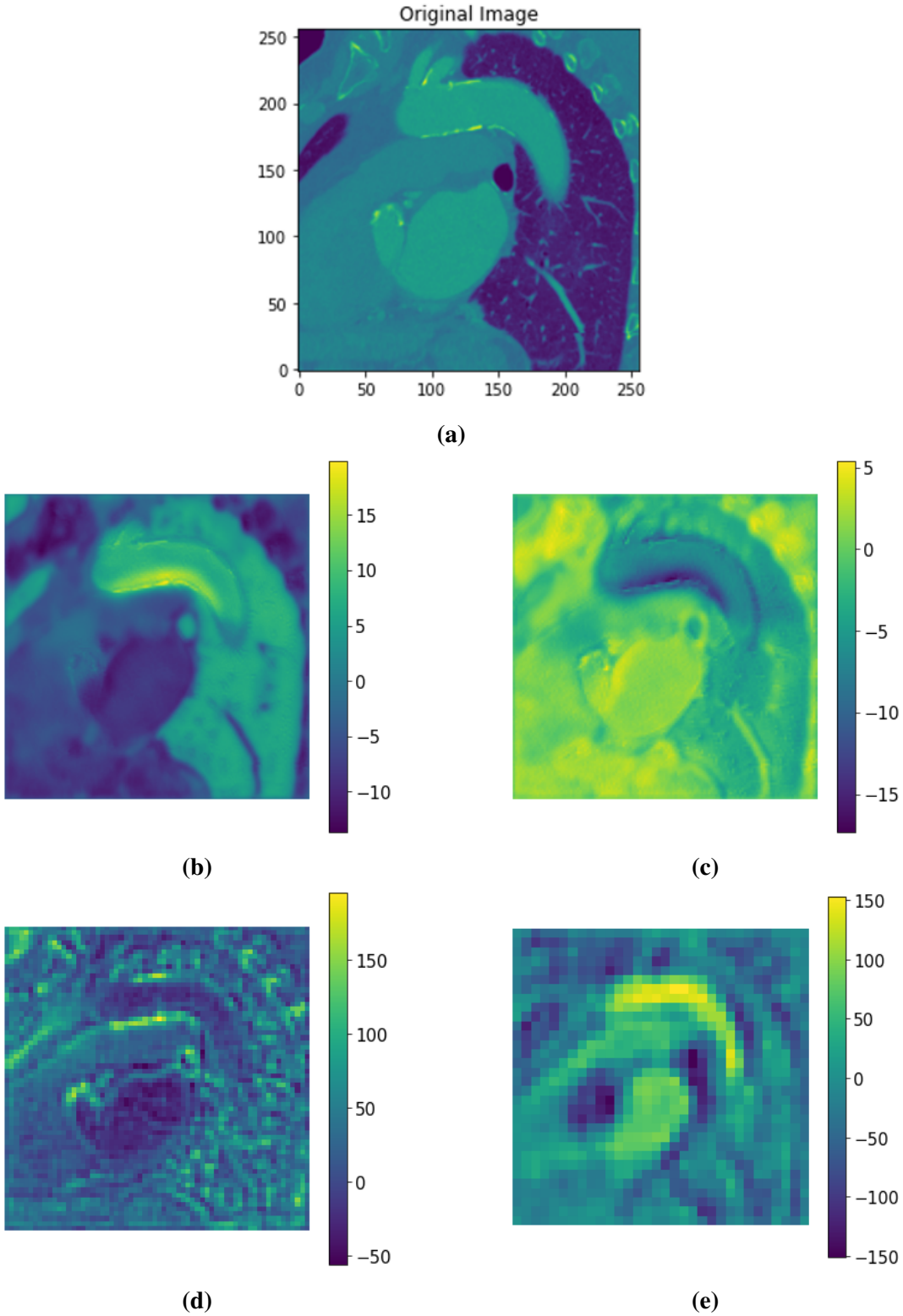


Figure 2.4: (a) Original image showing a calcified section of the aortic arch. (b–e) Feature maps extracted from different levels of a U-Net encoder, highlighting varying feature representations. (b) At higher resolution, the network focuses more on the calcified region with higher activation values. (c) Emphasizes edges and background structures. (d) At lower resolution, it captures both calcium deposits and edges. (e) The most abstract representation, primarily attending to the aortic lumen. Medical image provided by FlowReserve Labs S. L.

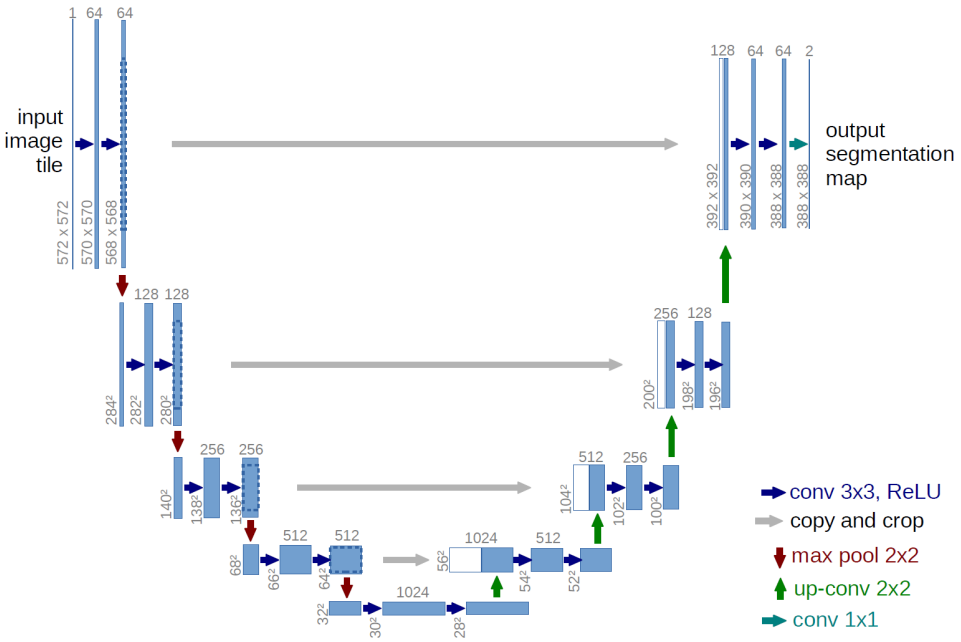


Figure 2.5: U-net architecture, shown here for a 32x32 pixel resolution at the lowest level. Each blue box represents a multi-channel feature map, with the number of channels labeled above. The x and y dimensions are noted at the bottom left of each box. White boxes indicate copied feature maps, and arrows illustrate the different operations. Figure from [76].

Finally, Figure 2.4e demonstrates the most abstract feature extraction, where activations are concentrated in the aortic lumen rather than in specific structures. This progression illustrates how CNNs learn increasingly complex and meaningful representations, allowing them to generalize from raw pixel data to higher-level patterns relevant for medical imaging tasks.

2.3.1.1 U-Net

The U-Net [76] architecture was originally designed for medical image segmentation, making it particularly suited for the automated analysis of anatomical structures in high-resolution images—a key reason for its selection in this work. Additionally, its design is especially advantageous for applications where training data is limited, as it is built upon the "fully convolutional network" (FCN) architecture proposed by Long, Shelhamer, and Darrell [77]. This allows U-net to achieve precise segmentations even with small datasets, a critical requirement in many medical imaging contexts. Numerous studies have demonstrated U-net's strong performance across a range of segmentation tasks, reinforcing its reliability and versatility. In this study, U-Net sets the baseline for the AI architectures explored, as subsequent models will build upon and adapt this core structure.

In Figure 2.5, we see the diagram of the U-Net architecture. On the left-hand side, the encoder path captures contextual information through a series of blocks composed of convolutional layers, activation functions, and max pooling operations that progressively reduce the dimensionality of the input. This process continues until we reach the bottleneck of the network. Subsequently, the upsampling process begins, mimicking the encoder architecture but utilizing upsampling layers instead of pooling.

A distinguishing feature of U-net is the absence of fully connected layers. Instead, it

uses only the valid part of each convolution (no padding), ensuring that the segmentation map contains pixels for which the full context is available in the input image. This design choice, along with its elegant U-shaped architecture comprising a contracting path to capture contextual information and a symmetric expanding path for precise localization, makes U-net highly effective in segmenting complex medical images [76].

Below, we highlight the essential components of this architecture.

- Encoding Path: The contracting path follows a typical convolutional network architecture designed to progressively capture the context of the input image. It consists of repeated application of two 3×3 convolutions without padding, each followed by a rectified linear unit (ReLU) activation function. Downsampling is achieved through a 2×2 max pooling operation with a stride of 2, effectively halving the spatial dimensions at each step. Importantly, at each downsampling stage, the number of feature channels is doubled, allowing the network to capture increasingly complex patterns. The repeated convolution and pooling operations gradually reduce the spatial dimensions while expanding the depth of the feature maps, ensuring robust feature extraction.

- Decoding Path: The expansive path mirrors the contracting path and aims to reconstruct the original resolution while preserving the learned contextual features. Each step begins with upsampling the feature map, followed by a 2×2 convolutional layer (also called transposed convolution) that halves the number of feature channels. The upsampled feature map is then concatenated with the correspondingly cropped feature map from the contracting path to preserve high-resolution details. This is followed by two 3×3 convolutions, each activated by ReLU, which refine the upsampled features and aid in precise reconstruction of the image's spatial structure. Cropping is necessary due to the reduction in feature map size from unpadded convolutions during the encoding path.

- Baseline (Bottleneck): The bottleneck layer acts as a bridge between the encoding and decoding paths. It processes the most compressed representation of the image, utilizing two 3×3 convolutions with ReLU activation to capture high-level semantic features. This layer serves as a critical transition, enabling the network to leverage global context before passing the features to the decoding path.

- Final Layer: At the end of the decoding path, a 1×1 convolutional layer is used to map the 64-component feature vector at each location to the desired number of output classes. This allows for precise pixel-level classification required for segmentation tasks.

In total, the U-net architecture comprises 23 convolutional layers. The careful balance between the contracting and expansive paths allows the network to excel in pixel-wise prediction, particularly when segmenting complex medical images, even with limited training data.

A key strength of U-net lies in its skip connections (represented by grey arrows in Figure 2.5), which connect corresponding layers in the encoding and decoding paths. By merging features from both paths, these connections help retain spatial information lost during downsampling, thus preserving both local and global context in the segmentation map. These skip connections allow U-net to capture the intricate relationships between different parts of the image, leading to highly accurate segmentation results.

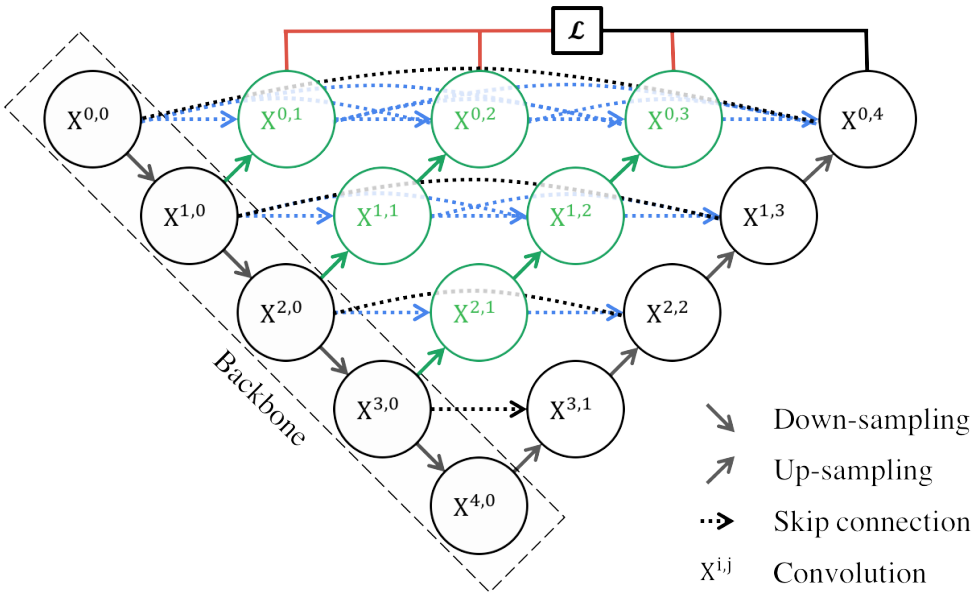


Figure 2.6: Illustration of the U-Net++ architecture. The network consists of an encoder-decoder structure connected by densely nested skip pathways. The upsampled feature maps are concatenated with the outputs of previous nodes in the skip pathway, reducing the semantic gap between the encoder and decoder. Figure from [78].

In this study, several modifications of the U-net architecture have been implemented, incorporating encoders based on architectures like U-Net++ [78], and the derived U-shaped VGG [79], ResNet [80], and Efficient-Net [81], as well as 3D and 2.5D adaptations of U-Net. These variants have been specifically tailored for the segmentation of coronary structures, and their architectural details will be further elaborated in the following sections.

2.3.1.2 U-Net++

U-Net++ [78], like its predecessor U-Net, was specifically designed for medical image segmentation, where precision is paramount. In medical applications, even a minor segmentation error can lead to an incorrect diagnosis, emphasizing the need for highly accurate models. U-Net++ addresses these challenges by introducing a more sophisticated encoder-decoder architecture that is deeply supervised and densely connected through nested skip pathways.

In Figure 2.6, we see the diagram of the U-Net++ architecture, where each convolution block is represented by a circle. The skip connections, depicted by dashed arrows, allow for a dense connection pattern between each convolution block in the encoder and all subsequent layers in the decoder. This contrasts with the original U-Net, which features direct skip connections only between corresponding layers. Additionally, the downsampling and upsampling processes are represented by solid arrows, where the encoder progressively reduces the spatial dimensions of the input through convolution and pooling operations, followed by an upsampling process in the decoder that mirrors the encoder architecture.

The redesigned skip pathways aim to reduce the semantic gap between feature maps in the encoder and decoder. This gap reduction is crucial because the closer the feature representations are in terms of semantics, the easier it becomes for the optimizer to learn accurate mappings

[78]. Unlike the original U-Net, where skip connections directly transfer feature maps from the encoder to the decoder, U-Net++ processes these feature maps through dense convolutional blocks, with the number of convolutional layers varying according to the pyramid level.

The effectiveness of these redesigned skip pathways lies in their ability to preserve fine-grained details crucial for accurate segmentation. Specifically, the computation for each node $x_{i,j}$ in the architecture is defined by Equation 2.9.

$$x_{i,j} = \begin{cases} H(x_{i-1,j}), & \text{if } j = 0 \\ H\left(\left[x_{i,k}\right]_{k=0}^{j-1}, U(x_{i+1,j-1})\right), & \text{if } j > 0 \end{cases} \tag{2.9}$$

where $H(\cdot)$ represents a convolution followed by an activation function, $U(\cdot)$ denotes an up-sampling operation, and $[\cdot]$ is the concatenation operation. The nested connections work as follows:

- Nodes at level $j = 0$ receive input only from the previous encoder layer.
- Nodes at level $j = 1$ take input from two sources: the current level of the encoder and the previous encoder level.
- Nodes at level $j > 1$ receive $j + 1$ inputs, including j outputs from previous nodes in the same skip pathway and one upsampled output from the lower-level skip pathway.

In the original paper, [78], all convolutional layers along a skip pathway $x_{i,j}$ use kernels of size 3×3, where the number of filters is defined as $k = 32 \times 2^i$. To enable deep supervision, a 1×1 convolutional layer followed by a sigmoid activation function is appended to each target node $\{x_{0,j} \mid j \in \{1,2,3,4\}\}$.

In total, this nested, dense connectivity allows U-Net++ to achieve more refined and precise segmentation results by enhancing the flow of both high-level semantic and low-level spatial information between the encoder and decoder.

2.3.1.3 VGG

Introduced in 2014 by the Visual Geometry Group (VGG), the VGG architecture [79] was originally designed for image classification, significantly advancing the field by demonstrating the impact of deeper neural networks. By increasing the network depth to 16 or 19 convolutional layers (hence VGG-16 and VGG-19), VGG achieved notable success, including top performance in the ImageNet competition. Its primary innovation lies in the use of small 3×3 convolutional filters, showing that deeper networks with smaller kernels can achieve superior accuracy compared to shallower networks with larger filters [79].

The original VGG-16 architecture consists of 16 weight layers: 13 convolutional layers followed by 3 fully connected layers. The network accepts images with a resolution of 224×224 pixels and can classify them into 1000 object categories [79, 82]. Despite its original classification purpose, VGG remains a widely used architecture due to its versatility and strong performance across various datasets.

The architecture of the original VGG-16 is illustrated in Figure 2.7, where parallelepiped shapes represent the convolutional blocks, which are color-coded in red and gray. The red parallelepiped denotes the initial convolutional layer designed to capture low-level features,

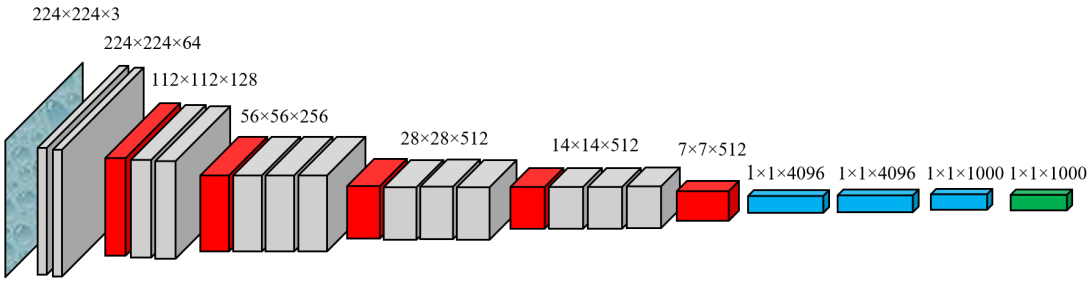


Figure 2.7: Illustration of the original VGG-16 architecture. The network consists of 13 convolutional layers using small 3×3 filters, followed by 3 fully connected layers. Each convolutional layer is activated by a ReLU function, and max-pooling layers are used for downsampling. The final fully connected layer outputs probabilities for $1,000$ object classes, as required by the ImageNet dataset. Image from Wikimedia Commons.

while the subsequent gray parallelepiped blocks represent additional convolutional layers that further extract increasingly abstract features from the input data.

Pooling layers are illustrated by the reduction in the dimensionality of these parallelepipeds, effectively downsampling the feature maps while preserving the most salient information. Following the convolutional and pooling layers, we have three fully connected layers represented in blue, which interpret the learned features. Finally, the output layer is depicted in green, where the model generates its predictions.

Below, we detail the key components that define this architecture.

- Convolutional Layers: Each convolutional layer in VGG uses small 3×3 filters with a stride of 1 and zero-padding to preserve the spatial resolution. To enhance non-linearity, each convolution is followed by a ReLU activation function. The network progressively doubles the number of feature channels, beginning with 64 and reaching up to 512 in deeper layers.

- Pooling Layers: Max pooling with a 2×2 kernel and a stride of 2 is used after each set of convolutions to downsample the feature maps by half, reducing the spatial dimensions while increasing the depth.

- Hidden Layers: All hidden layers employ the ReLU activation function to accelerate training and introduce non-linearity.

- Fully Connected Layers: In the original VGG implementation for classification, three fully connected layers follow the convolutional layers. The first two layers contain 4096 neurons each, while the final layer has 1000 neurons corresponding to the 1000 classes of ImageNet [82]. However, for segmentation purposes, these layers are omitted in favor of a decoder path.

For segmentation, the fully connected layers are replaced by a decoder network that mirrors the encoder. The decoder upsamples the feature maps using transposed convolutions to restore the original image resolution. Skip connections are added between the encoder and decoder at corresponding levels to preserve spatial information and enable the network to recover fine details lost during downsampling.

By leveraging the VGG encoder in a U-shaped architecture, we achieve precise segmentation maps that benefit from the rich feature extraction capabilities of the original VGG network. This adaptation makes VGG a powerful tool for image segmentation, particularly in domains such as medical imaging, where accurate pixel-level classification is critical.

2.3.1.4 ResNet

The ResNet architecture, introduced by He et al. in 2015 [80], addresses a fundamental challenge in deep learning: training very deep neural networks. Although it is widely accepted that deeper networks have the potential for greater representation power, their training can be severely hampered by issues such as vanishing or exploding gradients. These problems arise as gradients, during backpropagation, become too small (vanishing) or excessively large (exploding), causing either stalled learning or instability. ResNet mitigates these issues by introducing residual learning, allowing gradients to flow more effectively, even in very deep networks.

One of the main insights behind ResNet is the reformulation of layers to learn residual functions instead of direct mappings. Instead of expecting each layer to learn a new transformation from the input to the output, ResNet layers learn the difference, or residual, between the input and the desired output. Formally, if the input to a layer is x, instead of learning a function $H(x)$, the layer learns $F(x) = H(x) - x$, resulting in the final output $H(x) = F(x) + x$ [80]. This approach simplifies the optimization process, allowing the network to more easily refine the identity mapping if necessary.

The skip connections in ResNet allow gradients to bypass one or more layers during backpropagation, alleviating the vanishing gradient problem. As a result, ResNet enables the construction of very deep networks without performance degradation, which was a key limitation in earlier architectures like VGG (see Section 2.3.1.3).

To illustrate the structure of ResNet, we will focus on ResNet-34, a model with 34 convolutional layers. Each layer in the network follows a consistent pattern, illustrated in Figure 2.8. In the diagram, each convolutional layer is represented by a rectangle, where the kernel size and the number of feature maps are specified. The architecture follows two key design principles: (i) layers producing the same output feature map size maintain the same number of filters, and (ii) when the feature map size is halved, the number of filters is doubled to preserve the time complexity per layer. The network concludes with a global average pooling layer and a 1000-way fully connected layer with a softmax activation for classification. Regarding shortcut connections, solid lines are used when the input and output have the same dimensions, ensuring identity mappings. However, when the dimensions change (dotted line shortcuts in Figure 2.8), the shortcuts span feature maps of different sizes and are performed with a stride of 2 to accommodate the dimensional shift.

ResNet's architecture not only improves the ability to train deeper networks but also demonstrates superior generalization, making it highly effective across a range of computer vision tasks. In this work, ResNet will serve as a foundational encoder for image segmentation.

2.3.1.5 MobileNet

MobileNet is a neural network architecture designed for mobile and embedded vision tasks, where computational resources are limited. Introduced in 2017, MobileNet [83] addresses the need for lightweight models that can operate efficiently on devices with constrained processing

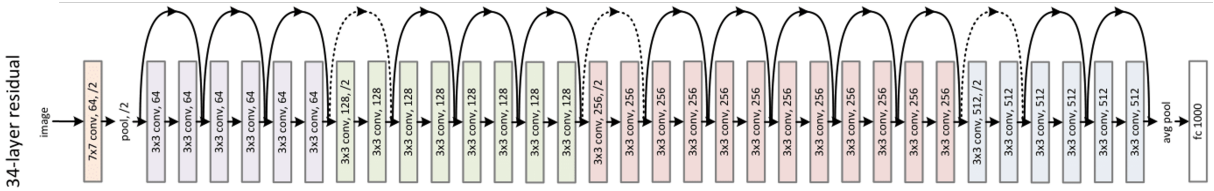


Figure 2.8: Architecture of ResNet-34. ResNet-34 consists of an initial 7×7 convolution followed by a max-pooling layer, and four stages of residual blocks with 3×3 convolutions. The residual blocks contain skip connections that facilitate gradient flow and enable the training of deep networks. Each stage increases the number of feature maps (64, 128, 256, and 512) while reducing spatial dimensions through convolutions with a stride of 2. Dotted skip connections indicate dimensional changes. Figure adapted from [80].

power and memory. Its primary focus is to optimize both latency and model size, making it ideal for real-time applications on mobile devices.

Unlike traditional architectures that prioritize accuracy at the cost of computational efficiency, MobileNet provides a balance by using an innovative design based on depthwise separable convolutions. This design enables developers to select a network configuration tailored to specific hardware constraints while maintaining competitive accuracy.

Depthwise separable convolutions split the standard convolutional operation into two distinct steps to significantly reduce computational overhead. In a conventional convolution, both filtering and combining of input features occur simultaneously. MobileNet decouples these operations by using:

- Depthwise Convolution: Applies a single filter to each input channel individually, performing spatial filtering independently for each channel. This drastically reduces the number of operations compared to a standard convolution.
- Pointwise Convolution: Uses a 1×1 kernel to combine the outputs from the depthwise convolution, creating new feature maps by linearly combining all channels. This step effectively integrates the spatial information from the previous layer.

Each depthwise separable convolution is followed by a Batch Normalization layer and a ReLU activation function, ensuring stability and non-linearity in the network's learning process. In addition, instead of the typical approach of reducing spatial dimensions through pooling layers, MobileNet reduces dimensionality through the stride in the depthwise convolution. This reduction not only preserves essential spatial information but also significantly decreases the number of computations. More information about the reduction of the computational cost can be found in Section A.1.

In this work, we utilize the MobileNetV2 architecture [84], which aims to be even more lightweight and computationally efficient than its predecessor, MobileNet. Similar to MobileNet, MobileNetV2 employs depthwise separable convolutions and pointwise convolutions, but it introduces an additional innovation: inverted residual connections. These connections are reminiscent of ResNet's residual connections (see Section 2.3.1.4), but they are applied between layers with a low number of channels, known as thin bottleneck layers. A complete review of this architecture is in Section A.1.1.

MobileNetV2 improves upon the original MobileNet by optimizing computational efficiency and memory usage through its inverted residual structure with linear bottlenecks.

Unlike MobileNet, which increases dimensionality as the network deepens—raising the cost of activations—MobileNetV2 expands dimensions only within its internal layers, keeping input and output representations compact. This design reduces activation costs, making MobileNetV2 more efficient in terms of memory and computation, ideal for mobile and embedded applications where resource constraints are crucial.

2.3.1.6 EfficientNet

EfficientNet [81], developed by Google researchers, represents a significant advancement in convolutional neural network (CNN) design by addressing the limitations of traditional scaling methods. Previous architectures like ResNet (Section 2.3.1.4) and VGG (2.3.1.3) predominantly focused on scaling depth—adding more layers to improve accuracy. While this approach enhanced performance to a certain extent, it also introduced challenges such as vanishing gradients and increased computational cost. In contrast, EfficientNet proposes a more holistic scaling strategy by simultaneously optimizing three dimensions: network depth, width, and input resolution, called compound scaling [81].

- Depth Scaling: Increasing depth, or the number of layers, enhances the network's ability to learn complex patterns, particularly when working with high-resolution images. More layers allow the network to retain and process detailed information. However, indiscriminate addition of layers can lead to diminishing returns due to increased computational complexity and saturation in accuracy gains.
- Width Scaling: Expanding the network's width by increasing the number of feature channels enables the model to capture a greater variety of features. This is particularly important for high-resolution images, where more feature maps are required to process the additional pixel information. However, excessive width can lead to redundant feature extraction, slower computation, and reduced model efficiency.
- Resolution Scaling: Enhancing the input resolution allows the network to capture finer details in the data, improving accuracy. Training on higher-resolution images enables the model to learn more intricate features, but this also increases the computational burden, necessitating a careful balance with depth and width.

By applying compound scaling (explained in detail in Section A.2), the authors derived a family of models from B0 to B7, each with increasing network capacity. Despite maintaining the same architecture, each subsequent model scales the number of layers (depth), channels (width), and input resolution differently, optimizing for specific computational budgets.

This balanced scaling approach enables EfficientNet to achieve superior accuracy with fewer parameters compared to traditional CNN architectures such as ResNet and VGG. By optimizing all dimensions simultaneously, EfficientNet effectively utilizes computational resources, providing models that are both accurate and computationally efficient [81].

2.3.1.7 Pix2pix

Pix2pix is a deep learning model designed for image-to-image translation, introduced by Isola et al. in 2017 [85]. It is based on Generative Adversarial Networks (GANs) and aims to learn a mapping between input and output images, making it highly effective for tasks such as

image colorization, style transfer, and semantic segmentation [85]. Unlike traditional GANs, which generate images from random noise, pix2pix takes an input image and translates it into a corresponding output image in a supervised learning manner. This framework allows the model to learn structured loss functions and generate high-quality, realistic images.

The pix2pix model consists of two main components: the generator and the discriminator, working in an adversarial setup. Both the generator and discriminator follow a convolution-BatchNorm-ReLU module structure [85].

The generator receives an input image and produces a translated version of it. Instead of a simple encoder-decoder structure, pix2pix uses a U-Net architecture, which features skip connections between corresponding layers in the encoder and decoder. This helps preserve fine details in the image.

The discriminator evaluates image pairs—either real (input + ground truth) or fake (input + generated image). Instead of classifying the entire image as real or fake, pix2pix employs a PatchGAN discriminator, which classifies small patches (e.g., 70×70 pixels) of the image independently. The final classification is obtained by averaging the outputs of all patches, making the discriminator more robust to local texture variations.

One of the key aspects that makes pix2pix effective is its training strategy, which balances two competing objectives: improving the generator while refining the discriminator.

- Fooling the Discriminator: The generator learns to create realistic images so the discriminator classifies them as real. This is achieved using adversarial loss, often implemented as binary cross-entropy (see Section 2.3.5).

- Minimizing Reconstruction Loss: The generator also ensures that the generated image closely resembles the ground truth, using, for instance, L_1 or L_2 reconstruction loss to maintain structural accuracy.

Figure 2.9 illustrates the architecture of the pix2pix model applied to a denoising task. The process begins with noisy input images, which are fed into the generator G, implemented as a U-Net. The generator learns to map these noisy images to denoised outputs, producing generated samples. Simultaneously, the ground truth (real clean images) is provided as a reference. Both the generated samples and the real clean images are then passed to the discriminator, which evaluates whether each input pair is real (from the dataset) or fake (generated by G). This is achieved through a PatchGAN discriminator, D. The discriminator's output contributes to two key loss functions: the discriminator loss, which measures its ability to distinguish real from fake images, and the generator loss, which helps G improve by learning to fool the discriminator. Through this adversarial setup, the generator refines its outputs to produce increasingly realistic denoised images.

Compared to VGG and ResNet, which are primarily classification networks, pix2pix is specifically designed for generative tasks, making it more effective in producing high-quality image transformations. The use of a PatchGAN discriminator provides a localized texture-based evaluation, allowing pix2pix to capture fine details more effectively than architectures like ResNet, which focus on global feature extraction. Additionally, the combination of adversarial loss and L_1 reconstruction loss enables pix2pix to balance realism with structural accuracy, whereas models like U-Net and VGG typically rely only on direct pixel-wise comparisons. These advantages make pix2pix a powerful tool for high-quality image synthesis in applications requiring both global coherence and local detail preservation.

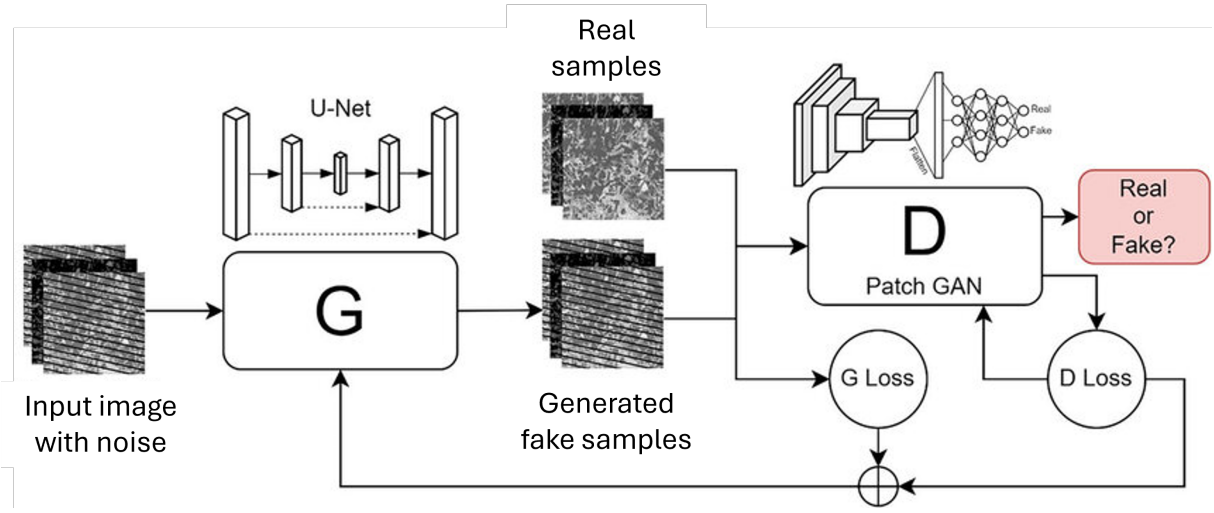


Figure 2.9: Diagram of the pix2pix architecture applied to a denoising task. The generator (G), implemented as a U-Net, receives noisy input images and generates denoised outputs. Both the generated and real (ground truth) images are fed into the PatchGAN discriminator (D), which determines whether an image is real or generated. The discriminator loss helps improve its classification accuracy, while the generator loss enables G to refine its outputs by learning to fool the discriminator, ultimately producing more realistic denoised images. Figure adapted from [86].

2.3.2 Transformer Based Architectures

The Transformer architecture, introduced in the paper Attention is All You Need by Vaswani et al. (2017) [87], revolutionized sequence processing in natural language processing (NLP) by replacing traditional recurrent mechanisms with self-attention. Unlike recurrent models, which process data sequentially and struggle to capture long-term dependencies, Transformers use self-attention to evaluate relationships between all elements in a sequence simultaneously. This design enables parallel processing and makes Transformers highly effective at capturing long-range dependencies, making them especially well-suited for complex, large-scale tasks where global context is crucial.

In the following, we will focus on the application of transformer architectures to the image domain, setting aside their use in natural language processing (NLP). The architectures included are Vision Transformers, Swin Transformer, SegFormer, and SwinIR. To maintain a smooth and engaging reading experience, technical details of these architectures are provided in Section A.3.

2.3.2.1 Vision Transformers (ViTs)

The Vision Transformer (ViT) [88] is a novel application of the Transformer architecture designed to handle image data instead of text. Traditionally, Transformers were developed for natural language processing tasks, where they process sequences of words or tokens. However, applying this architecture directly to images presents a challenge because images are composed of pixels, and using the original Transformer design would require that each pixel attend to every other pixel, resulting in a computationally expensive and impractical quadratic scaling of attention. This would make it difficult to process even moderately-sized images efficiently. To

address this, Vision Transformers bypass the pixel-by-pixel attention mechanism by dividing the image into smaller, fixed-size patches, treating each patch as a "token" similar to a word in text, which significantly reduces the sequence length and makes the model computationally feasible for image classification and other vision tasks [88].

To implement this, an image is first split into fixed-size patches. Each patch is then linearly embedded, and position embeddings are added to retain spatial information. The resulting sequence of patch embeddings is treated as a sequence of tokens, similar to how words are treated in NLP applications, and is fed into the standard Transformer encoder.

Formally, the image $x \in \mathbb{R}^{H\times W\times C}$ is reshaped into a sequence of flattened 2D patches, $x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where H and W are the height and width of the image, C is the number of channels, and $P\times P$ is the resolution of each patch (by default $P = 16$). The number of patches $N = HW/P^2$ is also the effective sequence length for the Transformer model. The flattened patches are then mapped to a fixed latent vector size D using a linear projection, producing the patch embeddings. These embeddings, along with the position embeddings, are provided as input to the Transformer, which processes them through self-attention and subsequent layers to generate the final output [88].

In summary, Vision Transformers represent a significant shift from traditional CNN-based approaches by directly applying the Transformer model to image patches, offering a promising alternative for image classification and other vision tasks, particularly when trained at scale.

2.3.2.2 Swin Transformer

The Swin Transformer [89] is a hierarchical and computationally efficient Vision Transformer (ViT) tailored for image analysis. Unlike conventional ViTs, which rely on global self-attention mechanisms, Swin Transformers address the critical scalability challenges that arise when processing high-resolution images.

Global self-attention, as employed in ViTs, scales quadratically with the number of image patches. This poses a significant challenge for high-resolution images: either the patches must be kept small, leading to an unmanageable number of patches and prohibitive computational costs, or the patches must be enlarged, which sacrifices resolution and fine-grained detail. Consequently, ViTs struggle to balance computational feasibility with the need for high-resolution processing.

The Swin Transformer overcomes these limitations through a carefully designed architecture that ensures both scalability and efficiency, while retaining the ability to capture fine-grained and multi-scale information. This is achieved through three key innovations:

1. Local Attention in Windows: Instead of calculating self-attention globally across all patches, Swin Transformers compute attention within fixed, non-overlapping regions (windows) of the image. This reduces the computational complexity to linear scaling with the number of patches, making it feasible for high-resolution images. An illustration can be seen in Figure 2.10, where the image is divided into patches (in gray), and the defined windows for local attention are highlighted in red. The figure demonstrates how the attention is computed within these windows, emphasizing the non-overlapping regions that allow for efficient processing while maintaining critical local context.

2. Shifted Window Mechanism: Successive layers shift the position of the windows to enable cross-window interactions. This design effectively captures broader contextual

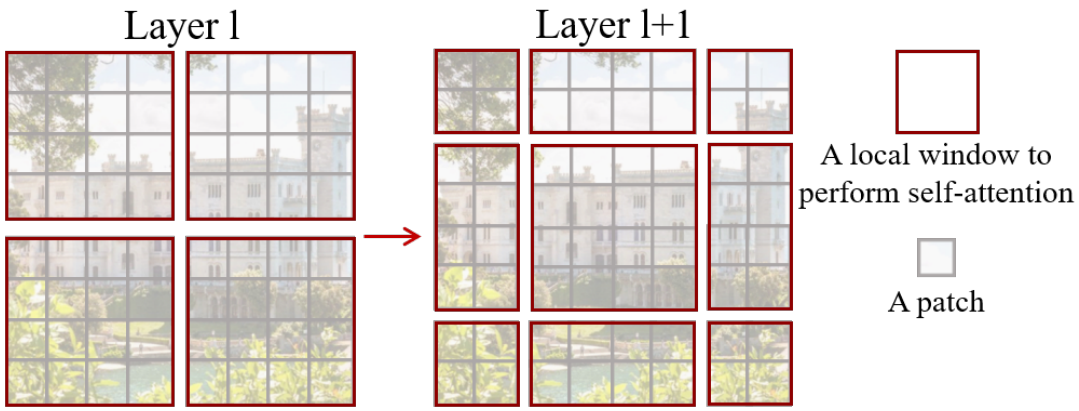


Figure 2.10: Illustration of the Swin Transformer windowing mechanism. On the left, the input image is divided into patches (represented as grey squares), with the red squares indicating the defined windows for local attention. On the right, the shifting of these windows is depicted, showcasing how they move across the image to capture different contextual information during the attention process. Figure from [89].

information without the need for full global attention.

3. Hierarchical Representation with Patch Merging: The spatial resolution of the image is progressively reduced across stages, while the feature dimensionality increases. This hierarchical approach not only mimics the multiscale representation of convolutional neural networks (CNNs) but also ensures computational efficiency as the model processes lower-resolution features in deeper layers.

Through these innovations, the Swin Transformer achieves a balance between computational efficiency and representational power, making it highly effective for tasks such as image classification, object detection, and semantic segmentation.

For a comprehensive understanding of the implementation details of the Swin Transformer architecture, including its specific layers, configurations, and optimization strategies, please refer to Section A.3.1. This section provides an in-depth exploration of how the architecture is structured and the methodologies employed to enhance its performance in various image processing tasks.

2.3.2.3 SegFormer

Semantic segmentation is a critical task in computer vision that requires pixel-wise classification of images. However, designing efficient architectures for segmentation presents unique challenges: computational resources are often limited, and available datasets tend to be much smaller than those used for tasks like image classification. These constraints necessitate models that are both computationally efficient and capable of generalizing well with limited data.

Traditional convolutional neural networks (CNNs) have been widely used for semantic segmentation, but their local receptive fields often fail to capture global context effectively. To address this, transformer-based architectures, such as SETR [90], have been proposed. SETR introduced the use of Vision Transformers (ViTs) (section 2.3.2.1) for semantic

segmentation, viewing the task from a sequence-to-sequence perspective. However, SETR exhibited limitations, including a lack of multi-scale features and inflexible output resolution, hindering its ability to adapt to varying input image sizes.

Pyramid Vision Transformer (PVT) [91] sought to address these issues by incorporating hierarchical representations, a characteristic inspired by CNNs. Building on these ideas, SegFormer [92] introduces a novel architecture consisting of two main components:

1. A hierarchical transformer encoder to extract multi-scale features, capturing both global and local information at varying resolutions.

2. A lightweight, all-MLP decoder that fuses these features to produce the final segmentation mask.

SegFormer eliminates positional encoding to enhance flexibility across resolutions and employs efficient self-attention to reduce computational complexity. This design enables state-of-the-art performance while remaining computationally efficient.

Hierarchical encoder The encoder of SegFormer is designed to extract multi-level features, similar to the feature hierarchies in CNNs and same idea as the Swin Transformer. This ensures that the decoder receives both high-resolution coarse features for capturing global information and low-resolution fine-grained features to enhance local detail.

The hierarchical encoder processes the image in stages, progressively reducing the spatial resolution while increasing the feature dimensionality [92]. Unlike the Swin Transformer, which employs non-overlapping patch merging (see Section 2.3.2.2), SegFormer introduces an overlapping patch merging mechanism to address limitations in preserving local continuity and boundary information. In the Swin Transformer, patches are extracted without overlap, which can disrupt local continuity and hinder the model's ability to capture fine-grained details. SegFormer improves upon this by allowing patches to overlap, ensuring smoother transitions and retaining more localized features around patch boundaries [92].

For detailed insights into the implementation of the SegTransformer architecture, including its design choices, layer configurations, and training procedures, refer to Section A.3.2.

2.3.2.4 SwinIR

The introduction of Transformers revolutionized data processing across various domains, including computer vision with the advent of the Vision Transformer (ViT) (Section 2.3.2.1). However, the quadratic computational cost with respect to patch size makes ViT impractical for high-resolution images. This limitation is addressed by SwinIR [93], a model that leverages the Swin Transformer for image restoration tasks. SwinIR has demonstrated superior performance in applications such as image denoising, JPEG compression artifact reduction, and super-resolution.

In this work, we employ SwinIR specifically for image denoising to reduce artifacts in medical imaging. Its architectural design is particularly well-suited for this task due to its ability to handle both local and global features efficiently.

The backbone of SwinIR is the Swin Transformer, and the architecture is organized into three key components.

1. Shallow Feature Extraction, which utilizes Convolutional Neural Networks (CNNs) to transition the input into an enhanced feature space.

2. Deep Feature Extraction, leveraging the Swin Transformer to extract high-level and fine-grained features using Residual Swin Transformer Blocks (RSTBs).

3. High-Quality Image Reconstruction, which employs CNNs to synthesize the final output image by aggregating shallow and deep features.

The detailed implementation of the SwinIR architecture is outlined in Section A.3.3.

The advantages of SwinIR lie in its ability to effectively capture fine details and contextual information through its residual architecture, making it particularly adept at removing noise while preserving image integrity. This specialization allows SwinIR to achieve superior performance in image denoising compared to the broader Swin Transformer, which may not be as finely tuned for this specific application.

2.3.3 Mamba Based Architectures

In recent years, the Mamba architecture has emerged as a groundbreaking innovation in sequence modeling, generating significant attention for its potential to rival Transformers. Introduced by Albert Gu and Tri Dao [94]. Mamba builds on the foundation of State-Space Models (SSMs) to achieve competitive accuracy with significantly improved efficiency.

SSMs have long been recognized for their ability to process sequences faster and with lower memory requirements than Transformers. However, their practical adoption was limited due to accuracy trade-offs. Mamba overcomes this limitation by introducing Selective State-Space Models (Selective SSMs), which maintain the speed and efficiency of SSMs while enhancing their predictive performance to be comparable to or better than Transformers.

This section provides an overview of the foundational concepts of SSMs and delves into the advancements introduced by Mamba, presenting its architectural innovations and their implications for sequence modeling.

2.3.3.1 Mamba Architecture

The Mamba architecture is an innovative framework designed to address the challenges of handling long sequences efficiently, combining the strengths of Transformers and State-Space Models (SSMs).

Traditional Transformers are powerful but suffer from quadratic scaling due to the self-attention mechanism, which computes relationships between all token pairs. This makes them resource-intensive as sequence lengths increase. In contrast, SSMs scale linearly: doubling the sequence length merely doubles memory and compute requirements. This efficiency is pivotal for tasks involving extremely long sequences.

SSMs resemble RNNs by processing tokens sequentially but overcome RNNs' inherent slowness. During training, SSMs leverage linear operations, enabling parallel computation of outputs for all tokens. This is achieved through precomputations using matrices like, which are combined into a single convolutional operation. This makes SSMs both fast and memory-efficient, though at the cost of flexibility in adapting to token-specific input variations.

To address SSMs' limitations, Selective SSMs introduce input-dependent transformations. Instead of using static matrices, they compute token-specific matrices using linear layers,

enhancing their ability to prioritize or ignore certain inputs—similar to attention mechanisms in Transformers. However, this input dependency complicates the use of precomputed convolutions, requiring alternative strategies.

Mamba overcomes the computational challenges of Selective SSMs by employing the parallel associative scan algorithm. This technique precomputes intermediate steps for sequential operations, akin to calculating prefix sums for fast array manipulations. Combined with hardware-specific optimizations, such as utilizing fast GPU SRAM for operations and efficient memory transfers, Mamba maintains linear scaling even for selective SSMs.

The explanation of the Spatial-Scale Modules (SSMs) used in the Mamba architecture can be found in Section A.4.1. This section delves into the design and functionality of SSMs, highlighting how they enhance the model's ability to process images at different spatial resolutions.

2.3.3.2 U-Mamba

U-Mamba [95] represents an innovative fusion of Convolutional Neural Networks (CNNs) and State Space Models (SSMs), designed to leverage their respective strengths for enhanced medical image segmentation. By combining CNNs' ability to extract localized features with SSMs' efficiency in modeling long-range dependencies, U-Mamba offers a balanced and highly capable architecture.

This hybrid model is built upon the foundation of the U-Net architecture (Section 2.3.1.1), with significant modifications to address the limitations of convolutional kernels in capturing global context. While U-Net primarily relies on symmetric encoder-decoder structures with skip connections to integrate multi-scale features, its receptive field is inherently limited to local regions. U-Mamba overcomes this by embedding Mamba blocks (Section 2.3.3.1) within its encoder-decoder framework, providing an effective means to process both local details and global contexts.

U-Mamba retains the encoder-decoder paradigm from U-Net, which consists of two major components:

- Encoder: Responsible for extracting hierarchical features at multiple scales.
- Decoder: Recovers spatial resolution and integrates low and high-level features for precise segmentation outputs.

This structure is augmented by the inclusion of Mamba blocks, allowing the network to efficiently model dependencies across longer sequences (see Figure 2.11).

A central innovation in U-Mamba is the hybrid CNN-SSM block, which combines:

- Residual Blocks: These include plain convolutional layers followed by Instance Normalization (IN) and Leaky ReLU activation. These Residual connections allow the network to bypass layers, mitigating the vanishing gradient problem and improving training stability.

  The Leaky Rectified Linear Unit (Leaky ReLU) is an activation function that modifies the standard ReLU by allowing a small, non-zero gradient for negative input values. It is defined following Equation 2.10.

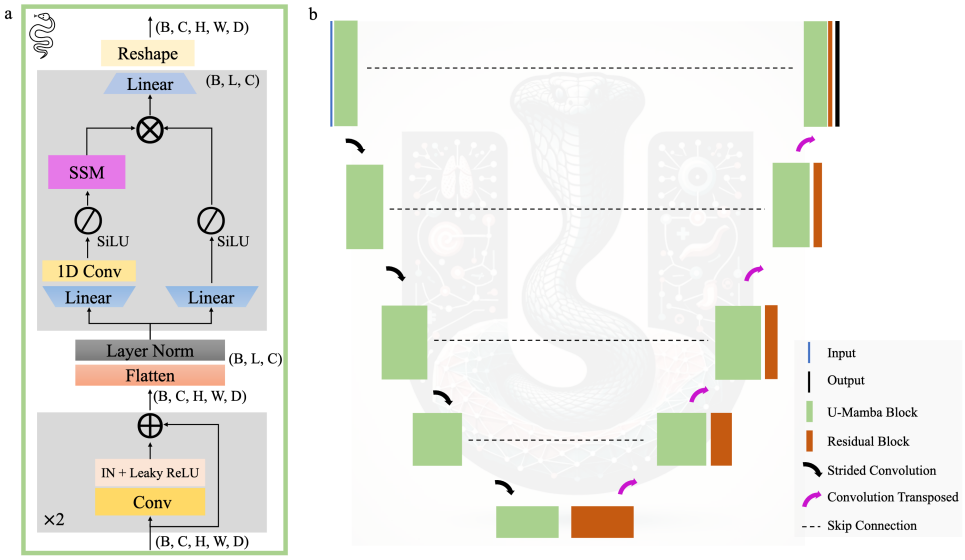


Figure 2.11: Overview of the U-Mamba architecture. The model integrates the encoder-decoder structure of U-Net with Mamba blocks, combining local feature extraction via convolutional layers with long-range dependency modeling through State Space Models (SSMs). Each encoder block consists of two Residual blocks followed by a Mamba block. Image features with a shape of (B,C,H,W,D) are flattened and transposed to (B,L,C), where $L = H \times W \times D$. The decoder reconstructs the segmentation map using transposed convolutions, residual blocks, and skip connections. A final $1 \times 1 \times 1$ convolutional layer and Softmax produce the segmentation probability map. Figure from [95].

$$f(x) = \begin{cases} x & \text{if } x > 0, \\ \alpha x & \text{if } x \leq 0, \end{cases} \tag{2.10}$$

where α is a small positive constant (commonly set to 0.01). Unlike standard ReLU, which outputs zero for all negative values, Leaky ReLU mitigates the "dying ReLU" problem where neurons can become inactive and stop learning. By maintaining a small slope for negative values, it ensures that all neurons contribute to learning throughout the network. However, Leaky ReLU is not without challenges, such as its reliance on the choice of α, which may need careful tuning for optimal performance.

- Mamba Blocks: Each block processes flattened and transposed image features as long sequences. The data passes through two parallel branches, as explained in Section A.4.1:
  - First Branch: Features are expanded with a linear layer, passed through a 1D convolution, activated with SiLU, and finally processed by an SSM layer for long-range modeling.
  - Second Branch: A simpler linear expansion followed by SiLU activation.

  The outputs of both branches are fused via a Hadamard product before being reshaped back to the original dimensions.

By integrating these components, the hybrid block captures local and global contexts simultaneously, a key requirement for accurate segmentation of complex biomedical images.

In addition, inspired by U-Net, U-Mamba retains skip connections that directly transfer encoder features to the corresponding decoder layers. This ensures that detailed local information is preserved while enhancing global context understanding.

Finally, The decoder, after processing hierarchical features, uses transposed convolutions to recover spatial resolution. A final $1 \times 1 \times 1$ convolution combined with a Softmax layer generates the segmentation probability map, ensuring precise delineation of structures in biomedical images.

2.3.4 2.5D and 3D Model Variants and Transfer Learning

The architectures discussed in Section 2.3 correspond to their 2D implementations, where CT volumes are processed slice by slice. However, these models can be adapted to handle 2.5D and 3D data, expanding their applicability to volumetric datasets commonly encountered in medical imaging. These methodologies are also leveraged in this work due to their flexibility and ability to capture richer spatial information.

This section explores the adaptations required to extend 2D architectures to their 2.5D (Section 2.3.4.1) and 3D (Section 2.3.4.2) counterparts, emphasizing their relevance and versatility in volumetric data processing. Additionally, we discuss transfer learning (Section 2.3.4.3), a crucial technique in deep learning that allows models pre-trained on large datasets to be adapted for other tasks, such as medical image segmentation.

2.3.4.1 2.5D Architectures

2.5D architectures offer a compromise between 2D and 3D processing by utilizing information from multiple planes of a volumetric image. In this approach, three channels corresponding to the axial, sagittal, and coronal planes are used as input, enabling the network to analyze spatial relationships across different anatomical perspectives. The output for each plane has the same spatial resolution as the input and represents the segmentation prediction in that plane.

Advantages: One of the key benefits of these architectures is their computational efficiency. By processing individual planes separately, they significantly reduce memory consumption and computational demands compared to fully three-dimensional (3D) models. This makes them more feasible for large-scale applications without sacrificing performance.

Additionally, these architectures provide enhanced spatial context by leveraging multiple anatomical planes. This approach improves the network's ability to capture and understand complex 3D spatial relationships while maintaining a relatively simple model structure.

Another advantage is their compatibility with existing frameworks. Since this method extends from well-established 2D architectures, it requires only minor modifications to input and output formats, making it straightforward to implement and integrate into existing deep learning pipelines.

Implementation Details: For 2.5D processing, the network is designed to accept three channels as input, each corresponding to one of the volumetric planes (axial, sagittal, and

coronal). The input tensor shape, therefore, changes to $(H, W, 3)$, where H and W represent the spatial dimensions. Each channel is processed independently, producing segmentation outputs for each plane. Since the outputs are per-plane predictions, an additional algorithm is required to merge these into a cohesive 3D segmentation. This merging step ensures that the final geometry integrates information from all three planes, providing a unified 3D representation of the segmented structure.

2.3.4.2 3D Architectures

3D architectures extend the network's operation to the volumetric domain by treating the entire image as a three-dimensional tensor. This approach enables the network to learn spatial relationships across all three dimensions (x, y, and z) simultaneously.

Advantages: A significant advantage of full 3D processing is its ability to provide comprehensive spatial analysis. By capturing the complete spatial context of volumetric data, these models can leverage richer structural information, potentially leading to higher accuracy in tasks such as segmentation and classification. This holistic approach ensures that spatial dependencies within the data are fully utilized, improving performance in complex medical imaging applications.

Furthermore, these architectures generate end-to-end volumetric output, meaning that the final result is a complete 3D segmentation or classification outcome. This aligns directly with the structure of the input data, preserving the volumetric integrity and enabling seamless integration with subsequent clinical or analytical workflows.

Challenges:

- High Computational Cost: 3D convolutions and volumetric data significantly increase memory and processing requirements.
- Data Scarcity: Training 3D networks often requires large datasets, as the number of learnable parameters is higher than in 2D or 2.5D models.

Implementation Details: To adapt a 2D network to a 3D version, the following changes are required:

- Replace 2D convolutional layers with 3D convolutions, where the kernel operates across all three dimensions.
- Modify pooling and upsampling layers to their 3D equivalents.
- Adjust the input tensor shape to (D, H, W, C), where D is the depth (number of slices), H is the height, W is the width, and C is the number of channels.

Table 2.1 summarizes the key differences between 2D, 2.5D, and 3D architectures in terms of input structure, spatial context, computational requirements, and typical use cases.

Table 2.1: Comparison of 2D, 2.5D, and 3D Architectures

	2D	2.5D	3D
Spatial Context	Single Slice In-Slice Only	Stack of Slices Partial Volumetric	Full Volume Full Volumetric
Comp. Cost	Low	Moderate	High
Output Type	2D Segmentation	2D Segmentation	3D Segmentation
Typical Use Cases	Thin Slices Small Memory Budgets	Moderate Memory Budgets Adjacent Slice Context	Full Volume Analysis High Accuracy Needs

2.3.4.3 Transfer Learning

Transfer learning is a machine learning technique that enables a model trained on a large dataset to be adapted for a different but related task. This approach is particularly beneficial in the medical imaging field, where annotated data is often scarce and expensive to obtain. By leveraging knowledge learned from large-scale datasets, models can generalize better, converge faster, and reduce the risk of overfitting, making them more efficient for medical applications.

In this study, we implement transfer learning by utilizing pre-trained encoders within U-shaped architectures. The encoders, responsible for extracting high-level features from images, are initialized with weights obtained from training on the ImageNet dataset [82]. ImageNet is a large-scale dataset consisting of millions of natural images across thousands of categories, including objects, animals, and landscapes. Despite the significant differences between natural images and medical scans, deep learning models can still adapt well, as the early layers of a neural network extract fundamental image features such as edges, textures, and shapes, which are transferable across domains.

To apply transfer learning effectively, we freeze the encoder's pre-trained weights, preserving the learned representations while training the decoder. The decoder is then optimized for medical image segmentation, ensuring the model is fine-tuned for the specific challenges of analyzing medical scans while benefiting from the robust feature extraction capabilities of pre-trained networks.

2.3.5 Loss functions

In machine learning, a loss function quantifies the deviation of a model's predictions from the true values or ground truth. It serves as the guiding metric during training, with the objective of minimizing this loss to improve the model's predictive performance. In what follows, we detail the loss functions utilized for both image generation (Section 2.3.5.1) and image segmentation tasks (Section 2.3.5.2).

2.3.5.1 Image Generation Loss Functions

For image generation, loss functions typically measure the difference between the generated and real images, ensuring that the synthetic outputs are both realistic and coherent. Common losses in this context include reconstruction-based losses like L_1 loss, which promotes sparsity and smoothness by minimizing the absolute differences between pixel values, Mean Squared Error (MSE), which focuses on pixel-wise accuracy, and Structural Similarity Index (SSIM), which enhances perceptual quality by considering structural information in the images. Below, we describe several loss functions tailored for medical image generation tasks, where $\hat{\mathbf{X}}$ and $\mathbf{X} \in \mathbb{R}^{d \times d}$ are the predicted and ground truth images values, respectively.

Weighted L_1 Loss This loss measures the absolute differences between predicted and true values, penalizing outliers and enhancing the robustness of artifact reduction.

$$L_1^{\mathbf{w}} = \|\hat{\mathbf{X}} - \mathbf{X}\|_1 \cdot \mathbf{w}, \tag{2.11}$$

where $\mathbf{w} \in \mathbb{R}^{d \times d}$ is a pixel-wise weight matrix.

Focal Frequency Loss (FFL) This loss focuses on high-frequency components, preserving fine details while suppressing artifacts [96].

$$\text{FFL}^{\beta,\alpha} = \frac{1}{d \cdot d} \sum_{u=0}^{d-1} \sum_{v=0}^{d-1} z(u,v) \cdot |F_{\hat{\mathbf{X}}}(u,v) - F_{\mathbf{X}}(u,v)|^2 \cdot \beta, \tag{2.12}$$

where

$$z(u,v) = |F_{\hat{\mathbf{X}}}(u,v) - F_{\mathbf{X}}(u,v)|^{\alpha}. \tag{2.13}$$

Here, $F_{\mathbf{X}}(u,v)$ represents the Fourier transform at frequency (u,v), α is the scaling factor, and β is a weight for spatial frequencies.

Mean Squared Error (MSE) This loss measures the average squared differences between predictions and targets, offering simplicity and easy interpretability.

$$\text{MSE} = \frac{1}{d \cdot d} \sum_{i=1}^{d} \sum_{j=1}^{d} (\hat{\mathbf{X}}_{\mathbf{ij}} - \mathbf{X}_{\mathbf{ij}}^{\mathbf{p}})^2, \tag{2.14}$$

where $\hat{\mathbf{X}}_{\mathbf{ij}}$ and $\mathbf{X}_{\mathbf{ij}}$ are the predicted and ground truth pixel values, respectively.

Structural Similarity Index (SSIM) This loss evaluates luminance, contrast, and structure, ensuring perceptual fidelity during artifact reduction [97].

$$\text{SSIM}(\hat{\mathbf{X}}, \mathbf{X}) = \frac{(2\mu_{\hat{\mathbf{X}}}\mu_{\mathbf{X}} + C_1)(2\sigma_{\hat{\mathbf{X}}\mathbf{X}} + C_2)}{(\mu_{\hat{\mathbf{X}}}^2 + \mu_{\mathbf{X}}^2 + C_1)(\sigma_{\hat{\mathbf{X}}}^2 + \sigma_{\mathbf{X}}^2 + C_2)}, \tag{2.15}$$

where μ and σ denote the mean and variance of pixel intensities, $\sigma_{\hat{\mathbf{X}}\mathbf{X}}$ is the covariance between images, and C_1 and C_2 are stability constants.

Multi-Scale Structural Similarity Index (MS-SSIM) MS-SSIM extends SSIM by computing structural similarity across multiple scales. It evaluates how structural information changes at different resolutions by averaging SSIM values across scales, offering a comprehensive measure of image similarity.

2.3.5.2 Image Segmentation Loss Functions

For image segmentation, loss functions aim to optimize pixel-wise classification, ensuring accurate delineation of objects within an image. These include cross-entropy loss, which is commonly used for categorical segmentation tasks, or Dice loss, which mitigates class imbalance by emphasizing overlap between predicted and ground-truth segmentations. Both critical for reliable segmentation outcomes [98]. Below, we describe several loss functions tailored for medical image segmentation tasks. Following the notation of the previous section, $\hat{\mathbf{X}}$ and $\mathbf{X} \in [0,1]^{d \times d}$ represent binary or probabilistic segmentation masks, respectively. To avoid overloading the notation, the limits of the summations are omitted, which allows for a more concise representation of the mathematical expressions. In this context, X_c represents the values of the mask for class c, while $\sum(X)$ denotes the summation of all the values in the mask, effectively aggregating the contributions across all classes.

Dice Loss : This metric evaluates the overlap between predicted and ground truth segmentations. It is effective in handling class imbalance by emphasizing the regions of interest. The formula for Dice Loss is:

$$DiceLoss = 1 - \frac{2\sum \hat{\mathbf{X}}\mathbf{X} + smooth}{\sum \hat{\mathbf{X}} + \sum \mathbf{X} + smooth}, \tag{2.16}$$

where $\hat{\mathbf{X}}$ and $\mathbf{X}$ represent the predicted and ground truth values, respectively, and *smooth* prevents division by zero. For multi-class segmentation, a weighted Dice Loss is computed by applying class-specific weights:

$$WeightedDiceLoss = \frac{1}{N}\sum_{c=1}^{N} w_c \left(1 - \frac{2\sum \hat{\mathbf{X}}_c \mathbf{X}_c + smooth}{\sum \hat{\mathbf{X}}_c + \sum \mathbf{X}_c + smooth}\right), \tag{2.17}$$

where N is the total number of classes and w_c is the weight for each class c.

Cross Entropy Loss (CE) : Commonly used in classification tasks, this loss penalizes the divergence between the predicted probability distribution and true labels. It is defined as:

$$CrossEntropyLoss = -\sum_{c=1}^{N} y_c \log(\mathbf{p}_c), \tag{2.18}$$

where y_c is the ground truth label, and $\mathbf{p}_c$ is the predicted probability for class c.

DiceCE Loss [99]: Combining Dice Loss and Cross Entropy Loss, this hybrid approach leverages the strengths of both methods. The Dice Loss improves overlap in segmentation,

while Cross Entropy stabilizes training:

$$DiceCELoss = DiceLoss + CrossEntropyLoss. \tag{2.19}$$

DiceFocal Loss (DiceFocal) [100, 101]: A combination of Dice Loss and Focal Loss, this approach is designed to address class imbalances and improve performance for difficult examples. The Focal Loss component is:

$$FocalLoss = -\alpha(1-\mathbf{p}_t)^{\gamma}\log(\mathbf{p}_t), \tag{2.20}$$

where α adjusts the weight of positive samples relative to negative samples, ensuring that underrepresented or harder-to-classify classes are not overshadowed by dominant ones. In segmentation tasks, where certain regions (e.g., background) may dominate, α ensures that more emphasis is placed on the minority regions, such as calcified plaques. On the other hand, γ controls the focus on hard-to-classify examples by reducing the relative contribution of well-classified samples to the overall loss. Higher values of γ assign greater weight to predictions with low confidence, effectively steering the model toward improving performance on difficult cases.

The combined loss is:

$$DiceFocalLoss = \lambda_{dice} \times DiceLoss + \lambda_{focal} \times FocalLoss, \tag{2.21}$$

where λ_{dice}, and λ_{focal} control the weighting of components.

Generalized Dice Loss (GD) [102]: This loss generalizes Dice Loss by incorporating class-specific weights to reduce the dominance of larger classes. It is expressed as:

$$GD = 1 - \frac{2\sum w_c \sum \hat{\mathbf{X}}_c \mathbf{X}_c}{\sum w_c \sum \hat{\mathbf{X}}_c + \sum w_c \sum \mathbf{X}_c}, \tag{2.22}$$

where w_c is the weight for class c, computed based on the inverse of its volume.

Generalized DiceFocal Loss (GDF) : Combining Generalized Dice Loss with Focal Loss, this function balances class distributions and prioritizes hard-to-classify samples. It is weighted using parameters λ_{gd} and λ_{focal}:

$$GeneralizedDiceFocalLoss = \lambda_{gd} \times GD + \lambda_{focal} \times FocalLoss. \tag{2.23}$$

Tversky Loss [103]: A generalization of Dice Loss, Tversky Loss introduces parameters α and β to control the trade-off between false positives and false negatives. It is defined as:

$$TverskyLoss = 1 - \frac{\sum \hat{\mathbf{X}}\mathbf{X}}{\sum \hat{\mathbf{X}}\mathbf{X} + \alpha \sum \hat{\mathbf{X}}(1-\mathbf{X}) + \beta \sum (1-\hat{\mathbf{X}})\mathbf{X}}, \tag{2.24}$$

where α and β adjust the trade-off between false positives and false negatives. A higher α gives more weight to false positives, and a higher β gives more weight to false negatives.

2.3.6 Evaluation Metrics for Model Performance

In this section, we describe the metrics used to evaluate the performance of the different methodologies, including image generation and image segmentation tasks. For image generation, we use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) (Section 2.3.6.1). For image segmentation tasks, we evaluate the models using Precision, Recall, Dice coefficient (or F_1-score), and Intersection over Union (IoU) (Section 2.3.6.2).

2.3.6.1 Image Generation Metrics

In this work, we evaluate the performance of different AI algorithms based on neural networks, whose task is to reduce artifacts caused by metal dental implants in CT images. Specifically, the evaluation aims to measure the similarity between the output of the algorithms and artifact-free images, in a multimodal technique, which is detailed in Chapter 3. The chosen evaluation metrics are the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM), both of which are widely used for assessing image quality in various domains.

The PSNR and SSIM are both objective image quality metrics, meaning they rely on numerical comparisons between the generated and reference (ground truth) images, as opposed to subjective methods that depend on human judgment. PSNR has been traditionally used to measure the quality of reconstructed images based on the amount of noise introduced, while SSIM is designed to capture perceptual similarity based on structural information, which makes it more sensitive to human visual system perceptions.

Peak Signal-to-Noise Ratio (PSNR) The PSNR measures the ratio between the maximum possible signal power and the power of the noise corrupting the signal, providing an objective measure of the quality of the reconstructed image. It is defined by Equation 2.25 [104].

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \tag{2.25}$$

where MAX is the maximum possible pixel value of the image, and MSE is the Mean Squared Error between the generated image g and the reference image f, defined by Equation 2.26.

$$\text{MSE}(f,g) = \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} (f_{ij} - g_{ij})^2, \tag{2.26}$$

with M and N being the image dimensions, and f_{ij} and g_{ij} representing the pixel values at position (i, j) in the reference and generated images, respectively. A higher PSNR value generally indicates better image quality, as it corresponds to a lower MSE and hence less noise.

PSNR is commonly used in applications where image fidelity is critical, such as image compression and denoising, and it has been found to perform well for noisy images [104]. However, PSNR has limitations, as it does not account for perceptual or structural information and can sometimes provide high values for images that are perceptually dissimilar, which is why SSIM is often used alongside PSNR.

Structural Similarity Index Measure (SSIM) SSIM is a metric designed to measure the perceptual similarity between two images by considering structural information, luminance, and contrast. It was introduced by Wang et al. [105] and has since been widely used for evaluating the quality of images, especially in scenarios where human visual perception plays a significant role. The SSIM is given by Equation 2.27 [105].

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{2.27}$$

where μ_x and μ_y are the mean pixel values of the images x and y, σ_x^2 and σ_y^2 are their variances, σ_{xy} is the covariance between the two images, and C_1 and C_2 are constants used to stabilize the division with weak denominators. The SSIM index ranges from -1 to 1, where 1 indicates perfect similarity, meaning the images are identical.

The SSIM metric evaluates the structural integrity of images and is especially sensitive to changes that affect the perceived quality of images, such as those caused by compression or noise [105]. It is considered a better reflection of human visual perception than PSNR, as it accounts for complex structural changes in the images, as image structure, luminance, and contrast.

Both PSNR and SSIM have their advantages and drawbacks, and the combination of these two metrics provides a more comprehensive understanding of image quality in terms of both noise reduction and structural similarity.

2.3.6.2 Image Segmentation Metrics

Image segmentation is a critical task in medical imaging, where the goal is to delineate specific structures accurately, in this case the aorta, coronary arteries, or calcified regions. Evaluating the performance of neural networks for segmentation involves comparing two masks: the predicted segmentation mask generated by the network and the ground truth mask provided by experts. The comparison is performed pixel by pixel, and several metrics are used to quantify the accuracy of the segmentation.

To calculate these metrics, it is important first to define the key terms:

- True Positives (TP): Pixels correctly predicted as belonging to the target class (e.g., the aorta or coronary arteries).
- False Positives (FP): Pixels incorrectly predicted as belonging to the target class when they do not belong to it.
- True Negatives (TN): Pixels correctly predicted as not belonging to the target class.
- False Negatives (FN): Pixels that belong to the target class but are incorrectly predicted as not belonging to it.

To visualize the classification performance, a confusion matrix (Table 2.2) is often used. This matrix summarizes the counts of TP, FP, TN, and FN in a clear and concise way.

Based on these definitions, the following metrics are used to evaluate segmentation performance. All metrics range between 0 and 1, where 0 represents the worst possible performance (e.g., no overlap or completely incorrect predictions), and 1 represents the best performance (e.g., perfect segmentation or prediction).

Table 2.2: Confusion Matrix for Binary Classification

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Precision Precision quantifies the proportion of correctly predicted positive pixels among all pixels predicted as positive. It indicates the model's ability to avoid false positives.

$$\text{Precision} = \frac{TP}{TP+FP}. \tag{2.28}$$

A Precision value of 1 means that no pixels are misclassified as belonging to the target class when they do not actually belong to it.

Recall (Sensitivity) Recall, also known as Sensitivity, measures the proportion of actual positive pixels correctly identified by the model. It focuses on avoiding false negatives.

$$\text{Recall} = \frac{TP}{TP+FN}. \tag{2.29}$$

A Recall value of 1 means that all target pixels were successfully segmented.

F1-Score (Dice Coefficient) The F1-Score, or Dice Coefficient, is the harmonic mean of Precision and Recall. It is particularly useful when the dataset is imbalanced, providing a balanced evaluation of both metrics.

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}. \tag{2.30}$$

In this case, value 1 indicates perfect segmentation.

Intersection over Union (IoU) Intersection over Union (IoU), also known as the Jaccard Index, measures the overlap between the predicted segmentation and the ground truth relative to their union.

$$\text{IoU} = \frac{TP}{TP+FP+FN}. \tag{2.31}$$

Similar to Dice, higher values represent better segmentation accuracy.

False Negative Rate (FNR) or miss rate This value ranges from $[0,1]$, measuring the proportion of false negatives out of the total positives in the ground truth (GT). Higher values indicate a greater portion of undetected vessels.

$$\text{FNR} = \frac{\text{FN}}{\text{FN}+\text{TP}}. \tag{2.32}$$

Critical Success Index (CSI) Similar to the F_1 score, this metric evaluates prediction accuracy but assigns less weight to true positives, making it more sensitive to prediction errors.

$$\text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}. \tag{2.33}$$

These metrics, combined with expert analysis, provide a comprehensive assessment of the segmentation models' ability to accurately delineate anatomical structures. Each metric has its strengths and applications, allowing for nuanced evaluations tailored to specific tasks.

2.4 Workflow and Implementation Pipeline

This section outlines the high-level structure and implementation of the solutions developed in this work. The workflow is organized into four main modules, each addressing a specific stage in the process, from dataset generation to final patient result reconstruction. Figure 2.12 illustrates the overall scheme, highlighting the interaction and flow between these modules. This modular approach ensures a streamlined and efficient pipeline, making it adaptable to different tasks while maintaining clarity and ease of use.

Dataset Generation The first component involves the creation of the dataset. All neural networks employed in this work rely on supervised learning, requiring both input data and corresponding ground truth labels for training. Python scripts were developed to handle the original data format (DICOM), with additional functionality to interact with the 3D Slicer API.

This integration with 3D Slicer provided several advantages: not only could preprocessing techniques be visualized directly, but the software's extensive toolkit could be used for tasks such as resampling, cropping, or aligning images. These preprocessing steps, detailed in subsequent chapters, ensure that the data meets the requirements of neural networks. The preprocessed images are ultimately exported in formats suitable for machine learning frameworks, such as arrays or tensors.

Model Training The second component centers around the training of the neural network model. The input to this module consists of the preprocessed images and their corresponding ground truth data, which are split into training and test sets. The output of this module is a trained model (network weights) ready for inference.

This module includes various configuration parameters, such as the architecture type, number of epochs, loss function, and other hyperparameters. These settings enable flexibility in designing and training different models suited for the specific task at hand.

Testing and Evaluation The third component evaluates the results of the training process. This module contains scripts to predict test set outputs and compute performance metrics such as Dice, IoU, Precision, and Recall (see Section 2.3.6).

Additionally, visualization tools are provided to compare the predicted outputs with the ground truth labels, offering insights into the model's performance and areas for improvement. These tools are crucial for both quantitative and qualitative evaluation of the model.

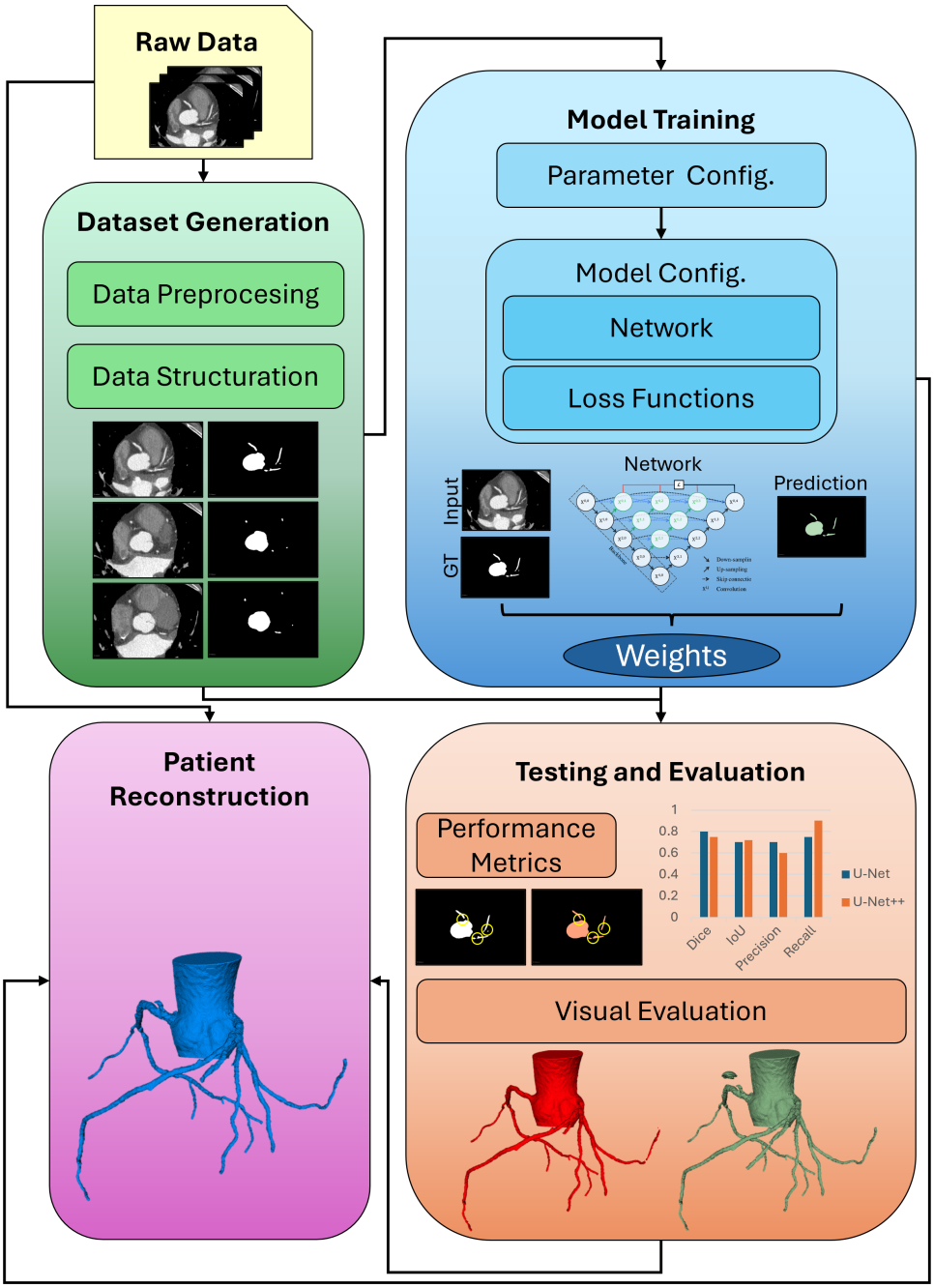


Figure 2.12: Overview of the workflow and implementation pipeline. The process is divided into four main modules: Dataset Generation, Model Training, Testing, and Patient Reconstruction. Each module is designed to perform specific tasks while seamlessly interacting with the others to ensure an efficient and modular implementation.

Patient Reconstruction The final component focuses on generating results for new patients. For instance, in the case of coronary segmentation, this module produces the final segmentation, including postprocessing steps like island removal or smoothing. Typically, the best-performing model from the testing and evaluation stage is selected for this task to ensure the highest accuracy and reliability. This module is also integrated into 3D Slicer, enabling users to generate results directly within the software. This integration minimizes the need for external tools and reduces friction for users, ensuring a seamless workflow from data preparation to final result visualization.

Chapter 3

Artifact Management Strategies in CT Imaging

Part of the text and figures for this chapter are reproduced from the following publications

(I) Serrano-Antón, B.[1,2,3], Rehman, M.[4,5], Martinel, N.[4], Avanzo, M.[5], Spizzo, R.[5], Fanetti, G.[5], P. Muñuzuri, A.[2,3], Micheloni, C.[4]. (2025). MAR-DTN: Metal Artifact Reduction Using Domain Transformation Network for Radiotherapy Planning. In: Antonacopoulos, A., Chaudhuri, S., Chellappa, R., Liu, CL., Bhattacharya, S., Pal, U. (eds) Pattern Recognition. ICPR 2024. Lecture Notes in Computer Science, vol 15311, pp. 143-159, 2025. Springer. DOI: https://doi.org/10.1007/978-3-031-78195-7_10

[1] FlowReserve Labs S.L., Santiago de Compostela, 15782, Galicia, Spain.
[2] Galician Center for Mathematical Research and Technology (CITMAga), Santiago de Compostela, 15782, Galicia, Spain.
[3] Group of Nonlinear Physics, University of Santiago de Compostela, Santiago de Compostela, 15782, Galicia, Spain.
[4] Machine Learning and Perception Lab, Università degli Studi di Udine, 33100, Udine (UD), Italy
[5] Centro di Riferimento Oncologico di Aviano IRCCS, 33081, Aviano (PN), Italy

3.1 Introduction

Throughout the introduction, the principles of CT imaging and some of its limitations have been outlined. One significant challenge, addressed in this chapter, is the presence of artifacts caused by metallic implants (see Section 1.2.2).

In cardiology, metal stents in coronary arteries often obscure the arterial lumen, making segmentation impossible and excluding affected patients from research studies. However, this issue extends beyond cardiology to other CT modalities used in clinical practice, such as kilovoltage computed tomography (kVCT) and megavoltage computed tomography (MVCT), which are essential for the diagnosis and treatment of head and neck cancer [106, 107, 108].

In radiation therapy, kVCT is primarily used during the treatment planning phase due to its superior soft-tissue contrast. This enables precise delineation of tumors and surrounding organs, crucial for maximizing the tumor dose while minimizing exposure to healthy tissues [109, 110]. Additionally, kVCT can track anatomical changes, such as tumor shrinkage or patient weight loss, throughout treatment, ensuring the plan remains accurate [109, 110].

Conversely, MVCT is mainly utilized for setup verification during each treatment session. Using the same energy level as the therapeutic beam, MVCT ensures accurate alignment of the radiation field with the target area [111, 109]. Its robustness against artifacts from high-density materials, such as metal implants, further enhances its utility [112]. However, MVCT images have lower soft-tissue contrast compared to kVCT, which may require supplemental imaging for precise tumor visualization.

Despite their individual strengths, both kVCT and MVCT face challenges with metal artifacts. kVCT's susceptibility to artifacts complicates tumor delineation near metal implants [112, 110], while MVCT's lower soft-tissue contrast reduces visualization precision. A combined approach leveraging the strengths of both modalities can mitigate these challenges, enabling better tumor localization and treatment precision.

This chapter focuses on reducing metal artifacts in kVCT imaging caused by metal implants in the dental region. To address this, we propose a supervised learning approach using neural networks, where the input is a kVCT scan, and the output is an MVCT scan with reduced artifacts. Our method specifically targets the artifact-affected regions while preserving soft tissue contrast. Future work aims to extend this approach to cardiology, improving imaging quality in the presence of metallic stents and implants.

3.2 Methodology

In this section, we outline the methodology employed in this study. First, we describe the imaging modalities used (Section 3.2.1) and the process of generating datasets (Section 3.2.2), ensuring high-quality and well-annotated data for training. Next, we present the overall workflow (Section 3.2.3), detailing the key steps in our approach. We then explore the neural network architectures implemented, ranging from traditional CNN-based models to more recent transformer-based solutions (Section 3.2.4). Additionally, we discuss the loss functions used to optimize training (Section 3.2.5), the implementation parameters (Section 3.2.6), and the evaluation metrics employed to assess model performance (Section 3.2.7).

3.2.1 Clinical Data

The CT images utilized in this study were obtained from 52 patients treated at the *National Cancer Institute (CRO) IRCCS*[1]. All patients underwent intensity-modulated radiotherapy (IMRT) for oropharyngeal or nasopharyngeal cancer.

Non-contrast-enhanced CT imaging was performed using a 32-slice scanner (Toshiba Aquilion LB, Toshiba Medical Systems Europe, Zoetermeer, the Netherlands) with the following parameters: 120 kVp, slice thickness ranging from 2 to 5 mm, and pixel size between 1.07 and 1.17 mm. In addition, patients underwent scanning with a helical tomotherapy system (Hi-Art II Tomotherapy System, Tomotherapy Inc., Madison, Wisconsin, USA), which employed a 6 MV radiotherapy linear accelerator (LINAC) capable of acquiring megavoltage CT (MVCT) images for daily patient setup verification. The imaging beam, generated by the same LINAC used for treatment delivery, had a nominal energy of 3.5 MV. MVCT images were acquired with a slice thickness between 2 and 5 mm and a pixel size of 0.75 mm.

[1] Centro di Riferimento Oncologico di Aviano IRCCS, Via F.Gallini 2, Aviano (PN), 33081, Italy

For each patient, both kVCT and MVCT images were collected. The kVCT images featured a matrix size of 512×512 in the axial plane, a pixel size of $1.074\,\text{mm} \times 1.074\,\text{mm}$, and a slice thickness of 2 mm. The MVCT images also had a matrix size of 512×512 in the axial plane, with a pixel size of $0.754\,\text{mm} \times 0.754\,\text{mm}$ and a slice thickness of either 2 mm or 4 mm.

3.2.2 Dataset Generation

The dataset generation process begins with pixel alignment of kVCT and MVCT images, a crucial step to ensure accurate supervised learning by maintaining spatial correspondence between inputs and outputs (Section 3.2.2.1). Following alignment, images undergo preprocessing and normalization to enhance contrast and stabilize intensity distributions, facilitating smoother convergence during training (Section 3.2.2.2). Finally, different datasets are constructed, distinguishing between images containing only metal artifacts and those with and without artifacts, allowing the network to learn both artifact reduction and general tissue contrast preservation effectively (Section 3.2.2.3).

3.2.2.1 Image alignment and Pre-processing

The primary objective of this step is to create a dataset with spatially aligned kVCT and MVCT images. Although both image volumes originate from the same patient and share the same reference system (with identical origin points), pixel-level misalignments occur due to slight variations in patient positioning and differences in image resolution. To achieve precise alignment, the following steps were performed using 3D Slicer (version 5.6.1) [64]:

1. Load the kVCT and MVCT volumes into 3D Slicer.
2. Resample the kVCT volume using the *Resample Scalar Volume* module with bSpline interpolation. This generates a resampled kVCT volume (*kVCTResampled*) matching the MVCT resolution.
3. Due to the fact that the previous step makes the *kVCTResampled* volume bigger in size, we crop the volume so we keep the original size 512×512 pixels. This is done using *Resample Scalar Vector* module of 3D Slicer.
4. Apply the *Elastix* module [113] to the MVCT volume to achieve precise alignment between the kVCT and MVCT datasets.

This alignment process is critical for the subsequent neural network training. By ensuring that the spatial dimensions and features of the kVCT and MVCT volumes are aligned, the network can establish correspondences between the input and ground truth data. This alignment enhances the network's ability to learn meaningful spatial features and improves the effectiveness of artifact reduction.

3.2.2.2 Pre-processing

Once the images have been aligned, the next step is to prepare and annotate the data for training with the neural network. This involves differentiating and annotating the body region

corresponding to each slice, allowing us to focus more on the head region, which is the area of interest for the task at hand. Additionally, we annotate the images to distinguish between those with and without artifacts by using a thresholding method. This helps in identifying and isolating the regions affected by artifacts. Finally, we normalize the images to ensure consistent input for the neural network, improving the training process and model performance. The following sections will detail each of these steps.

Region Categorization After alignment, the slices of each kVCT and MVCT volume are manually divided into three regions: head, neck, and body. In Figure 3.1a, an example of the sagittal plane of a MVCT (Megavoltage Computed Tomography) scan is shown. The regions labeled as head, neck, and body are represented between the horizontal lines. The head region is defined from the cranial cavity to the chin, while the neck region extends from the chin to the shoulders. The body region is excluded from the analysis, as the focus is on mitigating artifacts caused by metal implants in the teeth area.

Artifact Identification To differentiate slices with metal-induced artifacts from artifact-free slices, thresholds are applied to the intensity values. Artifacts in kVCT images are defined as intensities exceeding 2000 Hounsfield Units (HU), while for MVCT images, the threshold is set at 1000 HU. These thresholds were determined through visual inspection and recommendations from prior studies [114, 115].

Normalization The normalization process is essential for stabilizing neural network training and ensuring that intensity values across kVCT and MVCT images remain within a comparable range. To achieve this, we first set a lower threshold of -1000 HU to represent air, ensuring that all intensity values below this are clamped to -1000 HU. Similarly, upper thresholds are applied to kVCT (2000 HU) and MVCT (1000 HU) to prevent extreme values from distorting the images due to metal artifacts. Any pixel exceeding these upper limits is set to the corresponding threshold.

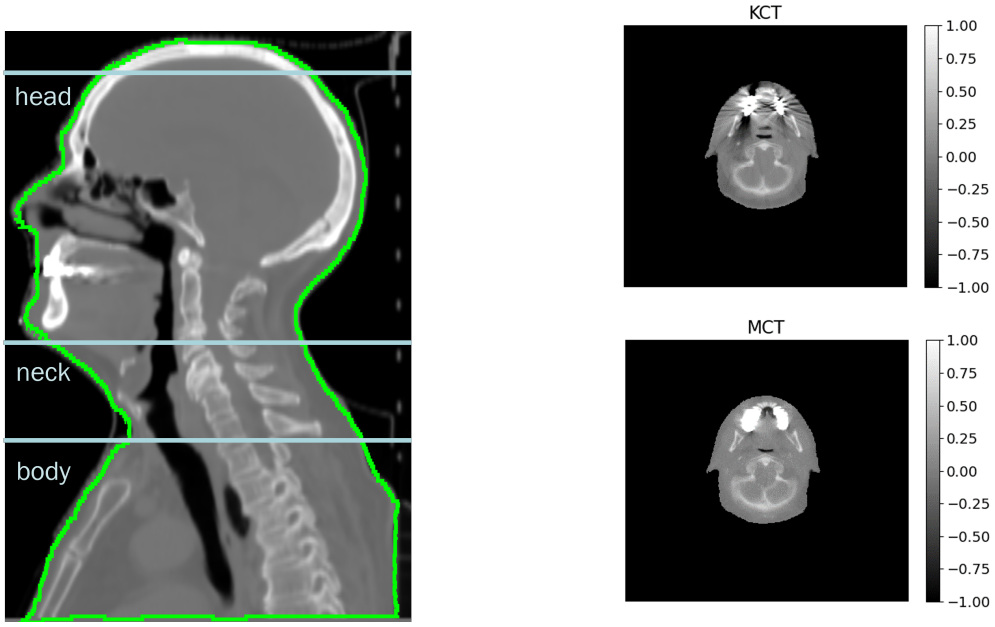
After applying these limits, we normalize the intensity values to the range $[-1, 1]$, which helps improve numerical stability and convergence during training. Finally, using clinician-provided body segmentations, as depicted in green in Figure 3.1a, background regions (outside the body) are set to -1, ensuring consistency across slices and reducing irrelevant intensity variations that could impact the model's performance. This process allows the network to focus on meaningful anatomical structures while minimizing the influence of background noise.

The result of this pre-processing step can be seen in Figure 3.1b, which shows, in the upper part, the kVCT (KCT), and in the lower part, the MVCT (MCT) after the pre-process step, displaying the same aligned slice. The background is assigned a value of -1, and the brightest region, corresponding to the metal in the teeth, is assigned a value of 1.

3.2.2.3 Dataset Organization

To train and evaluate the proposed model, two datasets are constructed:

- $\mathscr{D}_{\text{All}}$: This dataset includes all CT slices from the head to the neck region, encompassing both artifact-free and artifact-contaminated slices.



(a) **(b)**

Figure 3.1: (a) Sagittal view of an MVCT volume, where the green segmentation delineates the body region. (b) Top: kVCT slice exhibiting streak artifacts caused by the presence of a metal implant. Bottom: Corresponding MVCT slice, showing the absence of streak artifacts in the implant region. Both images have undergone preprocessing, including alignment and normalization, ensuring consistency for subsequent analysis. Image from [116].

- $\mathscr{D}_{\text{Art}}$: This dataset contains only CT slices affected by metal-induced artifacts. Artifact-contaminated slices constitute 14.78% of the total dataset.

Each dataset is further split into three subsets: 70% of the patients are allocated for training ($(\mathscr{D}^{Tr}_{\text{All}})$, $(\mathscr{D}^{Tr}_{\text{Art}})$), 20% for validation ($(\mathscr{D}^{Val}_{\text{All}})$, $(\mathscr{D}^{Val}_{\text{Art}})$), and 10% for testing ($(\mathscr{D}^{Ts}_{\text{All}})$, $(\mathscr{D}^{Ts}_{\text{Art}})$). In Table 3.1, the composition of each dataset for training, validation, and testing can be seen. Specifically, it shows the number of patients in each set and the number of slices (images) with and without artifacts.

Table 3.1: Number of patients and slices (images) in the acquired dataset.The head and neck region include the artifact slices since we work with artifacts caused by metallic dental implants.

Set	Number of patients	Number of slices of the head and neck regions	Number of slices with artifacts
Train	36	3858	560
Validation	10	1031	153
Test	6	580	96

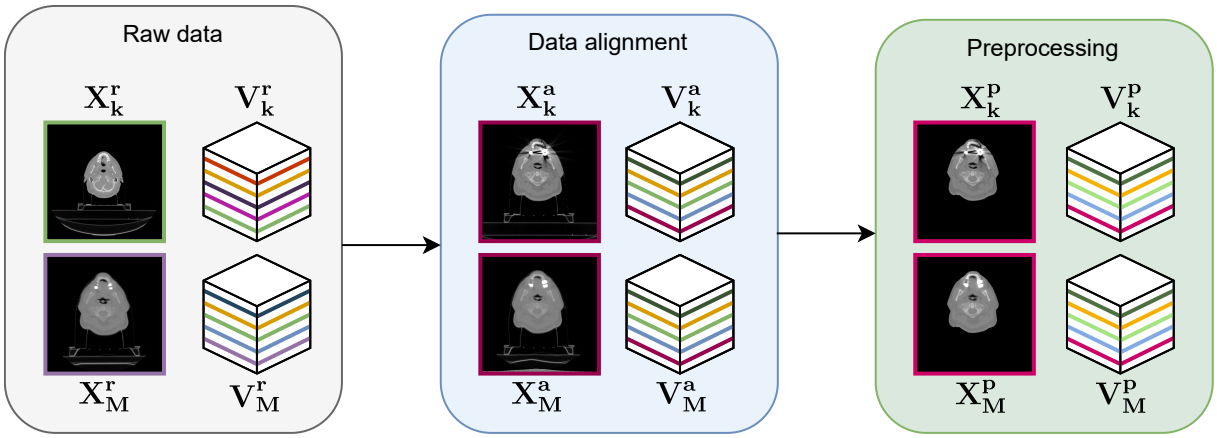


Figure 3.2: Steps followed for dataset generation. We start with raw and unaligned kVCT and MVCT volumes –slices (lines in the cube) do not correspond. Then, volumes are pixel-aligned and so the slices correspond. Finally, corresponding slices in kVCT and MVCT volumes are normalized and masked (Section 3.2.2). Image from [116].

3.2.3 Methodological Framework

To formalize the methodology, let $\mathbf{m}$ denote the imaging modality, where $\mathbf{k}$ corresponds to kVCT and $\mathbf{M}$ to MVCT. For a given slice ($\mathbf{X}$), $\mathbf{X}^{\mathbf{r}}_{\mathbf{m}} \in \mathbb{R}^{d\times d}$ represents the raw image ($\mathbf{r}$) with resolution $d = 512$. The full volume for a patient ($\mathbf{V}$), consisting of n slices, is denoted as $\mathbf{V}^{\mathbf{r}}_{\mathbf{m}} \in \mathbb{R}^{d\times d\times n}$.

After the alignment process described in Section 3.2.2, the raw volumes are transformed into aligned ($\mathbf{a}$) volumes $\mathbf{V}^{\mathbf{a}}_{\mathbf{k}}$ and $\mathbf{V}^{\mathbf{a}}_{\mathbf{M}}$ through Equation 3.1.

$$\mathbf{V}^{\mathbf{a}}_{\mathbf{k}}, \mathbf{V}^{\mathbf{a}}_{\mathbf{M}} = \text{alignment}(\mathbf{V}^{\mathbf{r}}_{\mathbf{k}}, \mathbf{V}^{\mathbf{r}}_{\mathbf{M}}). \tag{3.1}$$

Following alignment, each slice within these volumes undergoes preprocessing (see Section 3.2.2) to produce $\mathbf{X}^{\mathbf{p}}_{\mathbf{m}}$, following Equation 3.2.

$$\mathbf{X}^{\mathbf{p}}_{\mathbf{m}} = \text{preprocess}(\mathbf{X}^{\mathbf{a}}_{\mathbf{m}}), \quad \forall \mathbf{X}^{\mathbf{a}}_{\mathbf{m}} \in \mathbf{V}^{\mathbf{a}}_{\mathbf{m}}. \tag{3.2}$$

The preprocessed slices $\mathbf{X}^{\mathbf{p}}_{\mathbf{k}}$ and $\mathbf{X}^{\mathbf{p}}_{\mathbf{M}}$ form the dataset used in the experiments. A summary of the raw, aligned, and preprocessed data flow is illustrated in Figure 3.2.

The input to the proposed model is a preprocessed kVCT slice, $\mathbf{X}^{\mathbf{p}}_{\mathbf{k}}$, while the ground truth corresponds to the aligned and preprocessed MVCT slice, $\mathbf{X}^{\mathbf{p}}_{\mathbf{M}}$. The model's output is the domain-transferred kVCT-to-MVCT slice, denoted as $\hat{\mathbf{X}}_{\mathbf{M}} \in \mathbb{R}^{d\times d}$.

3.2.4 Network architectures

We develop a Metal Artifact Reduction using Domain Transformation Network (MAR-DTN), which draws inspiration from the U-Net framework [76] (see Section 2.3.1.1). The U-Net architecture has proven to be highly effective in various pixel-to-pixel tasks, including segmentation, denoising, and metal artifact reduction (MAR) in medical imaging [76, 117, 118]. The MAR-DTN architecture progressively refines spatial features using layers with 64, 128, and 256 output channels, ensuring robustness for diverse imaging challenges such

as segmentation and denoising

To evaluate the performance of MAR-DTN, we conduct a comparative analysis against three state-of-the-art methods. The first method is pix2pix, a Conditional Generative Adversarial Network (cGAN) introduced by Isola et al. [85]. Pix2pix employs a generator-discriminator framework (see Section 2.3.1.7), where the discriminator learns to distinguish between real MVCT images and those generated by the generator. Over successive epochs, the generator aims to produce increasingly realistic images to fool the discriminator. As an additional variant, we modify pix2pix by replacing its generator with MAR-DTN, resulting in a hybrid model referred to as custom-pix2pix.

We also include a transformer-based architecture, SwinIR [93], known for its robust performance in pixel-to-pixel image tasks. SwinIR is composed of three main modules: shallow feature extraction, deep feature extraction, and high-quality image reconstruction. The deep feature extraction module utilizes multiple Residual Swin Transformer Blocks (RSTB), each comprising Swin Transformer layers and residual connections (see Section 2.3.2.4). This hierarchical structure enables SwinIR to effectively capture local and global image features, making it well-suited for MAR tasks.

Finally, we incorporate the INet architecture [119], originally designed for medical image segmentation. INet maintains spatial information by avoiding downsampling and instead enlarges receptive fields through progressive increases in kernel sizes (e.g., from 3×3 to 7×7, and subsequently to 15×15). In our adaptation, the final activation layer is omitted, enabling INet to generate high-quality images rather than segmentation masks. By concatenating feature maps from all preceding layers, INet effectively fuses multilevel semantics, enhancing its optimization capability for generating artifact-free images.

This diverse selection of neural network architectures ensures a comprehensive evaluation of MAR-DTN's performance relative to contemporary methods, providing valuable insights into its strengths and potential applications.

3.2.5 Loss Functions

To effectively tackle the challenge of Metal Artifact Reduction (MAR) with neural networks, we selected a combination of loss functions that offer distinct advantages for enhancing image quality. The weighted L_1 loss ($\mathscr{L}_1^{\mathbf{w}}$) is utilized for its ability to emphasize important regions in the image, while the Focal Frequency Loss (FFL) ($\mathscr{L}_{\text{FFL}}^{\beta,\alpha}$) focuses on reducing frequency-related artifacts. Additionally, Mean Squared Error (MSE) ($\mathscr{L}_{\text{MSE}}$) provides a straightforward measure of pixel-wise differences, and the Structural Similarity Index (SSIM) ensures that perceptual quality is preserved. Finally, the Multi-Scale Structural Similarity Index (MS-SSIM) enhances the assessment of image similarity across multiple scales. Together, these loss functions are strategically chosen to address various aspects of MAR, optimizing both the fidelity and quality of the generated images. Detailed descriptions of each loss function can be found in Section 2.3.5.1.

3.2.6 Implementation Details

All networks were designed to accept input and generate output of size 512×512, corresponding to $\mathbf{X}_{\mathbf{k}}^{\mathbf{p}}$ and $\mathbf{X}_{\mathbf{M}}^{\mathbf{p}}$. Optimization was performed using the Adam optimizer [120], with a learning rate of 0.001 and weight decay of 5×10^{-4}.

The batch size was set to 4 for most networks, except SwinIR, which used a batch size of 2 due to memory constraints. Training was conducted for 20 epochs, with early stopping applied after 5 epochs of no improvement.

Data augmentation, following [121], included horizontal flipping ($p = 0.5$) and transformations like shifting, scaling, and rotation ($p = 0.8$). Parameters were set as follows: *shift_limit* = 0.0625, *scale_limit* = 0.1, and *rotate_limit* = 5. These augmentations introduced variability to mitigate data limitations and improve model generalization.

Models were trained on an Intel Xeon Server with 188GB of RAM and an Nvidia A100 GPU.

3.2.7 Evaluation Metrics

The chosen metrics for evaluation are the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM), which are particularly suited for image generation tasks. PSNR quantifies the fidelity of pixel intensities between generated and ground truth images, providing insight into reconstruction accuracy. SSIM, on the other hand, evaluates perceptual similarity by accounting for luminance, contrast, and structural patterns, reflecting human visual perception. Together, these metrics offer a balanced assessment, capturing both numerical accuracy and perceptual quality, which is vital for medical image artifact reduction. More information about this metrics can be found in Section 2.3.6.1.

Both metrics are calculated within the segmented body region, excluding background areas. To ensure the reliability of the results, we mask the loss computation to exclude noisy regions, focusing solely on meaningful anatomical structures within the segmentation. This approach reduces the influence of irrelevant background values, enhancing the accuracy of our evaluation.

3.3 Results

We begin by conducting an ablation study to identify the optimal parameter combination of loss functions, specifically $\mathscr{L}_1$ and $\mathscr{L}_{\text{FFL}}$ (Section 3.3.1). To ensure a fair and unbiased evaluation of the loss functions' impact on competitive networks, INet is excluded from this comparison. Including INet could skew the analysis and introduce biases, detracting from an accurate assessment of the loss functions' effects.

After identifying the best parameter configuration, we expand our analysis to include a comprehensive comparison of all networks using various loss function combinations (Section 3.3.2). Finally, we perform a thorough evaluation of the proposed approach in comparison to state-of-the-art methods to validate its efficacy in reducing metal artifacts in medical imaging (Section 3.3.3).

3.3.1 Ablation Study

To determine the optimal configuration of the loss functions for our MAR-DTN network, we first examine the impact of weighted $\mathscr{L}_1^{\mathbf{w}}$ loss on slices containing metal artifacts. Weight assignments are informed by clinician-provided body segmentations (see green body segment in Figure 3.1a). The weights $\mathbf{w}[i, j]$ are set to 0.1 outside the body segment, while for slices with artifacts, they vary within $\{1, 25, 50, 100\}$ inside the body segment. The chosen weight values are designed to progressively increase the model's focus on regions affected by metal artifacts.

For slices without artifacts, the weight is uniformly 1 within the body segment. In subsequent sections, we simplify the notation by using w to denote the weight within the body segment, e.g., $\mathscr{L}_1^{100}$ refers to a weight of 100.

We also investigate the parameters β and α of $\mathscr{L}_{FFL}^{\beta,\alpha}$, varying them across the set $\{0.5, 1, 1.5\}$ [96]. This exploration facilitates an understanding of how different parameter settings influence the network's capacity to mitigate artifacts while preserving image quality.

3.3.1.1 $\mathscr{L}_1^{\mathbf{w}}$ Analysis

To address the imbalance in the dataset, where only 6% of slices in $\mathscr{D}_{\text{All}}$ contain artifacts (see Table 3.1), the L_1 loss function was modified to assign greater weight to artifact-containing slices, as described in Section 3.2.5. This modification, denoted as $\mathscr{L}_1^{\mathbf{w}}$, prioritizes artifact reduction through weight adjustments within the body segment based on w values.

Figures 3.3a and 3.3b present the PSNR and SSIM metrics, respectively, for all networks trained using the weighted $\mathscr{L}_1^{\mathbf{w}}$ loss function with weight values $w \in \{1, 25, 50, 100\}$. The x-axis represents the different models evaluated, while the y-axis corresponds to the respective metric values. These figures compare performance across the full dataset, $\mathscr{D}_{\text{All}}$, (dots) and the subset of slices affected by metal artifacts, $\mathscr{D}_{\text{Art}}$, (bars), providing insight into how different weighting strategies influence artifact correction.

Despite varying w, the metrics exhibit limited variability. PSNR differences do not exceed 3 dB, and SSIM fluctuates by less than 10%. For MAR-DTN, increasing w beyond 25 improves artifact set results, whereas SwinIR achieves optimal performance when $w = 1$. Interestingly, for custom-pix2pix, results stabilize with $w > 1$, while for pix2pix, the impact of w is negligible. Notably, increasing w enhances artifact-specific results for custom-pix2pix, mirroring MAR-DTN's positive trend.

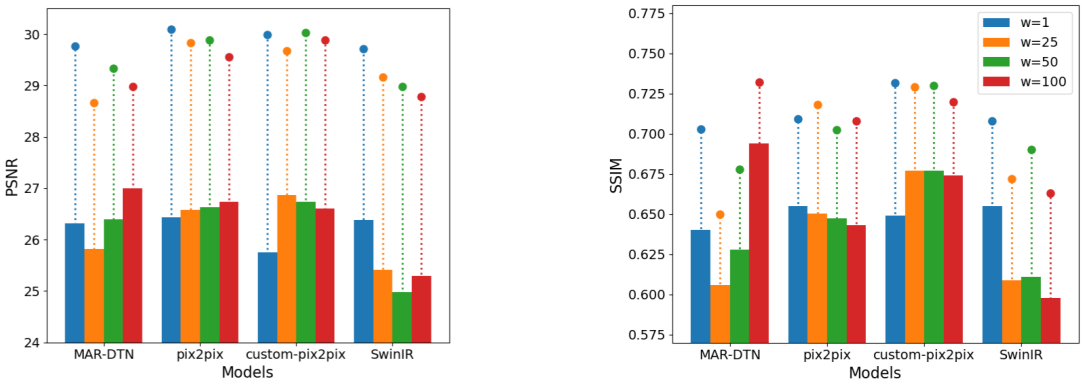
Examining $\mathscr{D}_{\text{All}}$, represented by dots in Figures 3.3a and Figure 3.3b, results are slightly lower when assigning higher weights to $\mathscr{D}_{\text{Art}}$. This trade-off is acceptable, given the study's focus on artifact reduction. Based on these findings, $w = 100$ is selected for subsequent experiments as it maximizes MAR-DTN performance and provides comparable outcomes for other networks.

3.3.1.2 $\mathscr{L}_{FFL}^{\beta,\alpha}$ Analysis

To identify the optimal parameter combination for $\mathscr{L}_{FFL}^{\beta,\alpha}$, the parameters β and α were varied within the set $\{0.5, 1, 1.5\}$.

Figures 3.4a and 3.4b present heatmaps representing the mean values of PSNR and SSIM, respectively, evaluated on the test dataset $\mathscr{D}_{\text{All}}^{Ts}$ after training the networks using the $\mathscr{L}_{FFL}^{\beta,\alpha}$ loss function with various combinations of the parameters α and β. In these heatmaps, the x-axis represents the values of α, while the y-axis represents the values of β. Each cell in the heatmap corresponds to the mean of eight values: the first four represent the PSNR or SSIM scores evaluated on artifact-affected slices from $\mathscr{D}_{\text{Art}}^{Ts}$, while the last four correspond to the same metric computed over the full dataset $\mathscr{D}_{\text{All}}^{Ts}$. These results are reported for the four networks studied: MAR-DTN, pix2pix, custom-pix2pix, and SwinIR, allowing us to analyze how different parameter choices influence model performance.

Similar to the $\mathscr{L}_1^{\mathbf{w}}$ study, the results exhibit limited variability, with PSNR differences below 2 dB and SSIM variations under 10%. Notably, increasing the value of α generally leads



(a) **(b)**

Figure 3.3: Bar plots with the mean metric values evaluated on the test dataset after training the networks using the $\mathscr{L}_1^w$ loss function. (a) PSNR. (b) SSIM. The dots represent the mean value of all slices in the dataset, while the bars represent the mean value of slices with artifacts. Values obtained using the four considered networks (MAR-DTN, pix2pix, custom-pix2pix and SwinIR) trained on $\mathscr{D}_{\text{All}}^{Tr}$. Image from [116].

to a decrease in performance metrics.

The best PSNR value, 27.81 dB, is achieved when $\alpha = 0.5$ and $\beta = 1$. For the same α value, the mean SSIM is 0.64, which is close to the highest observed SSIM value of 0.65. Based on these results, we conclude that the optimal parameter combination is $\beta = 1$ and $\alpha = 0.5$.

3.3.2 Comparative Performance Evaluation of Networks with Different Loss Function Combinations

This section presents a comparison of various network architectures when trained using different combinations of loss functions. We begin by fixing the optimal values for both $\mathscr{L}_1^w$ and $\mathscr{L}_{FFL}^{\beta,\alpha}$ parameters, followed by a performance comparison across networks.

Table 3.2 presents a comparative analysis of the performance results for different neural networks and loss function combinations. The table includes a check mark to indicate which sum of loss functions were used for training each network. For the pix2pix networks, it specifies the loss function applied to the generator. The dataset column indicates whether the network was trained and evaluated on the $\mathscr{D}_{\text{All}}$ or $\mathscr{D}_{\text{Art}}$ dataset. When the dataset is $\mathscr{D}_{\text{All}}$, the model is trained on $\mathscr{D}_{\text{All}}^{Tr}$ and tested on $\mathscr{D}_{\text{All}}^{Ts}$, whereas for $\mathscr{D}_{\text{Art}}$, the model is trained on $\mathscr{D}_{\text{Art}}^{Tr}$ and evaluated on $\mathscr{D}_{\text{Art}}^{Ts}$. The table then reports the PSNR and SSIM values obtained for the test sets. For the $\mathscr{D}_{\text{All}}$ dataset, the results are shown separately for artifact slices from $\mathscr{D}_{\text{Art}}^{Ts}$ and for the full dataset ($\mathscr{D}_{\text{All}}^{Ts}$) in parentheses. Underlined values highlight the highest performance for each network with specific loss function combinations, while the highlighted values indicate the best overall performance across all configurations.

The $\mathscr{L}_1^{100}$ loss function yields the highest performance on $\mathscr{D}_{\text{Art}}$, achieving a PSNR of 27.17 dB for MAR-DTN, and the second-best result for pix2pix at 26.31 dB. However, performance drops by almost 2 dB for both custom-pix2pix and SwinIR.

We also evaluated the $\mathscr{L}_{FFL}^{1,0.5}$ loss function independently. While it resulted in slightly lower performance than $\mathscr{L}_1^{100}$, with a PSNR reduction of up to 2 dB for pix2pix, custom-pix2pix still

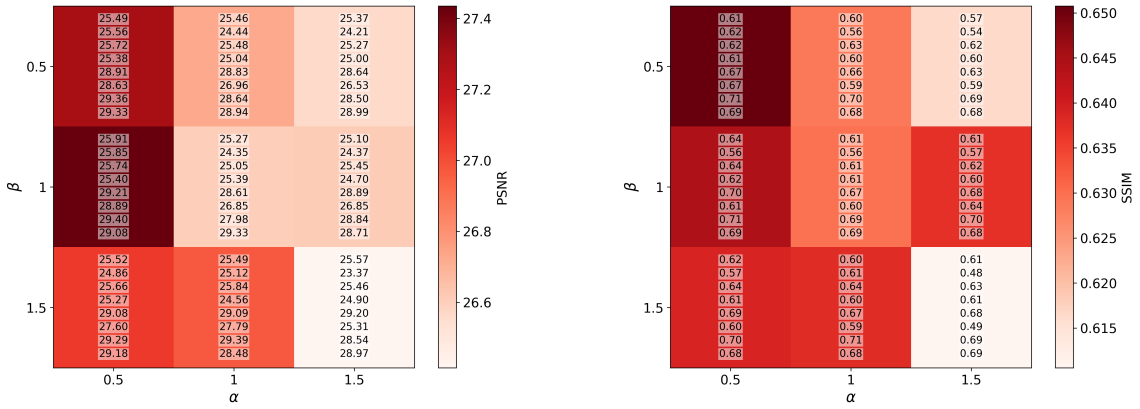


Figure 3.4: Heatmaps with the mean metric values evaluated on the test dataset after training the networks using the $\mathscr{L}_{FFL}^{\beta,\alpha}$ loss function with various combinations of the parameters α and β (x and y-axis, respectively). (a) PSNR. (b) SSIM. Each cell represents the mean of 8 values, the first 4 corresponding to the parameter value evaluated on $\mathscr{D}_{\text{Art}}^{Ts}$, and the last 4 corresponding to the parameter value evaluated on the $\mathscr{D}_{\text{All}}^{Ts}$, for each neural network in the study, MAR-DTN, pix2pix, custom-pix2pix, and SwinIR, respectively, trained on $\mathscr{D}_{\text{All}}^{Tr}$. Image from [116].

maintained a PSNR of 26.15 dB, which is comparable to other configurations. A significant improvement was seen when $\mathscr{L}_{MS-SSIM}$ replaced $\mathscr{L}_{SSIM}$, particularly with custom-pix2pix. The most complex loss function combination ($\mathscr{L}_1^{100} + \mathscr{L}_{MS-SSIM} + \mathscr{L}_{FFL}^{1,0.5}$) caused noise during training and did not outperform simpler loss functions. Nevertheless, SwinIR achieved the highest PSNR value of 26.42 dB.

Reconstruction examples can be seen in Figure 3.5. The first row displays the preprocessed kVCT and MVCT images, which serve as the ground truth. The first column of the figure lists the different loss functions used, while the subsequent columns show the corresponding results from each model. All networks were trained using the $\mathscr{D}_{\text{Art}}^{Tr}$ dataset. For each reconstructed slice, the PSNR and SSIM values are provided to evaluate the quality of the reconstruction.

When trained with $\mathscr{D}_{\text{All}}^{Tr}$, the inclusion of additional loss functions improved the results for certain architectures. The best combination for MAR-DTN and custom-pix2pix was $\mathscr{L}_1^{100} + \mathscr{L}_{SSIM}$, which reached a PSNR of 30.02 dB. For pix2pix and SwinIR, the optimal combination was $\mathscr{L}_1^{100} + \mathscr{L}_{MSE}$, resulting in PSNRs of 28.92 and 29.39 dB, respectively.

For slices with artifacts, the PSNR values were generally lower when trained on $\mathscr{D}_{\text{All}}^{Tr}$, a result of the unbalanced dataset, indicating the loss functions were not entirely effective at handling class imbalance.

Finally, the results obtained using INet are presented in the last column of Table 3.2, with a maximum PSNR of 12.67 dB for the artifact dataset. This demonstrates INet's underperformance compared to the other architectures. A similar trend was observed for SSIM, with a peak value of just 8%. Figure 3.5 illustrates INet's results, where the artifacts not only persist but gain higher contrast along with the image, with additional artifacts appearing and distorting other structures, such as visible streaks in the grid when using the loss combination $\mathscr{L}_1^{100} + \mathscr{L}_{MS-SSIM} + \mathscr{L}_{FFL}^{1,0.5}$.

Table 3.2: Comparative analysis for different networks and loss function combinations, indicated with a check mark which sum of loss functions have been used for training. For the pix2pix networks, it indicates the loss function of the generator. The dataset column indicates the dataset with which the network has been trained and evaluated; where dataset is $\mathscr{D}_{\text{All}}$ then model is trained on $\mathscr{D}_{\text{All}}^{Tr}$ and tested on $\mathscr{D}_{\text{All}}^{Ts}$, and in case of $\mathscr{D}_{\text{Art}}$ then model is trained on $\mathscr{D}_{\text{Art}}^{Tr}$, and tested on $\mathscr{D}_{\text{Art}}^{Ts}$. Finally, the remaining columns show the PSNR and SSIM values obtained for the test sets. Where the dataset is the $\mathscr{D}_{\text{All}}$, we report both on the performance obtained on artifact slices from within the $\mathscr{D}_{\text{Art}}^{Ts}$, and the mean of PSNR and SSIM on whole dataset $\mathscr{D}_{\text{All}}^{Ts}$ (in parentheses). Underlined values indicate the highest performance for each network with certain loss function combinations, while highlighted values indicate the highest overall performing model across all configurations. Table from [116].

Loss combination					Dataset	MAR-DTN		pix2pix [85]		custom-pix2pix		SwinIR [93]		INet [119]	
$\mathscr{L}_1^{100}$	$\mathscr{L}_{SSIM}$	$\mathscr{L}_{MS\text{-}SSIM}$	$\mathscr{L}_{MSE}$	$\mathscr{L}_{FFL}^{1.0.5}$		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
					$\mathscr{D}_{\text{Art}}$	27.17	0.69	26.31	0.64	25.24	0.68	25.46	0.61	11.61	0.04
✓					$\mathscr{D}_{\text{All}}$	26.99 (28.97)	0.71 (0.73)	26.36 (28.7)	0.63 (0.69)	26.61 (29.08)	0.67 (0.7)	25.29 (28.79)	0.59 (0.66)	12.02 (12.03)	0.04 (0.04)
					$\mathscr{D}_{\text{Art}}$	27.11	0.69	26.21	0.63	26.98	0.71	26.16	0.62	10.94	0.08
✓	✓				$\mathscr{D}_{\text{All}}$	27.09 (30.02)	0.69 (0.73)	26.39 (28.58)	0.64 (0.68)	27.13 (29.85)	0.68 (0.70)	24.90 (28.97)	0.68 (0.68)	12.01 (12.27)	0.03 (0.03)
					$\mathscr{D}_{\text{Art}}$	27.46	0.69	26.32	0.65	27.06	0.67	26.26	0.63	11.96	0.01
✓		✓			$\mathscr{D}_{\text{All}}$	27.08 (29.97)	0.69 (0.73)	26.37 (28.69)	0.58 (0.68)	27.04 (29.35)	0.64 (0.70)	26.25 (29.39)	0.63 (0.67)	12.67 (12.95)	0.05 (0.04)
					$\mathscr{D}_{\text{Art}}$	26.94	0.68	26.25	0.64	26.55	0.64	26.18	0.63	11.70	0.02
✓			✓		$\mathscr{D}_{\text{All}}$	27.11 (29.89)	0.69 (0.72)	26.42 (28.92)	0.68 (0.7)	26.98 (29.22)	0.64 (0.68)	25.24 (29.00)	0.60 (0.66)	11.23 (10.96)	0.01 (0.01)
					$\mathscr{D}_{\text{Art}}$	26.35	0.64	24.03	0.51	26.15	0.61	24.58	0.59	9.05	0.08
				✓	$\mathscr{D}_{\text{All}}$	26.52 (29.51)	0.66 (0.70)	25.85 (28.88)	0.56 (0.61)	26.02 (29.04)	0.60 (0.69)	25.39 (29.33)	0.61 (0.69)	10.03 (10.01)	0.02 (0.01)
					$\mathscr{D}_{\text{Art}}$	27.06	0.69	25.66	0.63	26.66	0.60	25.40	0.61	11.95	0.03
✓				✓	$\mathscr{D}_{\text{All}}$	26.99 (29.85)	0.69 (0.72)	25.99 (28.56)	0.59 (0.65)	26.48 (29.06)	0.62 (0.69)	25.55 (29.18)	0.60 (0.68)	11.54 (11.86)	0.08 (0.07)
					$\mathscr{D}_{\text{Art}}$	27.08	0.69	25.66	0.63	26.18	0.56	26.42	0.64	9.58	0.05
✓	✓			✓	$\mathscr{D}_{\text{All}}$	25.65 (28.65)	0.63 (0.69)	26.41 (28.74)	0.68 (0.64)	27.08 (28.04)	0.64 (0.68)	25.45 (28.79)	0.61 (0.66)	12.53 (12.21)	0.06 (0.05)

3.3.3 Comparison with State-of-the-Art Networks

We present a comparison between the proposed MAR-DTN architecture and other state-of-the-art networks considered in this study, as shown in Table 3.3. This table compares the best results obtained by MAR-DTN with those of other networks evaluated in the study. All networks were trained and evaluated on the complete dataset (both training and testing sets). The performance metrics considered are PSNR and SSIM, where the values in parentheses represent the average across the entire dataset, and the values above indicate the average for slices with artifacts.

MAR-DTN achieves the best performance on the *artifact* dataset, with a PSNR of 26.99 dB and an SSIM of 0.69 points. However, when evaluating the *all* dataset, custom-pix2pix outperforms MAR-DTN, reaching a PSNR of 29.88 dB while both custom-pix2pix and MAR-DTN achieve an SSIM of 0.73 points.

On the other hand, SwinIR exhibits a drop in performance across the *all* dataset, with a

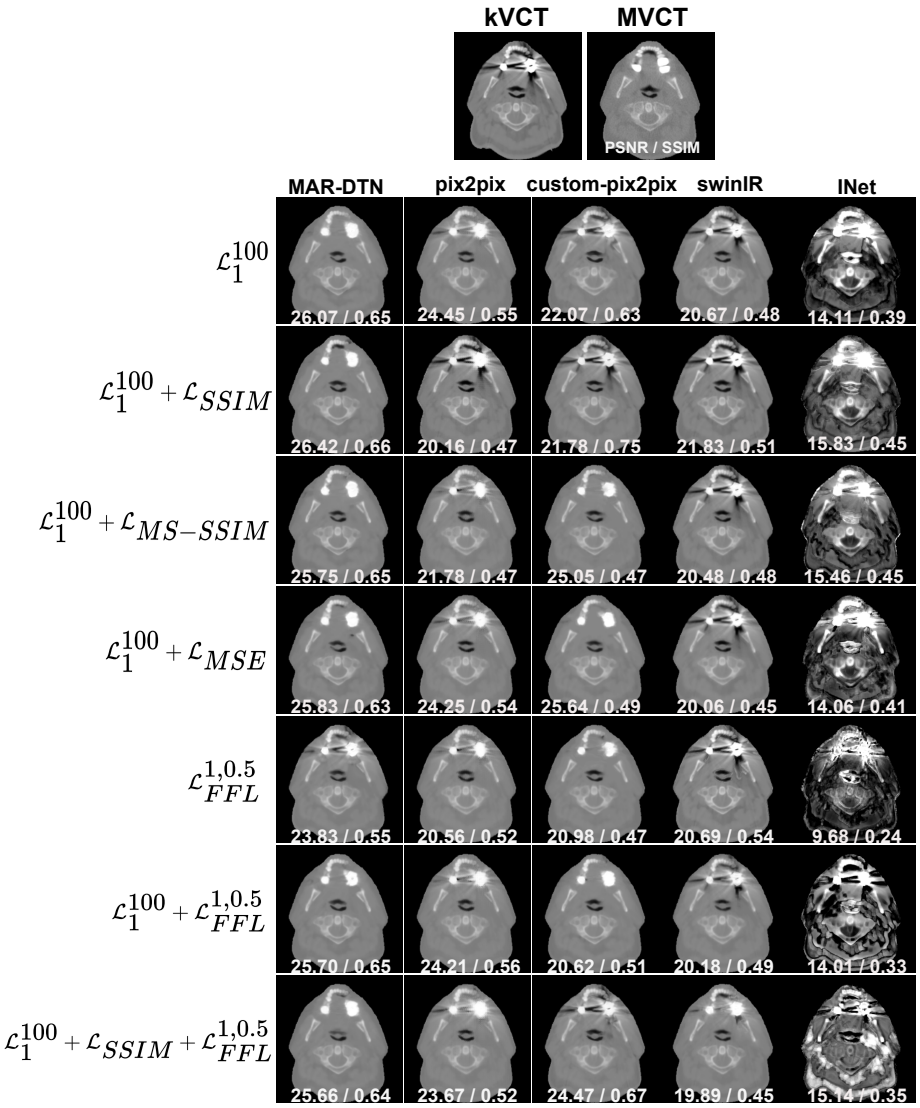


Figure 3.5: Reconstruction of a slice with artifacts by the different models and loss functions. First row shows preprocessed kVCT and MVCT (ground truth) images . First column indicates the loss function, and the following ones indicate the model used. Networks have been trained on the $\mathscr{D}_{\text{Art}}$. In each reconstructed slice, the PSNR and SSIM values are displayed. Image from [116].

PSNR decrease of 0.76 dB and an SSIM reduction of 0.07 points compared to MAR-DTN. This performance gap becomes more pronounced in the *artifact* dataset, where SwinIR's PSNR lags behind MAR-DTN by 1.7 dB.

Additionally, Table 3.4 provides technical details on the state-of-the-art networks, including the number of parameters, MACs, and the training and inference times.

Table 3.3: Comparative table between the best result of MAR-DTN and the state-of-the-art networks in the study. The networks have been trained and evaluated on *all* dataset (train and test, respectively). The metrics are PSNR and SSIM. In parentheses, the average value across the entire dataset is shown, with the average value for slices with artifacts above it. Table from [116].

Network	Loss Function	PSNR	SSIM
MAR-DTN [116]	$\mathscr{L}_1^{100} + \mathscr{L}_{SSIM}$	26.99 (28.98)	0.69 (0.73)
pix2pix [85]	$\mathscr{L}_1^1$	26.740 (29.55)	0.64 (0.71)
custom-pix2pix	$\mathscr{L}_1^1$	26.61 (29.88)	0.67 (0.73)
SwinIR [93]	$\mathscr{L}_1^1$	25.29 (28.79)	0.59 (0.66)

Table 3.4: Comparison of trainable parameters, number of multiplications and additions (MACs), training time computed for the $\mathscr{D}_{\text{All}}$ in 1 epoch and patient reconstruction time (in this case 170 slices) for state-of-the-art methods under study. Table from [116].

Network	Parameters (M)	MACs (G)	Training time (s)	Patient reconstruction time (s)
MAR-DTN	1.882	116.686	65.32	3.56
pix2pix	54.413	77.99	80.02	3.75
custom-pix2pix	4.646	123.277	67.42	4.25
SwinIR	1.614	425.034	2,774.76	47.27
INet	2.96	896.31	807.38	5.31

3.4 Discussion and Knowledge Transfer to Industry

Metal artifacts in CT imaging pose significant challenges to accurate diagnosis and clinical decision-making. Over the years, numerous methodologies have been proposed to mitigate these artifacts, aiming to enhance image quality and maintain diagnostic accuracy.

Among these, dataset-centric approaches, such as those introduced by Kaposi et al. [122], have facilitated advancements by providing annotated datasets for algorithm development. Other efforts have focused on leveraging deep learning techniques, ranging from image-to-image translation models like deep residual architectures [123] to interpretable dictionary-based networks [124]. Additionally, dual-domain methods, integrating both sinogram (a two-dimensional representation of the projections of an object taken at various angles during CT scanning) and image data [125, 126, 127], have demonstrated their efficacy in tackling the complexities of metal artifact reduction (MAR). They serve as a crucial intermediate step in the reconstruction of cross-sectional images, enabling the identification and correction of artifacts that can arise from metal objects in the imaging field [29].

Current solutions often involve computationally intensive models, limiting their practical implementation in real-time clinical settings. Moreover, the adaptability of existing techniques to specific clinical applications, such as cardiology, remains an underexplored area. Addressing these gaps, this work introduces MAR-DTN, a lightweight, dual-domain supervised learning framework optimized for MAR. The proposed method combines efficiency with state-of-the-art performance, paving the way for applications in resource-constrained medical environments.

Building on these advancements, our proposed MAR-DTN employs a dual-domain supervised learning framework, coupled with innovative loss function combinations, to address MAR challenges effectively. The performance of MAR-DTN was evaluated on two datasets: $\mathscr{D}_{\text{Art}}$, comprising artifact-affected images, and $\mathscr{D}_{\text{All}}$, containing both artifact and non-artifact images. These datasets provided a robust platform to test the adaptability and generalization of our method.

MAR-DTN consistently outperformed competing models across most configurations, particularly on $\mathscr{D}_{\text{All}}$. With the loss combination $\mathscr{L}_1^{100} + \mathscr{L}_{SSIM}$, MAR-DTN achieved the highest PSNR of 30.02 dB and an SSIM of 0.73, demonstrating its ability to minimize artifacts while preserving structural integrity. The performance remained competitive on the artifact-specific dataset $\mathscr{D}_{\text{Art}}$, achieving a PSNR of 27.46 dB and an SSIM of 0.69 with $\mathscr{L}_1^{100} + \mathscr{L}_{MS-SSIM}$.

Custom-pix2pix and pix2pix models delivered strong results, with custom-pix2pix marginally outperforming pix2pix in most cases. However, MAR-DTN consistently surpassed both models, particularly in PSNR values. SwinIR showed competitive SSIM values but lagged behind MAR-DTN, particularly on the artifact-inclusive dataset. INet, originally designed for segmentation tasks, exhibited the weakest performance, with a maximum PSNR of 12.67 dB and an SSIM of 0.08 points. Its limitations stem from its primary focus on segmentation rather than artifact reduction.

Compared to state-of-the-art MAR techniques, MAR-DTN demonstrates clear advantages. Dual-domain approaches like DuDoNet [125] have advanced MAR by integrating sinogram enhancement and image reconstruction, while iterative methods [128] and CNN-based frameworks [114] have leveraged MVCT images for improved kVCT artifact reduction. Despite these advancements, MAR-DTN achieves competitive results by optimizing loss function combinations and leveraging dual-domain processing, making it well-suited for complex artifact scenarios.

In summary, MAR-DTN emerges as a robust and effective solution for MAR, outperforming state-of-the-art models across artifact and mixed datasets. Its superior PSNR and SSIM values underscore its ability to enhance image quality while preserving structural details. Importantly, MAR-DTN is computationally light, making it highly suitable for real-time clinical applications where efficiency and speed are critical.

However, this study has limitations that must be acknowledged. The dataset size is relatively small, and the focus is solely on head CT images, which restricts the generalizability of the findings to other anatomical regions. As a preliminary step, this work sets the foundation for adapting the method to broader applications, particularly in the cardiology field, which is our primary objective. Future research must extend the dataset and explore the method's efficacy in complex cardiac imaging scenarios.

The Industrial PhD initiative emphasizes the direct transfer of knowledge from our research on metal artifact reduction (MAR) to practical applications in the medical imaging industry. A crucial outcome of this work has been the development of a comprehensive dataset that includes kVCT images with metal artifacts aligned with MVCT images without such distortions. This

fully aligned dataset serves as a vital resource for validating and refining machine learning models, establishing a strong foundation for further advancements within industrial settings.

To optimize the efficiency of this process, we have automated the dataset generation, which encompasses the alignment of volumes and the structured storage of images. This automation not only streamlines workflows but also guarantees consistency and accuracy, facilitating easier access to high-quality data for future investigations.

Additionally, our in-depth exploration of metal artifact formation from metallic implants has created a pathway for applying this knowledge to other imaging fields, such as cardiac imaging. The solutions developed for MAR are directly integrated into broader imaging practices, enhancing their effectiveness. By leveraging these insights, we aim to foster innovation across various imaging modalities, ultimately improving the quality of medical imaging and patient care.

Chapter 4

Automatic Coronary Artery Segmentation

Part of the text and figures for this chapter are reproduced from the following publications

(I) Serrano-Antón, B.[1,2,3], Otero-Cacho, A.[1,2,3], López-Otero, D.[4,5], Díaz-Fernández, B.[4,5], Bastos-Fernández, M.[4.5], Pérez-Muñuzuri, V.[3,6], González-Juanatey, J.R.[4,5], P. Muñuzuri, A.[2,3]. Coronary Artery Segmentation Based on Transfer Learning and UNet Architecture on Computed Tomography Coronary Angiography Images. In IEEE Access, vol. 11, pp. 75484-75496, 2023. IEEE. DOI: www.doi.org/10.1109/ACCESS.2023.3293090.

(II) Serrano-Antón, B.[1,2,3], Otero-Cacho, A.[1,2,3], López-Otero, D.[4,5], Díaz-Fernández, B.[4,5], Bastos-Fernández, M.[4.5], Massonis, G.[1], Pendón, S.[1], Pérez-Muñuzuri, V.[3,6], González-Juanatey, J.R.[4,5,7], P. Muñuzuri, A.[2,3]. Optimal Coronary Artery Segmentation Based on Transfer Learning and UNet Architecture. In: Wachinger, C., Paniagua, B., Elhabian, S., Li, J., Egger, J. (eds) Shape in Medical Imaging. ShapeMI 2023. Lecture Notes in Computer Science, vol 14350, pp. 55-64, 2023. Springer. DOI: https://doi.org/10.1007/978-3-031-46914-5_5.

(III) Serrano-Antón, B.[1,2,3], Insúa Villa, M.[1], Pendón Minguillón, S.[1], Paramés-Estévez, S.[1,2,3], Otero-Cacho, A.[1,2,3], López-Otero, D.[8,5], Díaz-Fernández, B.[4,5], Bastos-Fernández, M.[4,5], González-Juanatey, J.R.[4,5,7], P. Muñuzuri, A.[2,3]. Unsupervised clustering based coronary artery segmentation. In BioData Mining, vol. 18, no. 1, pp. 1-23, 2025. BioMed Central. DOI: https://doi.org/10.1186/s13040-025-00435-y.

[1] FlowReserve Labs S.L., Santiago de Compostela, 15782, Galicia, Spain.
[2] Galician Center for Mathematical Research and Technology (CITMAga), Santiago de Compostela, 15782, Galicia, Spain.
[3] Group of Nonlinear Physics, University of Santiago de Compostela, Santiago de Compostela, 15782, Galicia, Spain.
[4] Cardiology and Intensive Cardiac Care Department, University Hospital of Santiago de Compostela, Santiago de Compostela, 15706, Galicia, Spain.
[5] Centro de Investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), Madrid, 28029, Madrid, Spain.
[6] Institute CRETUS, Group of Nonlinear Physics, University of Santiago de Compostela, Santiago de Compostela, 15705, Galicia, Spain
[7] Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Santiago de Compostela, 15706, Galicia, Spain.
[8] Cardiology and Intensive Care Department, University Hospital of Pontevedra, Pontevedra, Galicia 36161, Spain

4.1 Introduction

The previous chapter focused on implementing a solution for metal artifact reduction in medical imaging, addressing a critical challenge in ensuring accurate visualization of anatomical structures. This analysis was crucial for enhancing the understanding of medical images, as artifacts can significantly impact the quality and reliability of segmentation. Building on this foundation, in this chapter, we shift our focus to the segmentation of coronary structures. While we do not directly employ the artifact reduction model from the previous chapter, the insights gained from that analysis contribute to a more robust interpretation of the imaging data, ultimately aiding in the development of accurate segmentation techniques.

The accurate segmentation of coronary arteries is a critical step in the diagnosis and treatment of Coronary Artery Disease (CAD). In Section 1.3.1 we discussed how Fractional Flow Reserve derived from Computed Tomography (FFR_{CT}) serves as a non-invasive alternative to conventional methods such as Invasive Coronary Angiography (ICA) and FFR measurement, which are both invasive and carry higher patient risks. Advances in medical imaging, particularly through techniques like coronary computed tomography angiography (CCTA), have revolutionized the visualization of coronary arteries [129]. These imaging modalities, enhanced with contrast agents, provide detailed anatomical views (Section 2.2.1). However, the need for precise 3D geometries extends beyond simple visualization; such reconstructions are pivotal for fluid dynamics simulations and comprehensive assessments of arterial pathologies, reducing reliance on invasive diagnostic procedures [43, 130, 131, 132].

Manual segmentation of these geometries, while feasible, is a labor-intensive and error-prone process. This underscores the importance of automation in this context, as automated methods not only save time but also enhance reproducibility and accuracy, both critical factors for clinical decision-making.

Several semi-automatic tools and platforms currently assist in this segmentation process, but they often require significant manual intervention and experimentation. Fully automatic methods, driven by Convolutional Neural Networks (CNNs), especially the U-Net architecture, have emerged as a cornerstone for medical image segmentation. Despite their success, these methods face challenges such as the need for large annotated datasets and computational resources. Transfer learning and pretraining offer potential solutions but are often limited by the scarcity of pre-trained models specific to medical imaging domains.

To address these limitations, this chapter explores two complementary approaches to coronary artery segmentation. The first involves using supervised learning with U-Net-based architectures, which are highly effective for structured image segmentation tasks (Section 4.2). The second approach introduces an unsupervised method for refining the initial segmentations generated by the neural networks. This refinement leverages high-quality images obtained through a distinct preprocessing pipeline, enhancing segmentation precision while mitigating the reliance on annotated datasets (Section 4.3). Together, these methods aim to establish a robust workflow for accurate and efficient coronary artery segmentation.

4.2 Supervised Segmentation Using U-Net Based Architectures

This section presents the first approach to supervised segmentation of the coronary artery tree using neural networks. The primary objective is to achieve comprehensive and accurate segmentations of the coronary arteries from the original CT images, while also identifying the limitations of the proposed methodology. To accomplish this, we evaluate various backbone architectures based on U-Net and investigate the impact of using pretrained encoders versus randomly initialized networks.

Additionally, the study examines the impact of dataset size by training with different numbers of patient samples. The inclusion of the aorta in the segmentation task is also analyzed, as its large size compared to coronary arteries may introduce dataset imbalance. Finally, the segmentation performance is assessed across different regions of the coronary vessels—proximal, mid, and distal segments—to determine the networks' capability to accurately capture spatial variations within the coronary tree.

4.2.1 Methodology

The methodology is divided into several steps, including dataset preparation, network architecture design, training strategy, and postprocessing techniques, each of which is detailed in the subsections below. Section 4.2.1.1 focuses on the dataset preparation, where we describe how the data was collected, preprocessed, and annotated. Section 4.2.1.2 presents the neural network architectures, providing an overview of the models employed in this study. The dataset used for model training is discussed in Section 4.2.1.3, while Section 4.2.1.4 covers the implementation details of the network. In Section 4.2.1.5, we discuss how the coronary tree was separated into three distinct regions: proximal, middle, and distal. Section 4.2.1.6 details the postprocessing techniques applied to refine the results. Finally, Section 4.2.1.7 outlines the evaluation metrics used to assess the performance of the models.

4.2.1.1 Dataset

The dataset employed in this study consists of CT images from 88 patients collected at the University Hospital in Santiago de Compostela (Spain). For each patient, Coronary Computed Tomography Angiography (CCTA) scans consist of 256 images with a resolution of 512×512 pixels, and an in-plane spacing of 0.38×0.38 mm.

Patients were selected based on the clarity of the images and the absence of significant calcium-related lesions. In cases where minor calcifications ($< 5\%$ of patients) were present, calcium was manually removed from the vessels while preserving the blood flow regions. Further details on this dataset can be found in Section 2.2.1.1.

To optimize computational efficiency, images were cropped to a size of 400×400 pixels to exclude non-essential areas. The number of slices per patient was reduced from 256 to 200. Pixel intensities, measured in Hounsfield Units, were rescaled to values between 0 and 255 for normalization purposes.

As this is a supervised learning task, each CT volume was manually annotated to segment both the aorta and coronary arteries, following the protocol described in Section 2.2.1.1. Two datasets were generated: one containing the segmentation of the complete coronary anatomy,

including both the aorta and coronary arteries (A+C.A), and another focusing solely on the coronary arteries (C.A). The segmentation in this dataset is highly detailed, including fine and distal vessel structures, ensuring precise and comprehensive annotations for robust training and evaluation.

4.2.1.2 Neural Network Architectures

In this study, four different network architectures based on the U-Net design were implemented. These architectures were selected to explore various configurations of 2D and 3D networks and to assess the benefits of transfer learning. The choice of models aims to balance segmentation performance, computational efficiency, and suitability for clinical deployment.

The first approach consists of a standard 2D U-Net (referred to as 2D in the following), serving as a baseline for comparison. This model, trained from scratch, allows evaluation of segmentation performance without any pretraining influence. The network consists of four downsampling blocks (see Section 2.3.1.1) and takes input images of size 512×512, which requires cropped images (400×400 pixels) to be padded with black borders. This ensures compatibility with architectures that may require pretraining. The model comprises $7,760,322$ parameters.

To investigate the benefits of transfer learning, a second 2D U-Net variation incorporates a MobileNetV2 encoder (see Section 2.3.1.5) pretrained on the ImageNet dataset [133] (referred to as 2D MB2-Pre in the following). This lightweight and computationally efficient architecture was chosen to reduce the complexity of the model while maintaining competitive accuracy. By freezing the encoder weights, only the decoder is trained, resulting in a total of $6,502,786$ parameters, of which $4,658,882$ are trainable and $1,843,904$ are fixed.

As a more advanced alternative, a third architecture replaces MobileNetV2 with an EfficientNet-B5 encoder (see Section 2.3.1.6). This architecture, referred to as 2D Eff-Pre in the following, was selected for its improved feature extraction capabilities while maintaining a balance between computational efficiency and segmentation accuracy. Given that EfficientNet employs compound scaling to enhance its performance, the B5 variant was specifically chosen based on input image size requirements. This architecture includes $37,468,673$ total parameters, where $9,125,905$ are trainable, and $28,342,768$ remain fixed due to pretraining on ImageNet [82]. Its selection is particularly relevant for clinical applications, as it offers a powerful yet efficient solution for deployment in real-world settings.

Finally, to leverage spatial context across slices, a 3D U-Net was implemented (referred to as 3D in the following). This model processes volumes of 16 consecutive CCTA slices as input and output, enabling it to capture inter-slice spatial relationships that are lost in 2D networks. The architecture consists of seven downsampling blocks and a total of $22,575,329$ trainable parameters. Unlike its 2D counterparts, the input image size remains 400×400 pixels, matching the cropped dimensions used for consistency across models. The inclusion of this 3D variant allows the exploration of volumetric segmentation benefits, which could be crucial for detailed coronary artery analysis.

By implementing and comparing these architectures, this study aims to assess the trade-offs between segmentation performance, computational complexity, and practical applicability, particularly in clinical environments where real-time or near-real-time execution is a key consideration.

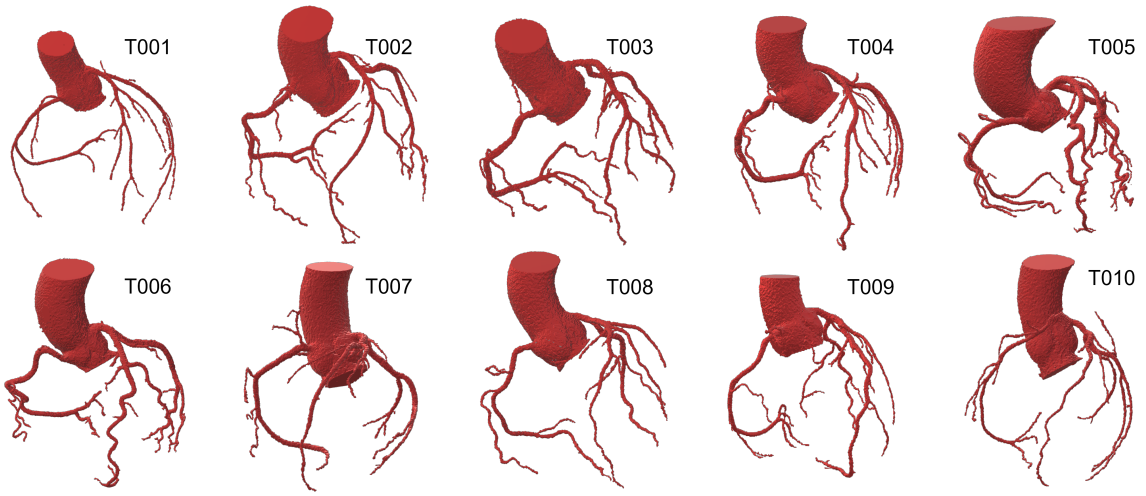


Figure 4.1: 3D coronary tree geometries of the 10 test patients of the study. Image from [117].

4.2.1.3 Training Dataset

A key objective of this study is to evaluate the relationship between data availability and segmentation performance. The dataset was divided into training, validation, and test sets using a train-test split strategy [134]. The test set consists of 10 patients with annotated segmentations (ground truth, GT), shown in Figure 4.1.

The number of training patients, N, was varied between 15 and 65 in increments of 10. The validation set was set at 20% of N, ensuring balanced representation. For example, with $N = 55$, the validation set included $55 \times 0.2 = 11$ patients. This study focuses on patient variation exclusively for the 2D U-Net with MobileNetV2 encoder and the 3D U-Net. These models were specifically chosen to examine how varying the number of patients in the dataset impacts segmentation performance, as computational limitations restricted broader analysis across other architectures.

4.2.1.4 Implementation Details

All models were implemented in Python (v3.7) using the TensorFlow Keras API (v2.6-tf) [135]. The binary cross-entropy loss function was used, weighted to address class imbalance by assigning five times more weight to the vessel class (see loss functions in Section 2.3.5). The Adam optimizer [120] was employed for training. The network architecture includes convolutional layers with a kernel size of 3, followed by MaxPooling with a stride of 2 and ReLU activation.

Each model was trained for 50 epochs, with early stopping applied to monitor validation loss and mitigate overfitting (patience = 5).

4.2.1.5 Separation of the Coronary Tree into Three Regions: Proximal, Middle and Distal

To assess the networks' performance, the coronary tree is divided into proximal, middle, and distal regions based on its length. Using 3D Slicer [64], a central curve is manually traced along the main vessels, such as the right coronary artery, left coronary artery, and circumflex

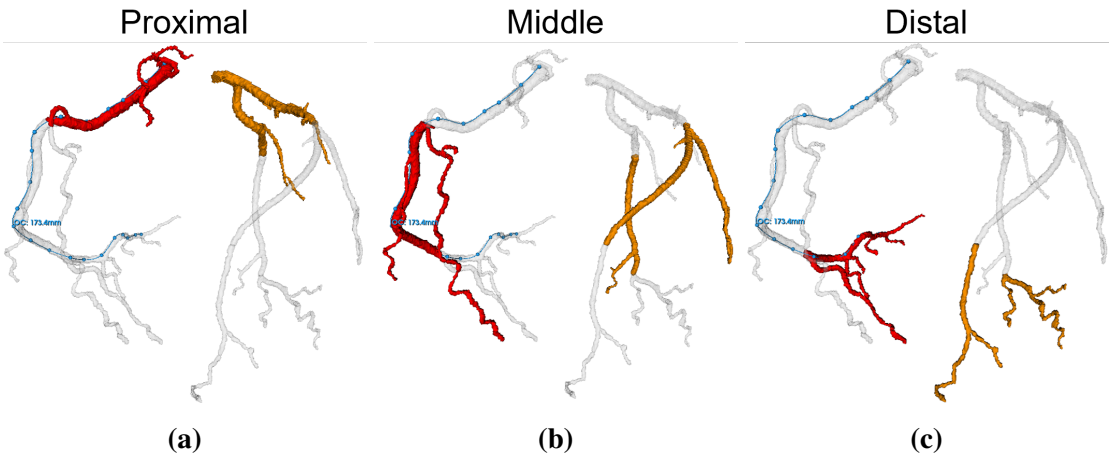


Figure 4.2: Manual segmentation of test patient T002. The right coronary tree (RCT) is highlighted in red, and the left coronary tree (LCT) is in orange. A blue curve represents the path along the RCT, with its length displayed in millimeters. (a) Proximal region of the coronary tree. (b) Middle region. (c) Distal region. Image from [136].

artery, extending from the ostium to the final bifurcation. The proximal region comprises 30% of the length from the ostium, the distal region includes the last 30%, and the middle region accounts for the central 40%. Figure 4.2 presents an example of the manual segmentation for test patient T002, illustrating the division of the coronary tree into different regions. The right coronary tree (RCT) is depicted in red, while the left coronary tree (LCT) is shown in orange. Additionally, a blue curve traces the path along the RCT, with its total length displayed in millimeters. Figure 4.2 is divided into three subfigures, each representing a specific region of the coronary tree: (a) proximal, (b) middle and (c) distal.

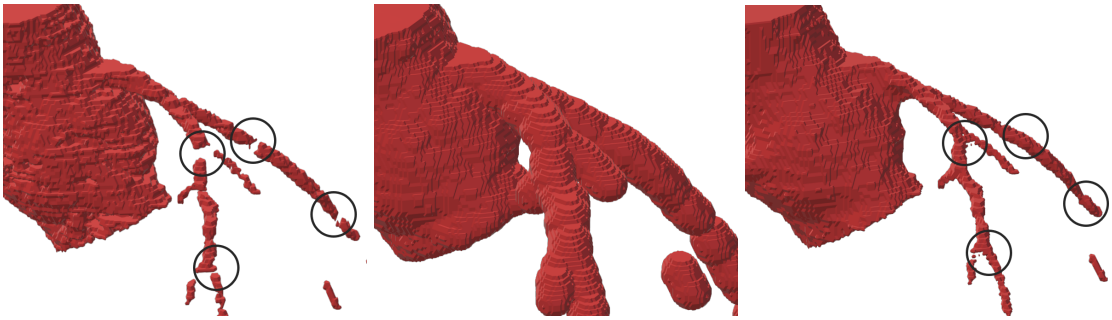
4.2.1.6 Postprocessing

To enhance segmentation quality and ensure cleaner coronary geometries, two postprocessing algorithms were applied. The first algorithm, denoted as *I*, removes small disconnected components (islands) with a volume of less than 50 voxels. The second algorithm, referred to as *G*, employs a grow-shrink technique with a width of 2mm to refine vessel connectivity. These postprocessing steps were tested in both *I*-*G* and *G*-*I* sequences.

Figure 4.3 illustrates the application of the Grow-Shrink algorithm to a predicted coronary geometry. In Figure 4.3a, the initial prediction contains disconnected vessel structures. The Grow step (Figure 4.3b) expands the segmented region, bridging small gaps, while the Shrink step (Figure 4.3c) refines the segmentation, resulting in a more connected and anatomically coherent vessel structure. This postprocessing approach improves the consistency of the final coronary tree segmentation.

4.2.1.7 Evaluation metrics

The network's performance was evaluated using the ten test patients described in Section 4.2.1.1. Standard metrics like precision, recall, and the F_1 score (or Dice Similarity Coefficient, DSC) were employed (see Section 2.3.6.2).



(a) (b) (c)

Figure 4.3: Example of the Grow-Shrink algorithm applied to a predicted coronary geometry. (a) Predicted geometry with disconnected vessels. (b) Application of the Grow step. (c) Application of the Shrink step, resulting in a connected vessel structure. Image from [117].

Additionally, a new metric is defined, the F_1^a score is a variant of the standard F_1 score designed to evaluate false positives (FPs) that are directly connected (i.e., belong to the same connected component) to the true coronary structure. This metric is computed by ignoring false positives that are not in contact with the ground truth vessel, ensuring that only relevant misclassifications are considered. Unlike isolated FPs, these connected FPs are particularly significant in clinical contexts such as FFR calculations, as they may erroneously expand the vessel geometry, affecting hemodynamic simulations. This metric helps quantify the extent of these inaccuracies and assesses how challenging it would be to clean the predicted segmentation to obtain a clinically usable coronary tree.

In addition to the previously described metrics, the evaluation of segmentation performance in the proximal, middle, and distal segments of the coronary arteries includes the False Negative Rate (FNR) and Critical Success Index (CSI) (see Section 2.3.6.2).

4.2.2 Results

In this section, we present the results of our supervised segmentation models, evaluating their performance across different experimental settings. First, in Section 4.2.2.1, we analyze how varying amounts of training data influence model learning and generalization. Then, Section 4.2.2.2 compares segmentation accuracy across different architectures. In Section 4.2.2.3, we assess the model's ability to detect and segment lesions, even though the training process was not explicitly designed for this task. Finally, Section 4.2.2.4 examines performance variations across different anatomical regions.

4.2.2.1 Impact of Dataset Size on Training Performance

To evaluate the impact of dataset size on segmentation performance, we analyzed the results obtained from two network architectures: the 2D MB2-Pre model and the 3D U-Net. These networks were trained on datasets of varying sizes (N) while reserving 20% of each dataset for validation. The results for test patients T001 and T003 are presented in Figures 4.4 and 4.5, respectively. Each figure illustrates predicted segmentations for the aorta and coronary arteries (A + C.A) as well as the coronary arteries alone (C.A), allowing a comparative assessment

of model performance across different dataset sizes. The last row in each figure displays the segmentation results after applying the IG post-processing algorithm (see Section 4.2.1.6).

For the 2D MB2-Pre network, an increase in vessel detection is observed alongside a rise in false positives. This issue is partially mitigated through post-processing with the IG algorithm, as can be seen in the final row of Figure 4.4, though residual structures unrelated to the coronary geometry remain attached. The 3D U-Net, in contrast, begins with segmented slices that lack detail but progressively improves, ultimately recognizing all major vessels while maintaining a cleaner and more connected segmentation. The results for T003, depicted in Figure 4.5, demonstrate that structures become increasingly cohesive while preserving the quality of segmented vessels.

A notable observation is the difference in training with and without the aorta. The 2D MB2-Pre network shows significantly fewer false positives when trained exclusively on coronary arteries (C.A). Moreover, also coronary artery recognition improves without the aorta, particularly for the 3D U-Net. However, the segmentation of T003 with the 2D MB2-Pre and C.A dataset highlights challenges in segmenting highly curved and intricate coronary geometries. In comparison, T001 presents a less complex geometry, which is segmented with greater accuracy.

Figure 4.6 presents the calculated metric values for each network and training configuration, displayed as bar plots. The x-axis represents the number of patients used for training (N), while the y-axis displays the corresponding mean parameter values. The results are reported for both the 2D MB2-Pre and 3D U-Net models, applied to datasets containing both the aorta and coronary arteries (A + C.A) as well as only the coronary arteries (C.A). The error bars indicate the mean and standard deviation computed across 10 test patients, using three independent training datasets of equal size (N), resulting in a total of 30 values per parameter. Below, we discuss each parameter individually.

F_1 Score The F_1 score, depicted in Figure 4.6.(A), measures the similarity between two binary masks, ranging from 0 (no similarity) to 1 (perfect match). When training with A+C.A, values are consistently high ($[0.89, 0.95]$), reflecting the aorta's dominance in volume and its accurate segmentation by the networks. The 2D MB2-Pre yields stable values (~ 0.7) for C.A training, independent of N, while the 3D U-Net demonstrates improvement (from 0.53 to 0.75) as N increases.

F_1^a Score Figure 4.6.(B) illustrates the F_1 score accounting only for false positives attached to the vessel. Results closely follow those of the standard F_1 score, with an increase of up to 0.2 in C.A training.

Recall Recall, which measures the proportion of the coronary tree segmented, is shown in Figure 4.6.(C). Both networks exhibit a general increase with N, with the 2D MB2-Pre achieving superior results. The influence of the aorta, evident in the elevated Recall values (> 0.8) for A+C.A training. However, Recall decreases in Figure 4.6.(D), which focuses solely on coronary arteries. Here, a dependence on N is observed for all cases. Notably, the 2D MB2-Pre trained on A+C.A achieves a steep increase from 0.57 to 0.93 when $N > 35$, surpassing C.A training for $N > 35$. The 3D U-Net exhibits similar trends, flattening at lower N when trained on A+C.A.

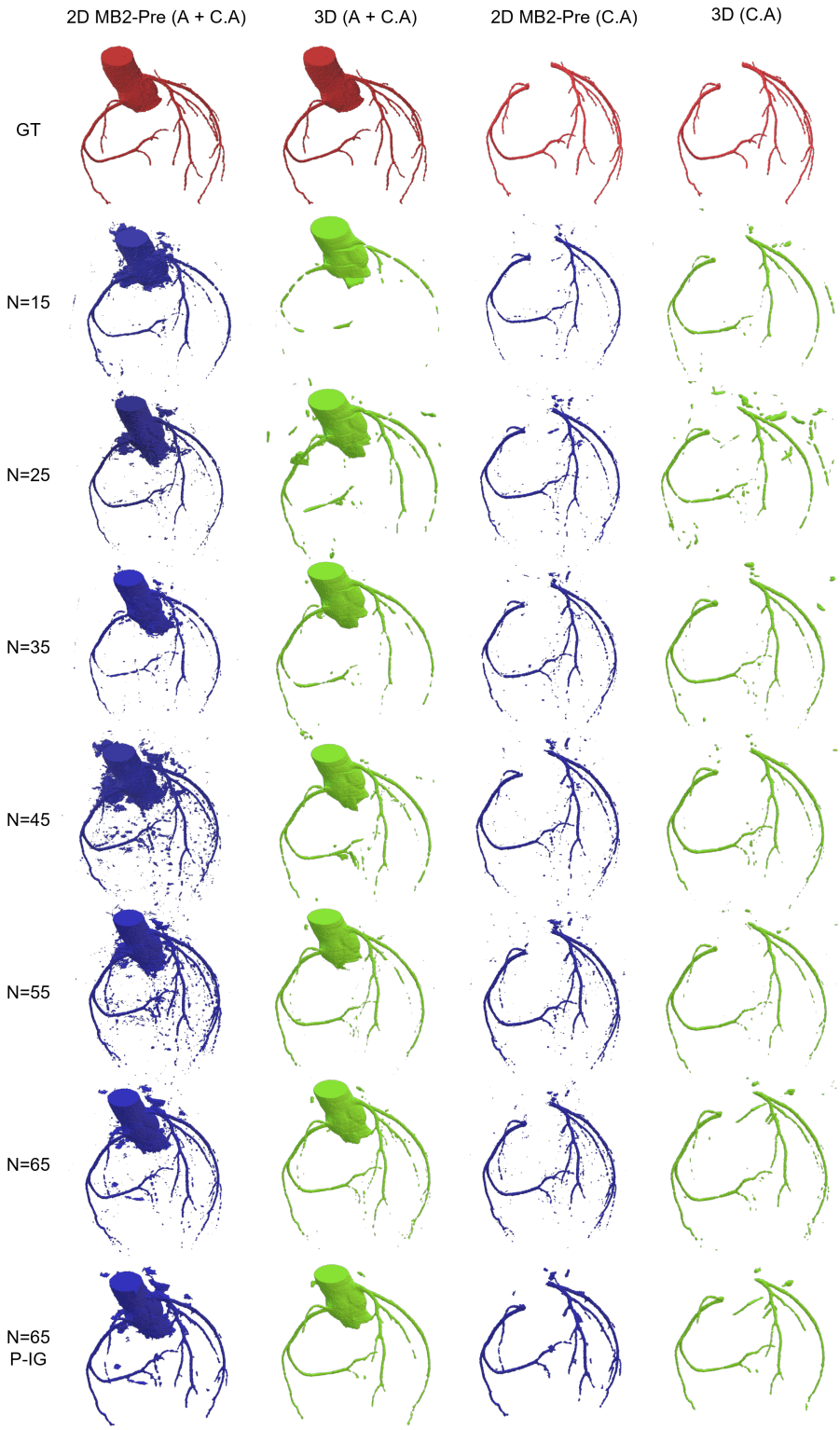


Figure 4.4: Example of predicted segmentations for test patient T001 using both the 2D MB2-Pre model and 3D U-Net, applied to the aorta and coronary arteries (A + C.A) and only the coronary arteries (C.A). The networks were trained with datasets of varying sizes (N) and validated using 20% of N. The final row displays the results after applying the IG algorithm post-processing, which first removes small islands and then applies the grow-shrink technique. Image from [117].

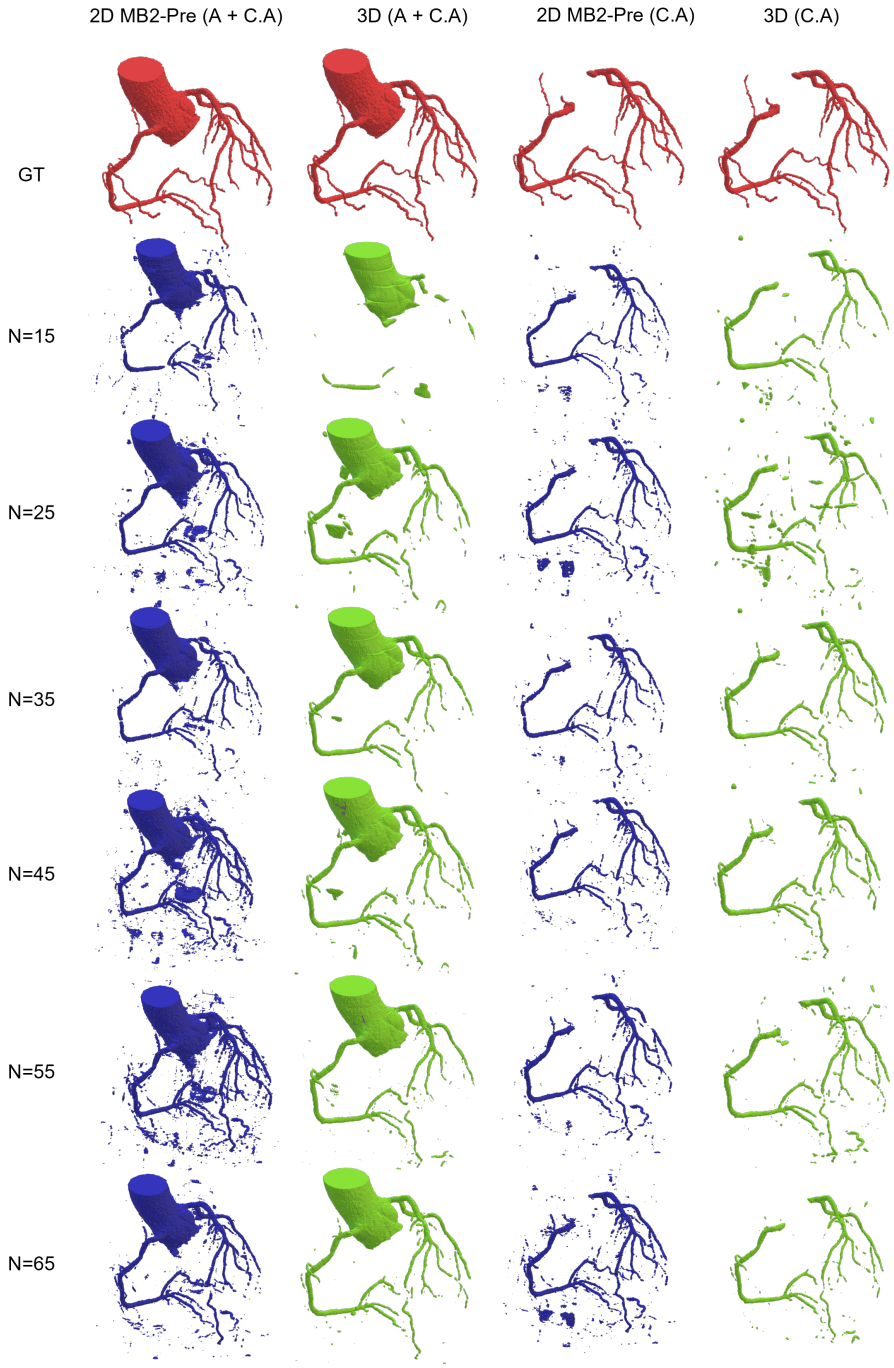


Figure 4.5: Example of predicted segmentations for test patient T003 using both the 2D MB2-Pre model and 3D U-Net, applied to the aorta and coronary arteries (A + C.A) and only the coronary arteries (C.A). The networks were trained with datasets of varying sizes (N) and validated using 20% of N. The final row displays the results after applying the IG algorithm post-processing, which first removes small islands and then applies the grow-shrink technique. Image from [117].

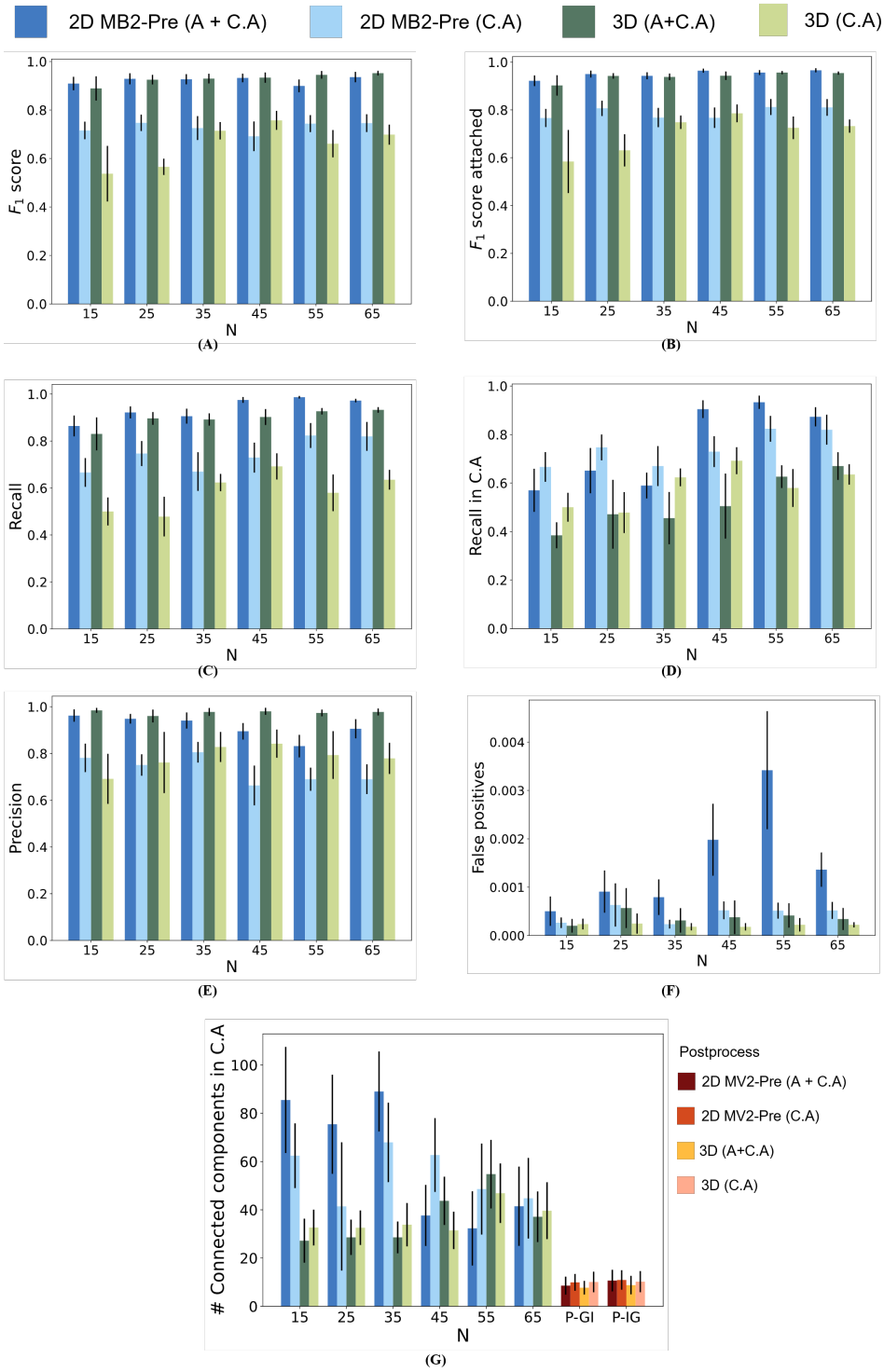


Figure 4.6: Parameter values obtained from the 2D MV2-Pre and 3D U-Net models applied to datasets of both aorta and coronary arteries (A + C.A) and coronary arteries alone (C.A). The X-axis represents the number of patients used for training (*N*), and the Y-axis displays the corresponding parameter values, mean (bar) and standard deviation (vertical line)-. A) F_1 score, B) F_1^a score, C) Recall, D) Recall specific to coronary arteries, E) Precision, F) False positive to background ratio, G) Number of connected components in coronary arteries. Image from [117].

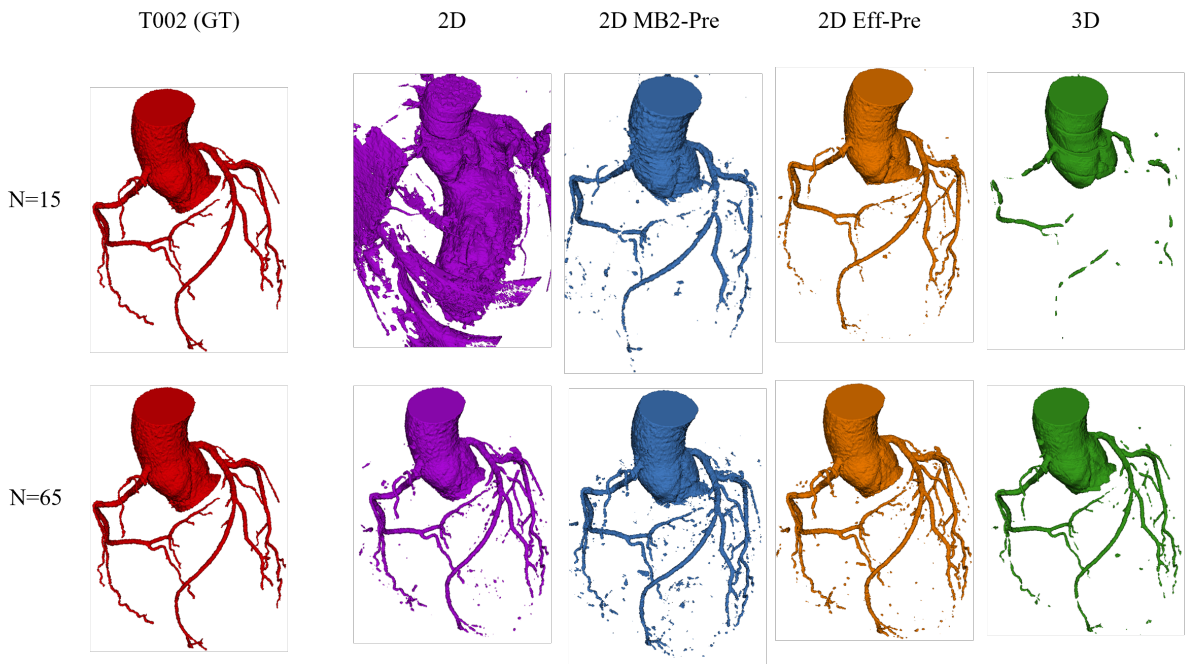


Figure 4.7: Segmentation results for test patient T002 using networks trained on aorta and coronary arteries (A + C.A). The networks include 2D, 2D MB2-Pre, 2D Eff-Pre, and 3D U-Net. Two sets of results are shown: the first row corresponds to training with $N = 15$ patients, and the second row shows results for $N = 65$ patients. Image from [117].

Precision Figure 4.6.(E) shows Precision, which decreases slightly with N for the 2D MB2-Pre trained on A+C.A, maintaining values above 0.8. For C.A training, a similar trend is observed, with values in the range $[0.66, 0.8]$. In contrast, the 3D U-Net shows stable values ($[0.96, 0.98]$) for A+C.A and an increasing trend for C.A, reaching 0.84 at $N = 45$.

False Positives (FP) and Connected Components (CC). As the number of training patients (N) increases, both False Positives (FP) and Connected Components (CC) decrease, showing a related trend (4.6.(F) and Figure 4.6.(G)). For $N > 35$, the 2D MB2-Pre trained with A+C.A experiences a steep drop in CC from > 80 to < 40, outperforming C.A training. Similarly, the 3D U-Net maintains CC below 50 for $N < 45$ and stabilizes at $30 - 46$ for $N \geq 55$. Post-processing further reduces CC to fewer than 15 at $N = 65$.

For all other parameters, post-processing does not significantly alter results, so only the original values are reported.

4.2.2.2 Results Across All Network Architectures

The results obtained from the four neural networks under study—2D, 2D MB2-Pre, 2D Eff-Pre, and 3D—are presented for training sets with $N = 15$ and $N = 65$ patients.

Figures 4.7 and 4.8 present segmentation results for test patient T002, showcasing predictions made by different networks trained on two configurations: aorta and coronary arteries (A + C.A) and coronary arteries alone (C.A), respectively. Similarly, Figures 4.9 and 4.10 illustrate segmentation results for test patient T005 under the same training conditions.

The evaluated networks include the baseline 2D U-Net, 2D MB2-Pre (with a MobileNetV2

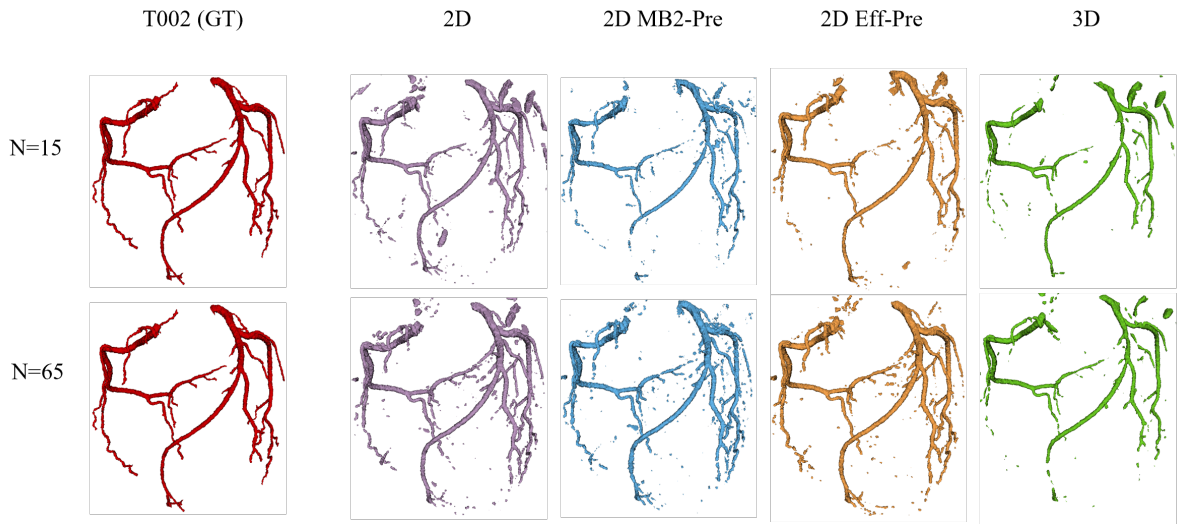


Figure 4.8: Segmentation results for test patient T002 using networks trained on coronary arteries (C.A). The networks include 2D, 2D MB2-Pre, 2D Eff-Pre, and 3D U-Net. Two sets of results are shown: the first row corresponds to training with $N = 15$ patients, and the second row shows results for $N = 65$ patients. Image from [117].

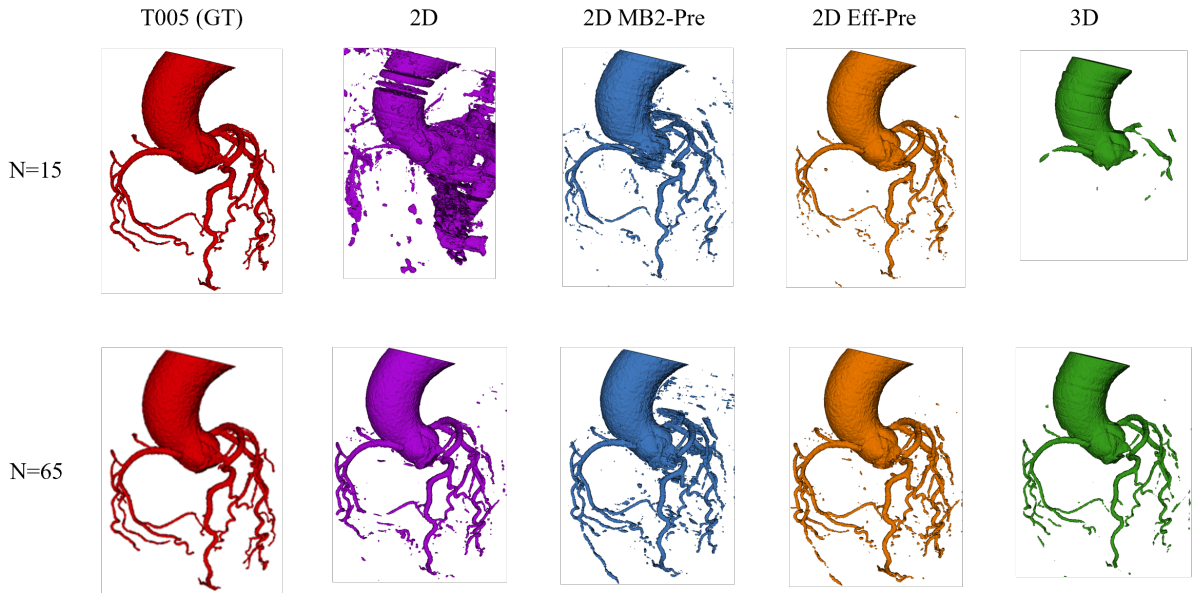


Figure 4.9: Segmentation results for test patient T005 using networks trained on aorta and coronary arteries (A + C.A). The networks include 2D, 2D MB2-Pre, 2D Eff-Pre, and 3D U-Net. Two sets of results are shown: the first row corresponds to training with $N = 15$ patients, and the second row shows results for $N = 65$ patients.. Image from [117].

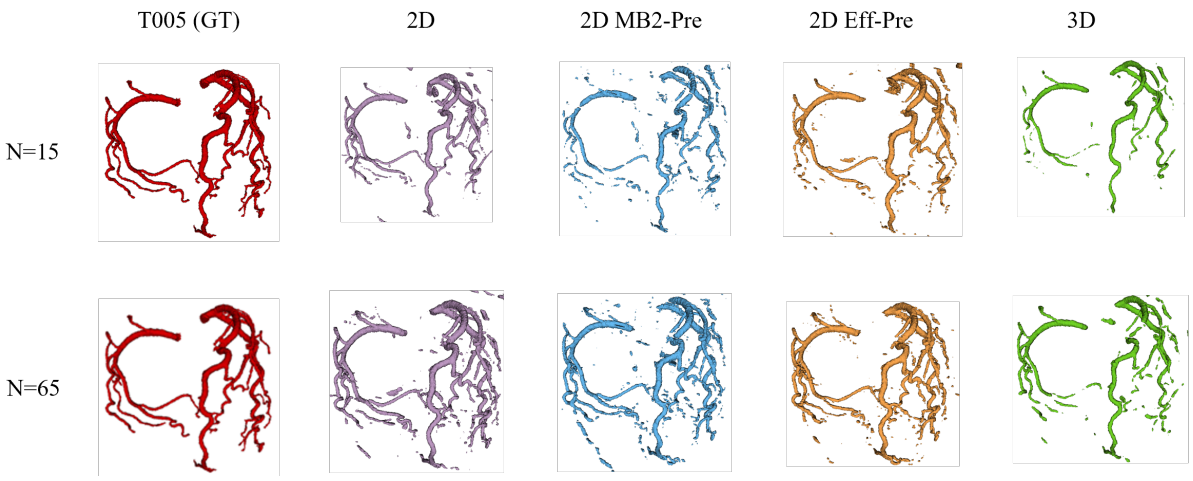


Figure 4.10: Segmentation results for test patient T005 using networks trained on coronary arteries (C.A). The networks include 2D, 2D MB2-Pre, 2D Eff-Pre, and 3D U-Net. Two sets of results are shown: the first row corresponds to training with $N = 15$ patients, and the second row shows results for $N = 65$ patients. Image from [117].

encoder), 2D Eff-Pre (with an EfficientNet-B5 encoder), and the 3D U-Net. Each figure provides two sets of results: the first row corresponds to models trained with a dataset of $N = 15$ patients, while the second row displays results obtained when training with $N = 65$ patients. This comparison allows us to assess how increasing the dataset size influences segmentation performance across different backbone architectures.

From these examples, we highlight the challenges faced by networks without pretraining (2D and 3D) in identifying coronary arteries when $N = 15$ and the training setup is A+C.A. These challenges are mitigated with a larger training dataset ($N = 65$). However, these models show significantly less difficulty when trained under C.A conditions, underscoring the impact of incorporating a large-volume structure such as the aorta during training.

Pretrained networks (2D MB2-Pre and 2D Eff-Pre) exhibit higher detail and precision with more training data, although this may lead to an increase in false positives (FP).

To support the observations made for test patients T002 and T005, Figures 4.11 and 4.12 present quantitative results computed across the entire test set for the training configurations A+C.A and C.A, respectively. These bar plots compare the performance of the networks of the study. The x-axis represents the number of patients in the training dataset (N), while the y-axis displays the corresponding parameter values, mean (bar) and standard deviation (vertical line).

Notably, we observe an increase in Recall with larger training datasets, along with a rise in FP (see Figure 4.12.(D) and 4.12.(F)).

Focusing on the general vascular tree structure, we note improved vessel connectivity with more training data across both configurations (see Figure 4.11.(G) and Figure 4.12.(G)). Furthermore, pretrained networks are less affected by the number of training samples, exhibiting increments of less than 12% in Dice score (see Figure 4.11.(A) and Figure 4.12.(A)).

4.2.2.3 Lession Evaluation

This section presents the segmentation results for two patients clinically diagnosed with a lesion. The performance of the 2D MB2-Pre and 3D U-Net models, trained with $N = 65$

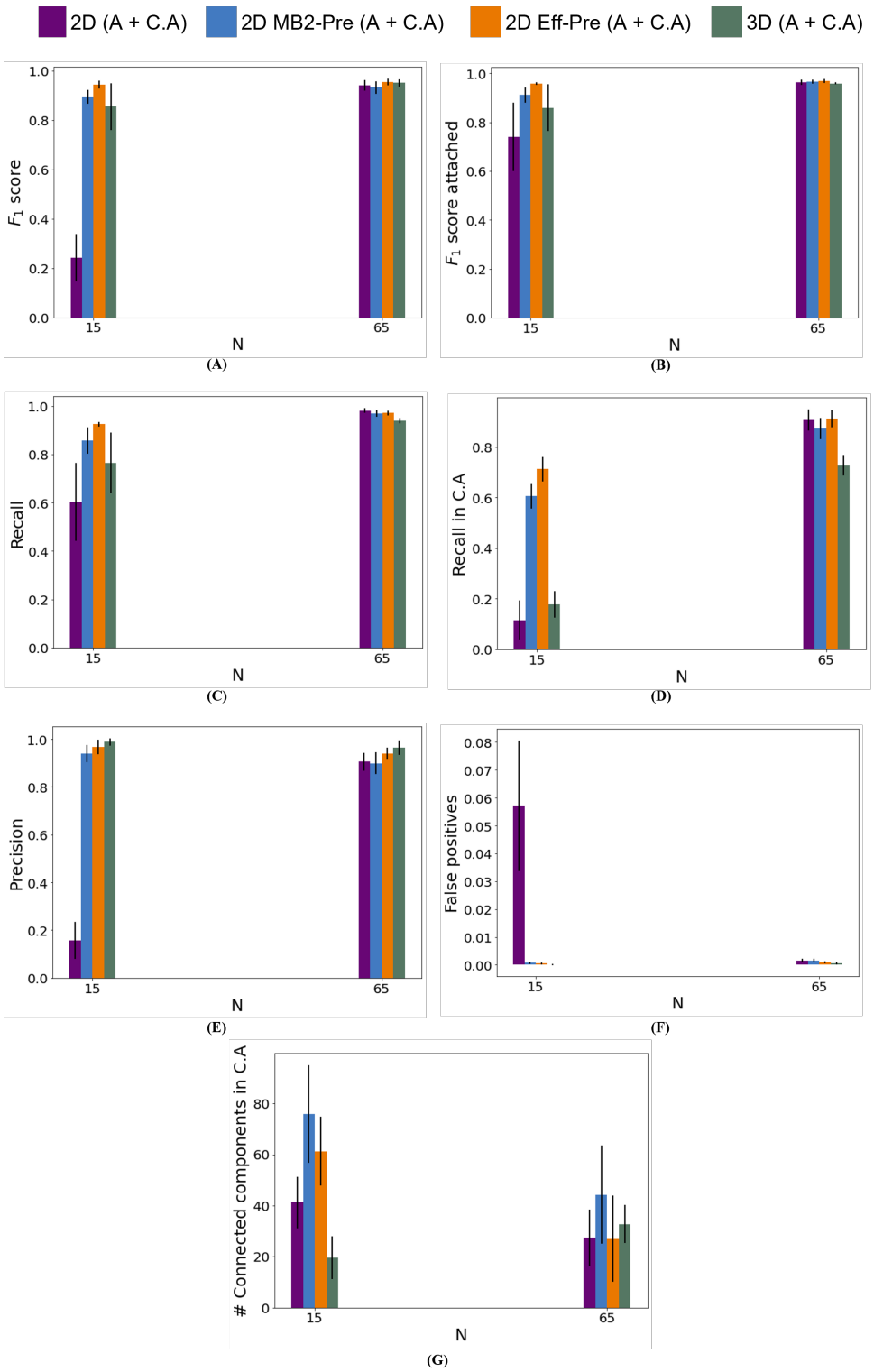


Figure 4.11: Parameter values for the 2D, 2D MB2-Pre, 2D Eff-Pre, and 3D networks trained on the aorta and coronary arteries (A + C.A) dataset. The X-axis represents the number of patients in the training set (N), while the Y-axis shows the value of the corresponding parameter, mean (bar) and standard deviation (vertical line). A) F_1 score. B) F_1^a score. C) Recall. D) Recall for coronary arteries. E) Precision. F) False positive to background class pixel ratio. G) Number of connected components in coronary arteries. Image from [117].

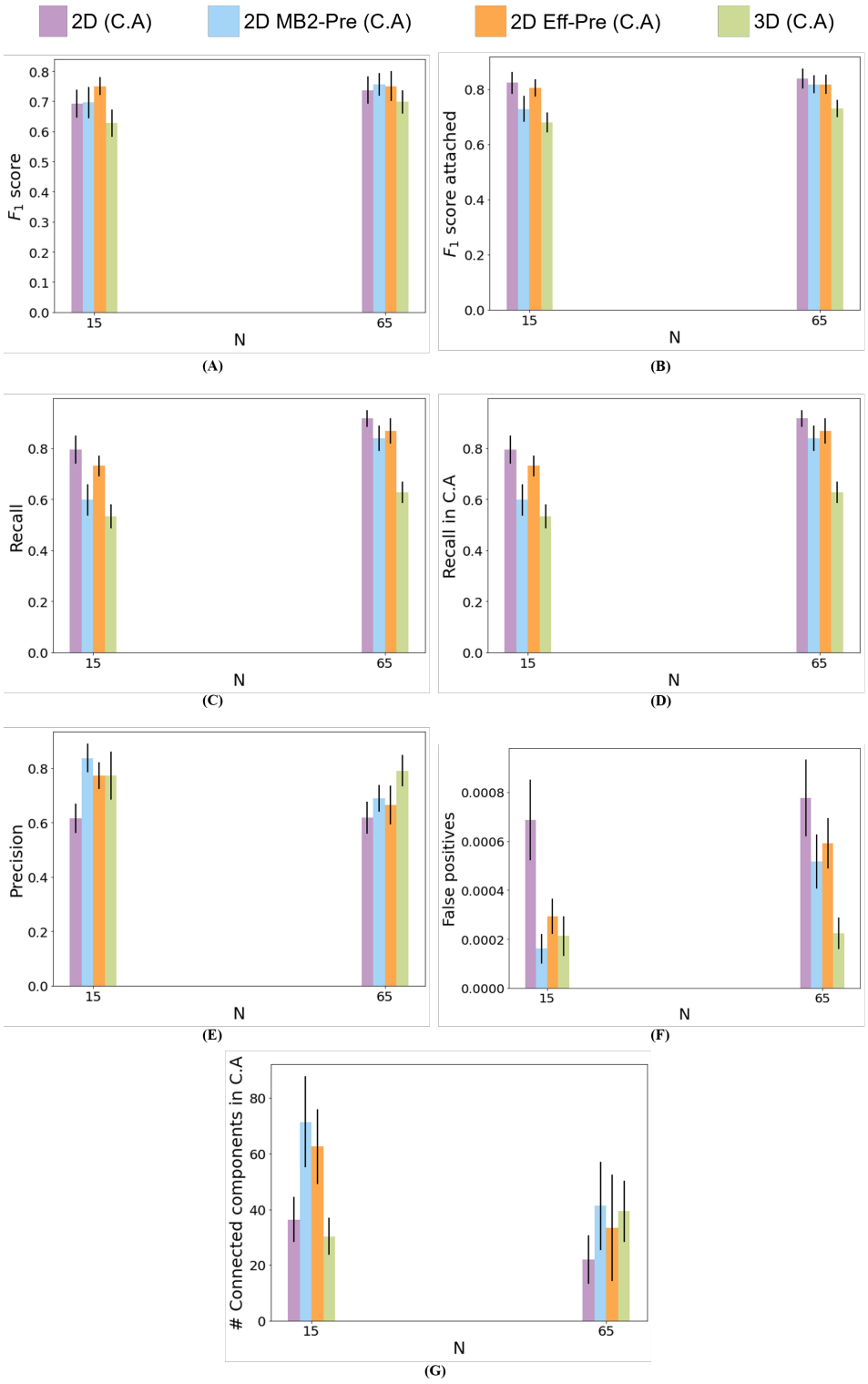


Figure 4.12: Parameter values for the 2D, 2D MB2-Pre, 2D Eff-Pre, and 3D networks trained on the coronary arteries (C.A) dataset. The X-axis represents the number of patients in the training set (N), while the Y-axis shows the value of the corresponding parameter, mean (bar) and standard deviation (vertical line). A) F_1 score. B) F_1^a score. C) Recall. D) Recall for coronary arteries. E) Precision. F) False positive to background class pixel ratio. G) Number of connected components in coronary arteries. Image from [117].

Table 4.1: Comparison of results between the segmentation predicted by the corresponding network and the manual segmentation for *lesion*1. The first row shows the difference between the network volume and the manual segmentation volume, the second row shows the percentage of volume at the intersection over the manual segmentation volume, the third, fourth and fifth rows show DSC, precision and recall parameters, respectively, between the network volume and the manual segmentation volume. Yellow shows the manual segmentation (ground truth) and blue shows the AI result. The corresponding networks are 2D MB2-Pre and 3D UNet, trained with $N = 65$, for aorta and coronary arteries ($A + C.A$) and coronary arteries alone ($C.A$). Table from [117].

*Lesion*1 Metrics	2D MB2 ($A + C.A$)	3D($A + C.A$)	2D MB2 ($C.A$)	3D ($C.A$)
Vol. Diff (mm^3)	5.04	−4.10	5.18	−21.62
Net overlap	88.70%	86.94%	90%	61.74%
DSC	0.85	0.90	0.86	0.75
Precision	0.82	0.93	0.83	0.97
Recall	0.89	0.87	0.90	0.62
Front view				
Side view				

patients, is evaluated for both aorta and coronary arteries ($A + C.A$) and coronary arteries alone ($C.A$).

Tables 4.1 and 4.2 summarize the metrics obtained for *lesion*1 and *lesion*2, respectively, comparing the network-predicted segmentation with the manual segmentation (ground truth). The first row displays the difference in volume between the predicted and manual segmentations, while the second row shows the percentage of volume at the intersection over the manual segmentation volume. The third, fourth, and fifth rows provide the Dice Similarity Coefficient (DSC), precision, and recall values, respectively, measuring the overlap between the network-generated and manual segmentations. For visual reference, yellow represents the manual segmentation, while blue denotes the AI-generated segmentation. These results allow for a quantitative and qualitative assessment of the model's ability to accurately segment lesions.

In all cases, the lesion is identified, as the segmentations narrow in the stenosis region. However, for the 3D model trained under C.A conditions on *lesion*1 (Table 4.1), the vessel is cut off, failing to segment the most constricted area. For the 2D MB2-Pre model, oversegmentation is observed for both lesions, as well as across the entire patient, leading to consistent segmentation throughout the vessel. This is illustrated in Figure 4.13a and Figure 4.13c, where the predicted lesions are shown in blue. On the other hand, the 3D U-Net model, trained with A+C.A, produces more accurate segmentations, as shown in Figure 4.13b and Figure 4.13d, where the predicted lesions are shown in green. In these figures, the ground truth segmentation is shown in red.

Table 4.2: Comparison of results between the segmentation predicted by the corresponding network and the manual segmentation for *lesion*2. The first row shows the difference between the network volume and the manual segmentation volume, the second row shows the percentage of volume at the intersection over the manual segmentation volume, the third, fourth and fifth rows show DSC, precision and recall parameters, respectively, between the network volume and the manual segmentation volume. Yellow shows the manual segmentation (ground truth) and blue shows the AI result. The corresponding networks are2D MB2-Pre and 3D UNet, trained with $N = 65$, for aorta and coronary arteries ($A+C.A$) and coronary arteries alone ($C.A$). Table from [117].

*Lesion*2 Metrics	2D MB2 ($A+C.A$)	3D ($A+C.A$)	2D MB2 ($C.A$)	3D ($C.A$)
Vol. Diff. (mm^3)	8.80	2.86	10.61	4.68
Net overlap	95.53%	86.58%	93.75%	88.22%
DSC	0.79	0.81	0.77	0.80
Precision	0.68	0.77	0.64	0.74
Recall	0.95	0.87	0.94	0.88
Front view				
Side view				

4.2.2.4 Evaluation in Proximal, Middle and Distal regions

Figure 4.14 illustrates the segmentation outcomes for proximal, middle, and distal regions obtained using different neural networks on patient T002. Upon initial observation, the proximal segments appear consistently well-reconstructed across all networks. However, significant variability is evident in the distal region results. To quantify these observations, the DSC, FNR, Sensitivity, and CSI metrics are analyzed for each region.

The bar plots in Figure 4.15 show the performance metrics for different models (2D, 2D MB2-Pre, 2D Eff-Pre, and 3D) in the segmentation of the proximal, middle, and distal regions of the coronary tree, distinguishing between the right and left branches.

In the proximal region, the DSC (Figure 4.15a) surpasses 0.8 for all 2D networks, whereas the 3D network achieves lower scores, particularly in the left branch. This trend correlates with the FNR (Figure 4.15c), revealing increased false negatives and reduced true positives, as evidenced by the lower Sensitivity (Figure 4.15d).

For the middle region, results mirror those seen in the proximal region but are slightly worse for the left coronary tree. This discrepancy may stem from the left artery's greater branching complexity compared to the right. Among the networks, the 2D Eff-Pre stands out, maintaining consistent DSC and CSI values (Figure 4.15a and Figure 4.15b) across both branches. Meanwhile, the 3D UNet achieves comparable DSC and CSI scores, even with lower vessel detection rates, suggesting fewer false positives and reduced vessel diameter overestimation.

The distal region presents a distinct pattern. DSC scores drop below 0.8 for 2D networks, with marked differences between the left and right branches (Figure 4.15b). The 3D UNet

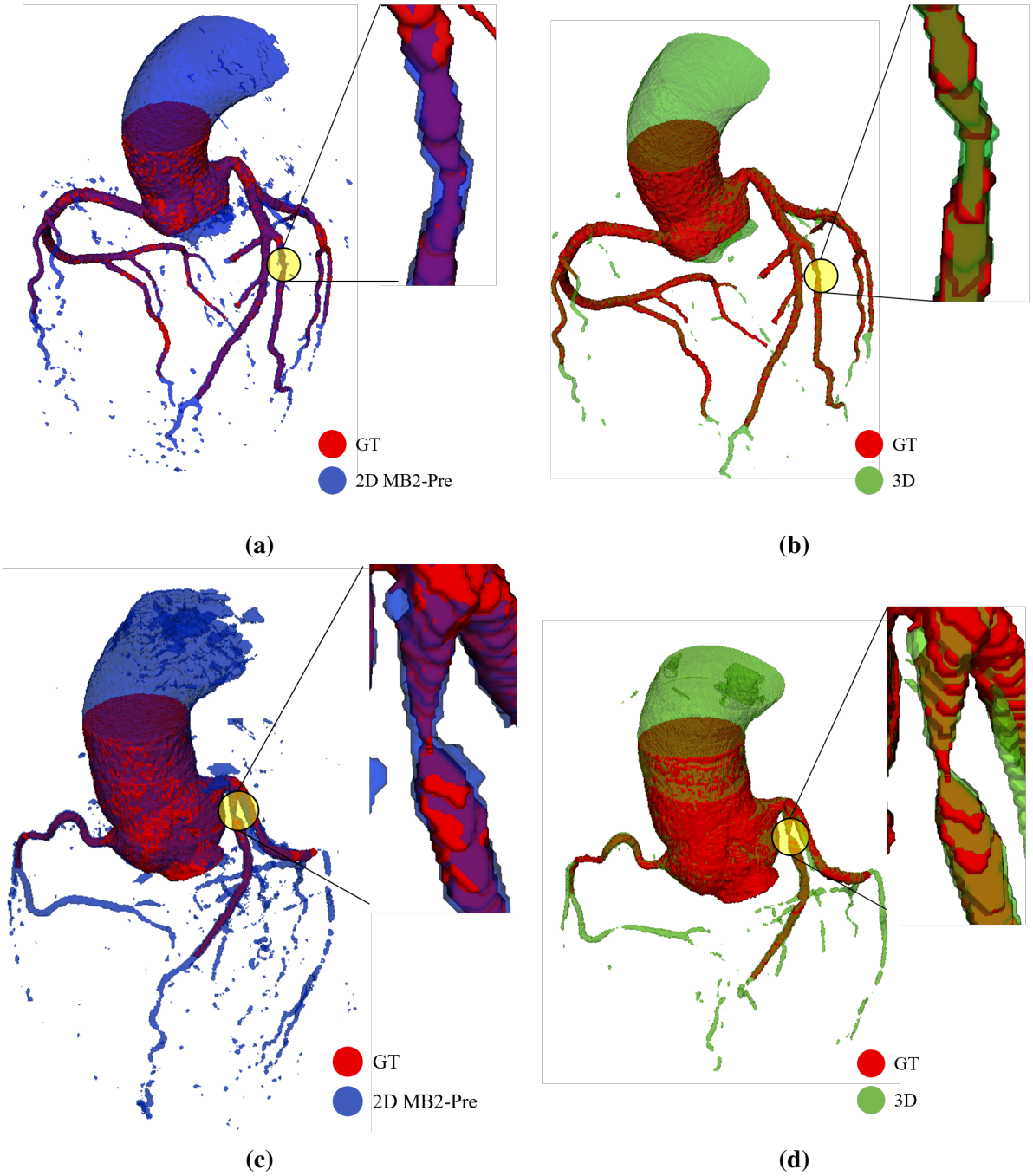


Figure 4.13: Segmentations of the coronary tree for *Lesion*1 and *Lesion*2 with $N = 65$ training patients. Ground truth segmentation is shown in red. (A) Prediction of Lesion 1 by the 2D MB2-Pre, shown in blue. (B) Prediction of Lesion 1 by the 3D, shown in green. (C) Prediction of Lesion 2 by the 2D MB2-Pre, shown in blue. (D) Prediction of Lesion 2 by the 3D, shown in green. Zoomed-in sections highlight the regions of the lesions. Images from [117].

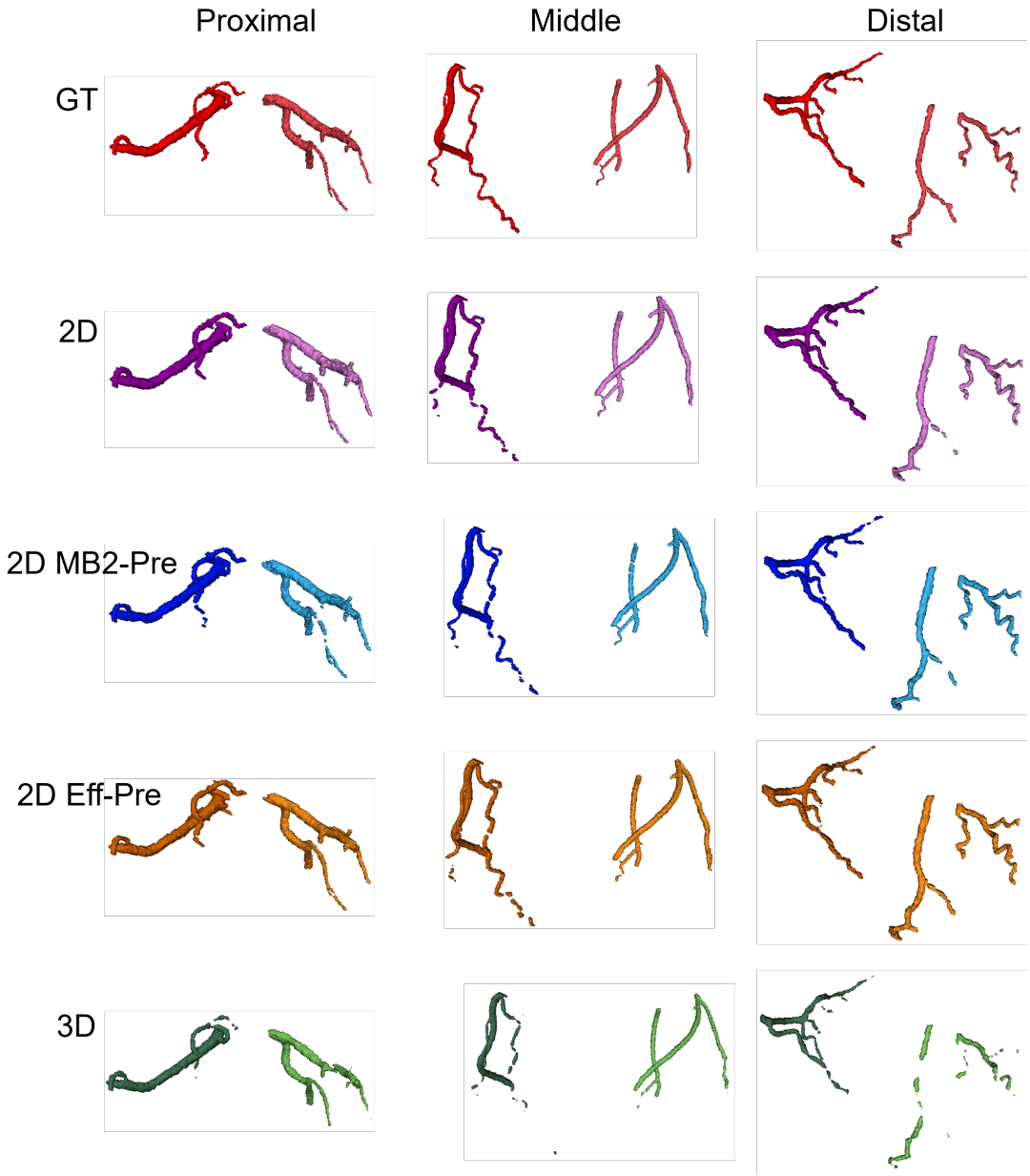


Figure 4.14: Segmentations of the coronary arteries of test patient T002. Columns represent the proximal, middle, and distal regions of the coronary arteries. Rows display the segmentation results of the following models: (1) manual segmentation (GT), (2) 2D, (3) 2D MB2, (4) 2D Eff-Pre, and (5) 3D. Image from [136].

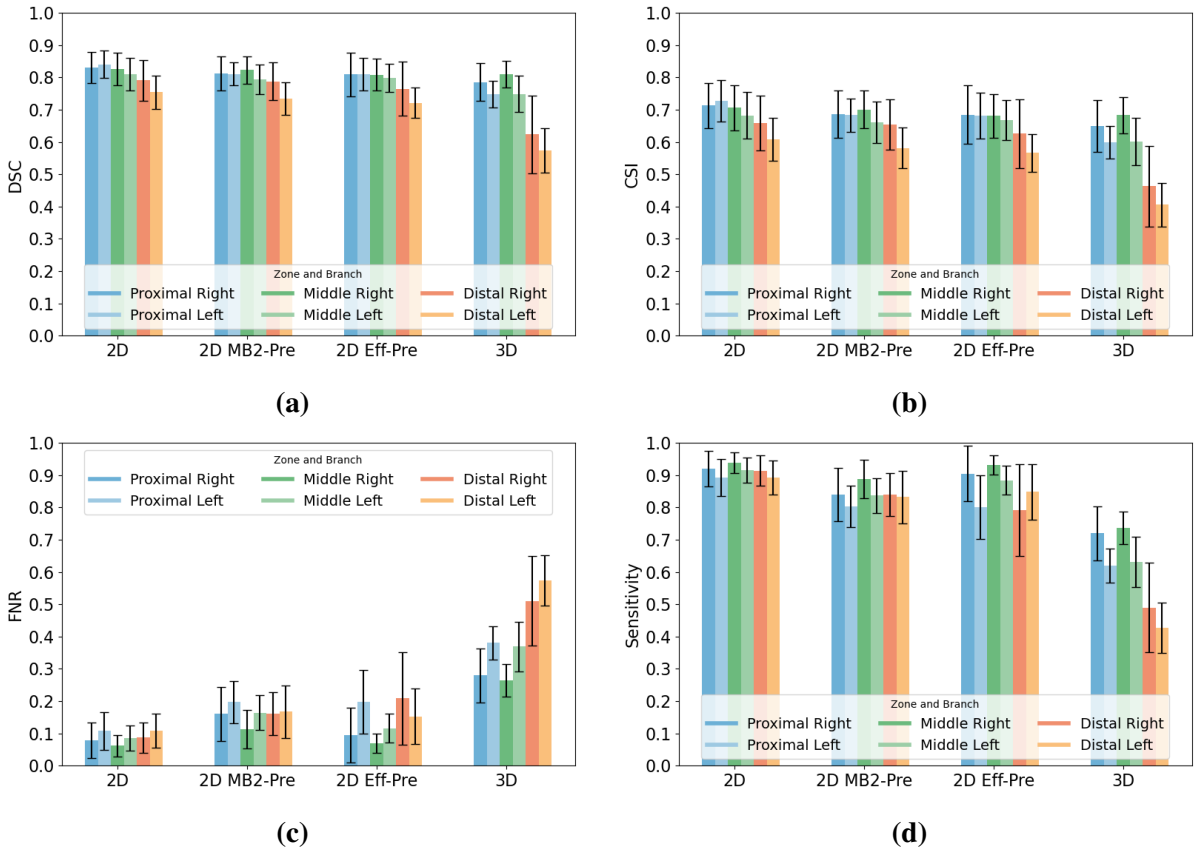


Figure 4.15: Bar plots showing the performance metrics for different models (2D, 2D MB2-Pre, 2D Eff-Pre, and 3D) in the segmentation of the proximal, middle, and distal regions of the coronary tree, distinguishing between the right and left branches. The x-axis represents the different models, and the y-axis shows the corresponding parameter value, mean (bar) and standard deviation (vertical line). The metrics include: (a) Dice Similarity Coefficient (DSC), (b) Critical Success Index (CSI), (c) False Negative Rate (FNR), and (d) Sensitivity.

struggles even more, identifying only 50% of the vessels (Figure 4.15d). Interestingly, the 2D U-Net demonstrates Sensitivity values exceeding 0.9, outperforming other 2D networks by up to 10%.

4.2.2.5 Computation time

The computational time required for the supervised segmentation models varies between approximately 6 to 9 minutes, depending on both the size of the patient dataset (number of slices) and the complexity of the model. These measurements were obtained by executing the segmentation algorithms within the 3D Slicer software, which may introduce additional processing delays. This slowdown is attributed to 3D Slicer's internal resource management, including the generation of segmentation nodes and the reading and saving of files.

The computations were performed on a workstation equipped with an Intel Core i5-8500T CPU and 32GB of RAM. Additionally, the segmentation process included the visualization of the resulting geometry within 3D Slicer, which contributed an extra 2 minutes to the total processing time across all methods.

4.2.3 Discussion and Knowledge Transfer to Industry

This section explores the feasibility of using various U-Net-based neural networks for coronary tree segmentation, a common technique in medical image segmentation. Rather than conducting a comprehensive comparison, the focus is on evaluating the accessibility and suitability of these architectures for the task at hand. Specifically, we address the following three key challenges:

1. Dataset Size and Pretraining Influence: The results demonstrate the significant impact of dataset size on the performance of AI models. Larger datasets consistently lead to better outcomes. Additionally, the influence of pretraining is particularly noteworthy. Models using pretrained weights, even on datasets unrelated to our domain, such as ImageNet [82], achieved competitive results with a very small dataset ($N = 15$). This highlights the importance of pretraining in medical imaging, where data collection and annotation by clinical experts remain challenging and resource-intensive.

   However, pretraining also has limitations. For instance, non-pretrained networks outperformed pretrained ones when using a larger dataset ($N = 65$). In this case, fixing the pretrained encoder weights may have restricted the network's ability to adapt to the specific features of the coronary artery images.

2. Performance of 3D U-Net: Among the implemented models, the non-pretrained 3D U-Net stood out for its ability to reduce false positives and produce more connected vessel structures. However, it struggled to recognize finer vessels, typically corresponding to bifurcations of secondary branches or distal regions. This limitation underscores the challenges inherent in segmenting thin and complex vascular structures.

3. Class Imbalance: Addressing class imbalance was crucial for successful training. Giving more weight to the vessel class rather than the background was necessary for the network to learn effectively.

The 2D MB2-Pre and 3D UNet networks demonstrate excellent and competitive performance compared to previous studies, achieving F_1 scores (DSC) of over 93% for both models in $A+C.A.$, and 75% in C.A., outperforming previous results of 91.20% and 88.80% [137]. The detailed segmentations used in this study, especially in C.A., may explain slightly lower performance in the distal regions, where thin vessels are challenging to segment.

The 2D MB2-Pre, also shows stability in performance across increasing data, while the 2D network (with no pre-training) struggles with fewer data ($N = 15$). On the other hand, the 2D Eff-Pre shows more realistic vessel shapes compared to 2D MB2-Pre, demonstrating better connected vessels. In contrast, the 3D UNet performs well with more data, producing clean structures but struggles with distal vessels.

As the size of the training dataset increases, the ability to recognize structures improves, especially for 2D networks, though this also leads to an increase in false positives. These false positives are usually disconnected from the vessels and can be removed through postprocessing. Interestingly, using the C.A. dataset for training results in fewer false positives, which is consistent with previous studies [137]. This reduction is likely due to the decreased impact of focal loss, as the absence of larger vessels like the aorta aids in noise suppression.

When analyzing performance across proximal, middle, and distal regions ($N = 65$, training with the C.A. dataset), distinct trends emerge. Proximal and middle regions show similar results, with 2D networks performing comparably regardless of pre-training. In the middle region, the left coronary tree proves more challenging due to its increased complexity, yet the 2D Eff-Pre performs well across both branches.

In the distal region, however, the 3D UNet struggles significantly with vessel detection, especially in narrower vessels and low-contrast areas. Surprisingly, the 2D U-Net outperforms other networks in this region, suggesting that pre-trained networks may encounter limitations when adapting to fine vascular structures.

In this study, the patients used for training and testing the networks did not have lesions diagnosed by clinical professionals, and cases with calcium were excluded. However, two patients with diagnosed lesions were tested using 2D MB2-Pre and 3D networks. Both approaches identified lesions, though 2D MB2-Pre tended to overestimate more. Training with the $A+C.A.$ dataset resulted in higher accuracy for the 3D U-Net.

The development of these methodologies within the context of an industrial PhD has provided substantial advantages, both for advancing academic research and addressing specific industrial needs. The benefits are particularly evident in two main areas: dataset generation and process automation. These contributions have improved efficiency, enabled integration into the company's workflows, and facilitated the broader application of these tools in the medical imaging industry.

A dataset has been generated consisting on medical images and manual segmentations of the aorta and coronary arteries, encompassing a total of 88 patients. This resource is highly valuable for advancing research in artificial intelligence and other analyses, as it enables new projects to be undertaken without the need to create a dataset from scratch. The creation of a dataset is particularly tedious, as the manual annotation of each patient can take between 1 to 2 hours, significantly increasing the time and resources required for developing segmentation models.

In addition, it is crucial that the solutions developed can be implemented within a production environment. To achieve this, the automation of numerous processes is necessary, including those required for dataset preparation, AI-driven segmentation inference, and result

visualization. This automation not only optimizes workflow but also ensures efficiency and consistency in data processing.

Alongside the development of algorithms, significant efforts have been made to create a user-friendly interface that allows users to load images, obtain segmentations via artificial intelligence, and visualize resulting geometries in 3D. This interface guides users step by step, facilitating geometry editing within the 3D Slicer software [64]. Notable automation features include:

- Importing and exporting images in DICOM or .npy formats.
- Generating 3D geometries from AI predictions in .stl format for seamless user interaction and editing within 3D Slicer.
- Saving 3D geometries ready for simulation environments.
- Developing macros to integrate geometries into simulation workflows.

Additionally, the reduction in time required to obtain segmentations is noteworthy (reducing the segmentation time to less than 10 minutes, including the aorta), resulting in increased productivity. With these improvements, personnel can perform less work and complete tasks in significantly less time, allowing for a more efficient approach to research and development in the field of medical imaging.

To sum up, even though we obtained full coronary trees with this methodology, the mean resolution of the images used for training and testing was 0.38×0.38 mm^2, which may be insufficient for accurately segmenting lesions with stenosis diameters between 0.45 and 0.9 mm. While interpolation could enhance resolution, it would require generating a new dataset and introduce significant computational and memory challenges. To overcome these limitations, the next section explores an unsupervised segmentation approach designed for higher-resolution images, eliminating the need for additional training data.

4.3 Unsupervised Clustering-Graph Segmentation

In the previous section, we explored supervised learning approaches using neural networks to achieve a complete coronary tree segmentation. While these methods excelled at identifying secondary branches and distal vessels, their effectiveness was inherently limited by the resolution of the training and testing images ($0.38 \times 0.38 \times 0.625$ mm^3). Smaller lesions with stenosis diameters between 0.45 and 0.9 mm require higher-resolution images for accurate segmentation.

Better resolution can be achieved by interpolating the original images. This process involves resampling to increase the pixel density, improving the granularity of anatomical features (see Section 2.2.1.1). However, this approach introduces significant challenges:

1. Memory Usage: Higher-resolution resampled images (600–800 pixels per side) significantly increase computational and memory demands compared to the original dimensions (e.g., 512×512 or 400×400).
2. Class Imbalance: Interpolation magnifies the disparity between the background and vessel classes, reducing the relative importance of the vessel class and potentially lowering segmentation precision.

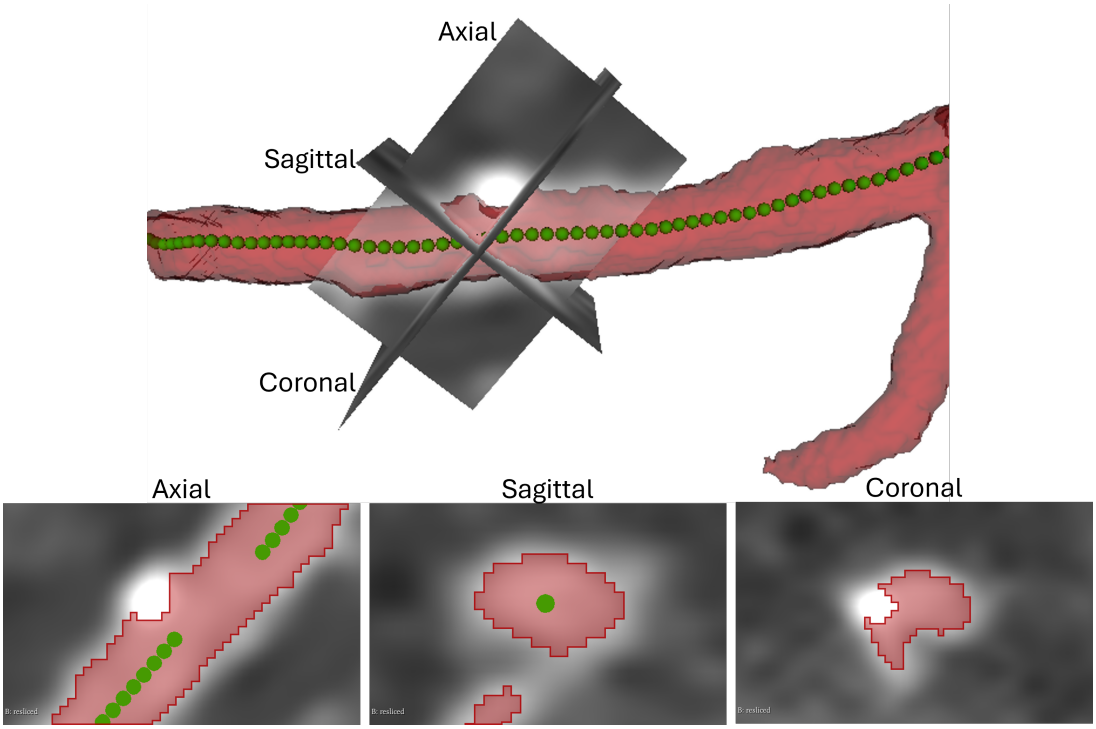


Figure 4.16: Segmentation of a vessel in the axial, sagittal, and coronal planes. The segmentation delineates the vessel lumen while excluding a visible calcium deposit. Green line represent the centerline. Voxel size is 0.25mm^3. Medical image provided by FlowReserve Labs S.L.

To address these issues, we develop an alternative segmentation strategy. Supervised neural networks provided an initial segmentation to extract the coronary artery centerlines. These centerlines were used to isolate specific views of each vessel point, as axial, sagittal and coronal planes illustrated in, Figure 4.16. This localized, focused approach mitigates the memory and class imbalance challenges.

However, resampling the original images necessitates re-annotating ground truth data, which is a labor-intensive process. To address this, we implement an approach combining unsupervised clustering algorithms with graph-based analysis. The clustering algorithms group similar data based on pixel intensity (in HU) and spatial distribution, while the graph structure provides a framework for connecting and interpreting clusters. This integrated strategy avoids the need for manual annotation, offering a scalable and computationally efficient solution for high-resolution coronary segmentation.

4.3.1 Methodology

We propose a novel approach utilizing a 2D clustering technique to achieve comprehensive 3D coronary artery segmentation. The method analyzes individual image slices to extract geometric features, spatial relationships, and inter-cluster connections, which are organized into a graph representation. This graph aids in automatically identifying the relevant vessel cluster for segmentation. The combined segmentations from all slices reconstruct the 3D geometry, followed by post-processing to refine the results.

The workflow involves:

1. Image extraction
2. Initial clustering (Ward's method)
3. Graph construction and background removal
4. Reapplication clustering algorithm (Ward's method)
5. Threshold selection
6. Coronary artery segmentation
7. Postprocessing

4.3.1.1 Dataset

The dataset for this study was provided by the Clinical University Hospital of Santiago de Compostela. It includes 22 patients with 30 clinician-diagnosed lesions (lesion set) and 10 patients without identified lesions (test set). The test set, also used in the previous section (see Section 4.2.1.3), was analyzed at the new resolution specified in this study. For additional details, please refer to dataset number 2 in Section 2.2.1.1.

Each image is 32×32 pixels, centered on the vessel, with a voxel resolution of $0.25 \times 0.25 \times 0.25$ mm^3. Vessel centerlines were extracted using 3D Slicer (v5.2.2) [64] with the Extract Centerline module, resampled at 0.25 mm intervals to maintain completeness across views. The dataset includes $46,307$ lesion-set images and $22,056$ test-set images.

Two segmentation algorithms are developed: 3Axis, which independently analyzes axial, sagittal, and coronal views, and Perp, which uses perpendicular cross-sections of the vessel for localized detail. These methods integrate directional and spatial information for improved vessel segmentation.

4.3.1.2 Ward's Clustering Method

The proposed segmentation method uses Ward's clustering algorithm [138], an agglomerative technique designed to minimize the variance within each cluster. The process begins with each data point (pixel of the image) as its own cluster, then iteratively merges the closest clusters to reduce the increase in variance. This continues until the desired number of clusters is formed. While initially expecting two primary clusters (vessel lumen and background), the complexity of coronary anatomy requires more clusters to distinguish various adjacent structures such as vessels, myocardium, fat, and calcium, all of which must be separated accurately for effective segmentation.

The clustering approach implemented in this study utilizes Agglomerative Ward clustering [138] with grid connectivity, specifically using the *grid_to_graph* function from *sklearn.feature_extraction.image* [139]. This method applies a grid structure to the data, where each pixel is connected to its neighboring pixels based on their relative positions in a 2D grid. This connectivity ensures that the algorithm captures the spatial relationships between adjacent pixels, enhancing the clustering's ability to preserve structure and produce meaningful groups for accurate segmentation.

The determination of the optimal number of clusters is the first critical step. Similar to other clustering methods like k-means, Ward's algorithm requires this parameter as input. Selecting the correct value is non-trivial and involved preliminary testing.

We performed a visual inspection to determine that 7 clusters provided the best differentiation of regions within and around the vessel, enabling the exclusion of non-relevant structures while isolating the vessel of interest.

4.3.1.3 Graph representation and background removal

The background removal algorithm and its application is demonstrated in Figure 4.17. In this figure, the original perpendicular image of the vessel after a bifurcation is shown in (a), and the result of the background removal algorithm applied to this image is presented in (b). The Ward's clustering extraction technique is applied to both the original and background-removed images, as seen in (c) and (d), respectively. The network extraction from the clusters is shown in (e) and (f), where the node numbers correspond to the clusters in (c) and (d), respectively. The x-axis in these plots represents the mean attenuation value (Hounsfield Unit, HU) of each cluster, while the y-axis represents the distance in pixels between the centroid of the cluster and the centerline point of the vessel. The edges between nodes represent the Euclidean distance between the centroids of adjacent clusters, defining the spatial relationships between them.

The clusters generated by Ward's algorithm are transformed into a graph where each node represents a cluster. Each node is assigned two coordinates: the x-coordinate reflects the distance from the vessel center to the centroid of the cluster, while the y-coordinate represents the mean Hounsfield Unit (HU) value of that cluster. The edges connecting the nodes are defined by the spatial adjacency of the clusters. If two clusters are neighboring, an edge is formed, with the edge's weight representing the Euclidean distance between the corresponding nodes (as seen in Figure 4.17e and Figure 4.17f). The graph is further refined by removing cycles using the Boruvka algorithm [140, 141], which computes the minimum spanning tree.

After constructing the graph, the vessel path is identified by selecting the clusters that correspond to the vessel, starting from the brightest cluster closest to the vessel center, which likely indicates the vessel lumen or calcium deposits. The path is traced by moving through the graph, following neighboring nodes with descending mean HU values. This process continues until nodes with an HU value below a predefined threshold (set to 100 HU by default) are encountered. Clusters not included in this ordered list are considered background and are removed by setting the corresponding pixels to a uniform value in the original image, effectively masking the vessel (see Figure 4.17b).

Masking the vessel is essential for overcoming a frequent issue when vessels are in proximity to other anatomical structures, such as neighboring vessels or the myocardium, often referred to as the *kissing vessel artifact* [142]. This situation causes Ward's algorithm to struggle with distinguishing between closely situated structures, which leads to a reduction in cluster resolution (see Figure 4.17a and Figure 4.17c). By masking the vessel, we isolate only the brightness variations within the vessel, thereby improving segmentation precision. Afterward, we reapply the Ward's algorithm to the cleaned image, as demonstrated in Figure 4.17b and 4.17d.

Once the background has been masked, Ward's clustering algorithm is reapplied to the masked image, and the vessel path is extracted following the same procedure as previously described.

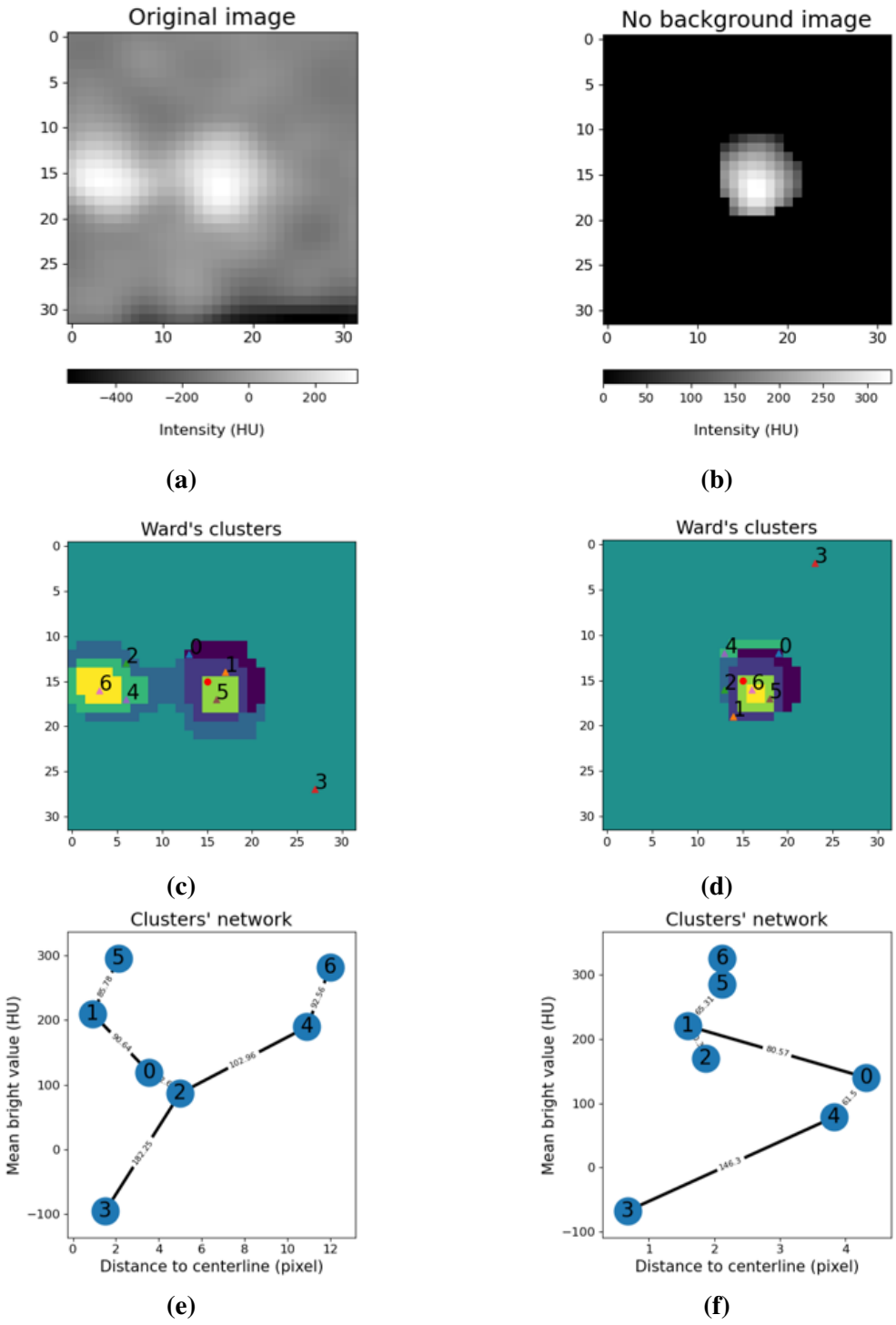


Figure 4.17: Background removal algorithm results. (a) Original perpendicular image of the vessel after a bifurcation. (b) Background removal algorithm result with input image (a). (c) Ward's clustering extraction of input image (a). (d) Ward's clustering extraction of input image (b). (e) Network extraction from clusters in (c). (f) Network extraction from clusters in (d). For both (e) and (f), the node number represents the cluster in (c) and (d), respectively. x-axis is the mean attenuation value of the cluster and y-axis is the distance in pixels between the centroid of the cluster and the centerline point. Voxel size is 0.25mm^3. The connection weight between nodes represents the Euclidean distance between the corresponding clusters. Image from [65].

4.3.1.4 Image segmentation

To segment the coronary artery lumen, a threshold is chosen based on the vessel path derived in Section 4.3.1.3. This path prioritizes the brightest clusters, representing the vessel's interior, while clusters further down represent the edges and background. Clusters within the brightness range of $[100, 600]$ HU are retained as they correspond to the coronary vessel [66, 67, 68, 69]. If there are more than three clusters, the third is selected; otherwise, the last is used. If no clusters meet these criteria, a default threshold of 150 HU is applied to ensure segmentation continuity. This thresholding approach avoids irregular segmentation shapes, maintaining a circular geometry consistent with the vessel structure.

The segmentation approach varies by method. For the perpendicular (Perp) method, segmentation is performed within a 2 mm spherical radius around the centerline point. In bifurcations, segmentation progresses forward along the blood flow, prioritizing areas with lower thresholds to ensure proper overlap and sequence.

In contrast, the three-plane (3Axis) method uses a flat, slice-specific brush with a 4 mm radius on axial, sagittal, and coronal planes, addressing the need for larger coverage as vessel shapes differ across views. Each plane is segmented individually before being combined, followed by a 1 mm median smoothing applied using the *Segment Editor* module in 3D Slicer to refine the final geometry [64].

4.3.1.5 Competing methods

To evaluate the performance of the Ward algorithm, we employ neural network-based algorithms as competing methods. This comparison is essential to assess whether an unsupervised model can outperform established supervised solutions in the segmentation task. Given that the Ward algorithm captures information from three orthogonal planes—axial, sagittal, and coronal—in its 3Axis version, or from a single perpendicular plane in its Perp version, we utilize both 2.5D and 3D deep learning architectures for comparison.

For the 2.5D models, we implement four well-known architectures as baseline comparisons: VGG-19 [79], ResNet-50 [80], EfficientNet-B2 [81], and U-Net++ [78] (see more architecture details in Section 2.3.1). The VGG-19, ResNet-50, and EfficientNet-B2 architectures are incorporated into a U-Net framework, where their pretrained encoders serve as the feature extraction backbone. Each encoder has been pretrained on the ImageNet dataset [82] to leverage learned representations from large-scale natural image data. This approach allows the networks to benefit from transfer learning, potentially enhancing performance in the medical imaging domain.

To ensure that the models capture as much anatomical context as possible while maintaining a manageable computational cost, we define the input dimensions as three slices (axial, sagittal, and coronal) with a resolution of 32×32 pixels.

In addition to the 2.5D architectures, we incorporate three 3D segmentation networks to compare our method against established reference models. Specifically, we evaluate two U-Net-based architectures and one transformer-based model. Following the methodology described in [143], we implement a basic 3D U-Net, which utilizes convolutional blocks composed of two fused convolutional layers. Recent studies [144] have demonstrated that integrating attention mechanisms in skip connections can improve segmentation performance in certain applications. To explore this effect, we include the U-Net-DR-LCT model [145], which leverages Local Contextual Transformers to refine feature extraction and

Table 4.3: Comparison of EfficientNet-B2, ResNet-50, VGG-19, U-Net++, 3D U-Net, 3D U-Net DR, and 3D Swin UNETR deep learning models in terms of their total number of parameters, including both trainable and non-trainable parameters. Table from [65].

Model	Nº of param. (M)	Trainable param. (M)	Non-trainable param. (M)
EfficientNet-B2	14.295	14.226	0.069
ResNet-50	51.605	51.505	0.099
VGG-19	29.062	29.058	0.004
U-Net++	9.163	9.163	0
3D U-Net	22.581	22.581	0
3D U-Net DR	10.700	10.700	0
3D Swin UNETR	15.703	15.703	0

employs Dense Residual blocks to preserve fine-grained structural details. This model was specifically designed for coronary artery segmentation, making it a relevant benchmark for comparison. Consequently, evaluating our proposed approach against this model also serves as a baseline for assessing performance improvements. Finally, considering the growing impact of transformer-based architectures in computer vision, we include Swin-UNETR [146], a state-of-the-art medical image segmentation model that utilizes self-attention mechanisms to enhance feature representation. More detail about the transformer architecture can be found in Section 2.3.2). For consistency in input representation across 3D models, we use $32 \times 32 \times 32$ voxel blocks as input data.

Details on encoder architectures and parameter counts are provided in Table 4.3.

Training Dataset The dataset consists of 32 patients, with coronary artery structures manually annotated by a medical imaging expert using images with isotropic resolution of 0.25 mm. We divide the data into 27 patients for training (85%) and 5 for testing (15%), ensuring a balanced evaluation [134]. A key challenge is the severe class imbalance, as only 5% of cases contain calcium plaques, and lesion regions represent less than 3% of the dataset.

Input images are centered on centerline points to maintain artery focus. The total centerline points is 52.4K. Given the scarcity of critical regions such as plaques and lesions, we apply targeted augmentations (rotations, flips, brightness changes, and zooms), increasing their representation to 25% of the training set and expanding the dataset to 65K images.

For preprocessing, we clip voxel intensities to $[-200, 1150]$ HU and apply min-max normalization [147] to scale values to $[0, 1]$, preserving relevant features while preventing incorrect calcium segmentation.

Training Setup We use Tversky loss [103] ($\alpha = 0.4$, $\beta = 0.6$) to penalize false positives and over-segmentation (see Section 2.3.5), training for 50 epochs with early stopping after 15 epochs without improvement. The Adam optimizer [120] is set to an initial learning rate of 0.001, adapting via ReduceLROnPlateau, that reduce learning rate when the metric has stopped improving. Augmentations are precomputed to speed up training.

Training runs on 64 Intel Xeon Ice Lake 8352Y cores with 1 NVIDIA A100 GPU.

4.3.1.6 Evaluation metrics

The evaluation metrics employed in this study include Dice coefficient, Intersection over Union (IoU), Precision, and Recall. These metrics are used to assess the segmentation performance and are described in detail in Section 2.3.6.2.

4.3.2 Results

This section presents the results obtained from the proposed methodology, structured into several key aspects. First, in Section 4.3.2.1, we detail the selection of the algorithm and parameter settings, providing insights into the optimization process and justifications for the chosen configurations. Next, in Section 4.3.2.2, we evaluate performance on the test set, assessing accuracy, robustness, and reliability across different cases. Section 4.3.2.3 focuses on the interpretability and clinical relevance of the developed methodology, analyzing how the results align with real-world clinical applications. Following this, in Section 4.3.2.4, we examine the lesion set, discussing the method's ability to identify and segment pathological structures. Finally, in Section 4.3.2.5, we provide an analysis of computation time, highlighting the efficiency of the approach and its feasibility for integration into clinical workflows.

4.3.2.1 Algorithm selection and Parameter Setting

To ensure the selected clustering algorithm met the study's requirements, several options were evaluated, focusing on their ability to handle varying brightness distributions, anatomical orientations, and vessel structures. The algorithms tested included KMeans [148], DBScan [149], standard Agglomerative Ward clustering [138], an enhanced version of Ward clustering (Ward*, our method) and Spectral Clustering [150]. All of them following consistent implementations:

- KMeans with 7 clusters.
- DBScan with eps=3.5 and min_samples=5.
- Agglomerative Ward clustering with 7 clusters (denoted as Ward).
- Agglomerative Ward clustering with grid connections and 7 clusters (denoted as Ward*, our method).
- Spectral Clustering (affinity=nearest_neighbors, random_state=0).

The primary criterion for evaluation was the ability to accurately delineate the concentric edges of vessels while maintaining consistency across examples.

In order to evaluate the performance of different clustering algorithms for vessel segmentation, we analyze their results across multiple anatomical planes. Each figure, from Figure 4.18 to Figure 4.20, presents a different vessel orientation, highlighting specific challenges in segmentation. The figures include the original image, the result after applying the background removal algorithm, and the segmentations produced by KMeans, DBSCAN, Ward, Ward* (our method), and Spectral Clustering.

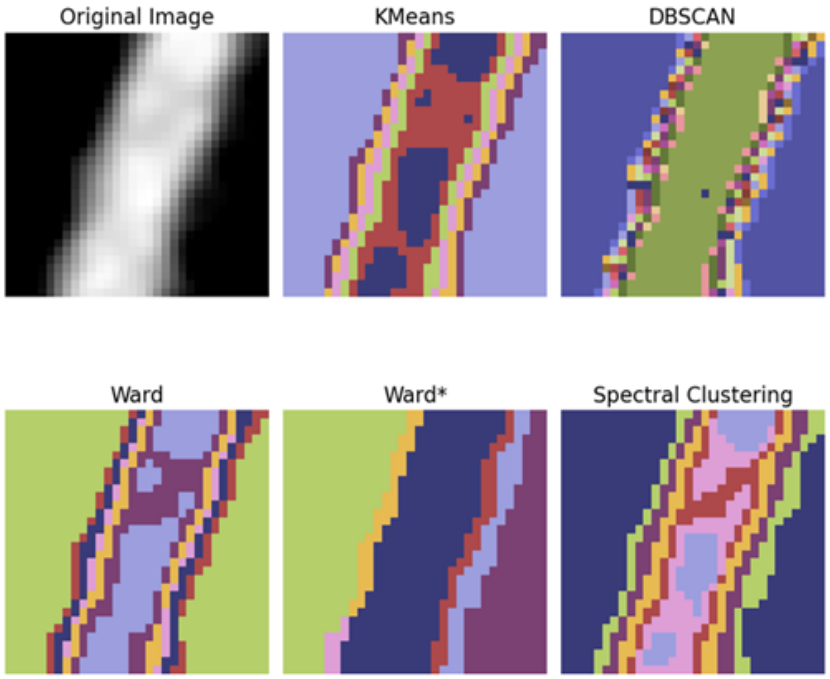


Figure 4.18: Comparison of segmentation results from different clustering algorithms—KMeans, DBSCAN, Ward, Ward* (our method), and Spectral Clustering—on the longitudinal view of a vessel in the sagittal plane. Voxel size is 0.25mm^3. The vessel maintains a smooth and uniform border, free of irregularities. Image from [65].

In the sagittal view shown in Figure 4.18, DBScan produced irregular boundary patterns, which failed to match the concentric shape of the vessel. In contrast, the Ward* method ensured consistent clustering across both the lumen and the boundary. KMeans, Ward and Spectral Clustering successfully obtain boundary clsuters but failed to generate compact clusters in the lumen.

Moving to the axial view in Figure 4.19, most algorithms successfully generated circular clusters within the lumen; however, only the Ward* method extended this uniformity to the boundary, demonstrating its superior performance.

Finally, in the coronal view presented in Figure 4.20, while DBScan created homogeneous clusters in this particular example, it lacked generalizability and struggled to maintain this performance in other cases from the same patient.

As a conclusion, Ward* was selected for its superior ability to produce homogeneous, compact clusters consistently, across varying conditions, ensuring reliable segmentation of vessel edges critical for this study's objectives.

As described in Section 4.3.1.2, the Ward's clustering algorithm requires the number of desired clusters as input. Selecting the appropriate number of clusters is not a straightforward task, so we tested several configurations to find the best fit.

To further evaluate the clustering performance, we conducted a visual inspection using images from patients outside the test and lesion datasets. Figure 4.21 illustrates an example of this analysis, showcasing the results of the Ward clustering algorithm in its second iteration, after background removal. The figure consists of several components. Panel (a) presents the original image along with its attenuation values in Hounsfield Units (HU). Panels (b) through

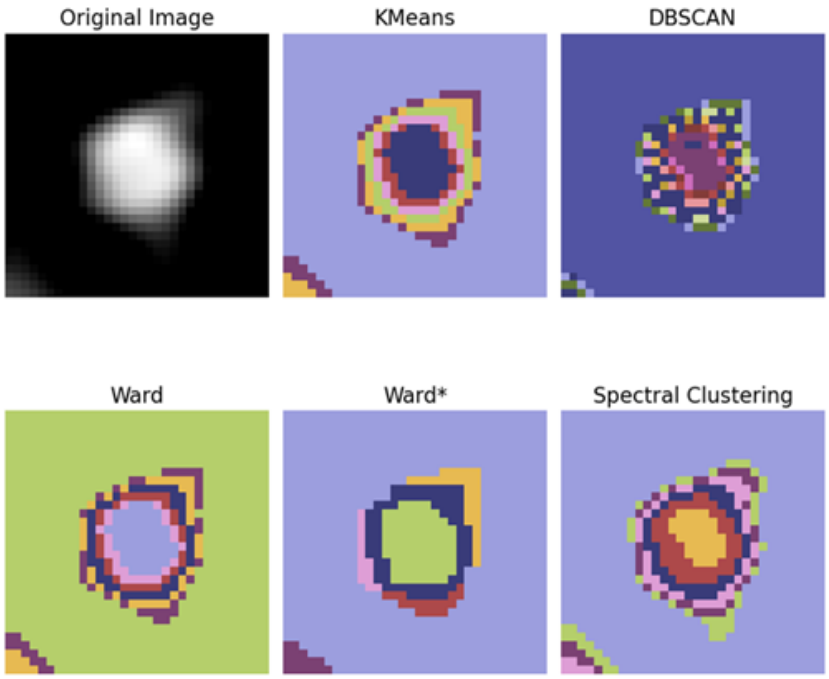


Figure 4.19: Segmentation outcomes from various clustering algorithms—KMeans, DBSCAN, Ward, Ward* (our method), and Spectral Clustering—applied to the axial view of the vessel. Voxel size is $0.25mm^3$. The section is nearly perpendicular to the centerline, highlighting its characteristic circular structure. Image from [65].

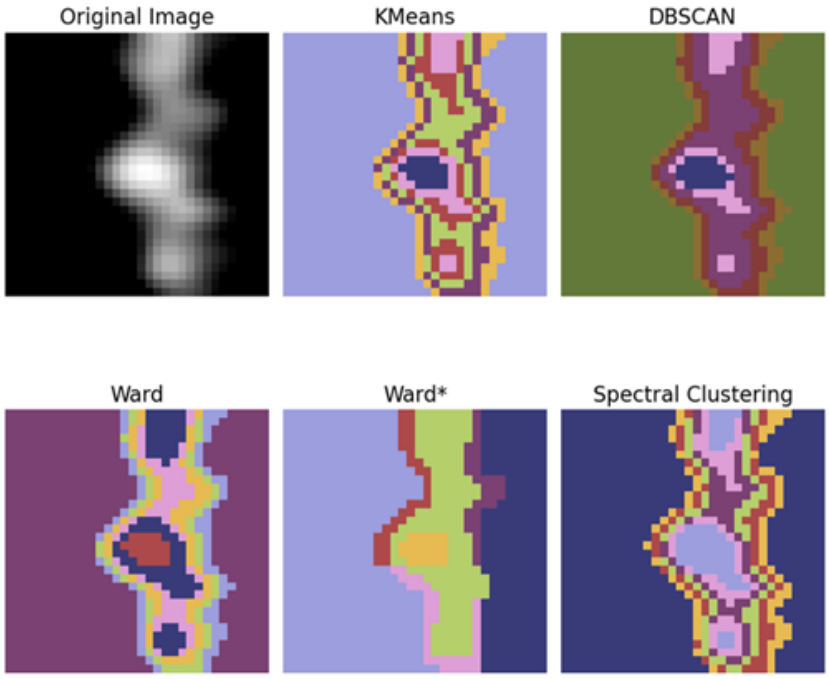


Figure 4.20: Segmentation results from KMeans, DBSCAN, Ward, Ward* (our method), and Spectral Clustering on the longitudinal view of the vessel in the coronal plane. Voxel size is $0.25mm^3$. This perspective reveals the presence of lesions and irregularities along the vessel's boundary. Image from [65].

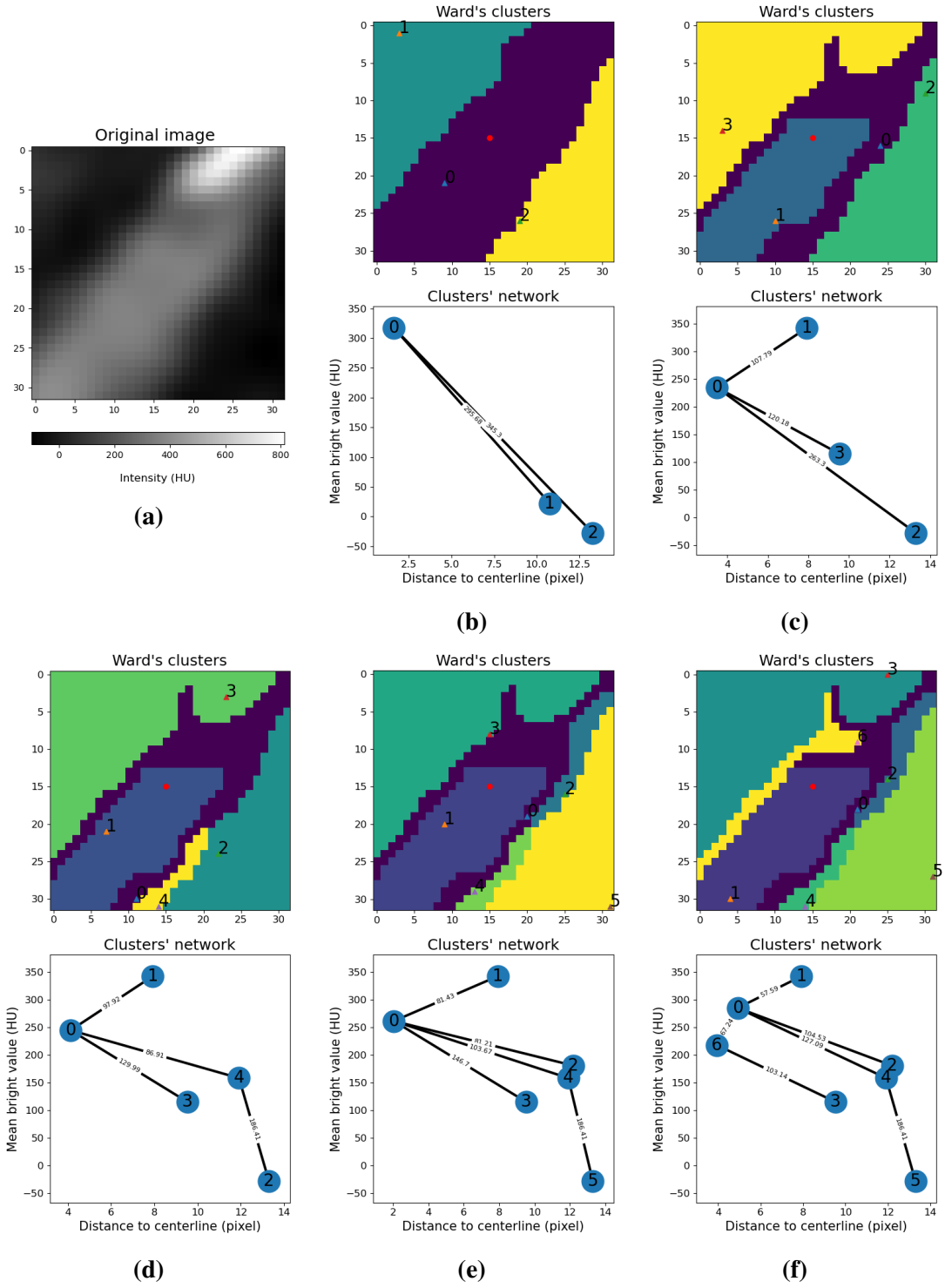


Figure 4.21: Results of the clusters obtained by the Ward algorithm in the second iteration (after removing the background of the image). (a) The original image and its attenuation values in Hounsfield Units (HU). Figures (b)-(f) show the clusters obtained by the Ward algorithm with different number of clusters (nC) at the top, and the graph associated with the clusters is shown at the bottom. X-axis is the mean attenuation value of the cluster and y-axis is the distance in pixels between the centroid of the cluster and the centerline point. The connection weight between nodes represents the Euclidean distance between the corresponding clusters. (b) $nC = 3$. (c) $nC = 4$. (d) $nC = 5$. (e) $nC = 6$. (f) $nC = 7$. Medical image voxel size is 0.25mm^3. Image from [65].

(f) display the clustering results for different numbers of clusters (nC), ranging from $nC = 3$ to $nC = 7$. Each panel contains two sections: the top section shows the segmented clusters, while the bottom section depicts the corresponding graph representation of the clusters.

In the original image showing a section of the vessel with calcium, we tested various number of clusters. With 3 clusters ($nC = 3$), the background and vessel (including calcium) were distinguishable. With $nC = 4$, the calcium was effectively removed and classified as background due to the background removal algorithm. For more precision at the vessel's edge, $nC = 7$ was determined to be optimal. Additional clusters would likely result in non-clinically relevant structures.

4.3.2.2 Test set

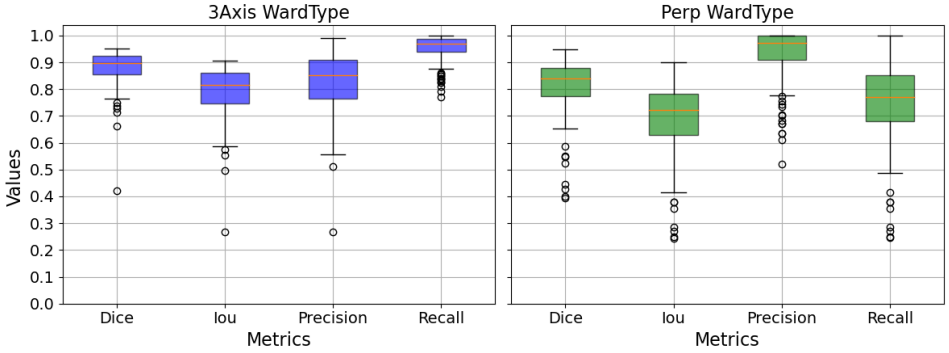
In this section, the performance of clustering algorithms and neural networks is compared using various metrics, particularly recall and precision. These two metrics provide insight into how well the vessels are recognized (recall) and how accurately they are segmented (precision). High recall with low precision may indicate oversegmentation, whereas low recall with high precision could point to undersegmentation. These insights help evaluate the strengths and weaknesses of the different methods.

Figure 4.22 presents a comparative analysis of the segmentation performance of clustering algorithms and 2.5D neural networks. The figure consists of box plots that illustrate the distribution of segmentation metrics—Dice, IoU, Precision, and Recall—across the test set, providing insight into the variability and accuracy of each approach. In these box plots, the central line represents the median, the box spans the interquartile range (IQR), and the whiskers extend to the minimum and maximum values within 1.5 times the IQR, with potential outliers displayed as individual points. The x-axis represents the segmentation model, while the y-axis indicates the corresponding parameter value. Panel (a) compares the clustering-based segmentation methods across different viewing planes. The 3Axis method aggregates results from axial, sagittal, and coronal views, while the Perp method evaluates segmentation in the perpendicular view. Panel (b) showcases the performance of the 2.5D neural network architectures, including EfficientNet, VGG, ResNet, and U-Net++, highlighting their segmentation accuracy across the same set of metrics.

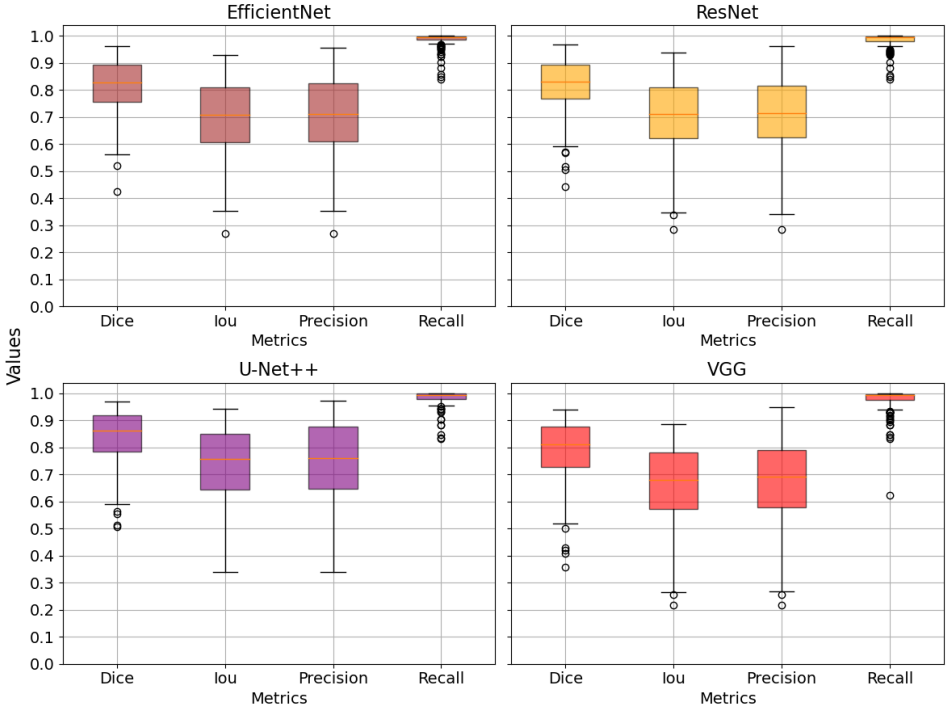
We begin by analyzing the Ward, 3Axis, and Perp methods, as shown in Figure 4.22a. A significant difference is observed in the recall metric: 3Axis yields a recall value that is 20% higher (mean: 0.95) than Perp, while its precision value is 10% lower (mean: 0.83). This suggests that 3Axis is more adept at identifying the majority of the vessel but is prone to a higher number of false positives, leading to oversegmentation. Regarding Dice and IoU, the 3Axis method outperforms Perp, with mean values of 0.88 and 0.79, respectively, compared to 0.81 and 0.7 for Perp. The Dice coefficient, emphasizing true positives, accounts for this performance difference, favoring methods that capture the vessel boundaries more accurately.

Next, we compare the performance of the Ward method with the 2.5D neural networks. A key observation is the high recall of the neural networks, which enables them to recognize nearly all of the vessel, including distal sections. However, their precision is lower than the Ward method, with the mean precision reduced by over 10%. This decrease in precision may be due to the networks struggling with difficult segmentation areas, indicating a tendency to oversegment vessels.

Additionally, the 2.5D neural networks show longer whiskers in the box plots, reflecting



(a)



(b)

Figure 4.22: Box plot displaying the mean parameter values, Dice, IoU, Precision and Recall, obtained by the segmentation algorithms in the test set. (a) Clustering segmentation algorithms in three views (axial, sagittal, and coronal), referred to as 3Axis, and the segmentation algorithm in the perpendicular view, referred to as Perp. (b) 2.5D neural network architectures, EfficientNet, VGG, ResNet and U-Net++. Image from [65].

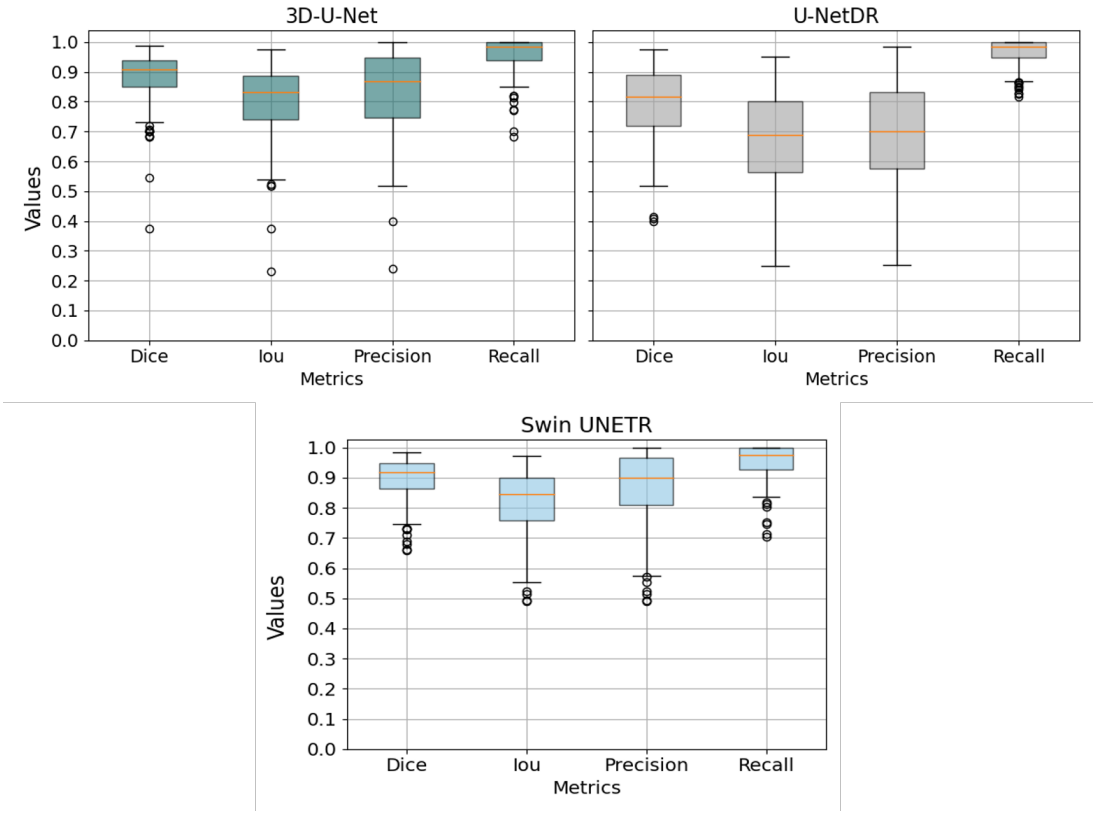


Figure 4.23: Box plot displaying the mean parameter values, Dice, IoU, Precision and Recall, obtained by the 3D segmentation algorithms in the test set. Image from [65].

more variability in their results.

To conclude, we assess the performance of the 3D neural networks in comparison to the clustering algorithms used in this study, as shown in Figure 4.23. The figure presents box plots illustrating the distribution of key segmentation metrics—Dice, IoU, Precision, and Recall—computed across the test set.

The 3D U-Net, which benefits from contextual information across adjacent slices, achieves a mean Dice coefficient of 0.88 (std: 0.0888), a result that closely mirrors the 3Axis method. In contrast, the 3D U-NetDR performs less effectively, obtaining a mean Dice coefficient of 0.64 (std: 0.23), highlighting its limitations. Notably, the Swin UNETR outperforms the 3Axis method with a Dice coefficient of 0.8978 (std: 0.0706), though this advantage comes with increased computational demands due to its transformer-based design.

In Figure 4.24, we present an example of the Ward segmentations, showing the 3D geometries corresponding to the ground truth, 3Axis, and Perp methods. These are depicted alongside a detailed plane from the area highlighted in pink within these geometries. The results are shown for three test patients, T003, T006, and T008, each displayed in separate rows. The first column shows the ground truth, represented in red, while the second and third columns show the segmentations using the 3Axis clustering method (in blue) and the perpendicular clustering method (in green), respectively. The final column presents the detailed plane, which is highlighted with an orange line in the 3D geometries, displayed as 2D images. This layout allows for a direct comparison of the segmentation accuracy and provides insights into the performance of each method across the different test patients.

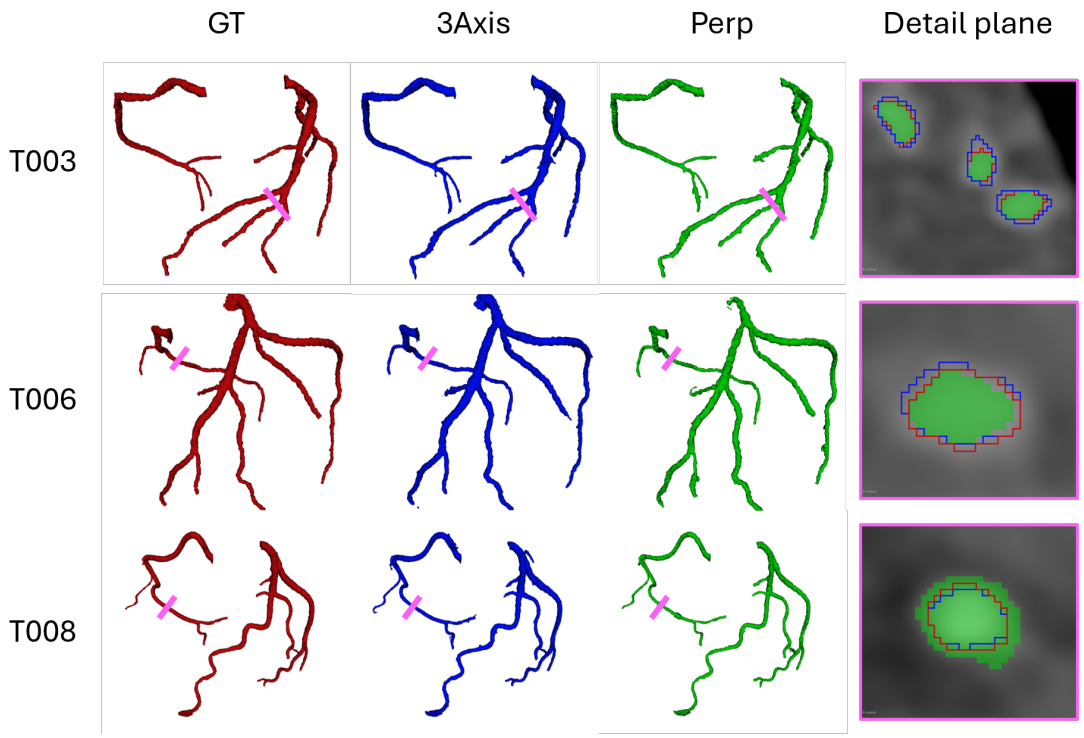


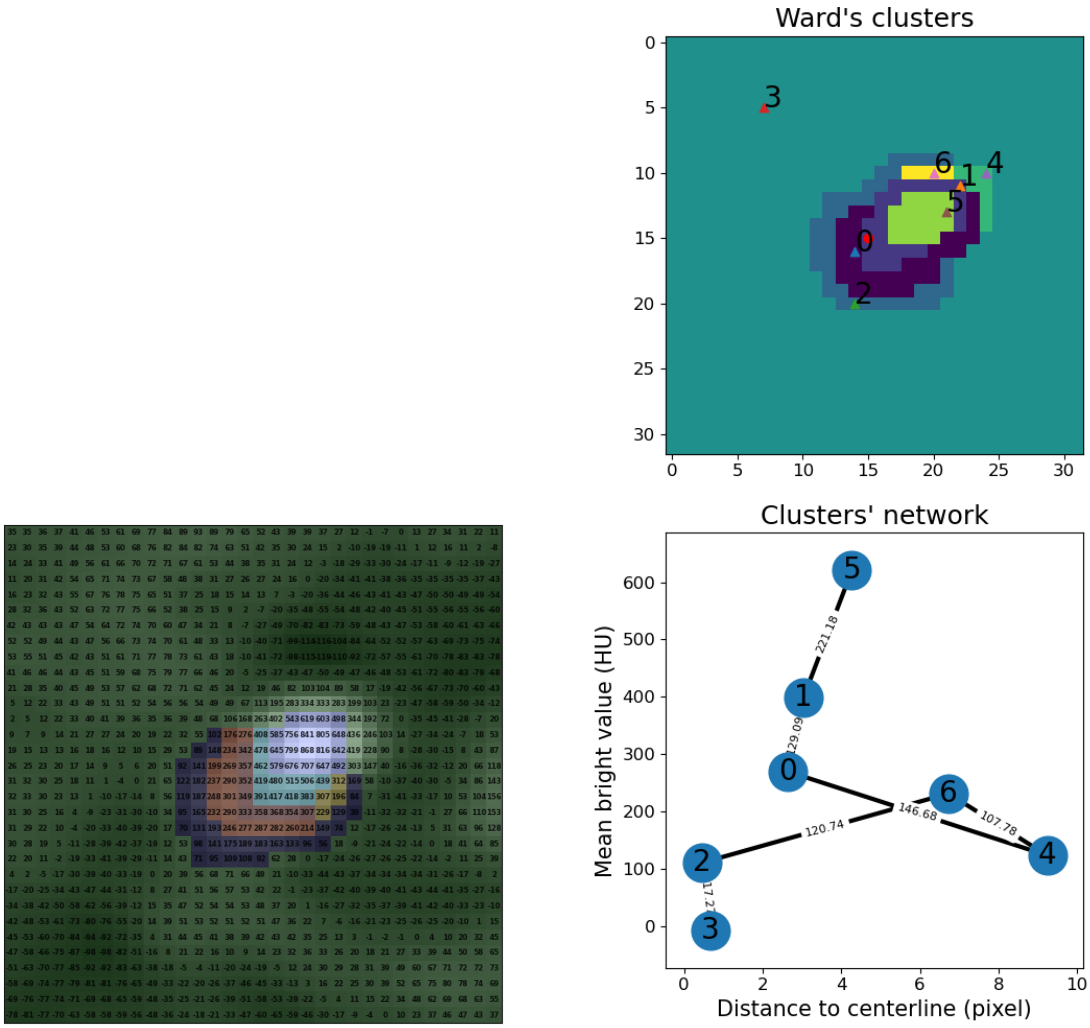
Figure 4.24: Results of the prediction for test patients T003, T006, and T008 (in rows). The columns represent the ground truth (red), segmentation using the 3Axis clustering method (blue), segmentation using the perpendicular clustering method (green), and the detail plane, depicted with an orange line in the previous geometries, in the 2D images. Medical image voxel size is 0.25mm^3. Image from [65].

In the detailed view, rows 1 and 2 reveal oversegmentation by the 3Axis algorithm, while row 3 highlights oversegmentation by the Perp algorithm. These inaccuracies may result from variations in patient brightness, which can affect the algorithm's ability to differentiate between vessel boundaries and surrounding tissue.

4.3.2.3 Interpretability and Clinical Relevance of the Developed Methodology

One significant advantage of this method lies in its transparency and interpretability. The clustering algorithm operates by analyzing pixel brightness values (measured in HU) and their spatial arrangement. Since brightness reflects tissue composition and density, this approach provides insights into the nature of the imaged structures. Beyond segmentation, it can be used to visualize suspicious regions or potential lesions, assisting clinicians in making informed decisions.

Figure 4.25 illustrates the results of the clustering methodology applied to vessel segmentation, following background removal. Figure 4.25a presents a detailed representation of the clusters, with pixel intensities shown in Hounsfield Units (HU). This provides an understanding of the distribution of tissue densities within the vessel, highlighting the background as the green cluster and calcium deposits as the violet cluster, which is a key feature for identifying pathological areas. Figure 4.25b shows the graph-based arrangement of these clusters, spatially organizing them from the vessel's interior to its outer edge, based on



(a) **(b)**

Figure 4.25: Outcomes of the clustering methodology applied to vessel segmentation.(a) Detailed representation of the clusters, where pixel intensities are expressed in Hounsfield Units (HU), offering insight into the distribution of tissue densities. (b) Graph-based arrangement of these clusters, which organizes them spatially from the vessel's interior to its edge. Medical image voxel size is 0.25mm^3. Image from [65].

decreasing brightness. This systematic organization is derived from the graph, and allows for a clear visualization of the vessel structure and pathological features.

This systematic arrangement enhances the ability to identify critical areas of interest, such as abnormalities or lesions, offering valuable diagnostic information and supporting early detection of potential issues.

4.3.2.4 Lesion set

In addition to evaluating the algorithms on the test set, we also assessed their performance on a dataset of 22 patients with 30 diagnosed lesions. The results are presented in Figure 4.26. This figure presents a box plot that displays the values of various segmentation metrics—Dice, IoU, Accuracy, Recall, and Precision—obtained by the clustering segmentation algorithms on the lesion dataset. Panel (a) shows the box plot for the values across the entire coronary tree, comparing the 3Axis method, which combines results from axial, sagittal, and coronal views, with the Perp method, which focuses on the perpendicular view. Panel (b) zooms in on a cube of 8mm centered around the lesion, providing the same metrics, to evaluate the performance specifically within the region of the lesions.

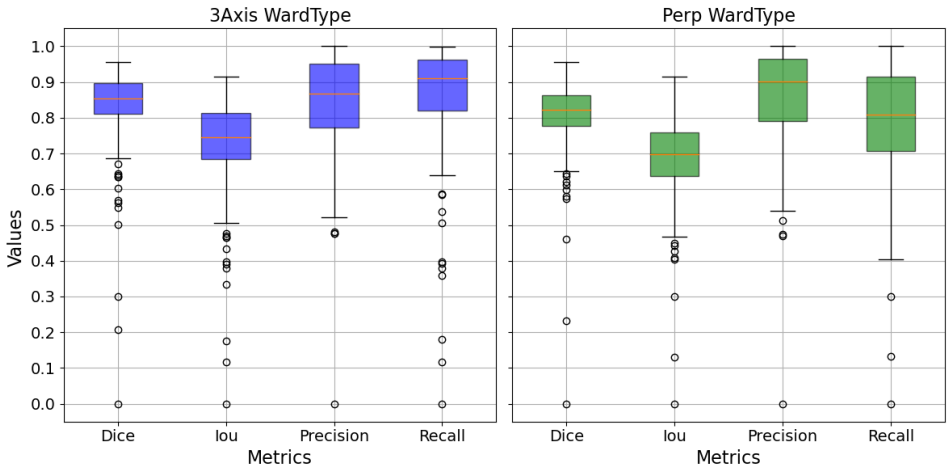
When comparing the complete patient results for the 3Axis and Perp (see Figure 4.26a), the 3Axis method outperforms with a higher mean Dice coefficient (0.8371 vs. 0.8094), suggesting better overall segmentation accuracy. However, Perp achieves slightly better mean Precision (0.8599 vs. 0.8431), while 3Axis shows superior mean Recall (0.8640 vs. 0.7976), meaning it is better at identifying the vessels but at the cost of a slightly higher number of false positives.

In Figure 4.26b, we examine the performance within a cube of 32 pixels (8 mm), centered on the lesion. Perp shows slightly better performance in Dice (0.7906 vs. 0.7704) and precision (0.7745 vs. 0.7369), while 3Axis maintains superior Recall (0.8731 vs. 0.8529). This indicates that the 3Axis method is more sensitive, accurately identifying a larger portion of the vessel.

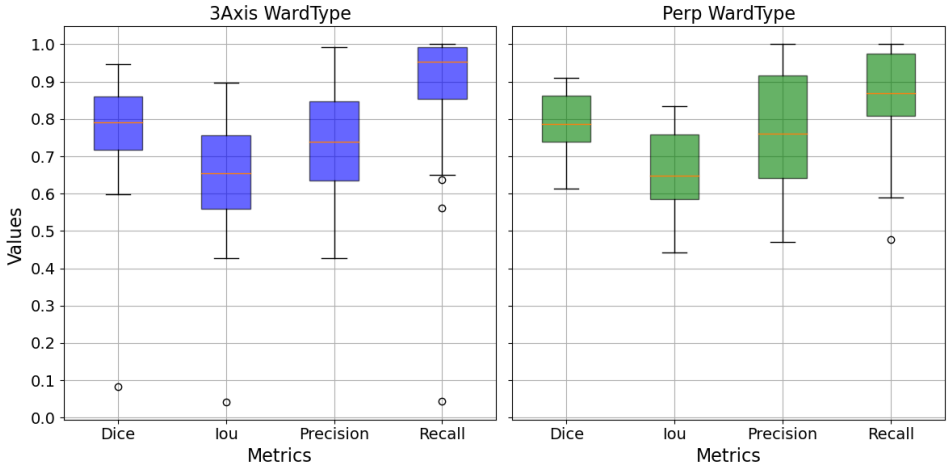
When comparing the results for the complete patient dataset to the lesion dataset, both methods show a decrease in performance for the lesion data. The mean Dice coefficients drop for both methods, with 3Axis decreasing from 0.8371 to 0.7704 and Perp dropping from 0.8094 to 0.7906. Recall values remain relatively high, although they decrease slightly for 3Axis (from 0.8640 to 0.8731). Precision decreases more significantly for both methods (up to a 10%), indicating that lesions present a more challenging segmentation task and lead to lower overall performance.

In addition to the numerical metrics, visualizing lesions is essential for a complete evaluation. Figure 4.27 display examples of lesion segmentation for two patients. This figure presents 3D geometries of a patient with a lesion, emphasizing the impact of segmentation accuracy on diagnostic outcomes. Panel (a) and (b) focus on a lesion located at a bifurcation, where poor segmentation can alter blood flow distribution, with ground truth shown in red and the lesion region detailed. Panels (c) and (d) show a lesion as a pronounced stenosis in a straight segment, where segmentation errors can significantly affect diagnostic parameters, such as Fractional Flow Reserve (FFR). Segmentation results from the 3Axis clustering algorithm are shown in blue, and from the Perp clustering algorithm in green, ground truth is also shown in red. The figures highlight the importance of accurate segmentation for proper clinical assessment, particularly in complex lesions.

Both algorithms effectively detect vessel narrowing (stenosis). In Figure 4.27a and Figure 4.27b, however, both methods tend to underestimate the vessel, leading to narrower



(a)



(b)

Figure 4.26: Box plot displaying the mean parameter values, Dice, IoU, Accuracy, Recall and Precision, obtained by the clustering segmentation algorithms in three views (axial, sagittal, and coronal), referred to as 3Axis, and the segmentation algorithm in the perpendicular view, referred to as Perp in the lesion set. (a) Mean values obtained in the whole coronary tree. (b) Mean values obtained in a cube of 8mm centered in the lesion. Image from [65].

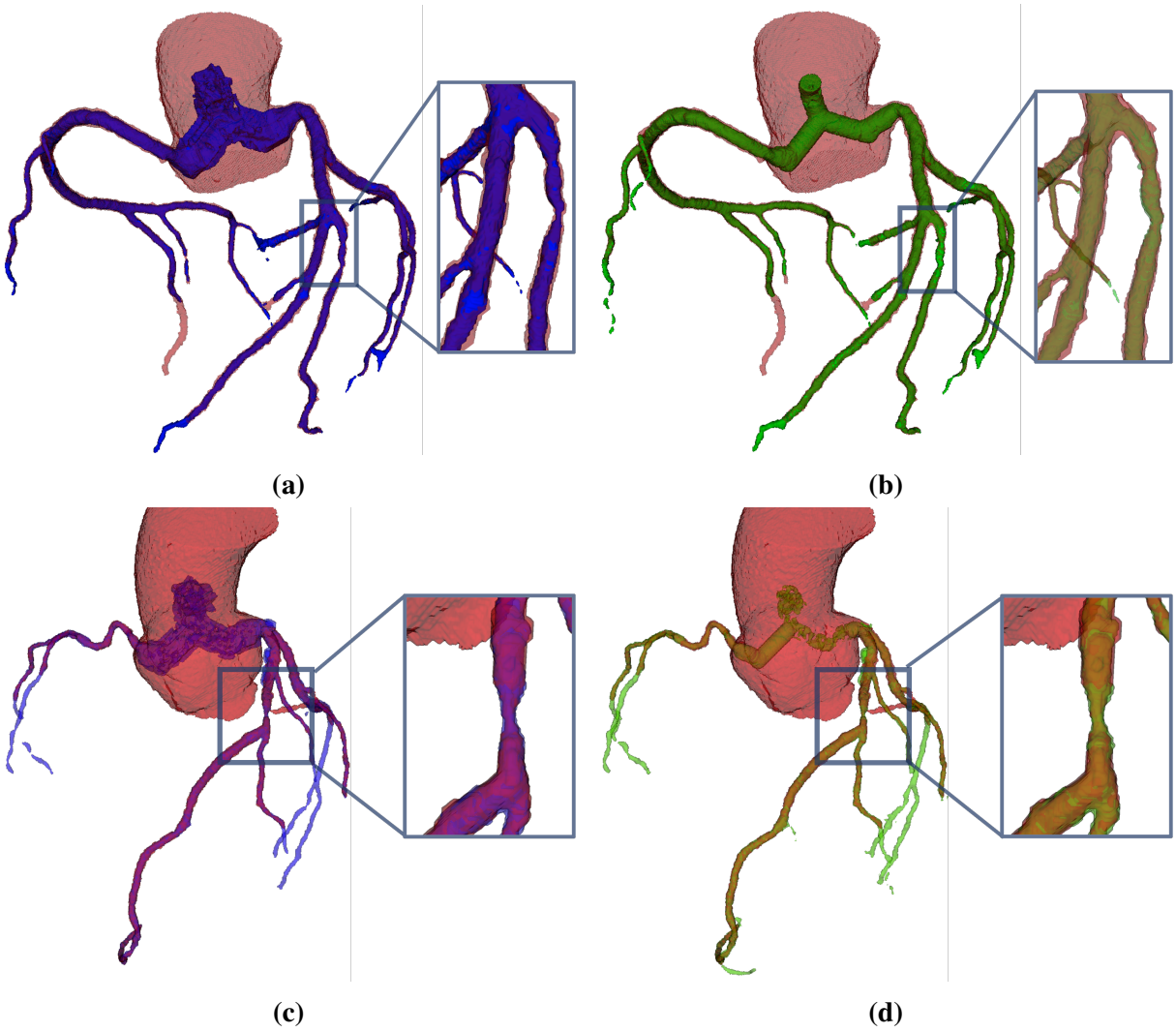


Figure 4.27: 3D geometries of a patient with a lesion, highlighting the impact of segmentation accuracy. In (a) and (b), the lesion is located at a bifurcation, where poor segmentation can alter blood flow distribution. In (c) and (d), the lesion is a pronounced stenosis in a straight segment, where segmentation errors can affect diagnostic parameters such as FFR. Ground truth is shown in red, with the lesion region detailed. Segmentation using the 3Axis clustering algorithm is shown in blue (a, c), while the Perp clustering algorithm is in green (b, d). Image from [65].

segmentations, with Perp slightly overestimating the lesion width. In contrast, Figure 4.27c and Figure 4.27d show both algorithms slightly overestimating the lesion width. Additionally, both methods mistakenly interpret fat, adhering to the vessel at the lesion's upper region, as part of the vessel due to its brightness in the image.

4.3.2.5 Computation time

In terms of speed, the Ward algorithm takes around 20 minutes to complete the segmentation task after centerline extraction. This is notably slower compared to the neural networks. The neural network methods, including the 2.5D networks, take approximately 7 minutes to segment and reconstruct 2.5D patches. For 3D methods, the 3D U-Net performs the segmentation in about 5 minutes, while U-NetDR and Swin UNETR take around 12 and 13 minutes, respectively.

The computations were carried out on a workstation with an Intel Core i5-8500T CPU and 32GB of RAM. The times were computed with the additional consideration of geometry visualization in 3D Slicer, which added approximately 2 minutes to the total processing time for all methods.

4.3.3 Discussion and Knowledge Transfer to Industry

In Section 4.3, we presented an unsupervised segmentation algorithm designed specifically for coronary artery segmentation, leveraging Ward's clustering method in combination with graph-based representations. This approach focuses on extracting meaningful structures from the imaging data without requiring annotated datasets, offering a practical and explainable alternative to supervised methods. The algorithm capitalizes on the inherent brightness distribution within the images and represents these features through graph structures, facilitating the identification of abnormalities and aiding clinical interpretation. Below, we discuss key aspects of this method, including its explainability and competitiveness with advanced neural network architectures.

1. Unsupervised Clustering Algorithm: The unsupervised clustering approach offers several distinct advantages. First, it eliminates the need for annotated datasets, significantly reducing the dependency on manual labeling. Second, its low computational complexity makes it accessible for broader applications. Third, and perhaps most importantly, the method is inherently explainable. Clustering algorithms provide valuable insights into the brightness distribution in the images, which can be structured and analyzed using graph-based representations. This explainability has significant clinical relevance. By identifying areas with abnormal brightness distribution, the method can aid in detecting potential lesions, vulnerable plaques, or other pathologies, enhancing its utility in clinical decision-making.

2. Competitiveness with Advanced Neural Networks: The clustering-based method demonstrated performance competitive with complex neural network architectures, including 2.5D and 3D convolutional networks and transformer-based models. Notably, the methodology leveraging centerline-based image extraction allows training on fewer patients, as it generates multiple images for each centerline point. This makes it a resource-efficient alternative while achieving comparable results to state-of-the-art approaches.

3. Limitations in Lesion Segmentation: While the algorithm successfully identifies regions with lesions and calcium deposits, its performance in accurately segmenting these areas is diminished compared to other regions. This highlights the need for further refinement and optimization of the method to improve its sensitivity and precision in these clinically significant areas, ensuring robust detection and characterization of pathological features.

We implement and evaluate two coronary artery segmentation techniques utilizing the Ward clustering algorithm. The first method employs images from the axial, sagittal, and coronal planes, referred to as 3Axis method, while the second method focuses on cross-sectional images, known as Perp method. We assessed both methods on two distinct datasets: the first comprised of 10 patients without lesions, serving as the test set, and the second included 22 patients with 30 clinically diagnosed lesions, designated as the lesion set.

The 3Axis method demonstrated superior performance, achieving a Dice score of 0.88 for the test set and 0.83 for the lesion set, compared to the Perp method's scores of 0.81 and 0.82, respectively. This trend indicates a clear advantage of the 3Axis approach across both datasets, aligning with previous studies [151, 152]. Notably, when compared to 2.5D neural network architectures such as EfficientNet, VGG, ResNet, and U-Net++, the 3Axis method also excelled in Dice coefficient performance. Among all 3D network strategies examined, only the transformer-based Swin UNETR managed to outperform the 3Axis Ward method, achieving a higher Dice score of 0.1.

While the 3D U-Net leverages spatial information from multiple slices to enhance segmentation accuracy, the 2.5D Ward method still provides remarkably competitive results. This finding underscores the effectiveness and reliability of clustering algorithms in accurately segmenting vascular lesions, even in the absence of the intricate frameworks often associated with 3D neural network models.

Moreover, our segmentation algorithm exhibits robustness in addressing the kissing vessel artifact without relying on information from adjacent slices, as the segmentation is conducted independently on a point-by-point basis. This independence contributes to the method's overall reliability and precision. Furthermore, our approach resembles the insights of Du et al. [153], which highlighted the significance of multi-objective clustering in coronary artery segmentation. However, their dependence on a toroidal model for vessel tracking led to a lower Dice coefficient, thus reinforcing the efficacy of our multi-plane imaging strategy.

Our clustering-based segmentation method not only delivers competitive performance but also stands on par with several state-of-the-art deep learning techniques developed for coronary artery segmentation [7]. Huang et al. [154] adopt a deep learning methodology that leverages vessel-centered input images along with an encoder-decoder architecture, featuring a 3D encoder and a 2D decoder. Their approach, trained on a small dataset of 29 patients (with only 5 for testing), achieves a Dice score of 0.86, which is slightly lower than the results of our clustering algorithms. Importantly, our method demonstrates competitive performance while requiring less computational complexity and a smaller amount of training data.

In contrast, Gu et al. [155] utilizes a V-Net architecture that processes full-image volumes ($128 \times 128 \times 60$) and is trained on a larger dataset of 70 patients (20 for testing). Their method records a higher Dice score of 0.91; however, it does not emphasize vessel-specific segmentation as effectively as our approach. Moreover, our algorithm excels in segmenting lesions, ensuring clinical relevance and accuracy in these challenging areas.

One notable advantage of our methodology is the clinically validated segmentation it achieves, leveraging precise manual segmentations as ground truth. While we recognize the

limitation posed by the dataset size, we evaluated 30 clinically diagnosed lesions, showcasing the high precision of our algorithms. The automatic identification of lesion regions presents significant challenges, as it often requires predicting their location and extent without prior knowledge, and some lesions may evade detection even by clinicians. Despite these obstacles, we consider the algorithm's performance commendable for a first iteration; even though segmentation may not be perfect in every instance, lesions are consistently recognized. Future work will focus on enhancing automatic detection and segmentation of these complex regions.

Overall, our research underscores the robustness of the 3Axis method across various clinical scenarios, including cases with and without lesions. Future investigations should aim to integrate additional imaging modalities and refine the algorithm to improve segmentation accuracy, particularly in challenging situations involving low-contrast or calcified lesions.

The methodology for unsupervised segmentation has also been integrated into the industrial workflow by utilizing the previous software solutions. Additionally, significant contributions have been made, including the generation of new datasets (also with lesions) with enhanced resolution (isotropic 0.25 mm) and the extraction of centerlines for each vessel. This improvement greatly aids in the research and development of more advanced algorithms based on segmentation, as well as the extraction of local parameters of the vessels, such as diameter, which is a crucial metric for clinicians. Concretely, the following features have been implemented:

- Extracting centerlines using the *Extract Centerline* module of 3D Slicer [64], which is essential for accurately representing the geometry of vascular structures.
- Automatically centering 2D image views on specific points, commonly centering on centerline points, to enhance the visualization of anatomical structures.
- Creating perpendicular vessel views and extracting parameters like section radius using modules such as *Cross-Section Analysis* and *Endoscopy* of 3D Slicer [64]. This visualization is vital during segmentation and enables also manual precise measurements of the vessel diameter.

These enhancements not only streamline the segmentation process but also provide valuable insights into vascular characteristics, supporting clinicians in making informed decisions based on accurate anatomical data.

Again, This automation of processes not only reduces the tedium but significantly accelerates the overall procedure. While manual segmentation can take 1–2 hours per patient, the automated process can complete in approximately 30 minutes. Importantly, the majority of this 30-minute timeframe is occupied by the computer computing predictions and generating visualizations, minimizing the active time required from the operator. This allows for more efficient use of resources and personnel.

Moreover, features such as perpendicular vessel views greatly enhance visualization, segmentation, and editing, improving overall efficiency and user experience. These developments demonstrate how the methodologies support both industrial goals and the resear in medical imaging technologies.

In this chapter, we have implemented different methods for coronary artery segmentation, progressing from the original images at their native resolution—where the primary challenge was distinguishing arteries from the background due to class imbalance—to an interpolated

dataset based on vessel segmentations centered around the centerline points, effectively mitigating this imbalance issue. Beyond the development of segmentation techniques, we also designed a modular software framework that enables both inference and visualization of the results. This comprehensive approach not only enhances coronary artery segmentation but also facilitates the extension of these techniques to other cardiovascular structures, from large vessels such as the aorta to smaller yet clinically significant features like calcium deposits.

Chapter 5

Aortic Calcium Segmentation in CTA

5.1 Introduction

In the previous chapter, we focused on segmentation techniques for the coronary lumen in CTA, aiming specifically to exclude calcifications that contribute to stenosis. This task presented several challenges, including class imbalance, as the coronary arteries occupy only a small portion of the overall image volume, making their precise delineation particularly difficult. Now, we shift our focus slightly to a different but related problem: the segmentation of aortic calcium. Unlike before, where calcifications were treated as structures to be excluded, here they become the primary target of our analysis. However, aortic calcium represents only a small fraction of the total aortic area, introducing similar class imbalance challenges. To address this, we will build upon the segmentation techniques previously explored, adapting them to the specific demands of this new task.

Atherosclerotic plaques are accumulations of lipids, calcium, and fibrous tissue within the arterial walls that can lead to vessel narrowing and reduced blood flow. These plaques pose significant health risks, including heart attacks and strokes, due to their potential to rupture or severely obstruct blood flow (see Section 1.1.3). When such plaques form on the aortic valve, they can impair its function by increasing rigidity and altering its morphology, leading to valve malfunction.

Due to their composition, atherosclerotic plaques are visible on CT scans as high-density regions. This visibility enables not only their identification but also their segmentation and quantification, providing valuable parameters such as plaque volume and calcium score [30, 156]. The calcium score, typically calculated using the Agatston method [56], is widely used in diagnosing aortic valve stenosis (AVS) and assessing coronary artery disease (CAD) (see Section 1.2.1 and Section 1.2.3). The Agatston method quantifies coronary calcium by identifying regions with an attenuation above a predefined threshold (typically 130 Hounsfield Units (HU)) and calculating a weighted score based on the area of calcification and its peak density. This score provides a standardized measure of calcium burden.

In clinical practice, the images used to compute the Agatston calcium score are non-contrast CT scans [56, 30] with limited view, providing only a partial view of the thoracic aorta. Since they primarily focus on the aortic valve, which may restrict comprehensive assessment of calcification across the other thoracic aorta regions as tubular aorta, aortic arch and descending aorta (see the morphology of the aorta in Section 1.1.1).

However, the accumulation of plaques in these regions, particularly the aortic arch, is of

significant interest due to its curved morphology, which may promote plaque buildup. This is especially critical in the context of transcatheter aortic valve implantation (TAVI), as the catheter navigates through the aortic arch and can inadvertently come into contact with calcified plaques. Such contact increases the risk of plaque dislodgement, which can lead to severe complications [61, 157]. The proximity of the aortic arch to the carotid arteries further heightens this concern, as dislodged plaque fragments can travel to the brain, potentially causing a cerebrovascular event. Therefore, precise knowledge of plaque composition and distribution in these regions is crucial for procedural planning. Identifying high-risk plaques in advance allows for the consideration of protective measures, such as the placement of embolic protection devices in the carotid arteries, which can help prevent stroke by capturing dislodged material during the intervention.

When patients are scheduled for a transcatheter aortic valve implantation (TAVI), they undergo a contrast-enhanced CT angiography (CTA) that provides a complete view of the thoracic aorta. In these images, calcium segmentation and scoring would be feasible and highly relevant, as it could identify high-risk regions for catheter passage. Despite CTA being widely used, there is no standard method for calcium segmentation or scoring in these images, highlighting the need for standardized protocols in this context [158, 31, 159].

The goal of this study is to develop a calcium score for distinct regions of the thoracic aorta using pre-TAVI CTA scans. Since CTA imaging involves variable thresholds for each patient, the segmentation algorithm requires validation. Initially, the method's accuracy was validated by comparing results with those from clinical practice.

Once validated, the segmentation process was automated, addressing its time-intensive nature and reliance on expertise. The unified framework segments the aorta and calcium across regions simultaneously using supervised learning. Multiple architectures, including U-Net derivatives, transformers, and UMamba, were explored to achieve multiregional segmentation effectively.

5.2 Methodology

In this sections, we present the methodology employed for segmenting and quantifying thoracic aorta calcifications, starting with manual segmentation and scoring approaches before progressing toward automated methods using neural networks.

We begin by detailing the clinical data utilized in this study (Section 5.2.1), followed by a description of the segmentation process (Section 5.2.2), which is adapted to both unenhanced CT and contrast-enhanced CTA. Given the anatomical complexity of the aorta and the varying distribution of calcium plaques, we introduce region-based segmentation techniques to improve accuracy. We then explore different scoring methods (Section 5.2.3), essential for clinical interpretation and risk stratification, and outline the validation metrics used to assess segmentation performance (Section 5.2.4).

The second half of the chapter focuses on the automation of this process, describing the CTA dataset used for training (Section 5.2.5), the neural network architectures implemented (Section 5.2.6), and the loss functions designed to address the challenges of class imbalance (Section 5.2.7). Finally, we discuss how we reconstructed the segmented 3D geometries (Section 5.2.8) and the evaluation metrics used to benchmark the performance of our models (Section 5.2.9), bridging the gap between traditional manual assessment and fully automated analysis.

5.2.1 Clinical data

The study includes 55 patients that underwent the TAVI procedure, each with two datasets: the unenhanced-set and the enhanced-set.

The unenhanced-set comprises 64-slice non-contrast CT scans with a slice spacing of 2.5 mm, commonly used for Agatston calcium scoring. The resolution is $0.488 \times 0.488 \times 2.5$ mm^3. For more information see Section 2.2.2.

The enhanced-set contains CTA (computed tomography angiography) images of the thoracic region, used to visualize aortic geometry pre-TAVI (transcatheter aortic valve implantation). These scans have a varying resolution, typically acquired at $0.621 \times 0.621 \times 0.625$ mm^3 for the x, y, and z dimensions, respectively. However, the resolution for the x and y dimensions can range from 0.594 mm to 0.914 mm, with a standard deviation of 0.0938 mm. For more information see Section 2.2.3.

For computational efficiency, images are cropped to $180 \times 256 \times 256$ voxels. This size was selected to accommodate all aortas in the dataset within the designated region, as well as any aorta captured in a CTA scan with a comparable resolution. Reducing the image dimensions improves processing efficiency while preserving clinical utility, as the original high-resolution images remain accessible for diagnostic purposes. The region of interest (ROI) is positioned approximately 25 mm above the top of the aortic arch. Horizontal placement is guided by the aorta's geometry, ensuring that the entire aorta is contained within the ROI, from the valve to the descending aorta.

5.2.2 Segmentation of Aorta and Calcium Plaques by Region

To compare the calcium scores measured in clinical practice using unenhanced CT with those obtained from contrast-enhanced CTA, it is essential to perform segmentation on both modalities. This allows for evaluating the consistency and potential discrepancies between the methods. The segmentation process focuses on identifying calcium in unenhanced CT (Section 5.2.2.1), following the clinical Agatston method, and adapting thresholding techniques for CTA to address the challenges posed by contrast variability (Section 5.2.2.2). These segmentations provide the foundation for the quantitative comparison of calcium scores across modalities.

5.2.2.1 Calcium Segmentation in Unenhanced CT

Calcium segmentation for unenhanced CT scans follows the clinical procedure outlined by the Agatston method (see Algorithm 1). For each axial slice, calcium is identified within the aortic leaflets exhibiting attenuation values above 130 HU. Additionally, only segments with an area exceeding 1 mm^2 are considered to ensure reliable identification of calcified regions.

5.2.2.2 Calcium Segmentation in CTA

Segmenting calcium in CTA images presents challenges due to variability in attenuation values caused by contrast agents. To address this, a patient-specific thresholding method is employed to adapt to the unique characteristics of each dataset while avoiding over- or under-segmentation.

Figure 5.1 illustrates the aorta and calcium segmentation workflow, detailing the different stages of the process. The original CTA image is shown in Figure 5.1a, where a region of

interest (ROI) is selected (green box) to ensure the aorta is fully captured while maintaining computational efficiency. Figure 5.1b highlights the manual ROI (yellow box) placed over the tubular region of the aorta, with blue regions representing areas where attenuation values fall within the predefined range. In Figure 5.1c, the segmented structures are depicted, where the green segment corresponds to the aorta from the valve to the descending aorta, and the yellow segment represents the calcified plaques. Additionally, the green cube marks the cropped region from the original image. Finally, Figure 5.1d presents the extracted aorta in green along with the blue centerline, which extends from the sinuses of Valsalva to beyond the arch. This visualization provides an overview of the segmentation process, emphasizing the differentiation between the aorta and calcifications.

In addition, the segmentation steps are explained below.

1. From the sagittal view of the cropped image (refer to Section 5.2.1 and Figure 5.1a), a region of interest (ROI) is selected within the lumen of the tubular aorta (see Figure 5.1b). The maximum and minimum attenuation values (*maxAttValue* and *minAttValue*, respectively) are determined within this ROI.

2. A visual inspection is conducted to confirm that *maxAttValue* encompasses the entire lumen. Adjustments are made if necessary, and an additional 10 HU is added to the threshold to reduce noise. A visual example and justification can be seen in Section B.1.

3. Thresholding is applied to segment the aorta and calcium. The aortic lumen's minimum threshold is determined from step 1, and calcium is segmented with a minimum threshold equal to *maxAttValue* and a maximum threshold equal to the highest HU in the volume. This process ensures consistent segmentation, as shown in Figure 5.1c. The segmented aorta is defined from the valve leaflets to the descending aorta, as indicated by the cropped region (see Figure 5.1b).

   These thresholds are applied across the entire volume, resulting in two separate segments: one for the aorta and one for the calcium. This ensures the segmentation of the complete structure. However, these segments may also include other anatomical structures, such as contrast-enhanced areas like the coronary arteries within the aorta segment or bones within the calcium segment, due to their similar intensity values. To resolve this, a manual post-processing step is implemented to eliminate any structures outside the thoracic aorta, ensuring that only the aorta and aortic calcium are retained in the segmentation.

4. A centerline is extracted using the *Extract centerline* module in the 3D Slicer software [64]. The endpoints are positioned just above the sinuses of Valsalva and beyond the aortic arch (see Figure 5.1d).

Regional Segmentation of Calcium In the segmentation process, after identifying the aortic calcium, it is categorized into four distinct regions based on anatomical landmarks [15] and planes defined relative to the aortic centerline. Figure 5.2 illustrates the planes used for this region definition. Figure 5.2a shows a plane that is perpendicular to the centerline and situated 5 cm from its origin. This plane plays a crucial role in distinguishing the tubular aorta region from the aortic arch. Figure 5.2b displays a plane that is parallel to the centerline, utilized to further divide the aortic arch into two sections: the lower and upper arch regions. These specific

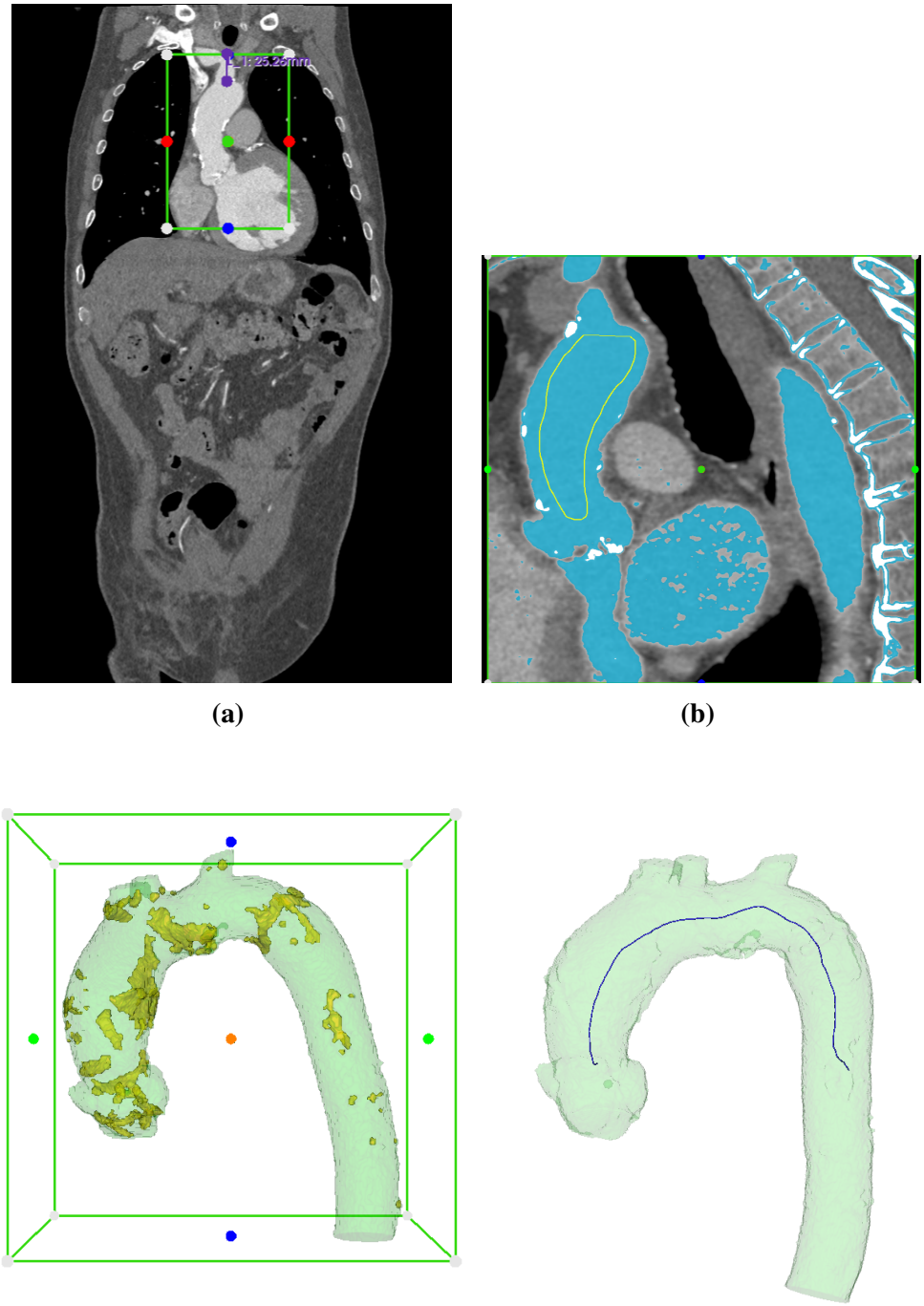


Figure 5.1: Aorta and calcium segmentation workflow. (a) Original-sized CTA image. The green box indicates the ROI of size $180 \times 256 \times 256$ voxels, which includes the aorta (see Section 5.2.1). (b) The yellow box shows the manual ROI over the tubular aorta region. The blue areas represent regions with attenuation values between the maximum and minimum attenuation values withing the pixels in the ROI. Medical image voxel size is $0.621 \times 0.621 \times 0.625$ mm^3. (c) The green segment represents the aorta, from the valve to the descending aorta. The yellow segment corresponds to the calcium. The green cube represents the ROI used in (a) to crop the original image. (d) The green segment indicates the aorta. The blue line represents the centerline, from the top of the sinuses of Valsalva to beyond the arch.

planes allow for a more precise classification of the aortic segments, enabling a detailed analysis of the aortic calcium distribution across the different anatomical regions, as further explained below.

- **Aortic Valve:** Calcium present exclusively on the leaflets is segmented. To improve accuracy, a plane perpendicular to the aortic valve can be used to delineate this region.
- **Tubular Aorta:** The region spans from the aortic sinuses (excluded) to the aortic arch. In this study, the tubular aorta is defined as extending from the sinuses to a plane perpendicular to the centerline, located 5 cm from its origin (see Figure 5.2a).
- **Superior Aortic Arch:** The region is defined as the area above the parallel planes along the centerline within the arch. The origins of supra-aortic arteries are excluded by cropping approximately 20 mm from their emergence on the aortic wall (see Figure 5.2b).
- **Inferior Aortic Arch:** This region lies below the parallel planes along the centerline within the arch (see Figure 5.2b) and connects the tubular aorta to the descending aorta.
- **Descending Aorta:** Extends from the aortic arch to the lower end of the descending aorta.

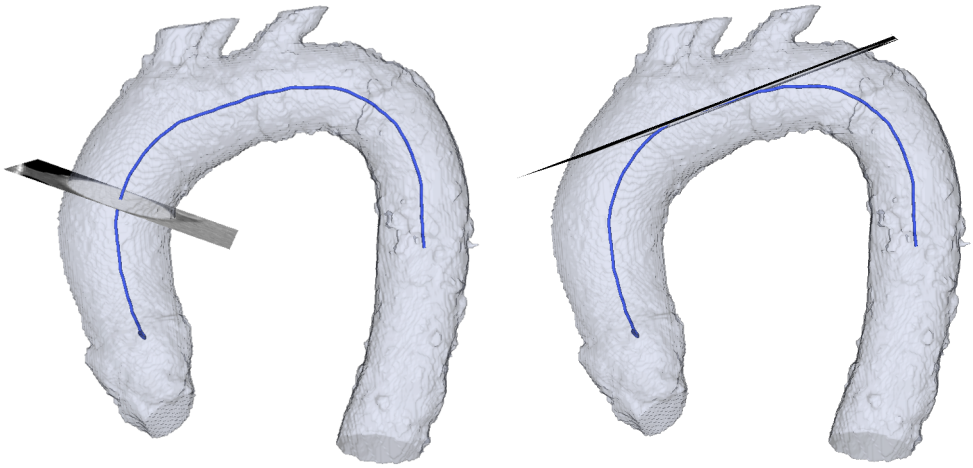
The decision to divide the aortic arch into superior and inferior regions is based on considerations related to blood flow dynamics and catheter positioning. In the curved structure of the arch, the superior portion experiences higher blood flow pressure, which can impact calcium deposition. Furthermore, during catheter-based procedures, the arch is more frequently in contact with the catheter due to the pressure required for navigation. This division facilitates a more accurate evaluation of potential risks associated with calcium displacement during interventions. Each region is distinctly defined, as it is determined by the pixel's position relative to a predefined plane. This approach ensures that each calcium voxel is assigned to only one region, eliminating any overlap or ambiguity. For example, in the aortic arch, a plane parallel to the centerline is used to divide the regions. A voxel can only be located either above or below this plane, but never in both at the same time. This strict spatial division guarantees a clear, consistent segmentation, ensuring that no calcium deposits are counted in multiple regions.

The segmentation process is semi-automated using specific software developed for 3D Slicer [64].

Consolidating step Due to the difference in slice spacing between the unenhanced CT images (2.5 mm) and the enhanced CTA images (0.625 mm), a reconstruction process is applied to standardize the spacing. This process combines four adjacent slices into one by averaging the attenuation values from the CTA images. As a result, a new CTA volume is created with a consistent slice spacing of 2.5 mm.

5.2.3 Scoring methods

In this study, two calcium scoring algorithms are considered: the Agatston method [56] and the calcium volume measurement. Another method, known as calcium mass scoring [160, 161, 162], is also documented in the literature. However, its implementation requires calibration



(a) **(b)**

Figure 5.2: Planes used to define the regions of the aorta. (a) Plane perpendicular to the centerline located 5 cm from its origin. This plane is used to separate the region of the tubular aorta from the arch. (b) Plane parallel to the centerline used to separate the lower arch region from the upper arch region.

through a phantom study, which was not feasible in our case due to the absence of essential data and limited access to the CT scanner's calibration settings.

Agatston Method The Agatston method [56], originally designed for unenhanced CT images with a slice spacing of 3 mm, calculates a calcium score based on the area and attenuation of calcified regions. This method was selected because it has been validated by numerous studies [159, 163, 164] and, most importantly, is the standard adopted in clinical practice, as outlined in the European guidelines for cardiovascular risk assessment [165]. As a result, it serves as our reference for calcium quantification. According to these guidelines, the slice thickness for coronary artery calcium (CAC) analysis should be 2.5 mm or 3 mm (the original spacing for which it was developed) to ensure consistency with existing CAC databases. Therefore, using 2.5 mm slice spacing poses no methodological concerns and remains appropriate for calcium scoring.

Algorithm 1 shows the pseudocode for this method. The variable *segmentedImages* contains the segmented calcium in the aortic leaflets, which is manually selected in each axial slice. Each calcium segment must have a minimum attenuation value of 130 HU and a minimum area of 1mm^2.

The following functions are utilized in the algorithm:

- $getCalcifiedRegions(image)$: Identifies the segmented pixels classified as calcium in an image.
- $calculateArea(calcifiedRegion)$: Computes the area of the input calcified region in mm^2.
- $getMaxAttenuationValue(calcifiedRegion)$: Retrieves the maximum brightness value in HU among all input pixels.

Algorithm 1 Agatston method

Input: Unenhanced CT images with segmented calcium in aortic leaflets (*segmentedImages*)

Output: calcium score in Agatston Units (*caScore*)

```
caScore ← 0
for image in segmentedImages do
    calcifiedRegionsPerSlice ← getCalcifiedRegions(image)
    for calcifiedRegion in calcifiedRegionsPerSlice do
        area ← calculateArea(calcifiedRegion)
        maxValue ← getMaxAttenuationValue(calcifiedRegion)
        if maxValue between 130 and 199 then
            densityScore ← 1
        else if maxValue between 200 and 299 then
            densityScore ← 2
        else if maxValue between 300 and 399 then
            densityScore ← 3
        else if maxValue ≥ 400 then
            densityScore ← 4
        end if
        caScore ← caScore + area ∗ densityScore
    end for
end for
return caScore
```

Volume Measurement The second scoring method involves measuring the calcium volume. This is achieved by calculating the total volume of segmented pixels, determined by multiplying the number of segmented pixels by the voxel dimensions (height, width, and depth). The final calcium volume score is obtained by summing the volume values across all segmented pixels.

This method was considered as an alternative due to its computational efficiency, direct correlation with physical properties, and its use in previous studies [161, 166].

A comparative table outlining the requirements, advantages, and limitations of the Agatston and volume methods is included in Section B.2.

5.2.4 Validation Metrics

In this study, we validate the calcium segmentation technique in contrast-enhanced computed tomography angiography (CTA) by comparing it with clinical calcium scoring methods. The validation is based on clinical data consisting of calcium scores measured using the Agatston method on non-contrast CT images. The comparisons are conducted in the following aspects:

- Agatston scoring validation: The Agatston score obtained from non-contrast CT images by clinicians is compared with the score calculated by a medical imaging expert from the same images.
- Comparison of Agatston scores across imaging modalities: The Agatston scores measured on non-contrast CT are compared with those obtained from CTA images.
- Comparison of calcium volume scoring across imaging modalities: The calcium volume score is computed for both non-contrast CT and CTA images to assess potential differences.

To ensure consistency across datasets, CTA images are resampled to a uniform slice thickness of 2.5 mm (see Section 5.2.2.2).

For statistical analysis, Pearson's correlation coefficient is used to assess the strength and direction of the linear relationship between two variables, such as calcium scores derived from different methods or imaging modalities. The coefficient ranges from -1 to $+1$, where $+1$ indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 indicates no linear correlation. A higher Pearson correlation suggests that the methods are consistent in their measurements.

Additionally, Bland-Altman plots are employed to assess the agreement between different calcium scoring techniques. The plot displays the differences between paired measurements on the y-axis and their average on the x-axis. This method helps to identify systematic biases (consistent differences between the methods) and assess the level of agreement across the range of measurements. If the plot shows that most data points fall within the limits of agreement (defined by ± 1.96 standard deviations), it indicates that the two methods are in good agreement. However, if the differences between the methods are large or systematically biased, it will be evident in the plot. The Bland-Altman method is particularly useful for detecting discrepancies between measurement techniques that might not be revealed by Pearson's correlation alone, especially when the data involves repeated measurements.

Table 5.1: Percentage of images containing calcium from different regions of the aorta, separated by train, validation, and test datasets.

Aortic Region	Train (%)	Validation (%)	Test (%)
Aorta	100.0	100.0	100.0
Valve	32.41	34.03	30.76
Tubular Aorta	29.19	34.57	27.05
Superior Arch	52.56	45.1	59.26
Inferior Arch	43.59	39.11	50.89
Descending Aorta	29.92	29.4	39.29

5.2.5 CTA Dataset for Automatic Segmentation

After validating the segmentation method, the contrast enhanced images, described in Section 5.2.1 and the manual calcium segmentation by regions, described in 5.2.2.2, were utilized to automate the segmentation of the aorta and calcium by regions in CTA scans. This step aims to streamline the process and enable accurate regional calcium quantification, building upon the validated manual and semi-automated segmentation results.

The training dataset consists of the 55 CTA scans described and segmented according to the methodology outlined in Section 5.2.2.2. These segmentations were organized into a labelmap containing six distinct classes: background (0), aorta (1), tubular aorta plaque (2), valve plaque (3), superior arch plaque (4), inferior arch plaque (5), and descending aorta plaque (6). To expose the model to the full aortic anatomy, slices were extracted from the sagittal plane [167, 168]. Images without aortic segments were excluded to minimize the dominance of background regions.

In Figure 5.3, the dataset generation process is illustrated. Figure 5.3a shows the initial segmentation results, where the aorta is represented in gray, the aortic calcium is highlighted in blue, and the aortic centerline is marked in red. Figure 5.3b provides a more detailed segmentation of the aortic calcium, with the different regions clearly defined: pink for the valve, red for the tubular aorta, yellow for the superior arch, orange for the inferior arch, and green for the descending aorta. Finally, Figure 5.3c displays the original cropped sagittal plane on the left, with pixel intensity measured in Hounsfield Units (HU). The middle image shows the preprocessed sagittal plane with the intensities normalized to a range of $[0,1]$, and the right image presents the corresponding label map. This figure effectively summarizes the key steps in the manual segmentation and pre-processing of the dataset.

The preprocessing pipeline included sigma clipping with bounds set to $[mean - 2 \times SD, mean + 4 \times SD]$ to mitigate the influence of outliers, followed by normalization to the $[0,1]$ range. The dataset was split into three subsets: training (70%, 38 patients, 3847 images), validation (20%, 11 patients, 1102 images), and testing (10%, 6 patients, 621 images). Table 5.1 summarizes the percentage of images representing each region in the training, validation, and test sets.

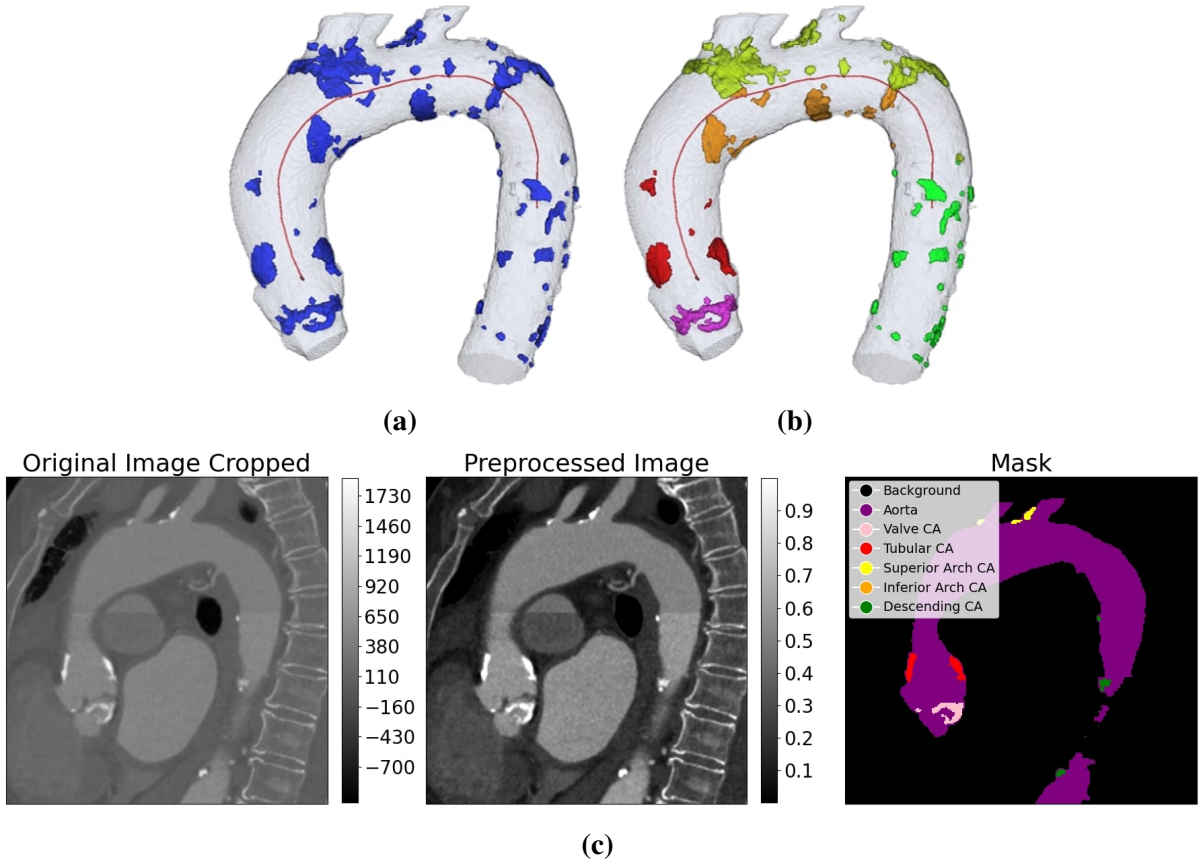


Figure 5.3: Results of manual segmentation and pre-processing. (a) Aorta segmentation (gray), aortic calcium segmentation (blue), and aortic centerline (red). (b) Aorta segmentation (gray), aortic centerline (red), and aortic calcium segmented by regions: pink for the valve, red for the tubular aorta, yellow for the superior arch, orange for the inferior arch, and green for the descending aorta. (c) The original cropped sagittal plane (left) with pixel intensity measured in Hounsfield Units (HU). The middle image shows the preprocessed sagittal plane with intensities normalized to the [0,1] range, and the right image displays the corresponding label map.

5.2.6 Neural Network Architectures

In this study, various neural network architectures were evaluated to segment calcium deposits and aortic regions in contrast-enhanced CT images. These architectures were selected based on their demonstrated strength in medical image segmentation tasks, as well as their ability to handle the complex and detailed anatomical structures typical in aortic imaging (see architecture implementation details in Section 2.3). Below is a justification for each architecture used in this study.

The first model tested was the 2D UNet with EfficientNet-B0 (EffB0) [81], a conventional 2D UNet architecture enhanced with an EfficientNet-B0 encoder. EfficientNet-B0 is known for its balance between computational efficiency and high performance in feature extraction, making it ideal for segmentation tasks that require both fine-grained spatial information and the extraction of semantic features, making it well-suited for handling the image size in our dataset. Additionally, a pre-trained variant, EffB0Pre, which uses an ImageNet-trained encoder [82], was included to assess if leveraging pre-trained models would improve segmentation accuracy, especially when limited training data is available.

The 2D UNet++ [78] with EfficientNet-B0 (EffB0++) was another architecture chosen for its refined capabilities in handling segmentation tasks. Unlike the traditional UNet, UNet++ incorporates dense skip connections between the encoder and decoder layers, which facilitates better feature reuse and improves the delineation of boundaries in segmentation. This feature is particularly beneficial for accurate segmentation of complex anatomical regions, such as the aorta and calcium deposits, where boundary precision is critical.

In the case of three-dimensional datasets, the 3D UNet with EfficientNet-B0 (3DEffB0) was employed. Extending the UNet architecture to three dimensions allows for the capture of volumetric relationships across adjacent slices, providing an advantage when segmenting intricate anatomical structures. This 3D variant is more suited to handle the complexity of the aorta and calcium deposit segmentation, which involve the interaction of structures across multiple slices. In addition, the pre-trained version, 3DEffB0Pre, and the 3D UNet++ variant, 3DEffB0++, were also evaluated to determine whether pre-training and dense skip connections could further enhance segmentation performance in volumetric datasets.

The 3DSegFormer [92] was selected as a transformer-based architecture designed specifically for volumetric medical image segmentation. The model utilizes multi-scale embeddings and attention mechanisms, which allow it to capture both global and local features of the input images. This capability is essential for a dataset like ours, which requires precise segmentation that balances the fine details of calcium deposits with computational efficiency. The SegFormer's ability to process large volumes of data and maintain high accuracy makes it a promising candidate for this task.

Lastly, the UMamba and 3D UMamba models [95] were included in the evaluation. UMamba is a convolutional network that employs State Space Sequence Models (SSMs), which are particularly effective in capturing long-range dependencies in medical images. This architecture surpasses traditional transformers in some applications, making it an ideal choice for segmenting large anatomical structures such as the aorta and its calcium deposits. The 3D variant, 3DUMamba, was also tested to address the volumetric nature of our dataset, ensuring that it could handle the segmentation of three-dimensional structures while retaining performance.

Each of these architectures was tested to evaluate their performance in segmenting calcium and aortic regions, with particular focus on their ability to process the complex

anatomical features presented in contrast-enhanced CT images. These models were chosen for their individual strengths, which together provided a comprehensive approach to accurately segmenting and analyzing the aorta and its calcium deposits.

Each model was evaluated for its ability to segment calcium and aortic regions, with particular attention to their effectiveness in processing the complex anatomical features present in contrast-enhanced CT images.

5.2.7 Loss Functions

To optimize model performance in segmentation tasks, several loss functions and their combinations were evaluated, focusing on improving boundary detection and addressing class imbalances. Additional details about these loss functions can be found in Section 2.3.5. The following loss functions are implemented:

- Dice Loss
- Cross Entropy Loss (CE)
- DiceCE Loss: Combined Dice and Cross Entropy Loss for overlap and class separation [99].
- DiceFocal Loss: Parameters were set to $\lambda_{dice} = 2$, $\lambda_{focal} = 1$, $\alpha = 1$, and $\gamma = 2$ to emphasize difficult examples [100, 101].
- Generalized Dice Loss (GD): Enhanced performance on smaller classes by weighted contributions [102].
- GeneralizedDiceFocal Loss (GDF): Combined GD and Focal Loss, with $\lambda_{gd} = 2$ and $\lambda_{focal} = 1$.
- Tversky Loss: Controlled false positives and negatives using $\alpha = 0.3$ and $\beta = 0.7$ [103].

For the loss functions Dice, CE, DiceCE, DiceFocal, and GDF, class weights ($w_c = [1, 1, 15, 20, 15, 15, 20]$) were assigned for background, aorta, valve, tubular aorta, superior arch, inferior arch, and descending aorta, respectively. The background and aorta classes were given lower weights because they are well-represented in the dataset and are relatively easier for the model to learn. In contrast, higher weights of 15 and 20 were assigned to the calcium regions to ensure the model gives sufficient attention to these smaller and more geometrically intricate areas, improving its segmentation capabilities in these clinically important regions. The tubular aorta and descending aorta were assigned higher weights due to their underrepresentation in the dataset (see Section 5.2.5), making them more challenging to segment.

5.2.8 3D Geometry Reconstruction

Reconstructing 3D geometries from sagittal slices can be approached using either 2D or 3D networks. For 2D networks, predictions are made slice by slice in the sagittal plane. In contrast, 3D networks predict overlapping volumetric blocks, necessitating a strategy to resolve inconsistencies between predictions.

To address this, a weighting kernel of size $256 \times 256 \times 32$ is applied. Peripheral slices in the kernel (first and last five slices) are assigned a weight of 0.3, slices from 6 to 10 and from 23 to 27 are weighted at 0.7, and central slices are weighted at 1.0. This weighting scheme adjusts the one-hot encoded predictions by emphasizing central slices and reducing the influence of edge slices, ensuring that the final pixel class corresponds to the highest weighted probability.

Following model predictions, segmentations are post-processed to retain only the largest connected component, removing extraneous smaller components. This step ensures that evaluation focuses exclusively on the aortic structures, which align with the training data and exclude irrelevant regions outside the aorta.

5.2.9 Evaluation Metrics

The Dice coefficient is the primary metric used to evaluate segmentation performance, quantifying overlap between predictions and ground truth. Complementary metrics, such as Intersection over Union (IoU), Precision, and Recall, are also calculated to assess various accuracy aspects. These metrics, previously detailed in Section 2.3.6.2, are applied to post-processed predictions and averaged across six test patients for different aortic and calcium regions. Additionally, a visual inspection is performed to provide qualitative insights into segmentation quality.

5.3 Results

In this section, we present the results of the aortic calcium segmentation and scoring process. The results are organized into several key areas, each addressing different aspects of the methodology. First, in Section 5.3.1, we discuss the manual segmentation of the aortic regions and evaluate the different attenuation values between CT and CTA. In Section 5.3.2, we validate the calcium scoring and segmentation in the valve region, including both Agatston scoring and volume scoring. We conclude the validation of the segmentation method by comparing it with existing calcium segmentation techniques in CTA, as detailed in Section 5.3.3.

Regarding the automation of the segmentation process, Section 5.3.4 presents the findings of an ablation study on different loss functions used in the automatic segmentation models. The performance of the segmentation is then broken down by aortic region in Section 5.3.5, highlighting any discrepancies and areas for improvement. Finally, Section 5.3.6 focuses on the 3D predicted geometries, providing a visual and quantitative assessment of the model's ability to recreate aortic structures. These results collectively demonstrate the effectiveness of the proposed methods in calcium segmentation and scoring, contributing to better understanding and prediction of aortic conditions.

5.3.1 Manual Region Segmentation and Attenuation Value Evaluation

To validate the method used for calcium score computation throughout the entire aorta, it is essential to first analyze the segmentation results of both aortic structures and calcium deposits in CT and CTA images. This evaluation serves as the foundation for accurately calculating calcium scores. By comparing the segmentation outcomes between the two imaging modalities, the computed scores can then be assessed for consistency and clinical relevance, ensuring the reliability of the approach before applying it to the intended analyses.

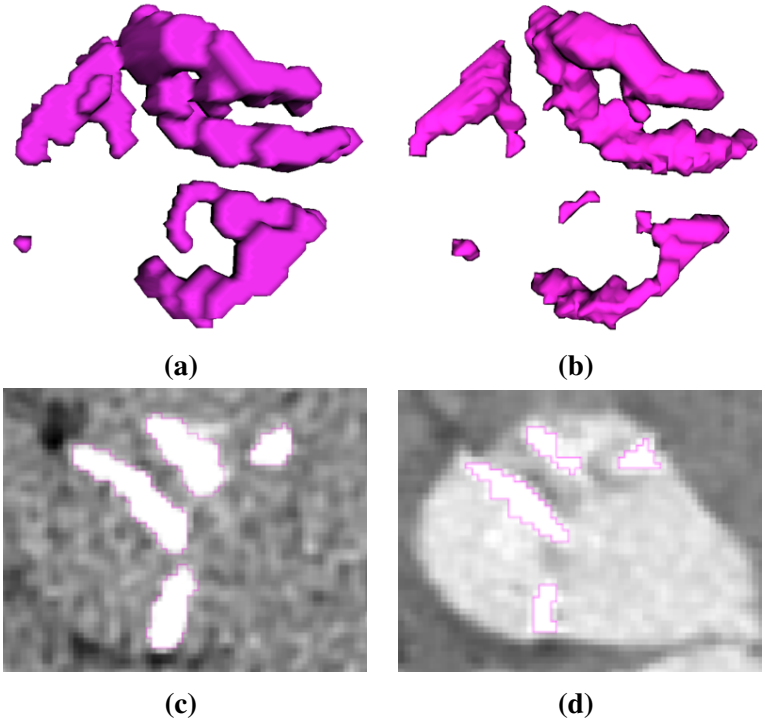


Figure 5.4: Example of calcium segmentation in the aortic valve. (a) Segmentation in unenhanced CT. Voxel size is $0.48 \times 0.48 \times 2.5$ mm^3. (b) Segmentation in CTA. (c) Unenhanced CT axial slice. (d) Enhanced CTA axial slice. Voxel size is $0.621 \times 0.621 \times 0.625$ mm^3.

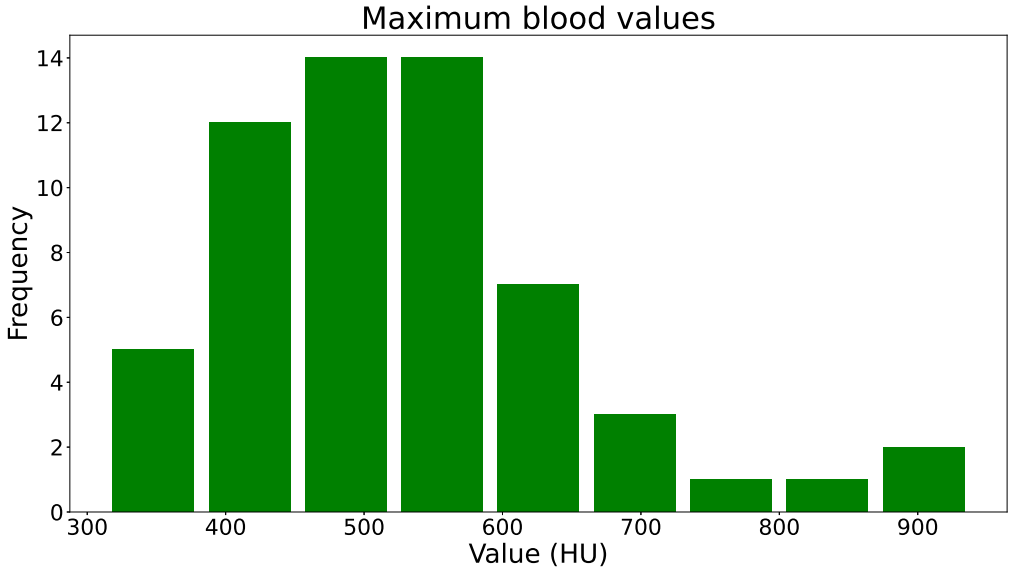
The segmentation of valve calcium follows the same process for both unenhanced and contrast-enhanced images, with the key difference being the threshold used for identification. In each axial slice, calcium on the aortic leaflets is segmented based on its specific threshold. When regions such as the tract are unclear, a perpendicular plane to the leaflets is used for better visualization. For contrast-enhanced images, calcium segmentation is performed for distinct regions of the aorta: aortic leaflets, tubular aorta, superior aortic arch, inferior aortic arch, and descending aorta.

Images acquired with and without contrast offer distinct perspectives on the lumen and aortic calcium. For example, Figure 5.4 highlights calcium segmentations of the valves for the same patient in both non-contrast (Figure 5.4a) and contrast-enhanced (Figure 5.4b) images. Additionally, axial slices of these images are shown in Figure 5.4c and Figure 5.4d.

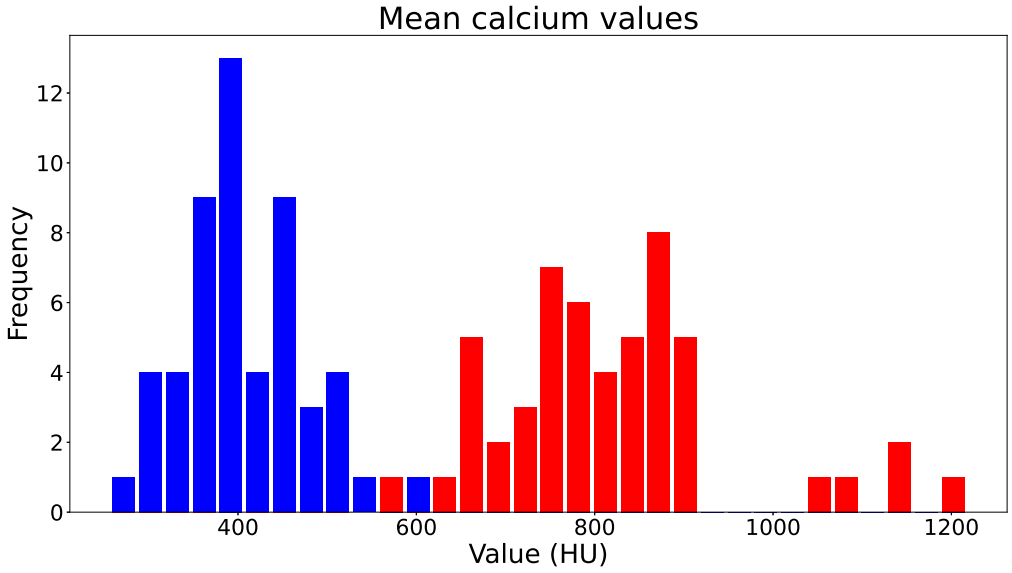
Contrast enhancement raises attenuation values for blood and calcium, making threshold values dependent on acquisition parameters, such as contrast volume and acquisition instant. Since the Agatston calcium score depends on attenuation values, an examination of the variation in Hounsfield Units (HU) between unenhanced CT and enhanced CTA images is conducted.

Figure 5.5a illustrates the maximum attenuation values of the aortic lumen in CTA images for the 55 patients analyzed, presented in the form of a histogram. The lumen is determined using the segmentation method described in Section 5.2.2.2. These values, which also set the minimum threshold for calcium segmentation, are consistently above 300 HU. Notably, 91% of patients have values exceeding 400 HU, with 78.18% falling within the range of 400 to 600 HU.

In Figure 5.5b, a histogram comparing the average calcium attenuation per patient between



(a)



(b)

Figure 5.5: Histograms of mean attenuation values (HU) per patient for the dataset of 55 patients. (a) Histogram of maximum attenuation values for blood in enhanced CTA images. (b) Histogram with mean calcium attenuation values for unenhanced CT (in blue) and enhanced CTA (in red).

unenhanced CT and CTA images is presented. In unenhanced CT, calcium values primarily range from 300 to 500 HU, peaking around 400 HU. In contrast, CTA images exhibit higher variability, with values typically between 600 and 900 HU. Only five patients have average calcium attenuation values exceeding 1000 HU in CTA, reflecting greater variability and a lack of dominant intervals compared to unenhanced CT.

5.3.2 Validation of Calcium Scoring and Segmentation in the Valve Region

To validate the calcium scoring methodology and segmentation process, we focus on the valve region, the only area routinely evaluated by clinicians. This approach ensures that both manual segmentation in CTA images and the derived calcium scores (Agatston and volume) are accurate. By confirming the method's reliability in this clinically significant region, we establish confidence in extending the approach to the entire aorta, providing a foundation for comprehensive aortic calcium assessment.

5.3.2.1 Agatston Scoring

To validate the reproducibility of the Agatston method, we assessed interobserver correlation by comparing calcium scores from two observers: Observer 1 (a clinician) and Observer 2 (an expert in medical image segmentation).

The results of the analysis using the Agatston method are illustrated in Figure 5.6. Figure 5.6a presents a scatter plot comparing the calcium scores obtained from unenhanced CT and CTA images. The x-axis represents the values obtained in clinical practice from the unenhanced images, while the y-axis indicates the Agatston calcium scores for both unenhanced images (with a threshold of 130 HU) as assessed by a segmentation expert (blue dots) and the enhanced reconstructed images (with a patient-dependent threshold), marked by the red dots.

Figure 5.6b shows a Bland-Altman plot comparing the Agatston calcium scores derived by clinicians using the unenhanced images and those obtained by a segmentation expert. In this plot, the x-axis represents the mean of the two scores, and the y-axis illustrates the difference between the scores obtained by the segmentation expert and those measured by the clinician.

Figure 5.6a demonstrates an excellent correlation (blue dots) with a Pearson's r^2 value of 0.99 (p-value: 1.07×10^{-49}) and a mean absolute error (MAE) of 147.23 Agatston units. The Bland-Altman plot (Figure 5.6b) shows no significant trends, with a small mean bias of -27.95 Agatston units.

Additionally, the calcium score derived from CTA images was compared to clinical segmentations on CT images (Figure 5.6a, red dots). This comparison, using a patient-specific threshold yielded a Pearson's r^2 value of 0.87 (p-value: 2.49×10^{-18}) and an MAE of 1138.88 Agatston units. Results indicate that calcium scores from CTA are consistently lower than those from unenhanced CT.

Figure 5.6c presents another Bland-Altman plot, this time comparing the Agatston calcium scores obtained by clinicians from unenhanced images and by the segmentation expert from the enhanced reconstructed CTA images. The x-axis again represents the mean of the two scores, while the y-axis displays the difference between the values obtained from the segmentation expert (using CTA) and the clinician's scores (using unenhanced CT).

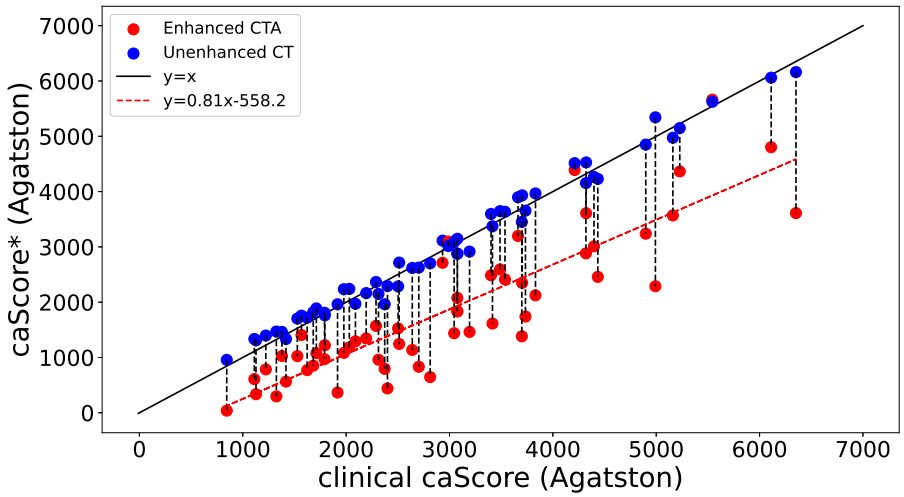


Figure 5.6: Agatston method results in aortic valve for enhanced and unenhanced images. (a) Scatter plot with the results of the calcium score. The x-axis represents the value obtained in clinical practice on the unenhanced images. The y-axis indicates the Agatston calcium score for unenhanced images (with threshold of 130HU) obtained by a segmentation expert (blue dots) and enhanced reconstructed images with patient-dependent threshold (red dots).

For CTA versus CT comparisons, the Bland-Altman plot (Figure 5.6c) reveals a mean deviation up to 40 times greater than that of interobserver CT evaluations, reflecting the systematically lower scores from CTA. However, no significant trends are observed here either.

Results using the original slice spacing of CTA images are available in Section B.3.1.

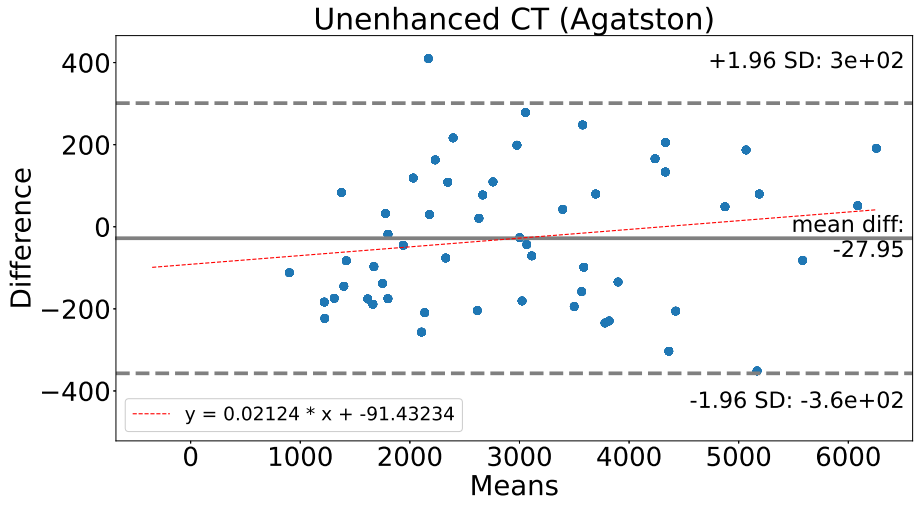


Figure 5.6: (b) Bland-Altman plot comparing the Agatston calcium score obtained by clinicians and by a segmentation expert using the unenhanced images. For each point, the x-axis represents the mean of the two scores, while the y-axis represents the difference between the value obtained by the segmentation expert and the value obtained by the clinician.

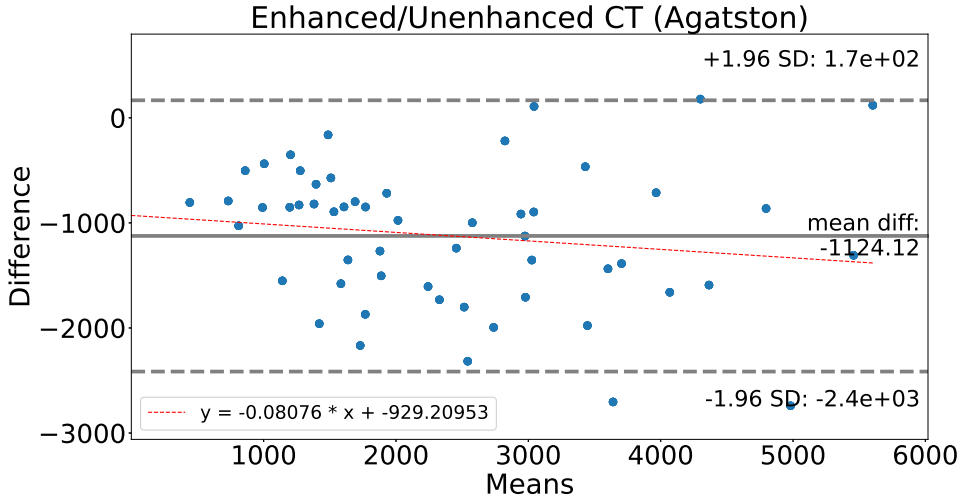


Figure 5.6: (c) Bland-Altman plot comparing the Agatston calcium score obtained by clinicians, using the unenhanced images, and by a segmentation expert using the enhanced reconstructed images. For each point, the x-axis represents the mean of the two scores, while the y-axis represents the difference between the value obtained by the segmentation expert (using CTA) and the value obtained by the clinician (using CT).

5.3.2.2 Volume Scoring

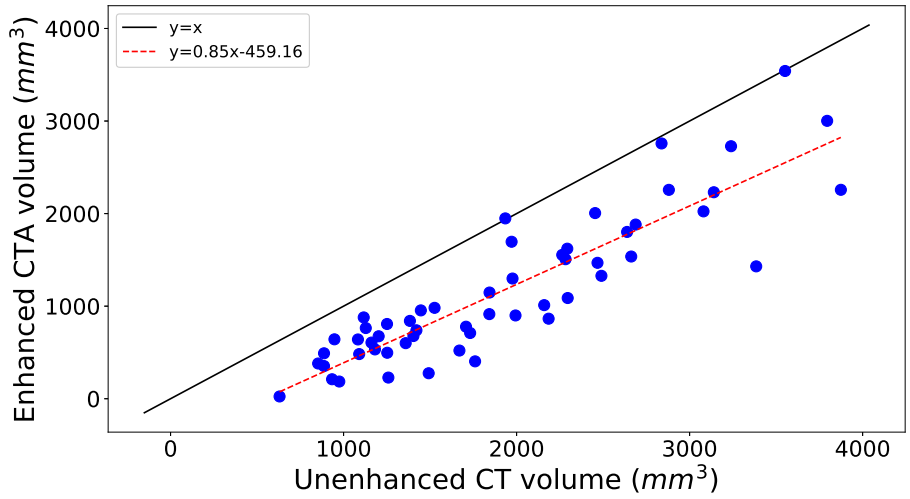
Unlike the Agatston method, calcium volume is not routinely documented in clinical practice. Therefore, our comparison of calcium volumes in the valve region is limited to evaluations performed by an expert in medical image segmentation, using both contrast-enhanced and unenhanced images, rather than clinicians.

The scatter plot illustrating the calcium volume results for the aortic valve in both enhanced and unenhanced images is shown in Figure 5.7a. In this plot, the x-axis represents the calcium volume obtained from unenhanced CT images, while the y-axis indicates the calcium volume score measured from contrast-enhanced CTA images. The distribution of points suggests a general underestimation of volume in the enhanced images compared to the unenhanced ones, though a strong correlation is observed, with a coefficient of $r^2 = 0.89$ (p-value: 3.35×10^{-20}).

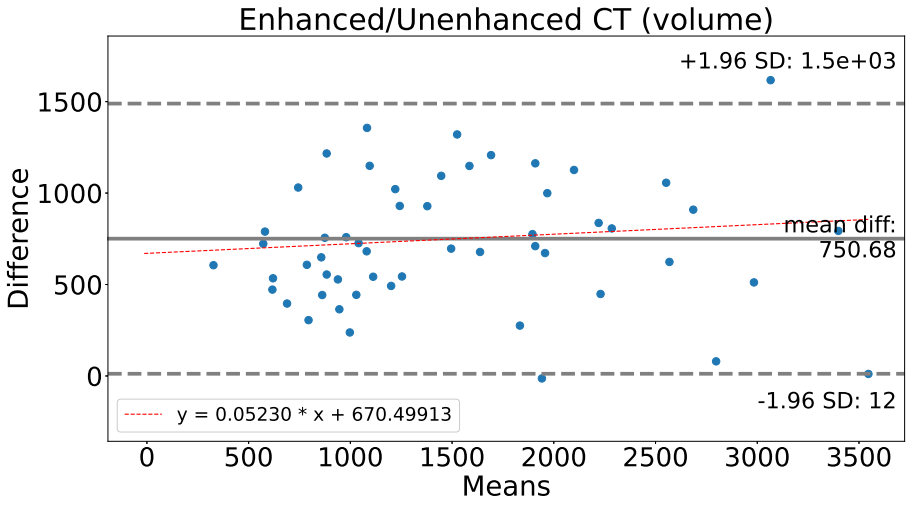
Additionally, Figure 5.7b presents a Bland-Altman plot comparing the calcium volume scores between the two imaging modalities. In this plot, the x-axis represents the mean of the scores from the unenhanced and enhanced images, while the y-axis depicts the difference between the values obtained from CTA and CT. The red line represents the correlation trend between the points. The plot reveals no significant trend, with a mean difference of 750.68 mm^3, and the data points are evenly distributed around this mean without indicating any systematic bias.

The results using the original slice spacing of the CTA images can be found in Section B.3.2.

Since CTA images encompass the entire aorta, the methodology used for calcium volume quantification in the valve region can be extended to other segments, including the tubular aorta, superior arch, inferior arch, and descending aorta. The segmentation procedure remains consistent across all regions, utilizing the same adaptive thresholding and regional separation techniques outlined in Section 5.2.2.2. As this approach has already been validated in the



(a)



(b)

Figure 5.7: Calcium volume results in aortic valve for enhanced and unenhanced images. (a) Scatter plot where the x-axis represents the value obtained on the CT unenhanced images and the y-axis indicates the volume calcium score on the CTA images. (b) Bland-Altman plot comparing the volume calcium score, using the unenhanced images and the enhanced images. For each point, the x-axis represents the mean of the two scores, while the y-axis represents the difference between the value obtained by using CTA and the value obtained using CT. Red line shows correlation line between points.

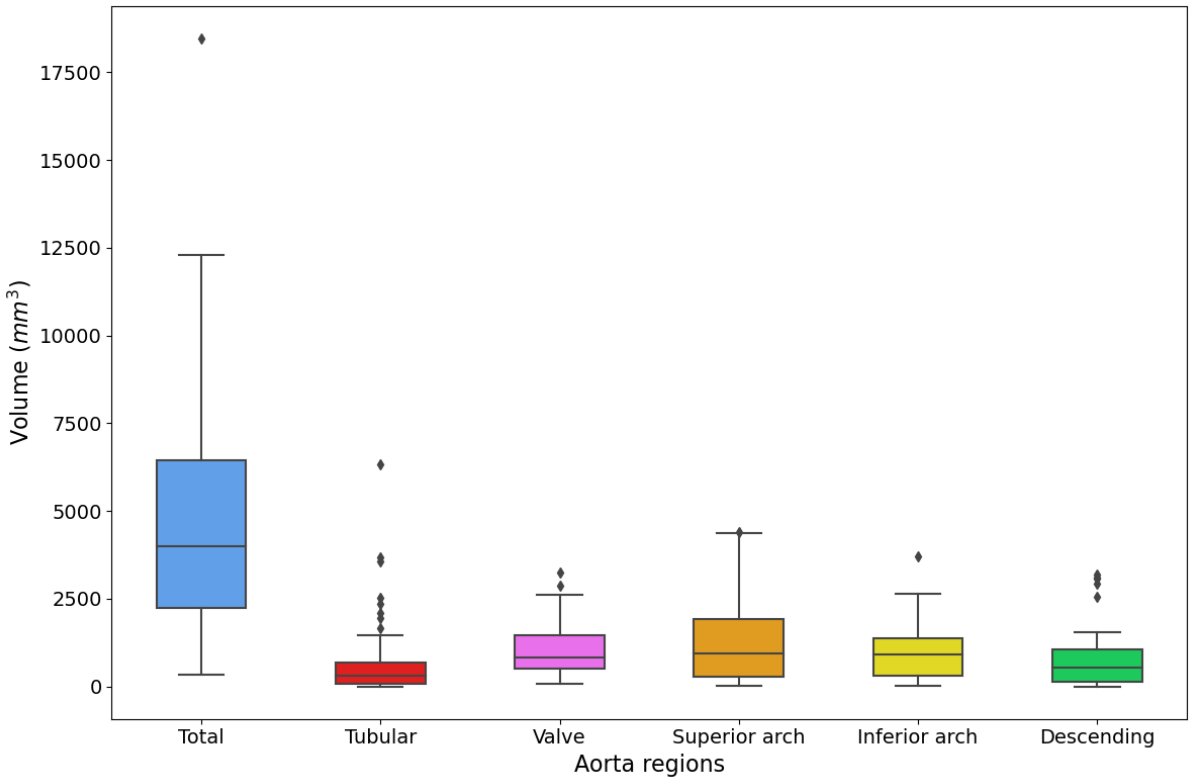


Figure 5.8: Box plot of the total calcium volume values (mm^3) for the enhanced set. From left to right on the x-axis: total calcium volume, tubular region, valve region, superior arch region, inferior arch region, and descending aorta region.

valve region, its application to additional aortic segments can be regarded as reliable. While no official guideline updates or external certifications specifically endorse this method, its uniform application across different regions and adherence to established clinical practices support its validity.

Figure 5.8 presents the box plot statistics for total calcium volume and regional volumes. The x-axis of the box plot represents the different aortic regions, listed from left to right: total calcium volume, tubular region, valve region, superior arch region, inferior arch region, and descending aorta region. As observed in the figure, the valve region, despite being the smallest, contains an average of 1079 mm^3 of calcium, which is approximately 300 mm^3 more than the tubular and descending aorta regions. Among all regions, the superior arch exhibits the highest calcium accumulation on average, reaching 1314 mm^3.

The calcium volume variability across different aortic regions is significant. The total calcium volume exhibits the highest variability, with a standard deviation of 3587.81 mm^3. The tubular region shows a standard deviation of 1150.67 mm^3, while the valve region has a standard deviation of 727.65 mm^3. The superior arch region has a standard deviation of 1218.06 mm^3, and the inferior arch and descending aorta regions show standard deviations of 802.95 mm^3 and 844.89 mm^3, respectively. These differences reflect the varying degrees of calcium accumulation throughout the aorta.

5.3.3 Comparison with Existing Calcium Segmentation Methods in CTA

In this section, we evaluate our proposed segmentation approach by comparing it with three existing techniques for calcium segmentation in CTA. These methods were chosen to represent a diverse set of strategies for calcium quantification.

The first method employs a fixed threshold of 350 HU for calcium segmentation, a widely used approach in the literature [162, 161]. This threshold is applied uniformly across the dataset, without accounting for variations in specific anatomical regions.

The second method, similar to ours, determines a segmentation threshold based on the mean attenuation value. However, instead of defining the threshold within a region of interest (ROI) inside the aorta, it considers the entire aortic structure. The threshold is set as mean + 4SD [158, 169], where the standard deviation (SD) is calculated from all attenuation values within the aorta. This method is referred to as MeanAorta + 4SDAorta.

The third method follows an iterative approach in which the threshold is increased incrementally by 25 HU, starting from 200 HU, until the borders of the three aortic valve cusps are delineated without including the blood pool. This strategy, referred to as the 200 + 25 HU iterative method, aims to refine calcium segmentation by progressively adjusting the threshold to better capture calcium deposits [31].

For the comparison, a subset of 16 patients was randomly selected from the original dataset of 55 patients (see Section 5.2.1). Calcium segmentation was performed for each patient using all four methods, allowing for a comparative evaluation of their performance and accuracy.

Figure 5.9 presents the numerical results of the calcium segmentation comparison, displayed across three vertically arranged plots. The x-axis represents the patient identifier, which remains consistent across all plots. The top plot shows the segmented calcium volume (in mm^3) for the three methods: AdaptativeThrROIAorta (our method), MeanAorta + 4SDAorta, and 200 + 25 HU iterative. The middle plot displays the minimum calcium segmentation threshold selected by each method. The fixed threshold method of 350 HU is excluded from these two plots because it tends to overestimate calcium volume in patients where the blood lumen is bright within this range, leading to incorrect segmentation of the lumen instead of calcium deposits. The bottom plot shows the mean and standard deviation of attenuation values in the aortic lumen, providing additional context for the comparison of segmentation approaches.

In Section 5.3.1, we emphasized the variability in aortic brightness across patients, which inherently necessitates an adaptable threshold for calcium segmentation. This explains why fixed-threshold methods are not suitable for CTA. In particular, the 350 HU threshold only produces reasonable results in patients with aortic lumen values below this threshold, such as P2, P9, or P10 (see the bottom plot in Figure 5.9). In other cases, the fixed threshold incorrectly segments the aorta instead of calcium, leading to its exclusion from the results. However, simply increasing the threshold is not a viable solution, as it would introduce a similar issue where either over-segmentation or under-segmentation occurs, compromising accuracy.

When analyzing the alternative methods, a strong similarity is observed between the AdaptativeThrROIAorta and 200 + 25 HU iterative approaches, as they yield comparable threshold values and calcium volumes. Conversely, the MeanAorta + 4SDAorta method generates higher thresholds, which in turn results in lower calcium volumes due to under-segmentation. Despite these differences, all methods exhibit similar trends in their results.

Another key observation is the high correlation between our method,

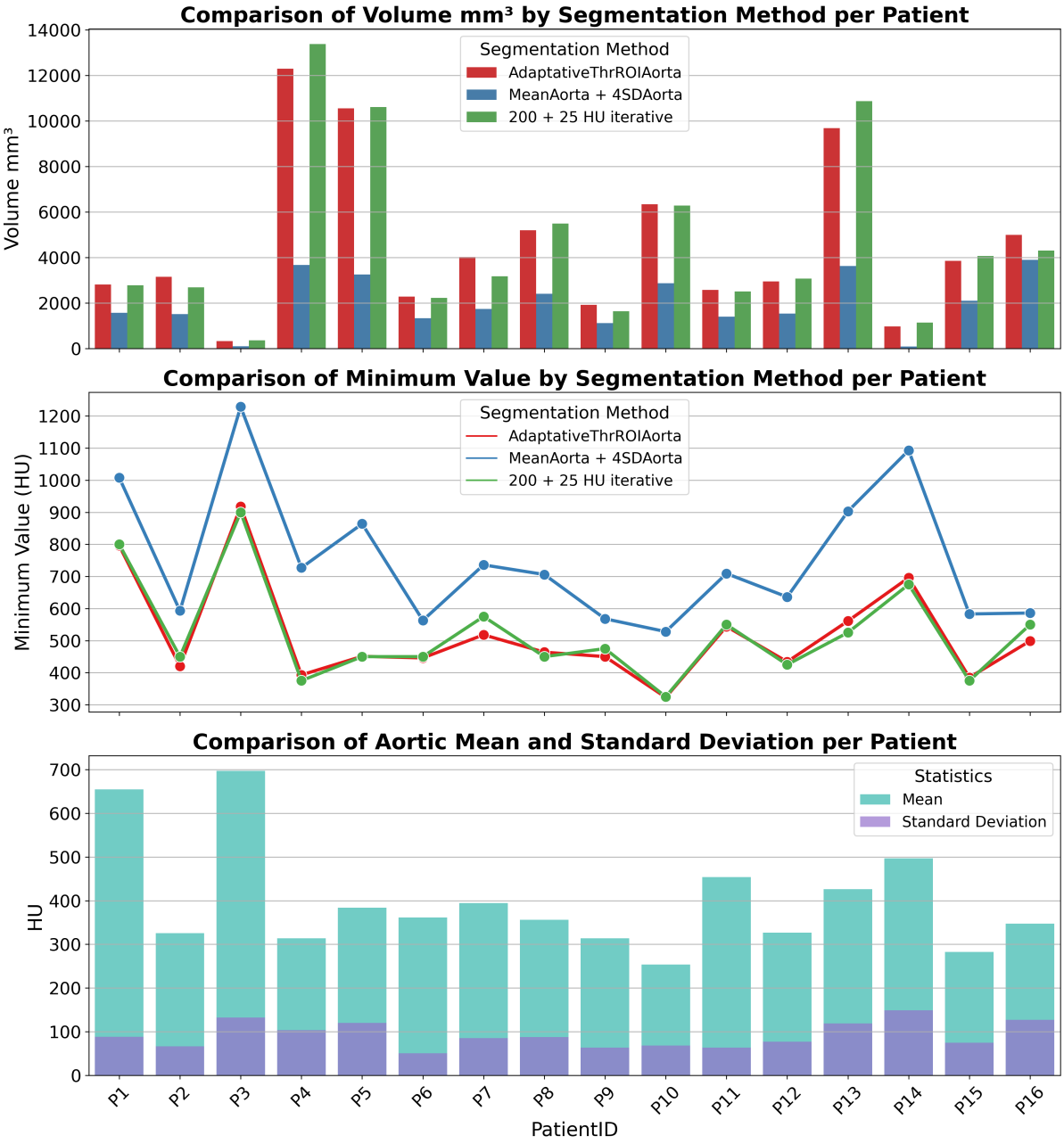


Figure 5.9: Comparison of calcium segmentation methods across three subfigures. In the superior subfigure, the total calcium volume in the thoracic aorta is shown (Y-axis: volume in mm^3). The middle subfigure displays the minimum threshold value for calcium segmentation (Y-axis: HU). In the inferior subfigure, the mean and standard deviation of attenuation values in the thoracic aorta, considering only the lumen, are presented (Y-axis: HU). The results in the superior and middle subfigures are shown for three methods: AdaptativeThrROIAorta (our method), MeanAorta + 4SDAorta, and 200 + 25 HU iterative.

AdaptativeThrROIAorta, and the sum of the mean attenuation value and standard deviation of the aorta, with a Pearson correlation coefficient of 0.9816. This suggests that the MeanAorta + SDAorta approach could be systematically employed as an alternative method, demonstrating that our approach is both robust and computationally efficient. Unlike methods requiring calculations over the entire aorta, our technique is applied within a localized region, making it more practical and suitable for clinical applications.

5.3.4 Automatic Segmentation. Loss Function Ablation Study

In the previous sections, we focused on generating and validating the segmentation methodology and, thus, the dataset. Now, we aim to automate the segmentation process of the aorta and its calcium regions in CTA images. To evaluate the performance of the models (Section 5.2.6), we experimented with different loss functions (see Section 5.2.7) specifically for segmenting aortic calcium, excluding the aorta itself as it is less critical for our task, since we segment the aorta for visualization purposes.

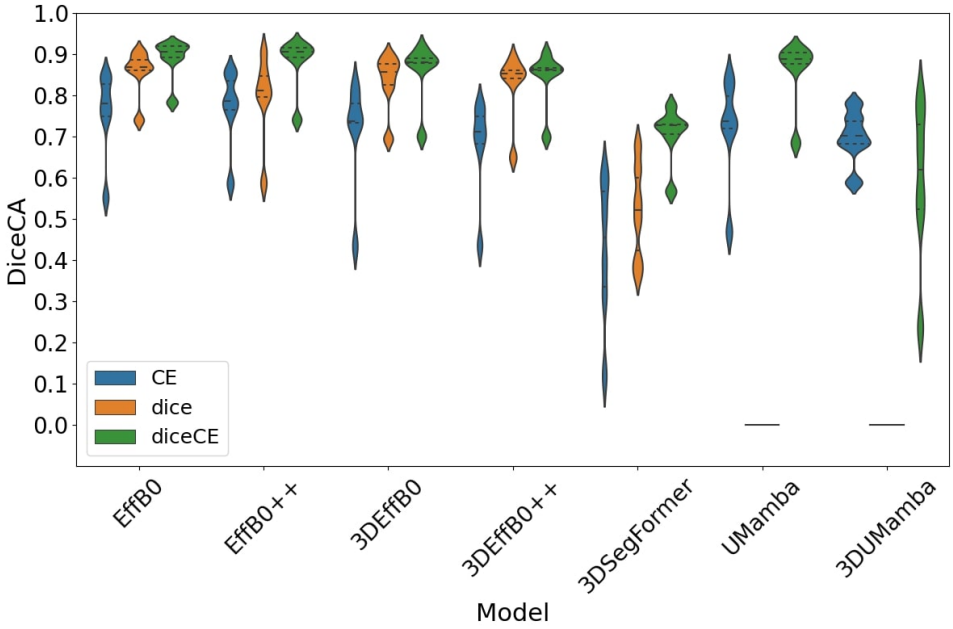
The performance of the models was primarily evaluated using the Dice coefficient (DiceCA) as the main metric. Figure 5.10a presents the results of this evaluation. A violin plot is used to visualize the distribution of the Dice metric values for the different loss functions applied. It combines aspects of both box plots and density plots to display the distribution, spread, and probability density of the data. The x-axis represents the various models, while the y-axis corresponds to the Dice coefficient values. The different loss functions used are indicated by different colors: blue for Cross Entropy (CE), orange for Dice, and green for the combination of Dice and Cross Entropy (DiceCE). The plot shows that the DiceCE loss consistently outperformed other loss functions, achieving DiceCA values around 0.9 for all models. This demonstrates the effectiveness of combining Dice and Cross Entropy losses in improving model performance, particularly in the segmentation of calcium regions.

Among the models tested, the 2D UNet with EfficientNet-B0 (EffB0) achieved the highest median DiceCA score of 0.888, exhibiting the best performance with minimal variability. In comparison, the 3D UNet (3DEffB0) scored slightly lower at 0.86, indicating that 2D models performed better for this task. The UNet++ variants, both 2D (EffB0++) and 3D (3DEffB0++), showed similar or slightly worse performance, suggesting that additional complexity did not lead to improved segmentation accuracy.

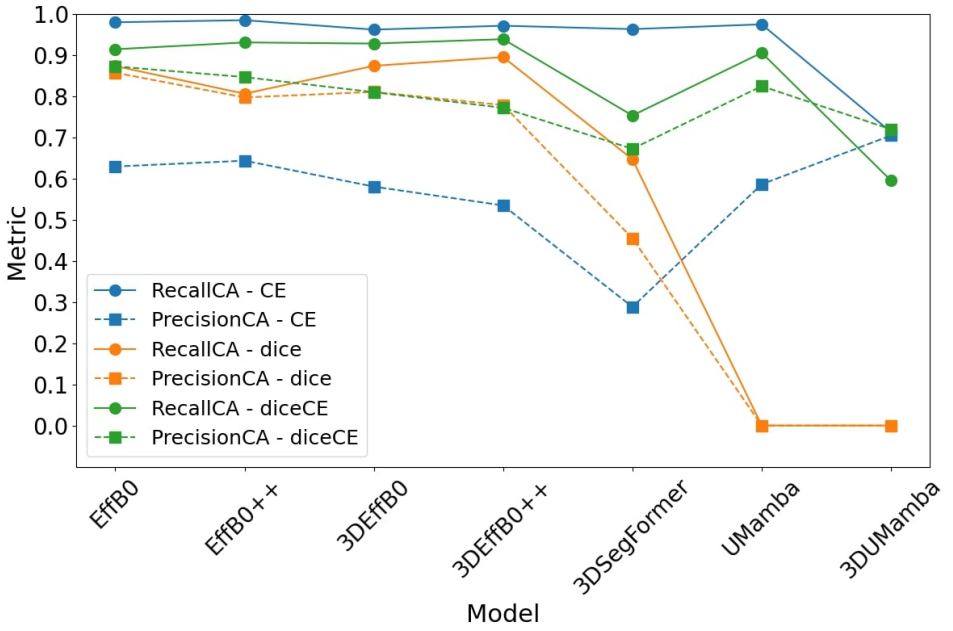
The 3DSegFormer model experienced a significant drop in performance, with DiceCA scores almost 0.2 points lower than the UNet models and higher variability. The UMamba architecture demonstrated variable results; the 2D UMamba with DiceCE loss outperformed 3DSegFormer by 0.1 Dice points, while the 3D variant performed poorly with a DiceCA score of 0.59, indicating potential for further optimization.

Across all models, the 2D architectures consistently outperformed their 3D counterparts, suggesting that simpler 2D models were better suited for this specific task. Other loss functions, such as DiceFocal Loss, GeneralizedDice Loss, and Tversky Loss, did not outperform DiceCE. DiceFocal Loss showed slight improvements (e.g., 0.85 for EffB0), while GeneralizedDice Loss was less effective, and Tversky Loss achieved a DiceCA score of 0.8 for 3DEffB0.

Figure 5.10b presents a scatter plot illustrating the precision and recall metrics for the models. The x-axis represents the models, while the y-axis shows the precision and recall values. The models trained with Cross Entropy (CE) loss are characterized by high recall (above 0.9) but low precision (below 0.65), with the exception of the 3D UMamba, where



(a)



(b)

Figure 5.10: Metrics evaluated on the test set, focusing on the calcium regions (excluding the aorta). The loss functions used are represented by different colors: blue for CE, orange for Dice, and green for DiceCE. (a) Violin plot of the Dice metric. (b) Precision and recall scatter plot.

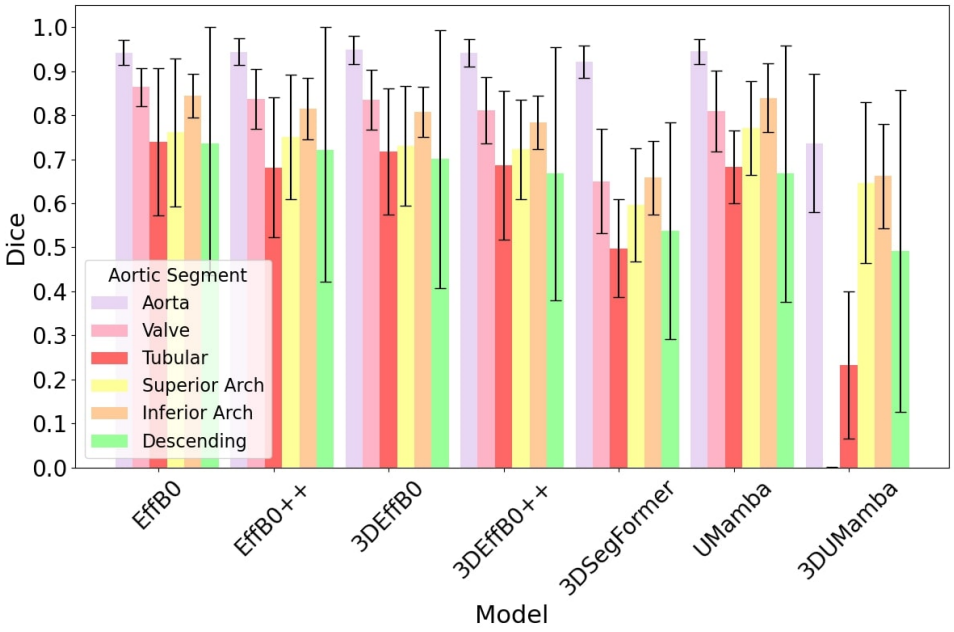


Figure 5.11: Bar plot showing the Dice metric computed on the test set for different regions of the aorta. The models, represented on the X-axis, are trained using the diceCE loss function. The colors represent different regions: purple for the aorta (including calcium), pink for the aortic valve region, red for the tubular aorta region, yellow for the superior aortic arch region, orange for the inferior aortic arch region, and green for the descending aorta region.

both precision and recall are more balanced, typically around 0.7. On the other hand, models trained with Dice and DiceCE losses demonstrate a more balanced precision-recall trade-off, with recall often inversely related to precision due to over-segmentation. This pattern highlights the challenges of balancing segmentation accuracy while avoiding over-segmentation in the calcium regions.

This study emphasizes the importance of selecting the right loss function to achieve the best segmentation performance, highlighting the trade-offs between recall and precision in automated segmentation tasks.

5.3.5 Metrics by Aortic Region

After evaluating the performance of different loss functions, DiceCE loss was selected to assess the segmentation of aortic calcium in various regions of the aorta. This approach allowed for a more detailed analysis of how the segmentation performed across different anatomical areas.

Figure 5.11 displays the Dice scores for segmenting both aortic calcium and the aorta across various models. A bar plot is used to represent the Dice metric for each region, where the x-axis corresponds to the different models and the y-axis represents the Dice scores. The bars indicate the mean Dice score for each model, while the vertical lines represent the standard deviation, providing a sense of the variability in performance across different test cases. The models, trained using the DiceCE loss function, are shown for various regions: the aorta (including calcium) in purple, the aortic valve in pink, the tubular aorta in red, the superior aortic arch in

yellow, the inferior aortic arch in orange, and the descending aorta in green.

The valve region exhibited the highest segmentation performance, with all models achieving scores above 0.8, except for 3DSegFormer, which showed lower performance. The aortic arch also yielded strong results, with Dice scores exceeding 0.75 for most models. The tubular region and descending aorta were more challenging to segment, though the EffB0 model produced the best results at 0.74. For overall aorta segmentation, all models (except 3DUMamba) performed well, with Dice scores above 0.9.

5.3.6 3D Predicted Geometries

A visual evaluation of the model predictions offers insight into the strengths and weaknesses of the segmentation process.

Figure 5.12 and Figure 5.13 show 3D reconstructions of the aortic geometry predicted by the models EffB0, 3DEffB0, 3DSegFormer, UMamba, and 3DUMamba for two test patients with different levels of calcium presence in the aorta. The first figure corresponds to a patient with a high presence of calcium across all aortic regions, while the second figure illustrates a patient with moderate or low calcium presence. In both cases, the ground truth (GT) is depicted, with selected regions highlighted: (A) valve, (B) transition tubular arch, (C) curved section of the arch, and (D) end of the descending aorta. The aortic regions are color-coded as follows: purple for the aorta (including calcium), pink for the aortic valve region, red for the tubular aorta region, yellow for the superior aortic arch region, orange for the inferior aortic arch region, and green for the descending aorta region. These reconstructions allow for a comparison of model predictions across different calcium levels, providing insights into their accuracy in representing the aortic geometry.

In the valve region (Zone A) of Figure 5.12, all models (except 3DUMamba) perform well, although EffB0 and 3DEffB0 occasionally confuse the calcium with the lower arch. In the transition zone (Zone B), EffB0 is the only model to accurately segment the regions, while UMamba struggles. In the curved arch section (Zone C), all models (except 3DUMamba) perform well. At the descending aorta end (Zone D), there is some confusion between the upper arch and the tubular aorta in EffB0, 3DSegFormer, and UMamba.

In the valve region (Zone A) of Figure 5.13, all models (except 3DUMamba) tend to over-segment, with 3DEffB0 showing the best performance. Small calcium regions in the tubular aorta are difficult to segment, and EffB0 fails. In the transition zone (Zone B), all models correctly identify the region as part of the upper arch. Zone C shows minor discrepancies, with EffB0 placing the separation point too high. Finally, in Zone D, all models segment the calcium correctly, except 3DUMamba, which confuses it with the tubular aorta.

5.3.7 Computation time

The segmentation of aortic calcium in CTA was evaluated in terms of processing time for both semi-automatic and AI-based automatic methods. The semi-automatic segmentation process, which relies on expert annotation and refinement, takes approximately 15 minutes per patient. In contrast, the automatic segmentation using deep learning models significantly reduces the processing time, enabling real-time or near real-time analysis.

Table 5.2 presents the mean inference time of the test set for different AI models, along with the number of parameters (in millions). The results show that 2D models, such as EffB0 and

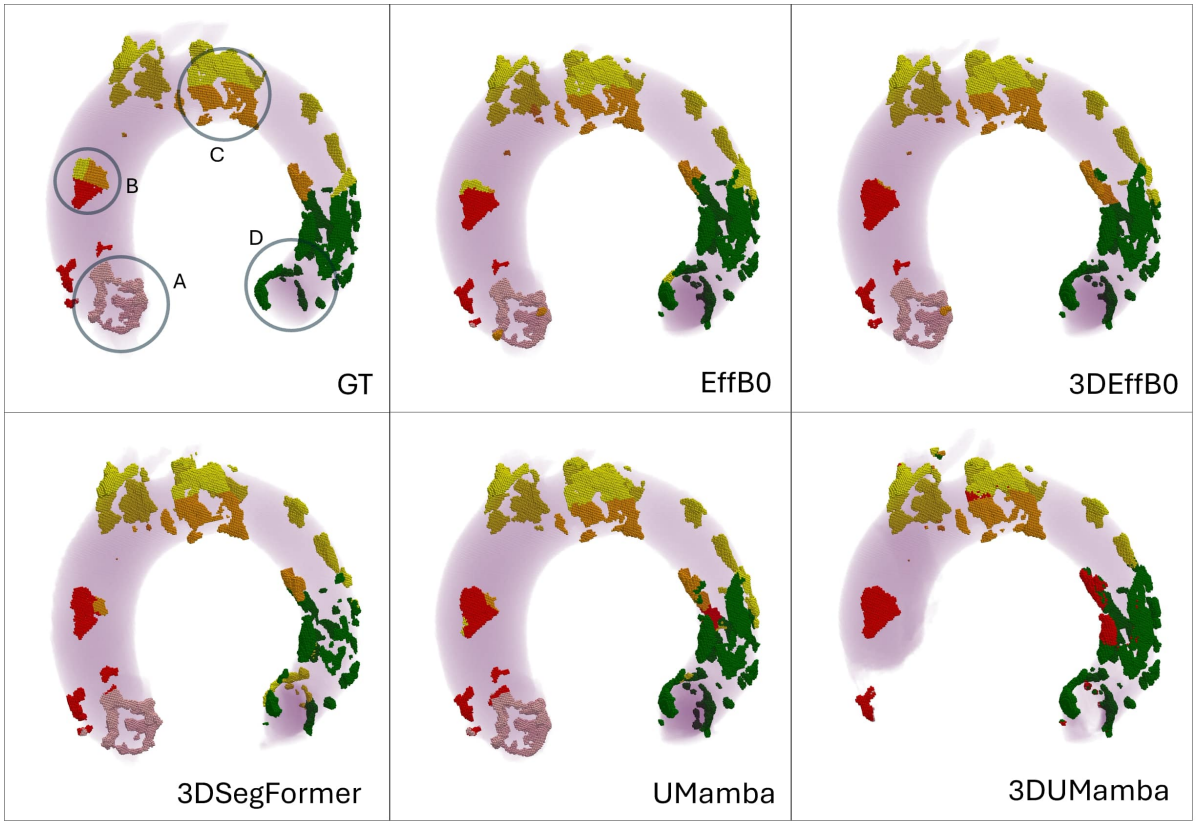


Figure 5.12: 3D reconstruction of the aortic geometry predicted by the models EffB0, 3DEffB0, 3DSegFormer, UMamba, and 3DUMamba for a test patient with a high presence of calcium in all aortic regions. Ground truth is represented as GT, where selected regions are highlighted: (A) valve, (B) transition tubular arch, (C) curved section of the arch, (D) end of the descending aorta. The colors represent different regions: purple for the aorta (including calcium), pink for the aortic valve region, red for the tubular aorta region, yellow for the superior aortic arch region, orange for the inferior aortic arch region, and green for the descending aorta region.

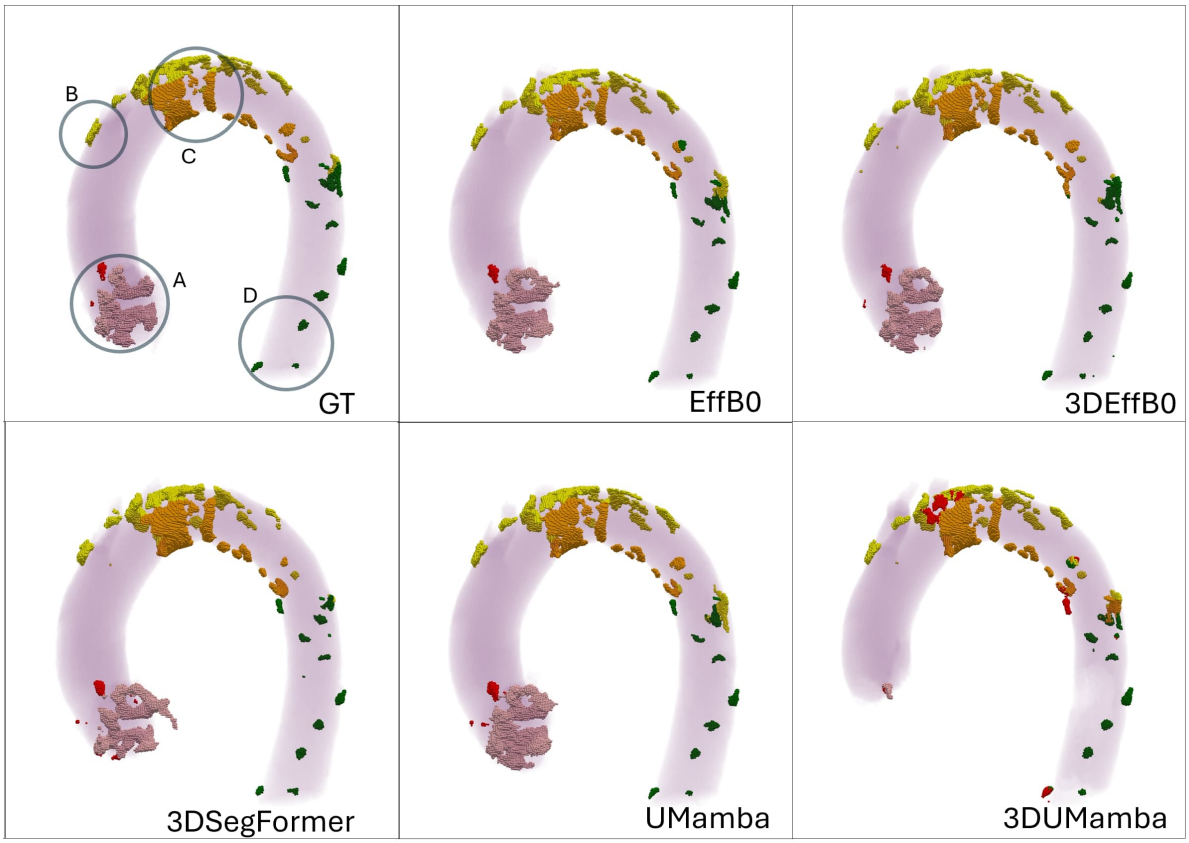


Figure 5.13: 3D reconstruction of the aortic geometry predicted by the models EffB0, 3DEffB0, 3DSegFormer, UMamba, and 3DUMamba for a test patient with a moderate or low presence of calcium in all aortic regions. Ground truth is represented as GT, where selected regions are highlighted: (A) valve, (B) transition tubular arch, (C) curved section of the arch, (D) end of the descending aorta. The colors represent different regions: purple for the aorta (including calcium), pink for the aortic valve region, red for the tubular aorta region, yellow for the superior aortic arch region, orange for the inferior aortic arch region, and green for the descending aorta region.

Table 5.2: Computation time for automatic aortic calcium segmentation in CTA across different models. The table presents the number of parameters (in millions) and the time required to process a single patient (in seconds).

Model	Nº of param. (M)	Time per patient (s)
EffB0	6.25	9.56
EffB0Pre	6.25	9.8
EffB0++	6.57	12
EffB0++Pre	6.57	12
3DEffB0	11.42	178
3DEffB0Pre	11.42	181
3DEffB0++	12.37	249
3DEffB0++Pre	12.37	253
3DSegFormer	4.49	1044
UMAMBA	2.85	3 (GPU)
3DUMAMBA	7.48	55 (GPU)

its variations, offer faster predictions, with times ranging from 9.56 to 12 seconds per patient. The 3D models, while providing more spatial context, require higher computational resources, with prediction times ranging from 178 to 1044 seconds. The 3DUMamba model, specifically optimized for GPU acceleration, achieves a significantly lower inference time of 55 seconds compared to other 3D architectures.

The computations were performed on a workstation equipped with an Intel Core i5-8500T CPU and 32 GB of RAM. However, UMamba was run on a system with 64 Intel Xeon Ice Lake 8352Y cores and an NVIDIA A100 GPU, as it required a Linux operating system.

These results highlight the trade-off between model complexity and inference time. While 3D models generally require more time due to their increased parameter count, optimized architectures such as UMamba and 3DUMamba leverage GPU acceleration to achieve significantly lower processing times, making them more suitable for real-time clinical applications.

5.4 Discussion and Knowledge Transfer to Industry

In this chapter, we address the dual objectives of segmenting the aorta and calcium in contrast-enhanced CT angiography (CTA) scans and automate this process using neural networks. The first part focuses on overcoming the inherent challenges of accurate calcium segmentation in CTA, including the variability in contrast levels and the impact on threshold-based methods. Once validated, the segmentation framework serves as the foundation for training neural network models using the dataset generated during the initial step. This discussion evaluates the results, compares them with existing literature, and highlights key findings and limitations.

Calcium scoring using the Agatston method is an established approach for evaluating

the severity of aortic stenosis and predicting stroke risk in TAVI patients. However, it has limitations. First, non-contrast-enhanced images used in this method fail to provide detailed anatomical visualization of the aorta. Second, the large slice spacing (typically 3 mm or 2.5 mm) increases partial volume effects, leading to mixed tissue representation within voxels.

An alternative is using contrast-enhanced images, offering higher-resolution visualization (less than 1 mm spacing) and anatomical detail of the aorta, enabling better assessment during TAVI.

When calculating calcium scores using contrast-enhanced images, one challenge is determining the appropriate calcium threshold due to the varying attenuation values in these images, as noted by Holcombe et al. (2022) [30]. This issue has been discussed in the literature, with Van et al. (2010) [166] using the Agatston method to calculate calcium scores in both contrast and non-contrast coronary artery images. In contrast images, an expert manually traces the calcium contour, bypassing the use of a threshold. The correlation between calcium scores from both image types is reported as $r = 0.78$ and $r = 0.62$ for two different observers, respectively.

Other studies, such as Yoon et al. (1997) [170], compare the Agatston method with volume or mass-based calcium scores. For example, Hong et al. (2002, 2003) [162, 161] use mass as a calcium score, calibrated via a phantom study, and argue that this approach has better reproducibility than the Agatston method. However, they employ a fixed calcium threshold of 350HU for segmentation in contrast-enhanced images.

Regarding aortic valve calcium, recent research by Holcombe et al. (2022) [30] advocates for using a patient-dependent threshold. For instance, Kim et al. (2018) [158] report a correlation of $r = 0.87$ between calcium scores in contrast and non-contrast images when using volume as a calcium score. This study also suggests that a patient-specific threshold (*aortic attenuation value* +4SD) improves results over a fixed threshold. Similarly, Alqahtani et al. (2017) [169] reports a high correlation of $r = 0.982$ using this approach. The upper threshold for detecting non-calcific tissue, proposed by Cartlidge et al. (2021) [31], is based on calibrating to blood pool attenuation starting at 200HU, and this method correlates CTA volume scores with CT Agatston scores with an $r = 0.68$.

Other studies, such as Bettinger et al. (2017) [171], Kofler et al. (2021) [172], and Angelillis et al. (2021) [32], also support the idea that computing a threshold based on the aortic attenuation value yields better results. In contrast, Eberhard et al. (2017) [159] developed a regression model to compute Agatston scores in CTA, reporting an intraclass correlation coefficient of 0.897 but noting an underestimation of calcium score by a factor of 2 when comparing CTA to CT.

In this study, we present a method for calculating the calcium score across the entire aorta in CTA, extending beyond the aortic leaflets. Our approach uses a threshold based on the maximum attenuation value of the aorta, allowing the method to adapt to each acquisition, in line with previous works [158, 169, 31, 171, 172, 32]. After segmenting the calcium, we compute the calcium score across various regions of the aorta, including the aortic leaflets, tubular aorta, aortic arch, and descending aorta.

To validate this method, we compare the calcium scores obtained from CTA images (reconstructed to match the slice thickness of CT images) with scores derived from the Agatston and volume methods in CT.

The results demonstrate excellent agreement between calcium scores computed by two different experts on the same unenhanced dataset using the Agatston method. The absolute

error of 147.23 Agatston units and a bias of −27.95 Agatston units are clinically acceptable, given the typical calcium score ranges observed in the aortic valve. Unlike coronary artery calcium (CAC) scoring, which follows broad clinical classification thresholds, aortic valve calcium (AVC) scoring operates on a different scale with significantly higher values. According to the 2021 ESC/EACTS Guidelines for the management of valvular heart disease [165], the probability of severe aortic stenosis (AS) based on calcium scoring is classified as follows: highly likely for men above 3000 and women above 1600 Agatston units, likely for men above 2000 and women above 1200 Agatston units, and unlikely for men below 1600 and women below 800 Agatston units.

Given these high thresholds, the observed bias of −27.95 Agatston units suggests only a minimal underestimation, which remains well within acceptable limits for clinical interpretation [163, 166]. Similarly, the absolute error of 147.23 Agatston units falls within the range of interobserver variability typically seen in clinical practice when two experts independently calculate the calcium score.

Regarding the CT-CTA comparison, we obtained correlations of $r = 0.87$ and $r = 0.89$ for the Agatston method and volume, respectively. These values are similar to those reported by other studies such as [158] and superior to those reported by [31].

The Bland-Altman mean difference of 750.68 mm^3 for the volume score suggests a systematic underestimation of calcium volume in contrast-enhanced CTA compared to non-contrast CT. Although this bias is evident, it does not necessarily diminish the clinical relevance of the method. The high correlation ($r^2 = 0.89$) indicates that calcium quantification remains consistent between imaging modalities, suggesting that despite a systematic offset, CTA-based measurements still offer reliable assessments. Unlike coronary artery calcium (CAC) scoring, which depends on well-defined threshold values, aortic calcium quantification does not rely on fixed clinical cutoffs. As a result, the observed underestimation is unlikely to have a significant impact on risk stratification. Nevertheless, future research could investigate the development of correction factors to refine CTA-derived scores, enhancing their alignment with non-contrast CT measurements and improving their applicability in clinical practice.

This underestimation is consistent with findings from Van der Bijl et al. (2010) [166], Hong et al. (2003) [161], and Cartlidge et al. (2021) [31]. Notably, the histogram analysis in Figure 5.5a shows that 91% of the maximum blood density values exceed 400HU, leading to an Agatston score consistently set to 4 for the majority of cases. This results in a calcium volume being calculated proportionally, as described in Algorithm 1.

As a key innovation, after validating the method for contrast-enhanced imaging, we apply the volume calcium score to quantify aortic calcium by region. This approach proves particularly useful prior to a TAVI procedure, as calcium may be dislodged not only from the aortic valve but also from the vessel walls during catheter insertion. The arch area, with its curvature and higher plaque accumulation, is especially vulnerable. In line with other studies like Guilenea et al. (2024) [62] and Otsuka et al. (2022) [157], we emphasize the importance of assessing calcium throughout the entire aorta to guide TAVI interventions effectively.

After validating the dataset, we moved on to automating the segmentation of the aorta and calcium by region in contrast-enhanced CT angiography (CTA) images. To achieve this, we explored various deep learning models, including widely known convolutional neural networks (CNNs), recent transformer-based architectures, and the UMamba model in both its 2D and 3D versions. Among these, the 2D UNet with EfficientNet-B0 (EffB0) backbone outperformed the others, achieving the highest Dice scores of 0.94 for the aorta and 0.888 for calcium

segmentation. This model's ability to capture fine details in individual slices made it especially well-suited for this task.

Although 3D models like 3D UNet and 3DSegFormer make use of volumetric data, they struggled with boundary precision. While the 3DSegFormer achieved a Dice score of 0.921 for the aorta, it performed less well in calcium segmentation compared to the EffB0 model. This suggests that CNN-based models, such as EffB0, are better suited for fine-grained tasks like calcium detection, where precise local feature capture is crucial.

Segmenting the tubular and descending aorta proved to be the most challenging due to their similar geometries and underrepresentation in the dataset, which accounted for around 30% of the slices. On the other hand, the aortic valve was well-segmented across most models, with EffB0 achieving Dice scores above 0.73, while 3DUMamba struggled in the leaflet regions.

The DiceCE loss function performed the best, striking a balance between overlap and pixel-level accuracy, compared to alternatives like CE, Generalized Dice, and Tversky losses. Additionally, incorporating class weights improved performance, especially in underrepresented regions, leading to more accurate calcium segmentation.

Despite the promising outcomes, the study is limited by the small dataset size, which could affect model robustness and generalizability. Expanding the dataset and establishing standardized benchmarks for aortic calcium segmentation could improve model performance and facilitate more reliable comparisons with other studies.

The knowledge transfer to industry in this project was immediate, as the research originated directly from a collaboration with the entity, which also provided financial support. This seamless integration between academia and industry ensured that the developed methodologies could be rapidly translated into practical applications.

As a result of this research, a fully annotated dataset comprising 55 CTA volumes was generated, including detailed segmentations of both the aorta and calcium deposits categorized by region. This dataset represents a valuable asset for further advancements in the field and potential future industrial applications.

Beyond validating a robust segmentation methodology, the research also led to the complete automation of the process, significantly reducing computation times. While semi-automatic segmentation required approximately 15 minutes per patient, the AI-based automatic approach drastically improved efficiency. Depending on the model, processing time ranged from just 3 seconds per patient (for UMAMBA on a GPU) to around 6 seconds for EffB0. These improvements in automation and efficiency highlight the potential for real-world deployment in clinical and industrial settings, optimizing workflows and reducing manual workload.

Conclusions

Various techniques for medical image processing and segmentation have been developed throughout this thesis to improve diagnostic accuracy and efficiency. These techniques aim to enhance the quality and interpretation of medical images in a more efficient and non-invasive way. In addition, artificial intelligence (AI) has played a crucial role in automating these processes, allowing for faster and more accurate results. This work is framed within an industrial PhD, emphasizing the immediate applicability of these techniques to real-world clinical and industrial settings. The research is structured into three main chapters, each addressing a specific challenge in medical imaging using different methodologies:

- Chapter 3: Focuses on reducing metal artifacts in CT images caused by metallic implants. This is essential for improving image quality and ensuring accurate clinical assessments.
- Chapter 4: Explores coronary artery segmentation in Coronary Computed Tomography Angiography (CCTA) to develop an automatic AI-driven coronary tree segmentation, providing the geometries necessary to compute clinical hemodynamic parameters as a non-invasive alternative to traditional methods for assessing coronary artery disease.
- Chapter 5: Focuses on the segmentation of thoracic aortic calcifications by region to generate a calcium score, which provides an estimate of the extent of calcification in the aorta. This information may help clinicians prior to a Transcatheter Aortic Valve Implantation (TAVI) procedure, as it can assist in assessing the need for carotid protection devices to prevent the dislodging of calcified plaque, which could potentially lead to cerebrovascular events.

The following points summarize the key findings, contributions, and future research directions.

1. Metal Artifact Reduction (MAR): A supervised learning approach based on a U-Net model was developed to reduce metal-induced artifacts in CT images, while preserving good contrast between different structures. The model is lightweight and has shown effective artifact elimination, thanks in part to the careful selection of the loss function. Although the approach has been tested with a limited number of patients, its potential extends beyond oncological imaging. This broader applicability is envisioned as part of future work.
2. Challenges in Coronary Segmentation: Contrast-enhanced CT images exhibit brightness variability across acquisitions, making fixed-threshold segmentation methods ineffective for structures like the lumen or calcium. This highlights the need for adaptive or learning-based approaches.

3. Applicability of CNNs in Medical Imaging and Dataset Size: Convolutional neural networks (CNNs) demonstrated strong performance, particularly when training data is limited. For metal artifact reduction and segmentation tasks, CNNs outperformed more advanced architectures like Transformers or UMamba under small-data conditions.

4. 3D Networks. Strengths and Limitations: While 3D neural networks capture volumetric context and produce more connected vessel segmentations, they struggle with recognizing finer structures, especially in distal regions. Their computational demands and need for large datasets present a trade-off between accuracy and efficiency.

5. Impact of Transfer Learning: Pretrained encoders significantly enhanced performance, demonstrating that models pretrained on large datasets, even those not related to medical imaging, can be effectively adapted for medical imaging tasks. This approach enabled the successful segmentation of entire coronary trees with as few as 15 training patients.

6. Advantages of Unsupervised Algorithms: Unsupervised algorithms present a notable advantage over supervised learning approaches by not requiring annotated datasets. These methods are particularly beneficial in contexts where labeled data is scarce or difficult to obtain, as is often the case with medical imaging datasets. Among unsupervised techniques, we have demonstrated that clustering methods can be a good alternative, effectively analyzing and segmenting medical images without the need for prior annotations.

7. Ward Clustering for Coronary Structures: The Ward clustering algorithm effectively differentiated coronary structures such as the lumen, calcium, and background, providing high-quality segmentations without requiring deep learning models.

8. Validation of calcium segmentation and scoring with Clinical CT: We propose and validate a methodology for the segmentation of thoracic aortic calcium by regions and its corresponding calcium score, using clinical CT data. The methodology captures the brightness variations that may occur in the images from different patients, ensuring its reliability and consistency with current diagnostic practices.

9. Automated Calcium Scoring: The algorithm is fully automated, performing all segmentations in a single step using a U-Net based model. This significantly improves efficiency by reducing manual intervention, offering potential applications in cardiovascular risk assessment, TAVI planning, and monitoring atherosclerosis progression.

10. Automation and Efficiency Gains: We automatized medical image analysis, leading to significant time savings in clinical workflows. For coronary artery segmentation, the process time has been reduced from 1-2 hours to under 30 minutes, including both segmentation and visualization. In the case of aortic calcium scoring, the time has decreased from around 15-20 minutes to less than one minute in the fastest cases. This reduction in processing time translates to improved efficiency, lower labor costs, and increased throughput, making these solutions highly valuable in a commercial setting.

11. Creation of High-Quality Annotated Datasets: A key outcome of this research was the development of annotated, high-quality medical imaging datasets, which are crucial for

advancing AI-driven diagnostic tools. These datasets serve as a foundation for further research, enabling more robust and generalizable models. In an industrial context, such datasets hold strategic value, fostering innovation, improving model accuracy, and potentially being leveraged for commercial AI solutions in medical imaging.

This thesis presents significant advancements in the automation of medical image segmentation, contributing to more accurate, efficient, and scalable analysis in clinical practice. By integrating artificial intelligence and cutting-edge processing techniques, this work enhances diagnostic precision, optimizes workflows, and paves the way for more personalized, data-driven decision-making. Beyond its direct applications in cardiology, the methodologies developed here have the potential to be expanded to other medical imaging domains, further improving patient care. This research not only addresses current challenges but also sets the foundation for future innovations in the field. The impact of this work extends beyond the academic setting, offering tangible benefits to healthcare professionals and, ultimately, to patients worldwide.

Appendix A

AI Architectures

In this appendix, we provide detailed information on several architectures discussed in the Methods section (Chapter 2). This supplementary material aims to streamline the main text by removing overly technical descriptions, making for a more engaging reading experience. This appendix serves as a comprehensive resource for those interested in the technical specifics, enhancing the overall understanding of the methodologies employed in this study.

A.1 MobileNet

In this section, we introduce the efficiency comparison of MobileNet, a lightweight architecture designed for mobile devices. MobileNet utilizes depthwise separable convolutions, which break down the convolution process into two steps: applying a single filter to each input channel (depthwise convolution) and then combining these outputs with a 1×1 convolution (pointwise convolution). This innovative approach significantly reduces computational costs and model size, allowing MobileNet to achieve high efficiency and speed without sacrificing accuracy, making it well-suited for resource-constrained environments.

Efficiency comparison By factorizing the convolution operation into these two steps, MobileNet achieves a substantial reduction in computational complexity [83]. We now compare the complexity of MobileNet with that of a standard convolution.

- For a standard convolution, the number of operations is given by Equation A.1, which accounts for sliding the kernel over every pixel in the input.

$$\text{Operations for standard convolution} = H \times W \times C_{in} \times C_{out} \times K^2, \tag{A.1}$$

  where H and W are the height and width of the input, C_{in} is the number of input channels, C_{out} is the number of output channels, and K is the kernel size.

- In depthwise separable convolution, we factorize the standard convolution into two parts. First depthwise convolution applies a $K \times K$ filter for each input channel. Then, pointwise convolution applies 1×1 convolution. The associated complexity of these two operations is given by Equation A.2 and Equation A.3, respectively.

$$\text{Operations for depthwise convolution} = H \times W \times C_{in} \times K^2, \tag{A.2}$$

$$\text{Operations for pointwise convolution} = H \times W \times C_{in} \times C_{out}. \tag{A.3}$$

Thus, the total number of operations for the depthwise separable convolution is given by Equation A.4.

$$\text{Total operations for MobileNet} = H \times W \times C_{in} \times (K^2 + C_{out}). \tag{A.4}$$

Now, we divide the total operations for standard convolution by the total operations for depthwise separable convolution to measure the efficiency gain:

$$\text{Efficiency factor} = \frac{\text{Operations for MobileNet}}{\text{Operations for standard convolution}}.$$

Substitute the formulas with both Equation A.1 and Equation A.4:

$$\text{Efficiency factor} = \frac{H \times W \times C_{in} \times (K^2 + C_{out})}{H \times W \times C_{in} \times C_{out} \times K^2}.$$

Simplifying the expression:

$$\text{Efficiency factor} = \frac{1}{C_{\text{out}}} + \frac{1}{K^2}.$$

With typical values such as $K = 3$ and $C_{out} \gg 1$, MobileNet can reduce operations by up to 8-9 times compared to a full convolution [84].

A.1.1 MobileNetV2

The core contribution of MobileNetV2 is the inverted residual with linear bottleneck module. This module starts with a low-dimensional, compressed input representation, which is first expanded to a higher dimension through a 1×1 convolution with a ReLU6 activation. A lightweight depthwise convolution follows, which filters features spatially while maintaining computational efficiency. Finally, the features are projected back to a low-dimensional representation through another 1×1 convolution without any non-linearity.

Applying the ReLU activation function in layers with low-dimensional representations can lead to significant information loss. ReLU sets all negative values to zero, which reduces the representational capacity when used in narrow layers [84]. To mitigate this issue, MobileNetV2 avoids using ReLU (or any activation) in the final projection layer, ensuring that the representational power is retained. Conversely, MobileNetV2 employs ReLU6, a variant of ReLU, which caps the maximum output value to 6 (see Equation A.5). This bounded non-linearity prevents excessively large activations, ensuring numerical stability and improved efficiency.

$$\text{ReLU6}(x) = \min(\max(0, x), 6). \tag{A.5}$$

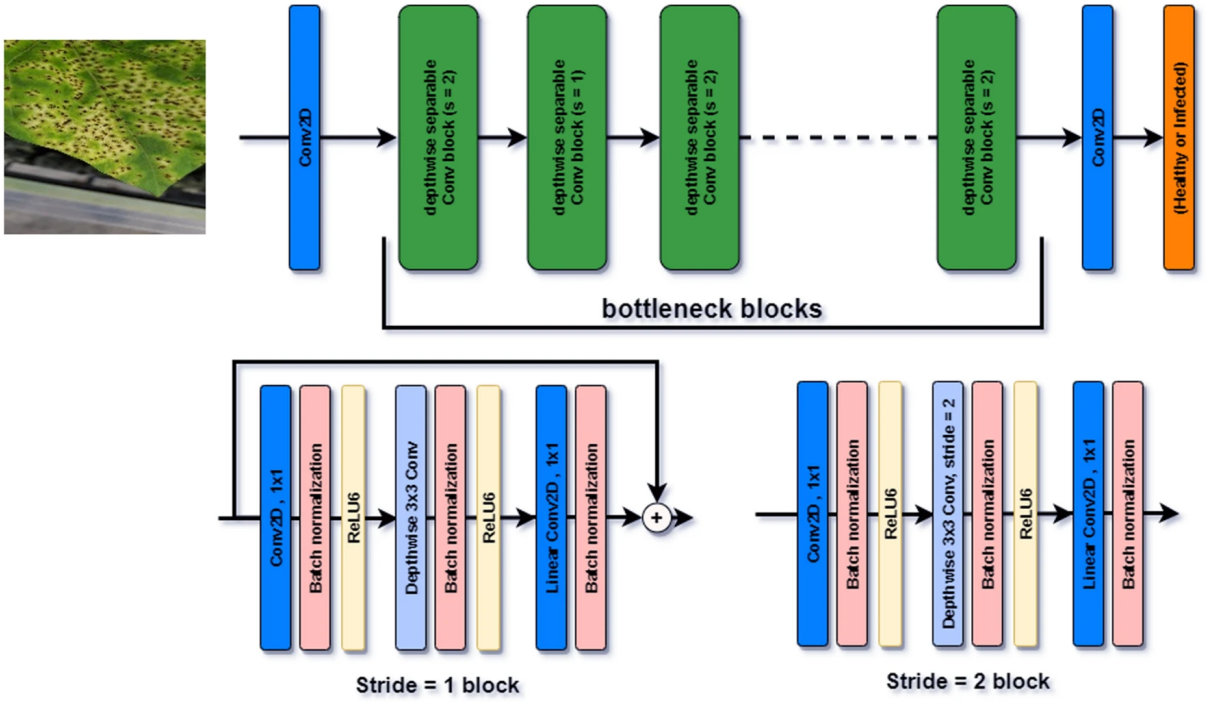


Figure A.1: Illustration of the two primary block types in MobileNetV2. The first block (left) shows a residual block with a stride of 1. The second block (right) represents a downsampling block with a stride of 2, where no residual connection is used due to differing input and output dimensions. Each block consists of three layers: (1) a 1×1 pointwise convolution with ReLU6 activation for expansion, (2) a depthwise convolution with ReLU6, and (3) a 1×1 pointwise convolution for projection, all followed by batch normalization. Image from [173].

The architecture of MobileNetV2, illustrated in Figure A.1 consists of two primary types of blocks: residual blocks with a stride of $s = 1$ and downsampling blocks with a stride of $s = 2$. Each block contains three layers, with every convolution followed by batch normalization to stabilize and accelerate training:

- First Layer: A 1×1 pointwise convolution with ReLU6 activation to expand the input to a higher dimension using an expansion factor t, where $t = 6$ in most experiments. For instance, if the input has 64 channels, the expanded output will have $64 \times 6 = 384$ channels.
- Second Layer: A depthwise convolution that applies a single filter to each channel separately, providing spatial filtering while keeping the computational cost low. Followed by a ReLU6 activation.
- Third Layer: Another pointwise 1×1 convolution to project the features back to a low-dimensional representation, but without any non-linearity to avoid information loss.

The inverted residual connection is applied only when the input and output dimensions are identical, enabling the input to be directly added to the output. This shortcut connection enhances gradient flow during backpropagation, improving training efficiency. However, when downsampling is required (e.g., using a stride of 2), the input and output dimensions differ,

and the residual connection is omitted to ensure proper spatial reduction while maintaining the network's lightweight, efficient design.

The term inverted residual arises from the fact that residual connections occur between compressed layers, reversing the typical structure seen in ResNet. Instead of linking layers with wide, high-dimensional representations, MobileNetV2 starts with a narrow, low-dimensional input, expands it to a higher dimension and then compresses it back to a low-dimensional form. This inversion creates a residual structure opposite to that of traditional architectures, where expansions happen after the skip connection.

A.2 EfficientNet

In this section, we explore the compound scaling method employed by EfficientNet. Compound scaling addresses the limitations of previous models by simultaneously scaling the depth, width, and resolution of the network, rather than adjusting these factors independently. This holistic approach enables EfficientNet to achieve better performance while maintaining a smaller computational footprint. By carefully balancing these dimensions, EfficientNet ensures that increases in capacity translate into meaningful improvements in accuracy, making it a powerful option for a variety of image classification tasks.

Compound Scaling: A Balanced Approach EfficientNet's scaling method introduces a compound coefficient ϕ to control how the network's depth (d), width (w), and resolution (r) are scaled. The scaling is performed according to the following equations:

$$d = \alpha^{\phi}, \tag{A.6}$$

$$w = \beta^{\phi}, \tag{A.7}$$

$$r = \gamma^{\phi}, \tag{A.8}$$

where α, β, and γ are constants determined through a small grid search, and ϕ is a user-specified coefficient that controls the total resources allocated for scaling. The authors impose the constraint:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2,$$

to ensure balanced scaling across all three dimensions. Additionally, the constraints $\alpha \geq 1$, $\beta \geq 1$, and $\gamma \geq 1$ guarantee that each dimension is scaled by at least its original size.

EfficientNet-B0 serves as the baseline model from which the rest of the EfficientNet family is derived. It is designed with a lightweight but powerful architecture that leverages inverted residual blocks and depthwise separable convolutions, similar to MobileNetV2, to ensure computational efficiency. It takes an input resolution of 224×224 and uses 16 inverted residual blocks This carefully designed architecture balances accuracy and computational efficiency, making it an ideal baseline for scaling. The architecture, described in Table A.1, includes:

- Stem Layer: A 3×3 convolution followed by batch normalization and ReLU activation, reducing the spatial dimensions while extracting basic features.
- Inverted Residual Blocks: These blocks consist of three layers—an expansion layer (pointwise 1×1 convolution), a depthwise convolution, and a projection layer (another

Table A.1: Architecture details of EfficientNet-B0, showing the configuration of each stage i, including the number of layers $\hat{L}_i$, input resolution $\langle \hat{H}_i, \hat{W}_i \rangle$, and channels $\hat{C}_i$. This baseline model serves as the foundation for the EfficientNet family, with subsequent versions (B1 to B7) scaling depth, width, and resolution according to the compound scaling formula (see Section A.2). Table from [81].

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

1×1 convolution). ReLU6 activation is applied after the expansion and depthwise layers, and batch normalization follows each convolution (see Section 2.3.1.5).

- Squeeze-and-Excitation (SE) Modules [174]: Embedded in the inverted residual blocks, SE modules recalibrate the feature maps by adaptively weighting each channel, enhancing the model's representational power without increasing computational cost significantly.
- Final Layers: A final pointwise convolution, global average pooling, and a fully connected layer with softmax activation for classification.

The parameters α, β, and γ, which control the scaling of depth, width, and resolution respectively, are determined through a comprehensive grid search to ensure an optimal balance between accuracy and efficiency [81]. Once fixed, these parameters provide a consistent scaling framework. The compound coefficient ϕ is then determined by the user to proportionally scale the network dimensions (see Section A.2).

A.3 Transformers

The Transformer follows an encoder-decoder architecture commonly used in sequence transduction tasks. The encoder transforms the input sequence into a set of continuous representations that encapsulate its contextual information. The decoder then takes this encoded information and sequentially generates an output sequence. This structure is essential for tasks like machine translation, where input and output sequences are often of different lengths. To understand how Transformers achieve this, we'll break down their main steps, following a sequence of words through the model.

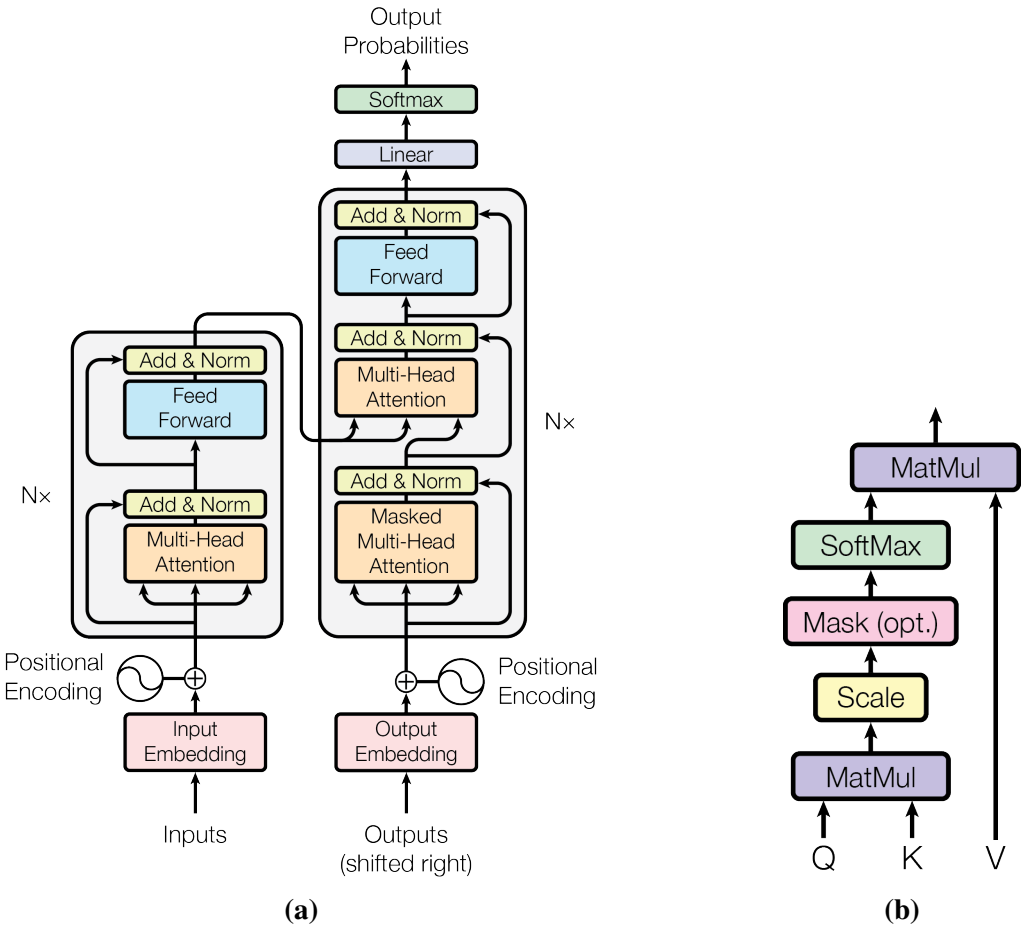


Figure A.2: (a) Transformer architecture as proposed in the original paper [87]. (b) Diagram illustrating the self-attention mechanism, as described in [87]. Images from [175].

Figure A.2 illustrates the architecture of the transformer model through two subfigures. In subfigure (a), the overall scheme of the transformer architecture is depicted, where individual layers are represented by bright-colored rectangles, signifying distinct components of the model. The gray rectangles represent layers that are repeated N times, indicating the modularity and depth of the architecture. Arrows connect these layers, demonstrating the flow of information throughout the network. Subfigure (b) provides a detailed view of the self-attention mechanism, highlighting how attention is computed over the query (Q), key (K), and value (V) vectors. This mechanism enables the model to weigh the importance of different input elements dynamically, facilitating effective information processing and enhancing the model's performance on various tasks.

1. Input Embeddings and Positional Encoding: The input sequence is tokenized into words or subwords and transformed into embedding vectors that capture each token's semantic meaning. As seen in Figure A.2a, this is the first step for the input. Since Transformers do not process data sequentially, Positional Encoding is added to these embeddings, using sine and cosine functions to convey information about each word's position. This ensures that the model retains information on word order, which is vital for sequence comprehension.

2. Self-Attention Mechanism: The self-attention mechanism enables the model to relate each word to every other word in the sequence, regardless of distance. For each word in the input sequence, the model generates three vectors: Query (Q), Key (K), and Value (V). These vectors are of the same dimensionality, and they are computed by linear transformations. This is, by multiplying the word embedding by three different weight matrices (which are learned during training).

   - Query (Q) represents the token being "attended to".
   - Key (K) represents all tokens that could be attended to by the current token.
   - Value (V) represents the content of each token in the sequence.

   The Transformer calculates self-attention by simultaneously processing all the Query (Q), Key (K), and Value (V) vectors in matrix form, making it efficient and enabling parallelization. Given an input sequence, the model creates matrices Q, K, and V, where each row in these matrices corresponds to a Query, Key, or Value vector for a token in the sequence.

   To compute self-attention over the entire sequence, the model calculates the dot product of Q with the transpose of K, giving a matrix of attention scores. Each element in this matrix reflects the relevance of one token to another across the sequence. These scores are then scaled by dividing by $\sqrt{d_k}$, where d_k is the dimensionality of the Query and Key vectors. Scaling helps stabilize the values, preventing large gradients that could hinder training.

   The softmax function is applied to these scaled scores, converting them into attention weights that sum to 1 along each row. These weights determine the influence each token should have on others in the sequence. Finally, the attention weights are used to compute a weighted sum of the V matrix, resulting in the final output matrix. An illustration of the full process can be seen in Figure A.2b.

   The entire self-attention process is then summarized by the following equation:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{A.9}$$

   This matrix formulation allows the Transformer to compute attention over all tokens at once, capturing global dependencies in the sequence efficiently.

   - Multi-Head Attention: Instead of relying on a single attention function, the Transformer uses multi-head attention to learn diverse relationships by applying multiple sets of Q, K, and V vectors. Each head captures different aspects of token relationships, enhancing the model's contextual understanding. Outputs from each head are concatenated and linearly transformed to form the final attention output for each layer.

3. Feed-Forward Neural Networks: After the attention layers, each token's output passes through a feed-forward neural network, consisting of two linear layers with a ReLU activation. This adds additional non-linear transformations that refine the representations for each token.

4. Residual Connections and Layer Normalization: Residual connections are added after each sub-layer (self-attention and feed-forward network) to retain original information and avoid vanishing gradients. Layer normalization stabilizes and accelerates training, ensuring convergence across the layers.

5. Stacking Layers: The Transformer encoder comprises several identical layers (typically 6-12), each reinforcing the self-attention and feed-forward transformations. Similarly, the decoder follows a layered structure.

The decoder (see right-hand side in Figure A.2a), which generates the final output sequence, mirrors many of the encoder's mechanisms but introduces key modifications:

- Masked Multi-Head Self-Attention: In the decoder, masked attention is applied to the output sequence, ensuring that each token only "sees" previous tokens or itself. This prevents future tokens from influencing current ones, which is crucial for autoregressive tasks where tokens are generated one at a time.

- Multi-Head Attention Over Encoder Outputs: The decoder also includes a second attention layer that attends to the encoder's output, allowing it to integrate contextual information from the input sequence. This interaction between the encoder and decoder enables the model to generate output tokens that align with the input sequence's context.

- Feed-Forward Network and Residual Connections: Each decoder layer has a feed-forward network and residual connections, mirroring the encoder. These layers refine the decoded representation and ensure stability throughout the sequence generation.

The Transformer architecture innovates sequence processing by replacing recurrence and convolution with self-attention. By processing all tokens in parallel and leveraging Q, K, and V vectors in its attention mechanism, the Transformer captures both short- and long-range dependencies efficiently. The multi-head attention enhances its ability to learn diverse patterns in the data, making it powerful for various sequence tasks. This has led to applications not only in NLP but also in computer vision, inspiring architectures like Vision Transformers (ViTs) that adapt these principles to image data.

A.3.1 SwinTransformer

To facilitate understanding, we will now explore the implementation of the Swin Transformer architecture, described in [89], step by step, using a concrete example. This detailed walkthrough highlights the functionality of each stage and how the architectural components interact.

Input Processing and Patch Splitting The Swin Transformer begins with an input image of dimensions $H \times W \times Z$, where H and W represent the height and width of the image, and Z denotes the number of channels (e.g., $Z = 3$ for RGB images). A commonly used input size for the Swin Transformer is $224 \times 224 \times 3$.

To prepare this image for processing:

- The image is divided into non-overlapping patches of size $P \times P$. For instance, if $P = 4$, each patch contains $P \times P \times Z = 4 \times 4 \times 3 = 48$ pixel values. These patches serve as the fundamental units of computation, referred to as "tokens." Figure A.3a illustrates how an image is partitioned into patches.

- Each patch is then projected into a C-dimensional feature space through a linear embedding layer. For example, with $C = 192$, the resulting sequence of features has dimensions:
$$\frac{H}{P} \times \frac{W}{P} \times C = \frac{224}{4} \times \frac{224}{4} \times 192 = 56 \times 56 \times 192.$$
This embedded feature map is the input to the first stage of the Swin Transformer, as shown in the initial step of Figure A.3b.

Stage 1: Local Attention with Windows In the first stage of the Swin Transformer, the 56×56 grid of tokens is divided into non-overlapping windows, each of size $M \times M$. An illustration of the window scheme can be seen in the left hand side of Figure A.3a. In our example, for $M = 7$, the feature map is partitioned into:

$$\frac{56}{7} \times \frac{56}{7} = 8 \times 8 = 64 \,\text{windows}.$$

Each window contains $M \times M = 7 \times 7 = 49$ tokens, significantly reducing the computational burden of self-attention compared to a global approach, which would need to consider all 3136 tokens simultaneously.

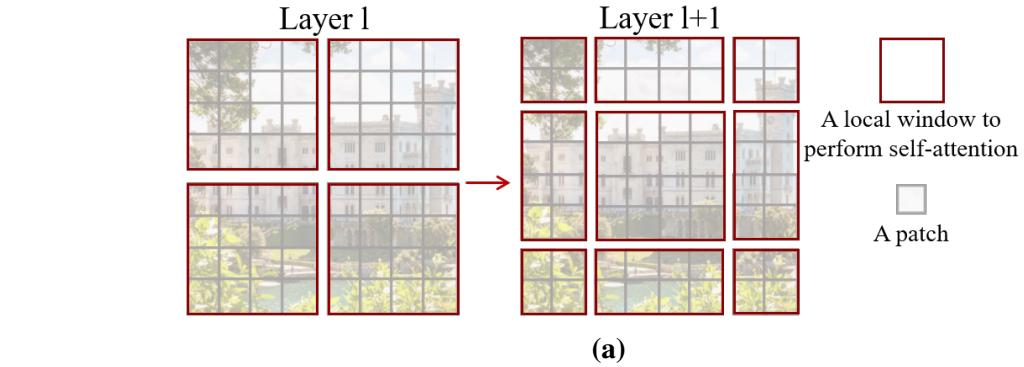
Within each window, the following operations are performed (see left hand side of Figure A.3c):

1. Self-Attention: Self-attention is computed exclusively among the 49 tokens within each window. This localized attention reduces computational complexity from quadratic in the number of global tokens to quadratic in the number of tokens per window.

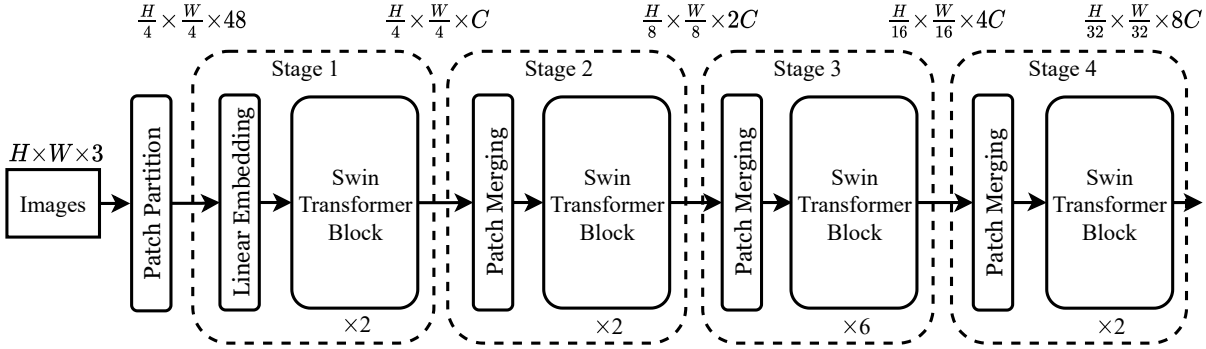
2. Query, Key, and Value Calculation: The tokens in each window are mapped into query, key, and value vectors through learned linear projections. Using these representations, the self-attention mechanism computes attention weights that determine how much each token contributes to updating others within the same window. The tokens are then updated through a weighted aggregation of the value vectors.

3. Layer Normalization and Feedforward Network (FFN): After the self-attention step, each token undergoes further processing through a feedforward network (FFN) to enhance its representation. Residual connections and layer normalization are applied to stabilize training and maintain the integrity of token features.

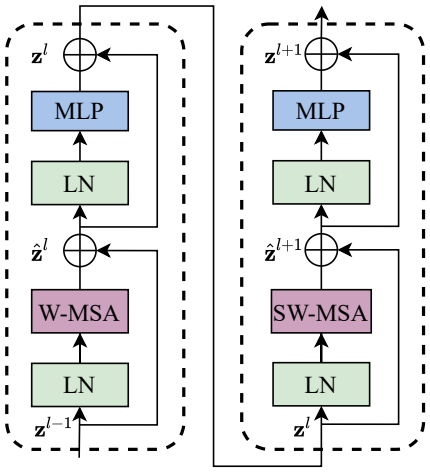
The output of this stage retains the same dimensions as the input: $56 \times 56 \times 192$, ensuring no loss of resolution while enabling more expressive local feature representations.



(a)



(b)



(c)

Figure A.3: Illustration of the Swin Transformer architecture. (a) The shifted window mechanism, which alternates between standard and shifted window partitions to enable cross-window connections. (b) The overall architecture of the Swin Transformer, showcasing the hierarchical encoder design with patch merging for progressive downsampling. (c) The Swin Transformer block, detailing the main components, including window-based multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA). Images from [89].

Shifted Window Mechanism While local window-based attention is computationally efficient, it has a significant limitation: tokens can only interact with others within the same window, restricting the model's ability to capture dependencies across windows. The shifted window mechanism overcomes this limitation by introducing overlap between neighboring windows in successive layers.

- Shift and Re-partitioning:
  - In the subsequent layer, the windows are shifted by half their size (e.g., $M/2 = 3$ for $M = 7$). This shift realigns the tokens such that some from neighboring windows are grouped together within the same shifted window. This is illustrated in Figure A.3a.
  - After shifting, the feature map is re-partitioned into a new set of windows, ensuring that tokens from adjacent regions can now interact. This step effectively connects the previously isolated regions without requiring global attention.
- Self-Attention in Shifted Windows: Self-attention is recomputed within these newly formed shifted windows. By alternating between regular and shifted window partitions, the Swin Transformer captures both fine-grained local details and broader cross-region dependencies. This ensures that tokens can progressively aggregate information from a larger receptive field across layers. The overall architecture illustrating this process is depicted in the right hand side of Figure A.3c.
- Residual Connections: To maintain consistency and facilitate effective gradient flow during training, the outputs of the shifted attention layer are combined with those of the preceding layer using residual connections. This design helps stabilize the learning process and preserve the integrity of learned features.

By combining efficiency and contextual richness, the shifted window mechanism enables the Swin Transformer to effectively process high-resolution images without the prohibitive cost of global self-attention.

Patch Merging and Hierarchical Representation After completing self-attention across regular and shifted windows, the hierarchical processing begins. At the end of each stage, a patch merging layer reduces the spatial resolution while increasing the feature dimensionality:

- Patch Concatenation: Features from 2×2 neighboring patches are concatenated, resulting in a vector of size $4C$.
- Linear Projection: A linear transformation reduces the $4C$-dimensional concatenated vector to $2C$, effectively increasing feature richness while reducing spatial resolution.

For example, at the end of Stage 1 (see Figure A.3b):

$$\text{New resolution: } \frac{56}{2} \times \frac{56}{2} = 28 \times 28, \qquad \text{Feature dimension: } 2 \cdot 192 = 384.$$

This process is repeated across four stages, progressively reducing the resolution and increasing the feature dimensions:

- **Stage 2**: 28×28 resolution, feature dimension 384.
- **Stage 3**: 14×14 resolution, feature dimension 768.
- **Stage 4**: 7×7 resolution, feature dimension 1536.

Final Feature Maps and Applications At the end of Stage 4, the feature map has dimensions $7 \times 7 \times 1536$. At this point, the feature map size matches the window size ($M = 7$), meaning that local attention now spans the entire feature map. For this reason, further downsampling is unnecessary. This rich representation can be adapted to various vision tasks:

- Image Classification: A global average pooling layer aggregates the features into a single vector, which is passed through a fully connected layer for classification.
- Object Detection and Segmentation: The multi-scale feature maps from all stages are used as inputs to downstream detection and segmentation frameworks, leveraging their hierarchical structure.

  In the case of semantic segmentation, the final feature map can be upsampled to match the original image resolution. This is typically done using a deconvolutional or upsampling layer, which helps recover fine-grained details lost during downsampling. The upsampled feature map is then passed through a segmentation head, which classifies each pixel into a predefined category.

A.3.2 SegFormer

The Segformer architecture represents a significant advancement in image segmentation tasks, combining the strengths of transformers with efficient semantic segmentation strategies. At its core, Segformer utilizes a lightweight design that leverages hierarchical feature extraction while maintaining computational efficiency. A key innovation in this architecture is the overlapping patch merging technique, which allows for the effective aggregation of features from neighboring patches. This approach not only enhances the model's ability to capture fine details but also reduces the risk of losing crucial contextual information during the segmentation process.

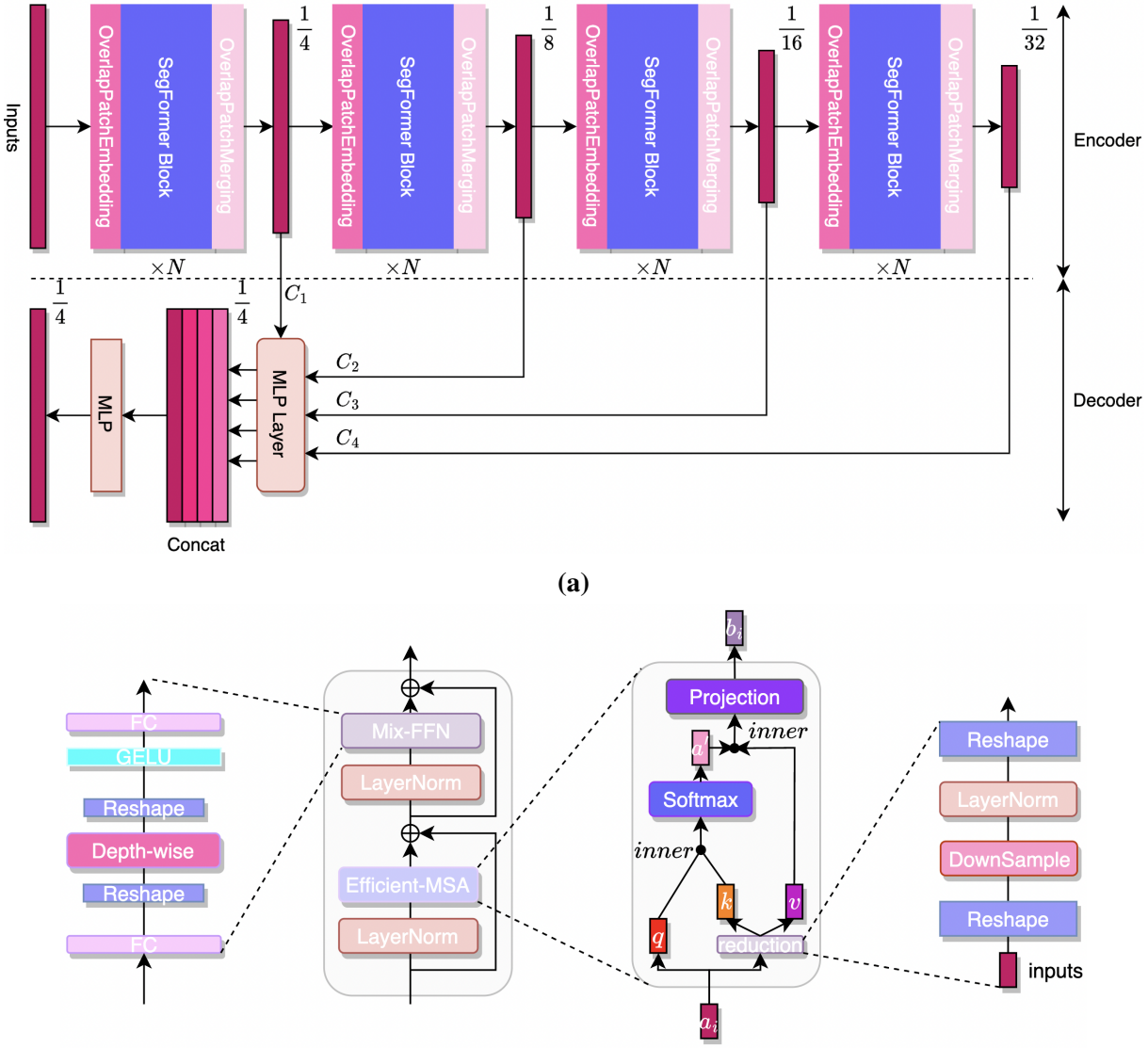
The overlapping patch merging mechanism operates by defining three key parameters: the kernel size (K), stride (S), and padding (P). These parameters determine the size of each patch, the step between patches, and the boundary padding, respectively. For example, using $K = 7$, $S = 4$, and $P = 3$ or $K = 3$, $S = 2$, and $P = 1$, SegFormer achieves overlapping patches while maintaining an effective reduction in spatial resolution [92]. This approach allows the model to process hierarchical feature maps, progressively reducing resolution while enriching feature complexity, such as:

$$F_1 : \frac{H}{4} \times \frac{W}{4} \times C_1 \quad \rightarrow \quad F_2 : \frac{H}{8} \times \frac{W}{8} \times C_2,$$

where F_1 and F_2 denote the feature maps at successive stages.

This overlapping patch merging not only improves the model's capacity to capture fine details but also provides flexibility through tunable K, S, and P parameters.

Figure A.4 illustrates the architecture of SegFormer. In subfigure (a), we see the complete model structure, where each module consists of three main components: the patch embedding



(a)

(b)

Figure A.4: Illustration of the SegFormer architecture. (a) The overall architecture of SegFormer, including a hierarchical transformer encoder and a lightweight All-MLP decoder. The encoder generates multi-scale features, while the decoder fuses them to produce the final segmentation mask. (b) Detailed structure of the SegFormer block, highlighting the Efficient Multi-Head Self-Attention (Efficient MSA) mechanism and the Mix-FFN module. Image from GitHub repository. Source: https://github.com/ACSEkevin/An-Overview-of-Segformer-and-Details-Description.

module, the SegFormer block, and the overlapping patch merging module. The outputs from the SegFormer blocks at different stages are then fed into an MLP-based decoder, which processes multi-scale features for segmentation. The arrows indicate the flow of information between different components, showcasing the efficient fusion of hierarchical representations. Full details of each component are explained in the following text.

Segformer Block The SegFormer block is the core component of the architecture, designed to efficiently capture global and local dependencies in an image. It consists of a lightweight self-attention mechanism, which enables effective feature extraction without the need for positional embeddings, making it scale-invariant. Additionally, a mix-feedforward network (Mix-FFN) enhances feature representation by applying depthwise separable convolutions, improving computational efficiency. This combination allows the SegFormer block to maintain a strong balance between accuracy and efficiency in semantic segmentation tasks. The details of its composition will be explained in the following text and are illustrated in Figure A.4b.

- Efficient Self-Attention: To reduce the quadratic complexity of traditional self-attention ($O(N^2)$), SegFormer employs a *reduction ratio* R. The key matrix K is reshaped from $N \times C$ to $\frac{N}{R} \times (C \cdot R)$, and then linearly projected back to $\frac{N}{R} \times C$. This reduces the complexity of the self-attention mechanism to $O\left(\frac{N^2}{R}\right)$, making it scalable for large images.
- Mix-FFN. Implicit Positional Encoding: The Mix-FFN replaces traditional positional encoding to ensure flexibility across different input resolutions while effectively incorporating positional information. Unlike positional encodings, which may require interpolation when the input resolution varies, the Mix-FFN integrates positional cues directly into the feed-forward network (FFN). This is achieved through the following components:

  1. 3×3 Depthwise Convolution: Positional information is encoded implicitly using a 3×3 depthwise convolution, which operates on spatially local regions of the input. Depthwise convolutions are used to reduce the number of parameters and improve efficiency.
  2. Feed-Forward Network: After the convolution, the features are processed through a standard MLP (multi-layer perceptron), which applies non-linearity and further refines the representation.

  The Mix-FFN can be mathematically formulated as Equation A.10 [92].

$$x_{\text{out}} = \text{MLP}(\text{GELU}(\text{Conv}_{3\times3}(\text{MLP}(x_{\text{in}})))) + x_{\text{in}}, \tag{A.10}$$

  where x_{in} is the feature from the preceding self-attention module. The use of a 3×3 convolution ensures that the model captures local positional dependencies while avoiding the inaccuracies associated with interpolated positional encodings. The Gaussian Error Linear Unit (GELU) is a smooth activation function defined as stated in Equation A.11 [176].

$$\text{GELU}(x) = x \cdot \Phi(x), \tag{A.11}$$

  where $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal distribution, $\Phi(x) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$, and $\text{erf}(x)$ is the error function. Intuitively,

GELU selectively retains positive inputs and attenuates negative ones, offering a smooth transition around zero. This property often improves model convergence compared to ReLU or other activation functions [176].

- Residual Connections and Layer Normalization: Each transformer block includes residual connections and layer normalization to stabilize training and improve gradient flow.

MLP decoder The SegFormer decoder is designed to fuse the hierarchical, multi-scale features generated by the encoder into a semantic segmentation mask. Unlike traditional decoders that rely on computationally intensive operations such as deconvolution or complex upsampling modules, SegFormer employs a lightweight All-MLP approach, making it efficient and scalable [92].

The decoder takes as input the multi-level features extracted by the encoder. Specifically, the hierarchical encoder provides features at four different resolutions: F_1, F_2, F_3, F_4, where:

$$F_i \in \mathbb{R}^{H_i \times W_i \times C_i}, \quad i = 1, \ldots, 4.$$

These features represent a spectrum of information, ranging from high-resolution global context (F_1) to low-resolution fine details (F_4).

Since the input features have different resolutions, the decoder first aligns them by upsampling all feature maps to a unified spatial size $H_1 \times W_1 = \frac{H}{4} \times \frac{W}{4}$, which corresponds to the resolution of F_1. This process includes:

- Linear Projections: Each feature F_i is passed through a linear layer to normalize the channel dimensions across all levels to a consistent size C, where C is the decoder's feature dimensionality.
- Upsampling: Lower-resolution feature maps (F_2, F_3, F_4) are upsampled to match the spatial resolution of F_1. After this step, all features have dimensions:

$$\hat{F}_i \in \mathbb{R}^{H_1 \times W_1 \times C}.$$

Once the features are aligned, the decoder fuses them using a lightweight multi-layer perceptron (MLP) structure:

- Concatenation: The upsampled features $\{\hat{F}_1, \hat{F}_2, \hat{F}_3, \hat{F}_4\}$ are concatenated along the channel dimension, forming a combined feature tensor:

$$F_{\text{concat}} \in \mathbb{R}^{H_1 \times W_1 \times (4 \cdot C)}.$$

- MLP Layers: This concatenated feature tensor is passed through a series of MLP layers, which learn to combine the multi-level features effectively. These layers are computationally efficient, avoiding the need for heavy operations like deconvolution.
- Downsampled Segmentation Mask: The output of the MLP layers is passed through a linear projection layer to produce the intermediate segmentation map:

$$M_{\text{seg}} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times K},$$

where K is the number of target classes. This lower-resolution mask provides the class probabilities for each patch at the $H/4 \times W/4$ scale.

To produce the final semantic segmentation output at the original image resolution $H \times W$:

- The intermediate segmentation map M_{seg} is upsampled back to $H \times W$ using bilinear interpolation.
- This operation ensures the decoder outputs a high-resolution segmentation mask:

$$M_{\text{final}} \in \mathbb{R}^{H \times W \times K}.$$

The simplicity of the All-MLP decoder ensures that SegFormer remains lightweight and computationally efficient, even with the additional step of upsampling the segmentation mask. This design choice strikes a balance between accuracy and computational cost, making SegFormer highly effective for semantic segmentation in resource-constrained environments. Figure A.4a illustrates the overall interaction between the encoder and decoder, highlighting their efficient and modular design.

A.3.3 SwinIR

In this section, we focus on three key components of the SwinIR architecture.

Figure A.5 illustrates the SwinIR architecture, which consists of three main components: shallow feature extraction, deep feature extraction, and image reconstruction. The deep feature extraction module (shown in subfigure (a)) is composed of multiple Swin Transformer layers (detailed in subfigure (b)) and a residual connection, which helps stabilize training and enhance feature representation.

Shallow Feature Extraction The shallow feature extraction phase serves as a preparatory step by transforming the low-quality (LQ) input image $I_{\text{LQ}} \in \mathbb{R}^{H \times W \times C_{\text{in}}}$ into a feature representation with C channels. This is achieved through a convolutional layer H_{SF} with a kernel size of 3×3:

$$F_0 = H_{\text{SF}}(I_{\text{LQ}}),$$

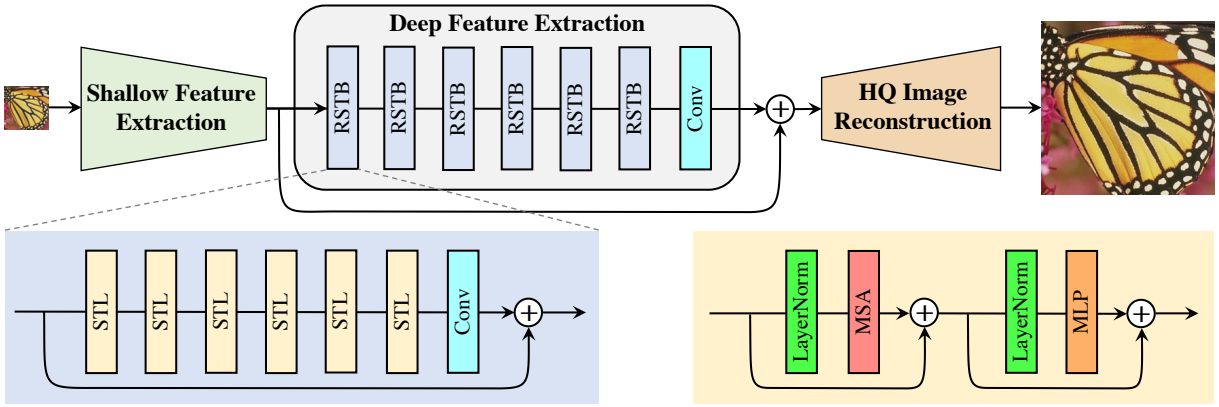
where F_0 is the shallow feature map. Including an early convolutional layer stabilizes training and facilitates faster convergence by introducing inductive biases from convolution operations.

Deep Feature Extraction The deep feature extraction module refines the shallow features F_0 to produce deeper feature representations $F_{\text{DF}} \in \mathbb{R}^{H \times W \times C}$:

$$F_{\text{DF}} = H_{\text{DF}}(F_0),$$

where $H_{\text{DF}}(\cdot)$ represents the deep feature extraction module. This module comprises K Residual Swin Transformer Blocks (RSTBs) and concludes with a 3×3 convolutional layer (see Figure A.5.(a)).

Each RSTB combines the Swin Transformer's efficient self-attention mechanism (see Figure A.5.(b)) with residual connections to enhance feature representation while maintaining computational efficiency (see Section 2.3.2.2). The concluding convolutional layer reintroduces convolutional biases, preparing the features for reconstruction.



(a) Residual Swin Transformer Block (RSTB)

(b) Swin Transformer Layer (STL)

Figure A.5: Illustration of the SwinIR architecture components. (a) Residual Swin Transformer Blocks (RSTB), which form the backbone of the deep feature extraction module, leveraging hierarchical feature representation and self-attention. (b) Swin Transformer Layer (STL), detailing the internal structure including the shifted window mechanism and Mix-FFN. Images from [93].

Image Reconstruction The reconstruction phase synthesizes the high-quality output image I_{RHQ} by aggregating shallow and deep features:

$$I_{\text{RHQ}} = H_{\text{REC}}(F_0 + F_{\text{DF}}),$$

where $H_{\text{REC}}(\cdot)$ denotes the reconstruction module.

Shallow features F_0 primarily retain low-frequency information, while deep features F_{DF} recover high-frequency details. The long skip connection between F_0 and F_{DF} ensures that low-frequency information bypasses the deep feature extraction module, enabling the latter to focus on recovering high-frequency content. This design also stabilizes training by reducing the burden on the deep feature extraction process.

For tasks that do not require upsampling, such as image denoising or JPEG compression artifact reduction, a single convolution layer suffices for the reconstruction step.

A.4 Mamba

In the following subsections, we first introduce the fundamentals of State Space Models, followed by a detailed explanation of the Mamba layer and its role in improving efficiency and expressiveness in deep learning architectures.

A.4.1 State-Space Models: The Foundation of Mamba

State-Space Models (SSMs) are linear, recurrent mechanisms used to process sequential data. They share structural similarities with Recurrent Neural Networks (RNNs) but are uniquely characterized by their reliance on linear transformations. SSMs model sequence dynamics by iteratively updating a hidden state representation using four primary matrices: Δ, A, B, and C. Each matrix serves a distinct role in transforming inputs and propagating

information through the sequence. The workflow of SSMs is illustrated in Figure A.6 and involves the following sequential steps [94]:

1. Discretization Step: The process begins with the parameter Δ, which modifies the entries of matrices A and B. This step is crucial for preparing the SSM to process input data.
   - A is updated to $\bar{A}$, dictating how much of the hidden state propagates forward.
   - B is updated to $\bar{B}$, determining how much input contributes to the hidden state.
   - The modification equations for A and B (Equation A.12 and Equation A.13) are parameterized by Δ, which is learned during model training [94].

$$\bar{A} = e^{\Delta A}, \tag{A.12}$$

$$\bar{B} = (\Delta A)^{-1}(e^{\Delta A} - I)\Delta B. \tag{A.13}$$

2. Linear RNN Step: After the discretization, the updated matrices $\bar{A}$ and $\bar{B}$ are used to process input tokens sequentially. The hidden state for token t, h_t, is computed by combining:
   - A transformation of the previous token's hidden state ($\bar{A} \cdot h_{t-1}$).
   - A transformation of the current input embedding ($\bar{B} \cdot x_t$).
   - These components are summed to form the hidden state for the current token (see Equation A.14).

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t. \tag{A.14}$$

   This mechanism enables SSMs to maintain a dynamic representation of the sequence.

3. Output Generation: To produce meaningful outputs, the hidden state is transformed using matrix C, which maps the internal representation to an output space (Equation A.15). This final representation can be used for tasks such as classification or prediction.

$$y_t = Ch_t. \tag{A.15}$$

A.4.2 Mamba Layer

We will now provide a detailed explanation of the complete architecture of a Mamba layer, which is comprised of the following fundamental components (see illustration in Figure A.7).

1. Dimensionality Expansion: A linear layer doubles the input token embedding dimensionality (e.g., from 64 to 128), allowing more expressive transformations.

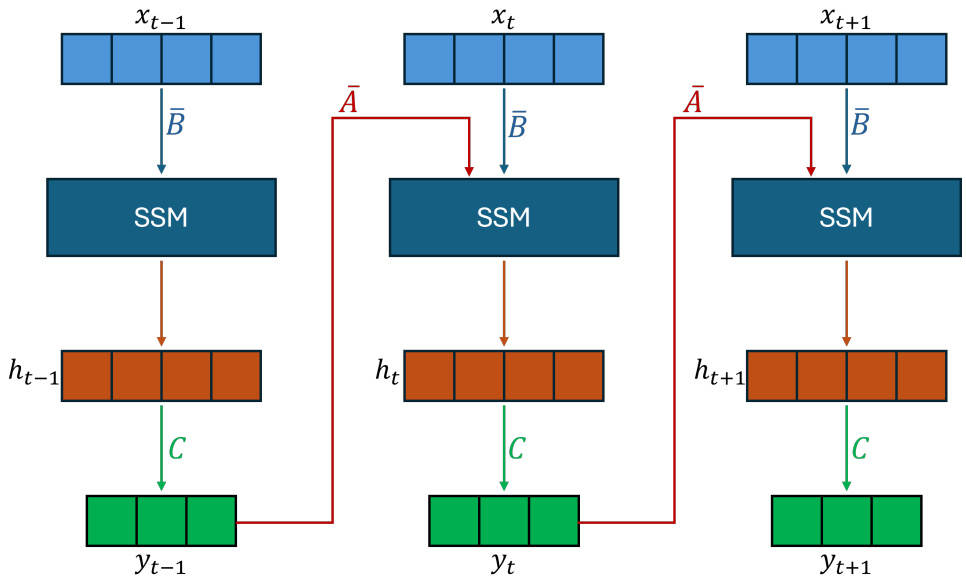


Figure A.6: State-Space Model (SSM) architecture, where h_i represents the hidden states, x_i the input tokens, and y_i the output. The matrices $\bar{A}$, $\bar{B}$, and C are the transformation matrices applied to the input tokens during processing, enabling efficient sequence modeling and computation of the hidden states and outputs.

2. 1D Convolution: A convolution layer rearranges information across dimensions using the SiLU activation function.

   SiLU (Sigmoid-weighted Linear Unit), also known as Swish, is an activation function increasingly utilized in deep learning for its smooth and non-monotonic properties. It is defined as:

   $$\sigma(x) = x \cdot sigmoid(x), \tag{A.16}$$

   where x represents the input, and the sigmoid function scales the output between 0 and 1. Unlike ReLU, SiLU is differentiable across its entire domain, facilitating smoother gradient propagation during backpropagation.

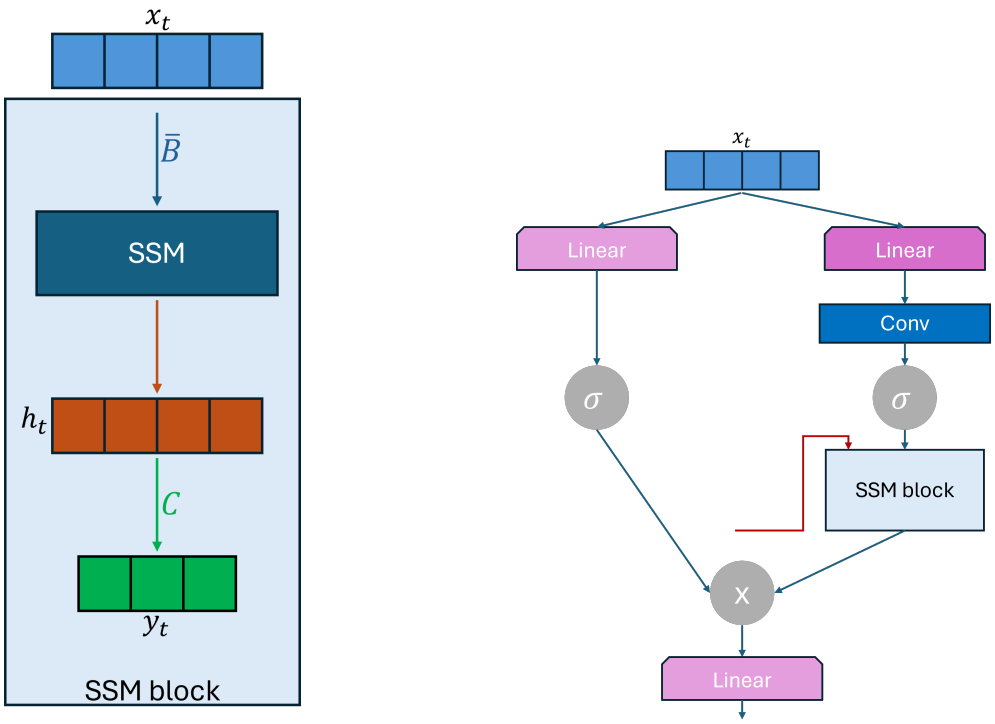
3. Selective SSM: Processes the output sequentially, incorporating token-specific matrices to focus on relevant information.

4. Gated Multiplication: The output of the selective SSM is multiplied with a transformed version of the input, capturing similarities between current tokens and prior context.

5. Dimensionality Reduction: A final linear layer reduces the dimensionality back to the original size.

These layers are stacked to form the complete Mamba model, mirroring the simplicity of Transformer architectures while avoiding their inefficiencies with long sequences.

Mamba bridges the gap between RNNs and Transformers, leveraging the efficiency of SSMs while addressing their rigidity through selective enhancements. Its design heralds the possibility of SSMs becoming a viable alternative to Transformers in large-scale sequence modeling, especially in applications where memory and speed are critical.



(a) **(b)**

Figure A.7: (a) Selective State-Space Model (SSM) block, illustrating the token-specific transformations and matrix operations that allow for flexible processing of inputs. (b) The complete Mamba architecture, showcasing the integration of the selective SSM with additional components, such as dimensionality expansion or reduction (Linear), convolutional layers (Conv), and gated multiplications (X), and the SiLU activation function (σ) to enable efficient sequence modeling and enhanced performance.

Appendix B

Supplementary Material for Calcium Segmentation in CTA

B.1 Threshold Adjustment for Lumen Segmentation Noise Reduction

The addition of +10HU was implemented to eliminate small segments of the lumen that do not correspond to calcium and may suggest that the threshold needs to be increased. This value was chosen because it represents the minimum HU difference at which we observe a significant reduction in these small lumen segments, without affecting the calcium volume in a meaningful way. This approach was not arbitrarily selected but was based on visual inspection and empirical observation. An example is shown in Figure B.1, where the calcium segmentation using the original threshold (524 HU) is depicted in blue, and the segmentation with a +10 HU increased threshold (534 HU) is shown in red. As seen, the red threshold results in fewer lumen artifacts, indicating that it is a more appropriate threshold for calcium segmentation.

Figure B.1: Comparison of calcium segmentation using two different thresholds: 524 HU (blue) and 534 HU (red). The increased threshold (534HU) reduces small lumen segments that do not correspond to calcium, as demonstrated by the reduced number of lumen artifacts in the red segmentation. This adjustment of +10HU helps eliminate minor noise without significantly altering the calcium volume, ensuring more accurate calcium segmentation.

B.2 Calcium Scoring Methods Comparison

The comparative Table B.1 provides an in-depth analysis of the two calcium scoring methods used in this study: the Agatston method and the volume-based method. This table outlines their key characteristics, such as the underlying principles, computational requirements, and application in clinical practice. Additionally, it highlights the advantages and disadvantages of each method. For instance, the Agatston method is widely accepted in clinical settings and provides a standard measure for assessing calcium burden, though it may be sensitive to slice thickness. On the other hand, the volume-based method offers a more direct and continuous measurement of calcium load, which is less sensitive to resolution changes but may lack the same level of clinical validation. This comparison allows for a clearer understanding of the trade-offs between the methods and aids in evaluating their respective suitability for different clinical applications.

Table B.1: Comparison of the Agatston and Volume Methods for Calcium Scoring. The table outlines the principal characteristics, advantages, and disadvantages of each method, focusing on their computational requirements, clinical applicability, and sensitivity to image resolution and slice thickness.

Criteria	Agatston Score	Calcium Volume
Input Requirements	- Non-contrast CT - Segmented calcium regions - Density threshold (>130 HU) - Weighted by peak HU value	- Segmented calcium regions - Voxel dimensions
Advantages	- Standardized in clinical guidelines - Widely used in clinical practice - Provides risk stratification with established thresholds	- Directly related to physical properties - Simple and reproducible computation
Limitations	- Dependent on image acquisition parameters - Influenced by partial volume effects - Not directly related to physical properties - Sensitive to variations in slice thickness and reconstruction kernel	- Does not account for calcium density variations - Lacks widely accepted clinical thresholds - Limited validation in routine clinical practice

B.3 Scoring method comparison

B.3.1 Agatston scoring method

Figure B.2 shows the results of the calcium score calculation for CTA images with the original slice spacing.

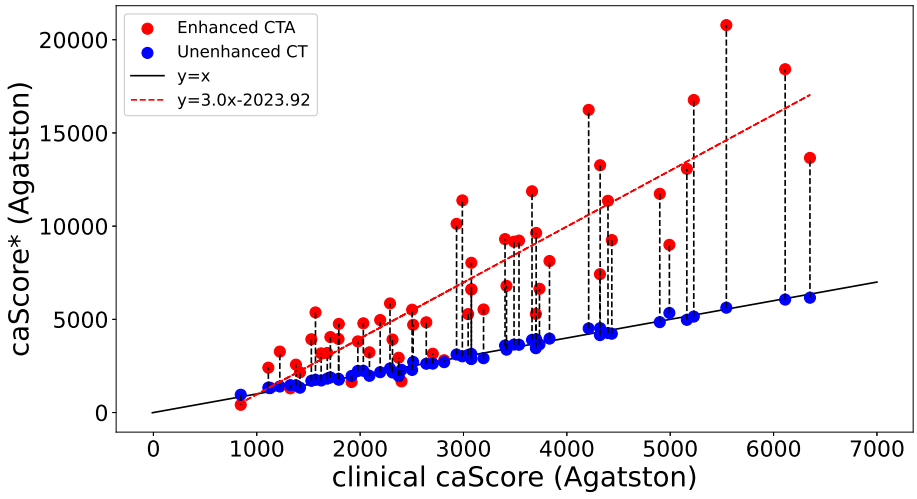
In Figure B.2.(a), each clinical segmentation (with CT images) on the x-axis is also associated with a calcium score value (in y-axis) of the CTA images (red dot), calculated using a patient-dependent threshold. Pearson's correlation between the calcium score measured on CTA and CT images has a r^2 value of 0.87 (p-value: $2.49e^{-18}$). The mean absolute error is 3982.5 Agatston units.

Bland-Altman plot of the red dots of Figure B.2.(a) can be seen in Figure B.2.(b). A clear pattern is observed: the higher the calcium score value, the greater the discrepancy between the two values. This shows that the Agatston method is not reproducible in contrast and non-contrast images for calcium score values higher than 1000 Agatston units.

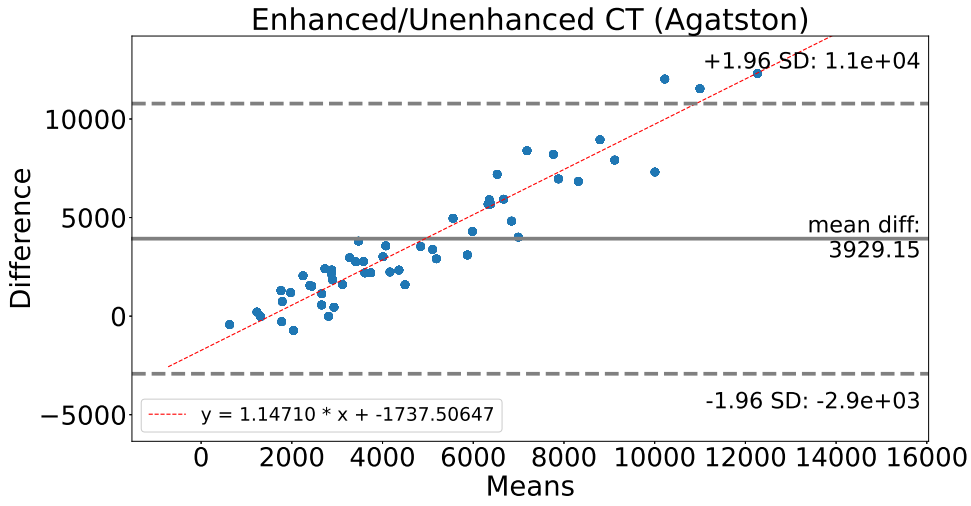
B.3.2 Volume scoring method

Analogous to that used for the Agatston method, the volume is calculated in images without and with contrast with the original slice spacing. Figure B.3.(a) shows the calculated volume values for images without contrast on the x-axis, and the volume for images with contrast on the y-axis. In this case the correlation value is $r^2 = 0.89$ (P-value: $3.35e^{-20}$). Mean absolute error of 829.33HU.

The Bland Altman plot is shown in Figure B.3.(b). On this occasion there is no clear trend in the data, unlike the Agatston method, which implies that the volume method is more stable in terms of calcium abundance.

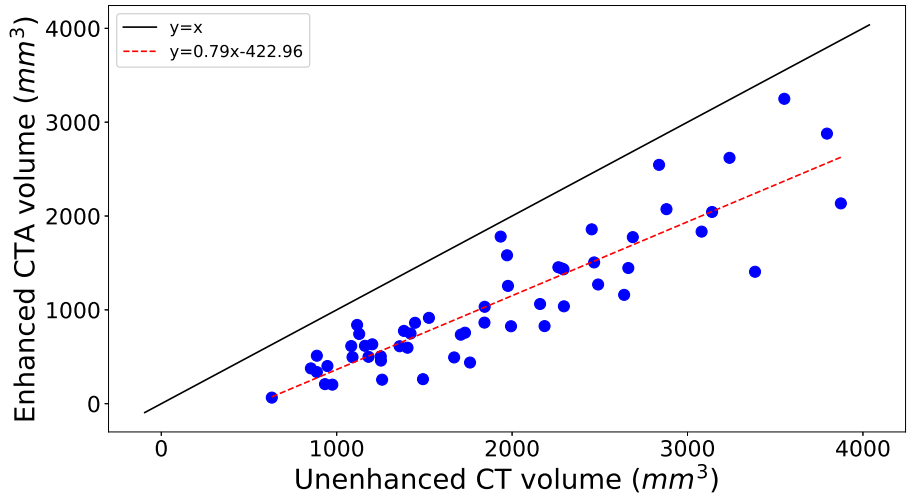


(a)

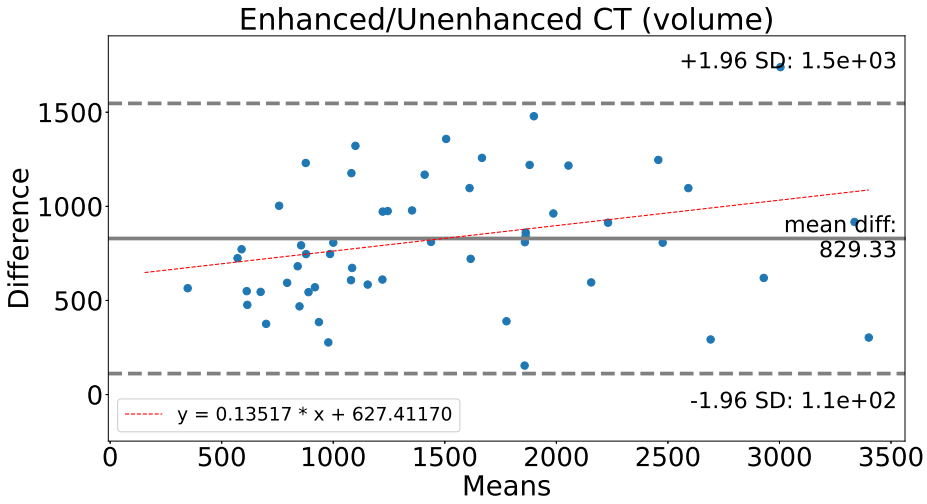


(b)

Figure B.2: (a) Scatter plot with the results of the calcium score for the Agatston method in the aortic leaflets. The x-axis represents the value obtained in clinical practice on the unenhanced images. The y-axis indicates the Agatston calcium score for unenhanced images (with threshold of 130HU) obtained by a segmentation expert (blue dots) and enhanced images with patient-dependent threshold (red dots). (b) Bland-Altman plot comparing the Agatston calcium score obtained by clinicians, using the unenhanced images, and by a segmentation expert using the enhanced images and the Agatston method with their respective threshold.



(a)



(b)

Figure B.3: (a) Scatter plot with the results of the volume score method in the aortic valve. The x-axis represents the value obtained on the unenhanced images. The y-axis indicates the volume calcium score for enhanced images. (b) Bland-Altman plot comparing the volume calcium score, using the unenhanced images and the enhanced images.

Abbreviations

List of abbreviations.

Abbreviation	Description
AA	Abdominal aorta
AI	Artificial intelligence
AVA	Aortic Valve Area
AVS	Aortic Valve Stenosis
B, H, W, C	Batch, Height, Width, Channel
CAC	Coronary Artery Calcification
CAD	Coronary Artery Disease
CAI	Coronary Artery Imaging
CCTA	Coronary Computed Tomography Angiography
CE	Cross Entropy
CMR	Cardiac Magnetic Resonance
CNN	Convolutional Neural Network
CT	Computed Tomography
CTA	Computed Tomography Angiography
CVD	Cardiovascular Disease
dB	Decibel
EACVI	European Association of Cardiovascular Imaging
ECG/EKG	Electrocardiogram
ESC	European Society of Cardiology
FFL	Focal Frequency Loss
FFN	Feed Forward Network
FFR	Fractional Flow Reserve
FP	False Positive
FN	False Negative
GD	Generalized Dice
GDF	Generalized Dice focal
GELU	Gaussian Error Linear Unit

Continued on next page

(Continued from previous page)

Abbreviation	Description
GT	Ground Truth
HU	Hounsfield Units
ICA	Invasive Coronary Angiography
IMRT	Intensity Modulated Radiotherapy
IoU	Intersection Over Union
kVCT	Kilovoltage Computed Tomography
LAD	Left Anterior Descending Artery
LCT	Left Coronary Tree
LQ	Low Quality
MAC	Multiply–Accumulate
MAR	Metal Artifact Reduction
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
MS-SSIM	Multi-Scale Structural Similarity Index
MSE	Mean Squared Error
MVCT	Megavoltage Computed Tomography
PCI	Percutaneous Coronary Intervention
PSNR	Peak Signal-to-Noise Ratio
RCT	Right Coronary Tree
ReLU	Rectified Linear Unit
ResNet	Residual Neural Network
RNN	Recurrent Neural Network
RSTB	Residual Swin Transformer Blocks
SAVR	Surgical Aortic Valve Replacement
SE	Stress Echocardiography
SiLU	Sigmoid-Weighted Linear Unit
SSIM	Structural Similarity Index
SSM	State Space Model
Swin	Shifted Window
TAVI	Transcatheter Aortic Valve Implantation
TA	Thoracic >orta
TEE	Transesophageal Echocardiography
TN	True negative
TP	True positive
TTE	Transthoracic Echocardiography

Continued on next page

(Continued from previous page)

Abbreviation	Description
VGG	Visual Geometry Group Network
ViT	Vision Transformer
XD	X-dimensional

List of figures

List of tables

List of publications

(I) Serrano-Antón, B., Rehman, M., Martinel, N., Avanzo, M., Spizzo, R., Fanetti, G., P. Muñuzuri, A., Micheloni, C.. MAR-DTN: Metal Artifact Reduction Using Domain Transformation Network for Radiotherapy Planning. In Antonacopoulos, A., Chaudhuri, S., Chellappa, R., Liu, CL., Bhattacharya, S., Pal, U. (eds) Pattern Recognition. ICPR 2024. Lecture Notes in Computer Science, vol 15311, pp. 143-159, 2025. Springer. Electronic ISSN: 1611-3349. https://doi.org/10.1007/978-3-031-78195-7_10.

Specific contribution in the publication

Generating the dataset and automating the process, implementing the models, analyzing the results, creating figures, and writing the original manuscript.

Quality indexes 2023

Impact factor: 0.606
CiteScore: 2.6
Quartile: Q2 (99/344 in COMPUTER SCIENCE (MISCELLANEOUS))

(II) Serrano-Antón, B., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., Pérez-Muñuzuri, V., González-Juanatey, J.R., P. Muñuzuri, A.. Coronary Artery Segmentation Based on Transfer Learning and UNet Architecture on Computed Tomography Coronary Angiography Images. In IEEE Access, vol. 11, pp. 75484-75496, 2023. Electronic ISSN: 2169-3536. DOI: www.doi.org/10.1109/ACCESS.2023.3293090.

Specific contribution in the publication

Collecting medical images, manually annotating the images, generating the dataset and automating the process, implementing the models, analyzing the results, creating figures, and writing the original manuscript.

Quality indexes 2023

Impact factor: 3.4
CiteScore: 9.8
Quartile: Q2 (87/250 in COMPUTER SCIENCE, INFORMATION SYSTEMS)

(III) Serrano-Antón, B., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., Massonis, G., Pendón, S., Pérez-Muñuzuri, V., González-Juanatey, J.R., P. Muñuzuri, A.. Optimal Coronary Artery Segmentation Based on Transfer Learning and UNet Architecture. In Wachinger, C., Paniagua, B.,

Elhabian, S., Li, J., Egger, J. (eds) Shape in Medical Imaging. ShapeMI 2023. Lecture Notes in Computer Science, vol 14350, pp. 55-64, 2023. Springer. Electronic ISSN: 1611-3349. DOI: https://doi.org/10.1007/978-3-031-46914-5_5.

Specific contribution in the publication

Generating the dataset, analyzing the results, creating figures, and writing the original manuscript.

Quality indexes 2023

Impact factor: 0.606
CiteScore: 2.6
Quartile: Q2 (99/344 in COMPUTER SCIENCE (MISCELLANEOUS))

(IV) Serrano-Antón, B., Insúa Villa, M., Pendón Minguillón, S., Paramés-Estévez, S., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., González-Juanatey, J.R., P. Muñuzuri, A.. Unsupervised clustering based coronary artery segmentation. In BioData Mining, vol. 18, no. 1, pp. 1-23, 2025. BioMed Central. Electronic ISSN: 1756-0381. DOI: https://doi.org/10.1186/s13040-025-00435-y.

Specific contribution in the publication

Collecting medical images, generating the dataset and automating the process, implementing the models, analyzing the results, creating figures, and writing the original manuscript.

Quality indexes 2023

Impact factor: 4.0
CiteScore: 7.9
Quartile: Q1 (8/66 in MATHEMATICAL & COMPUTATIONAL BIOLOGY)

(V) Serrano-Antón, B., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., González-Juanatey, J.R., P. Muñuzuri, A.. Contrast-enhanced computed tomography to measure Aortic calcium volume score. Submitted 2025.

Specific contribution in the publication

Collecting medical images, generating the dataset and automating the process, analyzing the results, creating figures, and writing the original manuscript.

(VI) Serrano-Antón, B., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., González-Juanatey, J.R., P. Muñuzuri, A.. Comparison of Deep Learning Architectures for Calcium Segmentation in Aortic Regions and Whole-Aorta in Contrast-enhanced computed tomography. Submitted 2025.

Specific contribution in the publication

Collecting medical images, manually annotating the images, generating the dataset and automating the process, implementing the models, analyzing the results, creating figures, and writing the original manuscript.

Copyright permissions

Figure 1.1. Image has been cropped from the original and is used under the Creative Commons Attribution 3.0 Unported license (OpenStax College, CC BY 3.0 https://creativecommons.org/licenses/by/3.0, via Wikimedia Commons).

Figure 1.2. Image has been modified from the original by Servier Medical Art (https://smart.servier.com/smart_image/aorta/) and is used under the Creative Commons Attribution 4.0 International license (https://creativecommons.org/licenses/by/4.0/).

Figure 1.3a. Image adapted and labeled from the original work by Patrick J. Lynch, medical illustrator, and further adapted by Mikael Häggström, M.D. Used under the Creative Commons Attribution-ShareAlike 3.0 Unported license. (Mikael Häggström, M.D., CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0, via Wikimedia Commons).

Figure 1.3b. Image adapted and cropped from the original work by OpenStax College. Used under the Creative Commons Attribution 3.0 Unported license. (OpenStax College, CC BY 3.0 https://creativecommons.org/licenses/by/3.0, via Wikimedia Commons).

Figure 1.5b. Case courtesy of Stefan Tigges, Radiopaedia.org, rID: 95338. https://radiopaedia.org/cases/95338?lang=us. Used under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported. https://creativecommons.org/licenses/by-nc-sa/3.0/.

Figure 1.7. Image from Medical gallery of Blausen Medical 2014, by Bruce Blaus, published in WikiJournal of Medicine 1 (2). DOI: 10.15347/wjm/2014.010. Used under the Creative Commons Attribution-Share Alike 4.0 International license. https://creativecommons.org/licenses/by-sa/4.0/.

Figure 1.8. Image from Blausen.com staff (29 August 2014). Medical gallery of Blausen Medical 2014. WikiJournal of Medicine 1 (2). doi: 10.15347/WJM/2014.010. Wikidata Q44276831. ISSN 2002-4436. Used under the Creative Commons Attribution 3.0 Unported license. https://creativecommons.org/licenses/by/3.0/.

Figure 1.9. Gharleghi, R., Chen, N., Sowmya, A., & Beier, S. Towards automated coronary artery segmentation: A systematic review. *Comput Methods Programs Biomed.* **225**, 107015 (2022). ISSN: 0169-2607. DOI: www.doi.org/10.1016/j.cmpb.2022.107015.

This article was published by Elsevier under the Creative Commons Attribution 4.0 license (www.creativecommons.org/licenses/by/4.0) in the Computer Methods and Programs in Biomedicine journal. No permission is therefore required. For further information, please refer to www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine/publish/open-access-options.

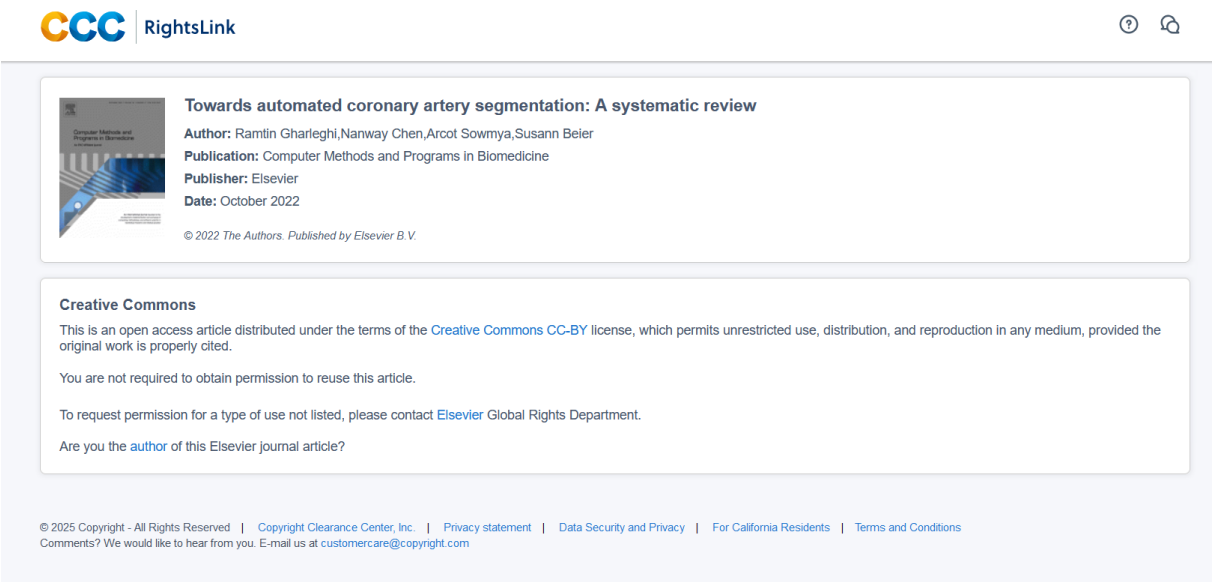


Figure 2.5. Ronneberger, O., Fischer, P., & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*, **18**, 234-241 (2015). DOI: www.doi.org/10.1007/978-3-319-24574-4_28.

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Mar 31, 2025

This Agreement between Belén Serrano Antón. Universidad de Santiago de Compostela ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5977121459644
License date	Feb 27, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	U-Net: Convolutional Networks for Biomedical Image Segmentation
Licensed Content Author	Olaf Ronneberger, Philipp Fischer, Thomas Brox
Licensed Content Date	Jan 1, 2015
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/ illustrations	1
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	no
Title of new work	PhD Thesis
Institution name	Universidad de Santiago de Compostela
Expected presentation date	May 2025
Portions	Fig 1
The Requesting Person / Organization to Appear on the License	Belén Serrano Antón. Universidad de Santiago de Compostela
Requestor Location	Belen Serrano Anton Rua de Jose María Suarez Nuñez, s/n, Santiago De Compostela, Galicia 15782 Spain
Billing Type	Invoice
Billing Address	Universidad de Santiago de Compostela Rua de Jose María Suarez Nuñez, s/n, Santiago De Compostela, Spain 15782
Total	**0.00 EUR**
Terms and Conditions	

Springer Nature Customer Service Centre GmbH Terms and Conditions
The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/ thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/ classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: *'Reproduced with permission from Springer Nature'.*

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Figure 2.6. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. U-Net++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings*, **4**, 3-11 (2018). DOI: www.doi.org/10.1007/978-3-030-00889-5_1.

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Mar 31, 2025

This Agreement between Belén Serrano Antón/Universidad de Santiago de Compostela ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5977130368286
License date	Feb 27, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	UNet : A Nested U-Net Architecture for Medical Image Segmentation
Licensed Content Author	Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh et al
Licensed Content Date	Jan 1, 2018
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	no
Title of new work	PhD Thesis
Institution name	Universidad de Santiago de Compostela
Expected presentation date	May 2025
Portions	Fig 1a
The Requesting Person / Organization to Appear on the License	Belén Serrano Antón/Universidad de Santiago de Compostela
Requestor Location	Belen Serrano Anton Rua de Jose María Suarez Nuñez, s/n, Santiago De Compostela, Galicia 15782 Spain
Billing Type	Invoice
Billing Address	Universidad de Santiago de Compostela Rua de Jose María Suarez Nuñez, s/n, Santiago De Compostela, Spain 15782
Total	**0.00 EUR**
Terms and Conditions	

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/ thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/ classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: *'Reproduced with permission from Springer Nature'.*

Springer Nature Customer Service Centre GmbH Terms and Conditions
The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany's choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Figure 2.7. The image is from Wikimedia Commons and is licensed under the Creative Commons Attribution-ShareAlike 4.0 License https://creativecommons.org/licenses/by-sa/4.0/. Source: https://en.wikipedia.org/wiki/File:VGG_neural_network.png.

Figure 2.8. He, K., Zhang, X., Ren, S., & Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778 (2016). DOI: www.doi.org/10.1109/CVPR.2016.90.

This article was published by the IEEE under the following license. No permission is therefore required. For further information, please refer to https://www.ieee.org/publications/rights/index.html.

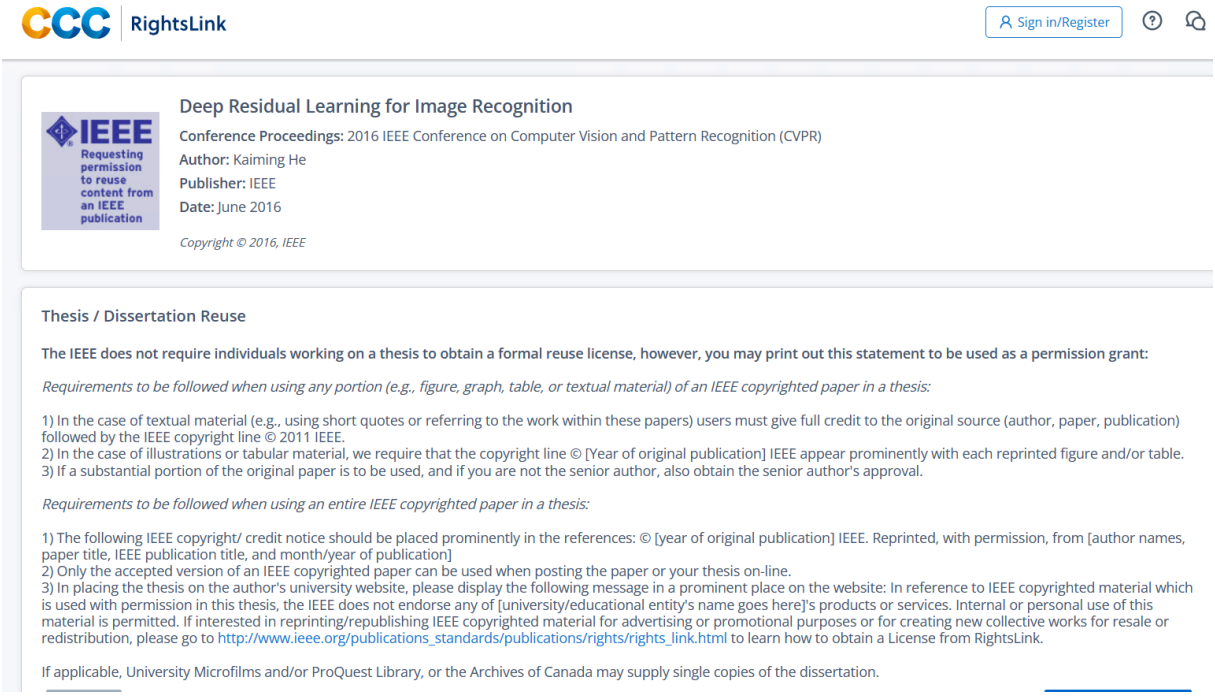


Figure 2.9. Adıyaman, H., Emre Varul, Y., Bakırman, T., & Bayram, B. Stripe Error Correction for Landsat-7 Using Deep Learning. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 1-13 (2024). DOI: https://doi.org/10.1007/s41064-024-00306-x.

This article was published under the Creative Commons Attribution 4.0 license (www.creativecommons.org/licenses/by/4.0). No permission is therefore required. For further information, please refer to https://www.springernature.com/la/open-science/policies/journal-policies/licensing-and-copyright.

Figure 2.10 and Figure A.3. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012-10022 (2021). DOI: www.doi.org/10.1109/ICCV48922.2021.01005.

This article was published under the following license. No permission is therefore required. For further information, please refer to https://www.ieee.org/publications/rights/index.html.

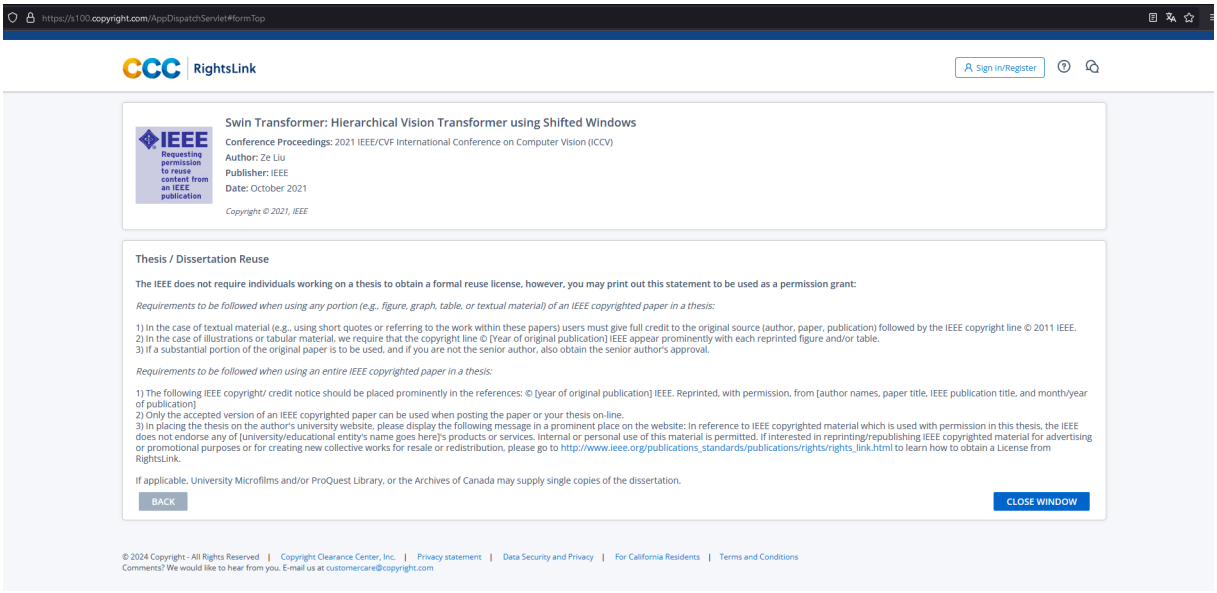


Figure 2.11. Ma, J., Li, F., & Wang, B. U-Mamba: Enhancing Long-Range Dependency for Biomedical Image Segmentation. *arXiv preprint arXiv:2401.04722* (2024). DOI: https://doi.org/10.48550/arXiv.2401.04722.

This article is available under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license (https://creativecommons.org/licenses/by-nc-nd/4.0/). No permission is therefore required. For further information, please refer to https://info.arxiv.org/help/license/index.html.

Figure A.1. Shahoveisi, F., Taheri Gorji, H., Shahabi, S., Hosseinirad, S., Markell, S., & Vasefi, F. Application of Image Processing and Transfer Learning for the Detection of Rust Disease. *Scientific Reports,* **13**(1), 5133 (2023). DOI: https://doi.org/10.1038/s41598-023-31942-9.

This article is published under the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/). No permission is therefore required. For further information, please refer to https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish.

Figure A.2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. Attention is All You Need. In *arXiv preprint arXiv:1706.03762* (2023). Available at: https://arxiv.org/abs/1706.03762.

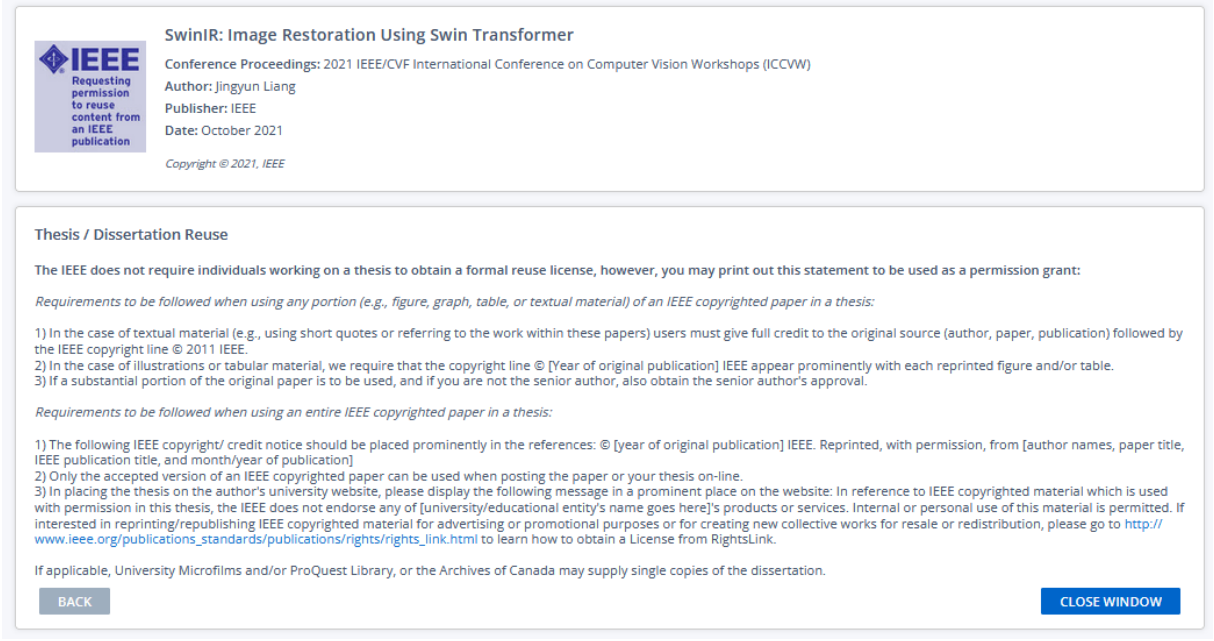
This article is published under the arXiv.org perpetual non-exclusive license, that grants permission to reproduce the tables and figures solely for use in journalistic or scholarly works, please refer to https://info.arxiv.org/help/license/index.html.

Figure A.4. Figure from https://github.com/ACSEkevin/An-Overview-of-Segformer-and-Details-Description under the MIT license.

Figure A.5. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* pp. 1833-1844 (2021).

This article is published under the following license. No permission is therefore required.

For further information, please refer to https://www.ieee.org/publications/rights/index.html.



SwinIR: Image Restoration Using Swin Transformer

Conference Proceedings: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)
Author: Jingyun Liang
Publisher: IEEE
Date: October 2021

Copyright © 2021, IEEE

IEEE Requesting permission to reuse content from an IEEE publication

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

Article 1. Serrano-Antón, B., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., Pérez-Muñuzuri, V., González-Juanatey, J.R., & P. Muñuzuri, A. Coronary artery segmentation based on transfer learning and UNet architecture on computed tomography coronary angiography images. *IEEE Access* **11**:75484-75496 (2023). ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3276768.

This article was published by IEEE under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license (https://creativecommons.org/licenses/by-nc-nd/4.0/). Therefore, no permission is required. For further information, please refer to https://www.ieee.org/publications/rights/copyright-policy.html.

Article 2. Serrano-Antón, B., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., Massonis, G., Pendón, S., Pérez-Muñuzuri, V., González-Juanatey, J. R., & P. Muñuzuri, A. Optimal coronary artery segmentation based on transfer learning and UNet architecture. In *International Workshop on Shape in Medical Imaging* (pp. 55-64). Cham: Springer Nature Switzerland (2023, October).

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Mar 31, 2025

This Agreement between Belén Serrano Antón/Universidad de Santiago de Compostela ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5977491238965
License date	Feb 28, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Optimal Coronary Artery Segmentation Based on Transfer Learning and UNet Architecture
Licensed Content Author	Belén Serrano-Antón, Alberto Otero-Cacho, Diego López-Otero et al
Licensed Content Date	Jan 1, 2023
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	PhD Thesis
Institution name	Universidad de Santiago de Compostela
Expected presentation date	May 2025
The Requesting Person / Organization to Appear on the License	Belén Serrano Antón/Universidad de Santiago de Compostela
Requestor Location	Belen Serrano Anton Rua de Jose María Suarez Nuñez, s/n, Santiago De Compostela, Galicia 15782 Spain
Billing Type	Invoice
Billing Address	Universidad de Santiago de Compostela Rua de Jose María Suarez Nuñez, s/n, Santiago De Compostela, Spain 15782
Total	**0.00 EUR**

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/ thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/ classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: *'Reproduced with permission from Springer Nature'.*

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany's choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? E-mail us at customercare@copyright.com.

Article 3. Serrano-Antón, B., Rehman, M., Martinel, N., Avanzo, M., Spizzo, R., Fanetti, G., P. Muñuzuri, A., & Micheloni, C. MAR-DTN: Metal artifact reduction using domain transformation network for radiotherapy planning. In *International Conference on Pattern Recognition* (pp. 143-159). Springer, Cham (2025).

This article will be published by Springer under the following permission.

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Mar 31, 2025

This Agreement between Belén Serrano Antón ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5961201495437
License date	Feb 03, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	MAR-DTN: Metal Artifact Reduction Using Domain Transformation Network for Radiotherapy Planning
Licensed Content Author	Belén Serrano-Antón, Mubashara Rehman, Niki Martinel et al
Licensed Content Date	Jan 1, 2025
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	PhD Thesis
Institution name	Universidad de Santiago de Compostela
Expected presentation date	May 2025
The Requesting Person / Organization to Appear on the License	Belén Serrano Antón
Requestor Location	Belen Serrano Anton Rua de Jose María Suarez Nuñez, s/n, Santiago De Compostela, Galicia 15782 Spain
Billing Type	Invoice
Billing Address	Universidad de Santiago de Compostela Rua de Jose María Suarez Nuñez, s/n, Santiago De Compostela, Spain 15782
Total	**0.00 EUR**
Terms and Conditions	

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/ thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/ classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature*'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany's choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? E-mail us at customercare@copyright.com.

Article 4. Serrano-Antón, B., Insúa Villa, M., Pendón-Minguillón, S., Paramés-Estévez, S., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., González-Juanatey, J.R., & Muñuzuri, A.P. Unsupervised clustering-based coronary artery segmentation. *BioData Mining* **18**(1):1-23 (2025). Publisher: BioMed Central.

I declare that the rest of the figures shown in this thesis were either of my own making or reproduced from the publications stated above.

Funding information

This research has been supported by the Xunta de Galicia (Grant No. 2021-PG036-1), the Spanish Ministerio de Ciencia e Innovación (Grants No. PID2022-138322OB-I00, PID2022-141626NB-I00, MCIN/AEI/https://doi.org/10.13039/501100011033), European Union Next Generation EU/PRTR (Grant No: DIN2020-011068) and Interreg VI-A Spain - Portugal (Project 0330_NEW_HEART_1_E). All these programs are co-funded by ERDF (EU).

Bibliography

[1] Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., Bonny, A., Brauer, M., Brodmann, M., Cahill, T. J., Carapetis, J., Catapano, A. L., Chugh, S. S., Cooper, L. T., Coresh, J., Criqui, M., DeCleene, N., Eagle, K. A., Emmons-Bell, S., Feigin, V. L., Fernández-Solà, J., Fowkes, G., Gakidou, E., Grundy, S. M., He, F. J., Howard, G., Hu, F., Inker, L., Karthikeyan, G., Kassebaum, N., Koroshetz, W., Lavie, C., Lloyd-Jones, D., Lu, H. S., Mirijello, A., Misganaw Temesgen, A., Mokdad, A., Moran, A. E., Muntner, P., Narula, J., Neal, B., Ntsekhe, M., Moraes de Oliveira, G., Otto, C., Owolabi, M., Pratt, M., Rajagopalan, S., Reitsma, M., Ribeiro, A. L. P., Rigotti, N., Rodgers, A., Sable, C., Shakil, S., Sliwa-Hahnle, K., Stark, B., Sundström, J., Timpel, P., Tleyjeh, I. M., Valgimigli, M., Vos, T., Whelton, P. K., Yacoub, M., Zuhlke, L., Murray, C. & Fuster, V. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Stud. *Journal of the American College of Cardiology* **76**, 2982–3021 (2020).

[2] Timmis, A., Vardas, P., Townsend, N., Torbica, A., Katus, H., De Smedt, D., Gale, C. P., Maggioni, A. P., Petersen, S. E., Huculeci, R., , Kazakiewicz, D., de Benito Rubio, V., Ignatiuk, B., Raisi-Estabragh, Z., Pawlak, A., Karagiannidis, E., Treskes, R., Gaita, D., Beltrame, J. F., McConnachie, A., Bardinet, I., Graham, I., Flather, M., Elliott, P., Mossialos, E. A., Weidinger, F., Achenbach, S. & European Society of Cardiology. European Society of Cardiology: cardiovascular disease statistics 2021. *European Heart Journal* **43**, 716–799 (2022).

[3] Timmis, A., Aboyans, V., Vardas, P., Townsend, N., Torbica, A., Kavousi, M., Boriani, G., Huculeci, R., Kazakiewicz, D., Scherr, D., , Karagiannidis, E., Cvijic, M., Kapłon-Cieślicka, A., Ignatiuk, B., Raatikainen, P., De Smedt, D., Wood, A., Dudek, D., Van Belle, E., Weidinger, F. & ESC National Cardiac Societies. European Society of Cardiology: the 2023 Atlas of Cardiovascular Disease Statistics. *European Heart Journal* **45**, 4019–4062 (2024).

[4] Sacramento-Pacheco, J., Sánchez-Gómez, M. B., Gómez-Salgado, J., Novo-Muñoz, M. M. & Duarte-Clíments, G. Prevalence of Cardiovascular Risk Factors in Spain: A Systematic Review. *Journal of Clinical Medicine* **12**, 6944 (2023).

[5] García-Ortiz, L., Barreiro-Perez, M., Merchan-Gómez, S., Ignacio Recio-Rodriguez, J., Sánchez-Aguadero, N., Alonso-Dominguez, R., Lugones-Sanchez, C., Rodríguez-Sanchez, E., Sanchez, P. L. & Gómez-Marcos, M. A. Prevalence of coronary atherosclerosis and reclassification of cardiovascular risk in Spanish population

by coronary computed tomography angiography: EVA study. *European Journal of Clinical Investigation* **50**, e13272 (2020).

[6] Writing Committee Members, Isselbacher, E. M., Preventza, O., Hamilton Black III, J., Augoustides, J. G., Beck, A. W., Bolen, M. A., Braverman, A. C., Bray, B. E. & Brown-Zimmerman, M. M. 2022 ACC/AHA Guideline for the Diagnosis and Management of Aortic Disease: A Report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice Guidelines. *Journal of the American College of Cardiology* **80**, e223–e393 (2022).

[7] Gharleghi, R., Chen, N., Sowmya, A. & Beier, S. Towards automated coronary artery segmentation: A systematic review. *Computer Methods and Programs in Biomedicine* **225**, 107015 (2022).

[8] Serruys, P. W., Hara, H., Garg, S., Kawashima, H., Nørgaard, B. L., Dweck, M. R., Bax, J. J., Knuuti, J., Nieman, K., Leipsic, J. A., Mushtaq, S., Andreini, D. & Onuma, Y. Coronary Computed Tomographic Angiography for Complete Assessment of Coronary Artery Disease: JACC State-of-the-Art Review. *Journal of the American College of Cardiology* **78**, 713–736 (2021).

[9] Rajiah, P. & Abbara, S. Ct coronary imaging–a fast evolving world. *QJM: An International Journal of Medicine* **111**, 595–604 (2018).

[10] Nedadur, R., Wang, B. & Tsang, W. Artificial intelligence for the echocardiographic assessment of valvular heart disease. *Heart* **108**, 1592–1599 (2022).

[11] Hahn, L. D., Baeumler, K. & Hsiao, A. Artificial intelligence and machine learning in aortic disease. *Current Opinion in Cardiology* **36**, 695–703 (2021).

[12] James, T. N. Anatomy of the coronary arteries in health and disease. *Circulation* **32**, 1020–1033 (1965).

[13] InformedHealth.org. How does the blood circulatory system work? (2023). URL https://www.ncbi.nlm.nih.gov/books/NBK279250/. Updated 2023 Nov 21.

[14] MedlinePlus. Circulation of blood through the heart (2022). URL https://medlineplus.gov/ency/imagepages/19387.htm. Reviewed 2022 Oct 5.

[15] di Gioia, C. R. T., Ascione, A., Carletti, R. & Giordano, C. Thoracic Aorta: Anatomy and Pathology. *Diagnostics* **13**, 2166 (2023).

[16] Dagenais, F. Anatomy of the Thoracic Aorta and of Its Branches. *Thoracic Surgery Clinics* **21**, 219–227 (2011).

[17] Ho, S. Y. Structure and anatomy of the aortic root. *European Journal of Echocardiography* **10**, i3–i10 (2009).

[18] Cleveland Clinic. How does the blood flow through your heart? (2023). URL https://my.clevelandclinic.org/health/articles/17060-how-does-the-blood-flow-through-your-heart. Reviewed 2023 Nov 8.

[19] Loukas, M., Groat, C., Khangura, R., Owens, D. G. & Anderson, R. H. The normal and abnormal anatomy of the coronary arteries. *Clinical Anatomy: The Official Journal of the American Association of Clinical Anatomists and the British Association of Clinical Anatomists* **22**, 114–128 (2009).

[20] Loukas, M., Sharma, A., Blaak, C., Sorenson, E. & Mian, A. The clinical anatomy of the coronary arteries. *Journal of Cardiovascular Translational Research* **6**, 197–207 (2013).

[21] Dodge Jr, J. T., Brown, B. G., Bolson, E. L. & Dodge, H. T. Lumen diameter of normal human coronary arteries. influence of age, sex, anatomic variation, and left ventricular hypertrophy or dilation. *Circulation* **86**, 232–246 (1992).

[22] Saremi, F. & Achenbach, S. Coronary Plaque Characterization Using CT. *American Journal of Roentgenology* **204**, W249–W260 (2015).

[23] Baumann, S., Renker, M., Meinel, F. G., Wichmann, J. L., Fuller, S. R., Bayer, R. R., Schoepf, U. J. & Steinberg, D. H. Computed Tomography Imaging of Coronary Artery Plaque : Characterization and Prognosis. *Radiologic Clinics* **53**, 307–315 (2015).

[24] Libby, P. Current Concepts of the Pathogenesis of the Acute Coronary Syndromes. *Circulation* **104**, 365–372 (2001).

[25] Laal, M. Innovation Process in Medical Imaging. *Procedia-Social and Behavioral Sciences* **81**, 60–64 (2013).

[26] Hussain, S., Mubeen, I., Ullah, N., Shah, S. S. U. D., Khan, B. A., Zahoor, M., Ullah, R., Khan, F. A. & Sultan, M. A. Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review. *BioMed Research International* **2022**, 5164970 (2022).

[27] Flower, M. A. *Webb's Physics of Medical Imaging* (CRC press, 2012).

[28] Smith-Bindman, R., Miglioretti, D. L. & Larson, E. B. Rising Use Of Diagnostic Medical Imaging In A Large Integrated Health System. *Health Affairs* **27**, 1491–1502 (2008).

[29] Hsieh, J. *Computed Tomography: Principles, Design, Artifacts, and Recent Advances* (SPIE Press, 2003).

[30] Holcombe, S. A., Horbal, S. R., Ross, B. E., Brown, E., Derstine, B. A. & Wang, S. C. Variation in aorta attenuation in contrast-enhanced ct and its implications for calcification thresholds. *PLoS One* **17**, e0277111 (2022).

[31] Cartlidge, T. R., Bing, R., Pawade, T. A., Doris, M. K., Kwiecinski, J., Guzzetti, E., Adamson, P. D., Massera, D., Lembo, M., Peeters, F. E., Couture, C., Berman, D. S., Dey, D., Slomka, P., Pibarot, P., Newby, D. E., Clavel, M.-A. & Dweck, M. R. Contrast-enhanced computed tomography assessment of aortic stenosis. *Heart* **107**, 1905–1911 (2021).

[32] Angelillis, M., Costa, G., De Backer, O., Mochi, V., Christou, A., Giannini, C., Spontoni, P., De Carlo, M., Søndergaard, L., Miccoli, M. & Petronio, A. S. Threshold for calcium volume evaluation in patients with aortic valve stenosis: correlation with agatston score. *Journal of Cardiovascular Medicine* **22**, 496–502 (2021).

[33] Suetens, P. *Fundamentals of Medical Imaging* (Cambridge university press, 2017).

[34] Cademartiri, F., Mollet, N. R., Runza, G., Bruining, N., Hamers, R., Somers, P., Knaapen, M., Verheye, S., Midiri, M., Krestin, G. P. & de Feyter, P. J. Influence of intracoronary attenuation on coronary plaque measurements using multislice computed tomography: observations in an ex vivo model of coronary computed tomography angiography. *European Radiology* **15**, 1426–1431 (2005).

[35] Dalager, M. G., Bøttcher, M., Andersen, G., Thygesen, J., Pedersen, E. M., Dejbjerg, L., Gøtzsche, O. & Bøtker, H. E. Impact of luminal density on plaque classification by CT coronary angiography. *The International Journal of Cardiovascular Imaging* **27**, 593–600 (2011).

[36] Maffei, E., Martini, C., Arcadi, T., Clemente, A., Seitun, S., Zuccarelli, A., Torri, T., Mollet, N. R., Rossi, A., Catalano, O., Messalli, G. & Cademartiri, F. Plaque imaging with ct coronary angiography: effect of intra-vascular attenuation on plaque type classification. *World Journal of Radiology* **4**, 265 (2012).

[37] Fezzi, S., Huang, J., Lunardi, M., Ding, D., Ribichini, F. L., Tu, S. & Wijns, W. Coronary physiology in the catheterisation laboratory: an A to Z practical guide. *AsiaIntervention* **8**, 86 (2022).

[38] Rioufol, G., Dérimay, F., Roubille, F., Perret, T., Motreff, P., Angoulvant, D., Cottin, Y., Meunier, L., Cetran, L., Cayla, G., Harbaoui, B., Wiedemann, J.-Y., Van Belle, É., Pouillot, C., Noirclerc, N., Morelle, J.-F., Soto, F.-X., Caussin, C., Bertrand, B., Lefèvre, T., Dupouy, P., Lesault, P.-F., Albert, F., Barthelemy, O., Koning, R., Leborgne, L., Barnay, P., Chapon, P., Armero, S., Lafont, A., Piot, C., Amaz, C., Vaz, B., Benyahya, L., Varillon, Y., Ovize, M., Mewton, N. & Finet, G. Fractional Flow Reserve to Guide Treatment of Patients With Multivessel Coronary Artery Disease. *Journal of the American College of Cardiology* **78**, 1875–1885 (2021).

[39] Bolognese, L. & Reccia, M. R. Computed tomography to replace invasive coronary angiography? The DISCHARGE trial. *European Heart Journal Supplements* **24**, I25–I28 (2022).

[40] Koo, B.-K., Erglis, A., Doh, J.-H., Daniels, D. V., Jegere, S., Kim, H.-S., Dunning, A., DeFrance, T., Lansky, A., Leipsic, J. & Min, J. K. Diagnosis of ischemia-causing coronary stenoses by noninvasive fractional flow reserve computed from coronary computed tomographic angiograms: results from the prospective multicenter DISCOVER-FLOW (Diagnosis of Ischemia-Causing Stenoses Obtained Via Noninvasive Fractional Flow Reserve) study. *Journal of the American College of Cardiology* **58**, 1989–1997 (2011).

[41] Pijls, N., Van Son, J., Kirkeeide, R. L., De Bruyne, B. & Gould, K. L. Experimental basis of determining maximum coronary, myocardial, and collateral blood flow by pressure measurements for assessing functional stenosis severity before and after percutaneous transluminal coronary angioplasty. *Circulation* **87**, 1354–1367 (1993).

[42] Rajiah, P., Cummings, K. W., Williamson, E. & Young, P. M. CT Fractional Flow Reserve: A Practical Guide to Application, Interpretation, and Problem Solving. *Radiographics* **42**, 340–358 (2022).

[43] Otero-Cacho, A., López-Otero, D., Insúa Villa, M., Díaz-Fernández, B., Bastos-Fernández, M., Pérez-Muñuzuri, V., Muñuzuri, A. P. & González-Juanatey, J. R. Validation of a new model of non-invasive functional assessment of coronary lesions by computer tomography fractional flow reserve. *REC: CardioClinics* **59**, 35–45 (2024).

[44] Saraste, A. & Knuuti, J. ESC 2019 guidelines for the diagnosis and management of chronic coronary syndromes: Recommendations for cardiovascular imaging. *Herz* **45**, 409–420 (2020).

[45] Otero-Cacho, A., Insúa Villa, M., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., Pérez-Muñuzuri, V., Muñuzuri, A. P. & González-Juanatey, J. R. Influence of the pressure wire on the fractional flow reserve calculation: CFD analysis of an ideal vessel and clinical patients with stenosis. *Computer Methods and Programs in Biomedicine* **255**, 108325 (2024).

[46] Neumann, F.-J., Sousa-Uva, M., Ahlsson, A., Alfonso, F., Banning, A. P., Benedetto, U., Byrne, R. A., Collet, J.-P., Falk, V., Head, S. J., Jüni, P., Kastrati, A., Koller, A., Kristensen, S. D., Niebauer, J., Richter, D. J., Seferović, P. M., Sibbing, D., Stefanini, G. G., Windecker, S., Yadav, R., Zembala, M. O. & ESC Scientific Document Group. 2018 ESC/EACTS Guidelines on myocardial revascularization. *European Heart Journal* **40**, 87–165 (2018). URL https://doi.org/10.1093/eurheartj/ehy394. https://academic.oup.com/eurheartj/article-pdf/40/2/87/29005222/ehy394.pdf.

[47] Byrne, R. A., Rossello, X., Coughlan, J. J., Barbato, E., Berry, C., Chieffo, A., Claeys, M. J., Dan, G.-A., Dweck, M. R., Galbraith, M., Gilard, M., Hinterbuchner, L., Jankowska, E. A., Jüni, P., Kimura, T., Kunadian, V., Leosdottir, M., Lorusso, R., Pedretti, R. F. E., Rigopoulos, A. G., Rubini Gimenez, M., Thiele, H., Vranckx, P., Wassmann, S., Wenger, N. K., Ibanez, B. & ESC Scientific Document Group. 2023 ESC Guidelines for the management of acute coronary syndromes: Developed by the task force on the management of acute coronary syndromes of the European Society of Cardiology (ESC). *European Heart Journal* **44**, 3720–3826 (2023). URL https://doi.org/10.1093/eurheartj/ehad191. https://academic.oup.com/eurheartj/article-pdf/44/38/3720/56728087/ehad191.pdf.

[48] Marano, R., Rovere, G., Savino, G., Flammia, F. C., Carafa, M. R. P., Steri, L., Merlino, B. & Natale, L. Ccta in the diagnosis of coronary artery disease. *La radiologia medica* **125**, 1102–1113 (2020).

[49] Min, J. K., Berman, D. S., Budoff, M. J., Jaffer, F. A., Leipsic, J., Leon, M. B., Mancini, G. J., Mauri, L., Schwartz, R. S. & Shaw, L. J. Rationale and design of the DeFACTO (Determination of Fractional Flow Reserve by Anatomic Computed Tomographic AngiOgraphy) study. *Journal of Cardiovascular Computed Tomography* **5**, 301–309 (2011).

[50] Pawade, T., Clavel, M.-A., Tribouilloy, C., Dreyfus, J., Mathieu, T., Tastet, L., Renard, C., Gun, M., Jenkins, W. S. A., Macron, L., Sechrist, J. W., Lacomis, J. M., Nguyen, V., Galian Gay, L., Cuéllar Calabria, H., Ntalas, I., Cartlidge, T. R. G., Prendergast, B., Rajani, R., Evangelista, A., Cavalcante, J. L., Newby, D. E., Pibarot, P., Messika Zeitoun, D. & Dweck, M. R. Computed Tomography Aortic Valve Calcium Scoring in Patients With Aortic Stenosis. *Circulation: Cardiovascular Imaging* **11**, e007146 (2018).

[51] Rana, M. Aortic Valve Stenosis: Diagnostic Approaches and Recommendations of the 2021 ESC/EACTS Guidelines for the Management of Valvular Heart Disease –A Review of the Literature. *Cardiology and Cardiovascular Medicine* **6**, 315 (2022).

[52] Rader, F., Sachdev, E., Arsanjani, R. & Siegel, R. J. Left Ventricular Hypertrophy in Valvular Aortic Stenosis: Mechanisms and Clinical Implications. *The American Journal of Medicine* **128**, 344–352 (2015).

[53] Mittal, T. & Marcus, N. Imaging diagnosis of aortic stenosis. *Clinical Radiology* **76**, 3–14 (2021).

[54] Dweck, M. R., Loganath, K., Bing, R., Treibel, T. A., McCann, G. P., Newby, D. E., Leipsic, J., Fraccaro, C., Paolisso, P., Cosyns, B., Habib, G., Cavalcante, J., Donal, E., Lancellotti, P., Clavel, M.-A., Otto, C. M. & Pibarot, P. Multi-modality imaging in aortic stenosis: an EACVI clinical consensus document. *European Heart Journal-Cardiovascular Imaging* **24**, 1430–1443 (2023).

[55] Beyersdorf, F., Vahanian, A., Milojevic, M., Praz, F., Baldus, S., Bauersachs, J., Capodanno, D., Conradi, L., De Bonis, M., De Paulis, R., Delgado, V., Freemantle, N., Gilard, M., Haugaa, K. H., Jeppsson, A., Jüni, P., Pierard, L., Prendergast, B. D., Sádaba, J. R., Tribouilloy, C., Wojakowski, W. & ESC/EACTS Scientific Document Group. 2021 esc/eacts guidelines for the management of valvular heart disease: developed by the task force for the management of valvular heart disease of the european society of cardiology (esc) and the european association for cardio-thoracic surgery (eacts). *European journal of cardio-thoracic surgery* **60**, 727–800 (2021).

[56] Agatston, A. S., Janowitz, W. R., Hildner, F. J., Zusmer, N. R., Viamonte Jr, M. & Detrano, R. Quantification of coronary artery calcium using ultrafast computed tomography. *Journal of the American College of Cardiology* **15**, 827–832 (1990).

[57] Ito, S. & Oh, J. K. Aortic Stenosis: New Insights in Diagnosis, Treatment, and Prevention. *Korean Circulation Journal* **52**, 721–736 (2022).

[58] Melidi, E., Latsios, G., Toutouzas, K., Vavouranakis, M., Tolios, I., Gouliami, M., Gerckens, U. & Tousoulis, D. Cardio-anesthesiology considerations for the trans-catheter aortic valve implantation (TAVI) procedure. *Hellenic Journal of Cardiology* **57**, 401–406 (2016).

[59] Stortecky, S., Buellesfeld, L., Wenaweser, P. & Windecker, S. Transcatheter aortic valve implantation: the procedure. *Heart* **98**, iv44–iv51 (2012).

[60] Ruparelia, N. & Prendergast, B. D. TAVI in 2015: who, where and how? *Heart* **101**, 1422–1431 (2015).

[61] Craiem, D., Chironi, G., Casciaro, M. E., Graf, S. & Simon, A. Calcifications of the Thoracic Aorta on Extended Non-Contrast-Enhanced Cardiac CT. *PLoS One* **9**, e109584 (2014).

[62] Guilenea, F. N., Casciaro, M. E., Soulat, G., Mousseaux, E. & Craiem, D. Automatic thoracic aorta calcium quantification using deep learning in non-contrast ECG-gated CT images. *Biomedical Physics & Engineering Express* **10**, 035007 (2024).

[63] Andreini, D., Collet, C., Leipsic, J., Nieman, K., Bittencurt, M., De Mey, J., Buls, N., Onuma, Y., Mushtaq, S., Conte, E., Bartorelli, A. L., Stefanini, G., Sonck, J., Knaapen, P., Ghoshhajra, B. & Serruys, P. Pre-procedural planning of coronary revascularization by cardiac computed tomography: An expert consensus document of the Society of Cardiovascular Computed Tomography. *Journal of Cardiovascular Computed Tomography* **16**, 558–572 (2022).

[64] Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J. V., Pieper, S. & Kikinis, R. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging* **30**, 1323–1341 (2012).

[65] Serrano-Antón, B., Insúa Villa, M., Pendón-Minguillón, S., Paramés-Estévez, S., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., González-Juanatey, J. R. & P Muñuzuri, A. Unsupervised clustering based coronary artery segmentation. *BioData Mining* **18**, 1–23 (2025).

[66] George, R. T., Arbab-Zadeh, A., Cerci, R. J., Vavere, A. L., Kitagawa, K., Dewey, M., Rochitte, C. E., Arai, A. E., Paul, N., Rybicki, F. J., Lardo, A. C., Clouse, M. E. & Lima, J. A. C. Diagnostic performance of combined noninvasive coronary angiography and myocardial perfusion imaging using 320-MDCT: the CT angiography and perfusion methods of the CORE320 multicenter multinational diagnostic study. *American Journal of Roentgenology* **197**, 829–837 (2011).

[67] Nagao, M., Kido, T., Watanabe, K., Saeki, H., Okayama, H., Kurata, A., Hosokawa, K., Higashino, H. & Mochizuki, T. Functional assessment of coronary artery flow using adenosine stress dual-energy CT: a preliminary study. *The International Journal of Cardiovascular Imaging* **27**, 471–481 (2011).

[68] Steigner, M. L., Mitsouras, D., Whitmore, A. G., Otero, H. J., Wang, C., Buckley, O., Levit, N. A., Hussain, A. Z., Cai, T., Mather, R. T., Smedby, Ø., DiCarli, M. F. & Rybicki, F. J. Iodinated contrast opacification gradients in normal coronary arteries imaged with prospectively ECG-gated single heart beat 320-detector row computed tomography. *Circulation: Cardiovascular Imaging* **3**, 179–186 (2010).

[69] Choi, J.-H., Min, J. K., Labounty, T. M., Lin, F. Y., Mendoza, D. D., Shin, D. H., Ariaratnam, N. S., Koduru, S., Granada, J. F., Gerber, T. C., Oh, J. K., Gwon, H.-C. & Choe, Y. H. Intracoronary transluminal attenuation gradient in coronary CT angiography for determining coronary artery stenosis. *JACC: Cardiovascular Imaging* **4**, 1149–1157 (2011).

[70] Slicer Wiki. Registration: Resampling — Slicer Wiki (2010). URL https://www.slicer.org/w/index.php?title=Registration:Resampling&oldid=16013. Online; accessed 24-February-2025.

[71] Hecht, H. S., Cronin, P., Blaha, M. J., Budoff, M. J., Kazerooni, E. A., Narula, J., Yankelevitz, D. & Abbara, S. 2016 SCCT/STR guidelines for coronary artery calcium scoring of noncontrast noncardiac chest CT scans: a report of the Society of Cardiovascular Computed Tomography and Society of Thoracic Radiology. *Journal of Cardiovascular Computed Tomography* **11**, 74–84 (2017).

[72] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).

[73] O'Shea, K. & Nash, R. An Introduction to Convolutional Neural Networks. *arXiv Preprint:1511.08458* (2015).

[74] Dubey, S. R., Singh, S. K. & Chaudhuri, B. B. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing* **503**, 92–108 (2022).

[75] Dumoulin, V. & Visin, F. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285* (2016).

[76] Ronneberger, O., P.Fischer & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351 of *LNCS*, 234–241 (Springer, 2015). URL http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a. (available on arXiv:1505.04597 [cs.CV]).

[77] Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440 (2015).

[78] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4*, 3–11 (Springer, 2018).

[79] Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv Preprint:1409.1556* (2014).

[80] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).

[81] Tan, M. & Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML*, 6105–6114 (PMLR, 2019).

[82] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255 (2009).

[83] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. & Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv Preprint:1704.04861* **126** (2017).

[84] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510–4520 (2018).

[85] Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976 (2017).

[86] Adıyaman, H., Emre Varul, Y., Bakırman, T. & Bayram, B. Stripe Error Correction for Landsat-7 Using Deep Learning. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science* 1–13 (2024).

[87] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. & Polosukhin, I. Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. & Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017). URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[88] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)* (2021). URL https://openreview.net/forum?id=YicbFdNTTy.

[89] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022 (2021).

[90] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. & Zhang, L. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6877–6886 (2021).

[91] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P. & Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision ICCV)*, 548–558 (2021).

[92] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M. & Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Proceedings of*

the 35th International Conference on Neural Information Processing Systems (NeurIPS), NIPS '21 (Curran Associates Inc., Red Hook, NY, USA, 2021).

[93] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L. & Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 1833–1844 (2021).

[94] Gu, A. & Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv Preprint:2312.00752* (2023).

[95] Ma, J., Li, F. & Wang, B. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. *arXiv Preprint:2401.04722* (2024).

[96] Jiang, L., Dai, B., Wu, W. & Loy, C. C. Focal Frequency Loss for Image Reconstruction and Synthesis. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 13899–13909 (2021).

[97] Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**, 600–612 (2004).

[98] Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X. & Martel, A. L. Loss odyssey in medical image segmentation. *Medical Image Analysis* **71**, 102035 (2021).

[99] Taghanaki, S. A., Zheng, Y., Zhou, S. K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D. & Hamarneh, G. Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics* **75**, 24–33 (2019).

[100] Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**, 318–327 (2020).

[101] Zhu, W., Huang, Y., Zeng, L., Chen, X., Liu, Y., Qian, Z., Du, N., Fan, W. & Xie, X. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical physics* **46**, 576–589 (2019).

[102] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Jorge Cardoso, M. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, 240–248 (Springer, 2017).

[103] Salehi, S. S. M., Erdogmus, D. & Gholipour, A. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In *International workshop on Machine Learning in Medical Imaging*, 379–387 (Springer, 2017).

[104] Horé, A. & Ziou, D. Image Quality Metrics: PSNR vs. SSIM. In *Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR)*, 2366–2369 (IEEE, 2010).

[105] Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**, 600–612 (2004).

[106] Abe, K., Kadoya, N., Ito, K., Tanaka, S., Nakajima, Y., Hashimoto, S., Suda, Y., Uno, T. & Jingu, K. Evaluation of the MVCT-based radiomic features as prognostic factor in patients with head and neck squamous cell carcinoma. *BMC Medical Imaging* **23**, 102 (2023).

[107] Martin, S. & Yartsev, S. kVCT, MVCT, and hybrid CT image studies—Treatment planning and dose delivery equivalence on helical tomotherapy. *Medical Physics* **37**, 2847–2854 (2010).

[108] Branchini, M., Fiorino, C., Dell'Oca, I., Belli, M., Perna, L., Di Muzio, N., Calandrino, R. & Broggi, S. Validation of a method for "dose of the day" calculation in head-neck tomotherapy by using planning ct-to-MVCT deformable image registration. *Physica Medica* **39**, 73–79 (2017).

[109] Korreman, S., Rasch, C., McNair, H., Verellen, D., Oelfke, U., Maingon, P., Mijnheer, B. & Khoo, V. The European Society of Therapeutic Radiology and Oncology–European Institute of Radiotherapy (ESTRO–EIR) report on 3D CT-based in-room image guidance systems: a practical and technical review and guide. *Radiotherapy and Oncology* **94**, 129–144 (2010).

[110] Jackowiak, W., Bąk, B., Kowalik, A., Ryczkowski, A., Skórska, M. & Paszek-Widzińska, M. Influence of the type of imaging on the delineation process during the treatment planning. *Reports of Practical Oncology & Radiotherapy* **20**, 351–357 (2015).

[111] Forrest, L. J., Mackie, T. R., Ruchala, K., Turek, M., Kapatoes, J., Jaradat, H., Hui, S., Balog, J., Vail, D. M. & Mehta, M. P. The utility of megavoltage computed tomography images from a helical tomotherapy system for setup verification purposes. *International Journal of Radiation Oncology*Biology*Physics* **60**, 1639–1644 (2004).

[112] Timmerman, R. D. & Xing, L. *Image-guided and adaptive radiation therapy* (Lippincott Williams & Wilkins, 2012).

[113] Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. elastix: A Toolbox for Intensity-Based Medical Image Registration. *IEEE Transactions on Medical Imaging* **29**, 196–205 (2009).

[114] Kim, H., Yoo, S. K., Kim, D. W., Lee, H., Hong, C.-S., Han, M. C. & Kim, J. S. Metal artifact reduction in kV CT images throughout two-step sequential deep convolutional neural networks by combining multi-modal imaging (MARTIAN). *Scientific Reports* **12**, 20823 (2022).

[115] Liugang, G., Hongfei, S., Xinye, N., Mingming, F., Zheng, C. & Tao, L. Metal artifact reduction through MVCBCT and kVCT in radiotherapy. *Scientific Reports* **6**, 37608 (2016).

[116] Serrano-Antón, B., Rehman, M., Martinel, N., Avanzo, M., Spizzo, R., Fanetti, G., P. Muñuzuri, A. & Micheloni, C. MAR-DTN: Metal Artifact Reduction Using Domain Transformation Network for Radiotherapy Planning. In Antonacopoulos, A., Chaudhuri, S., Chellappa, R., Liu, C.-L., Bhattacharya, S. & Pal, U. (eds.) *Pattern Recognition*, 143–159 (Springer Nature Switzerland, Cham, 2025).

[117] Serrano-Antón, B., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., Pérez-Muñuzuri, V., González-Juanatey, J. R. & P. Muñuzuri, A. Coronary Artery Segmentation Based on Transfer Learning and UNet Architecture on Computed Tomography Coronary Angiography Images. *IEEE Access* **11**, 75484–75496 (2023).

[118] Wang, J., Zhao, Y., Noble, J. H. & Dawant, B. M. Conditional Generative Adversarial Networks for Metal Artifact Reduction in CT Images of the Ear. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, 3–11 (Springer, 2018).

[119] Weng, W. & Zhu, X. INet: Convolutional Networks for Biomedical Image Segmentation. *IEEE Access* **9**, 16591–16603 (2021).

[120] Kinga, D. & Adam, J. B. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, vol. 5 (2015).

[121] Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M. & Kalinin, A. A. Albumentations: Fast and Flexible Image Augmentations. *Information* **11**, 125 (2020).

[122] Kaposi, P., Youn, T., Tóth, A., Frank, V., Shariati, S., Szendrői, A., Magyar, P. & Bérczi, V. Orthopaedic metallic artefact reduction algorithm facilitates CT evaluation of the urinary tract after hip prosthesis. *Clinical Radiology* **75**, 78.e17–78.e24 (2020).

[123] Huang, X., Wang, J., Tang, F., Zhong, T. & Zhang, Y. Metal artifact reduction on cervical CT images by deep residual learning. *Biomedical Engineering OnLine* **17**, 1–15 (2018).

[124] Wang, H., Li, Y., He, N., Ma, K., Meng, D. & Zheng, Y. DICDNet: Deep Interpretable Convolutional Dictionary Network for Metal Artifact Reduction in CT Images. *IEEE Transactions on Medical Imaging* **41**, 869–880 (2021).

[125] Lin, W.-A., Liao, H., Peng, C., Sun, X., Zhang, J., Luo, J., Chellappa, R. & Zhou, S. K. DuDoNet: Dual Domain Network for CT Metal Artifact Reduction. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10504–10513 (2019).

[126] Lee, D., Park, C., Lim, Y. & Cho, H. A Metal Artifact Reduction Method Using a Fully Convolutional Network in the Sinogram and Image Domains for Dental Computed Tomography. *Journal of Digital Imaging* **33**, 538–546 (2019). URL https://doi.org/10.1007/s10278-019-00297-x.

[127] Yu, L., Zhang, Z., Li, X., Ren, H., Zhao, W. & Xing, L. Metal artifact reduction in 2D CT images with self-supervised cross-domain learning. *Physics in Medicine & Biology* **66**, 175003 (2021). URL https://dx.doi.org/10.1088/1361-6560/ac195c.

[128] Ni, X., Shi, Z., Song, X., Tang, T., Li, S., Hou, Z., Zhang, W., Wang, W. F., Chen, F., Li, J., Yang, G., Li, R. & Wang, X. Metal artifacts reduction in kV-CT images with polymetallic dentures and complex metals based on MV-CBCT images in radiotherapy. *Scientific Reports* **13**, 8970 (2023).

[129] Schoepf, U. J., Becker, C. R., Ohnesorge, B. M. & Yucel, E. K. CT of Coronary Artery Disease. *Radiology* **232**, 18–37 (2004). URL https://doi.org/10.1148/radiol.2321030636.

[130] Yang, L., Xu, P. P., Schoepf, U. J., Tesche, C., Pillai, B., Savage, R. H., Tang, C. X., Zhou, F., Wei, H. D., Luo, Z. Q., Wang, Q. G., Zhou, C. S., Lu, M. J., Lu, G. M. & Zhang, L. J. Serial coronary CT angiography–derived fractional flow reserve and plaque progression can predict long-term outcomes of coronary artery disease. *European Radiology* **31**, 7110–7120 (2021). URL https://doi.org/10.1007/s00330-021-07726-y.

[131] Kurata, A., Fukuyama, N., Hirai, K., Kawaguchi, N., Tanabe, Y., Okayama, H., Shigemi, S., Watanabe, K., Uetani, T., Ikeda, S., Inaba, S., Kido, T., Itoh, T. & Mochizuki, T. On-site computed tomography-derived fractional flow reserve using a machine-learning algorithm—clinical effectiveness in a retrospective multicenter cohort—. *Circulation Journal* **83**, 1563–1571 (2019).

[132] Gao, Z., Wang, X., Sun, S., Wu, D., Bai, J., Yin, Y., Liu, X., Zhang, H. & de Albuquerque, V. H. C. Learning physical properties in complex visual scenes: An intelligent machine for perceiving blood flow dynamics from static CT angiography imaging. *Neural Networks* **123**, 82–93 (2020).

[133] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255 (Ieee, 2009).

[134] Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, vol. 2 (Springer, 2009).

[135] Chollet, F. Keras. https://keras.io (2015).

[136] Serrano-Antón, B., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., Massonis, G., Pendón, S., Pérez-Muñuzuri, V., González-Juanatey, J. R. & P. Muñuzuri, A. Optimal Coronary Artery Segmentation Based on Transfer Learning and UNet Architecture. In *International Workshop on Shape in Medical Imaging*, 55–64 (Springer, 2023).

[137] Cheung, W. K., Bell, R., Nair, A., Menezes, L. J., Patel, R., Wan, S., Chou, K., Chen, J., Torii, R., Davies, R. H., Moon, J. C., Alexander, D. C. & Jacob, J. A Computationally Efficient Approach to Segmentation of the Aorta and Coronary Arteries Using Deep Learning. *IEEE Access* **9**, 108873–108888 (2021).

[138] Ward Jr, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58**, 236–244 (1963).

[139] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[140] Borůvka, O. O jistém problému minimálním. *Práce Moravské přírodovědecké společnosti* **3**, (36)–58 (1926). URL http://hdl.handle.net/10338.dmlcz/500114. Zentralblatt ID: JFM 57.1343.06.

[141] Nešetřil, J., Milková, E. & Nešetřilová, H. Otakar Borůvka on Minimum Spanning Tree Problem: Translation of Both the 1926 Papers, Comments, History. *Discrete Mathematics* **233**, 3–36 (2001).

[142] Tomandl, B., Hastreiter, P., Eberhardt, K., Rezk-Salama, C., Nimsky, C. & Buchfelder, M. The "Kissing Vessel"-artifact: A problem occurring in the visualization of intracranial aneurysms using volume rendering and virtual endoscopy. In *Radiology*, vol. 213, 311–311 (1999).

[143] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 424–432 (Springer, 2016).

[144] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B. & Rueckert, D. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv Preprint: 1804.03999* (2018).

[145] Wang, Q., Xu, L., Wang, L., Yang, X., Sun, Y., Yang, B. & Greenwald, S. E. Automatic coronary artery segmentation of CCTA images using UNet with a local contextual transformer. *Frontiers in Physiology* **14**, 1138257 (2023).

[146] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R. & Xu, D. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 272–284. Springer (Springer International Publishing, 2022).

[147] Patro, S. & Sahu, K. K. Normalization: A preprocessing stage. *arXiv Preprint: 1503.06462* (2015).

[148] MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, vol. 5, 281–298 (University of California press, 1967).

[149] Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, vol. 96, 226–231 (1996).

[150] Ng, A., Jordan, M. & Weiss, Y. On Spectral Clustering: Analysis and an algorithm. In Dietterich, T., Becker, S. & Ghahramani, Z. (eds.) *Advances in Neural Information Processing Systems*, vol. 14 (MIT Press, 2001). URL https://proceedings.neurips.cc/paper_files/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf.

[151] Liu, L., Yao, Y., Sun, N. & Han, G. Fully automated segmentation of coronary lumen based on the directional minimal path and image fusion. In *Proceedings of the 2017 6th International Conference on Computer Science and Network Technology (ICCSNT)*, 439–442 (2017).

[152] Huang, Y., Yang, J., Sun, Q., Ma, S., Yuan, Y., Tan, W., Cao, P. & Feng, C. Vessel filtering and segmentation of coronary CT angiographic images. *International Journal of Computer Assisted Radiology and Surgery* **17**, 1879–1890 (2022).

[153] Du, H., Shao, K., Bao, F., Zhang, Y., Gao, C., Wu, W. & Zhang, C. Automated coronary artery tree segmentation in coronary CTA using a multiobjective clustering and toroidal model-guided tracking method. *Computer Methods and Programs in Biomedicine* **199**, 105908 (2021).

[154] Huang, K., Tejero-de Pablos, A., Yamane, H., Kurose, Y., Iho, J., Tokunaga, Y., Horie, M., Nishizawa, K., Hayashi, Y., Koyama, Y. & Harada, T. Coronary Wall Segmentation in CCTA Scans Via a Hybrid Net with Contours Regularization. In *Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 1743–1747 (IEEE, 2020).

[155] Gu, J., Fang, Z., Gao, Y. & Tian, F. Segmentation of coronary arteries images using global feature embedded network with active contour loss. *Computerized Medical Imaging and Graphics* **86**, 101799 (2020).

[156] Achenbach, S., Delgado, V., Hausleiter, J., Schoenhagen, P., Min, J. K. & Leipsic, J. A. SCCT expert consensus document on computed tomography imaging before transcatheter aortic valve implantation (TAVI)/transcatheter aortic valve replacement (TAVR). *Journal of Cardiovascular Computed Tomography* **6**, 366–380 (2012).

[157] Otsuka, K., Ishikawa, H., Kono, Y., Oku, S., Yamaura, H., Shirasawa, K., Hirata, K., Shimada, K., Kasayuki, N. & Fukuda, D. Aortic arch plaque morphology in patients with coronary artery disease undergoing coronary computed tomography angiography with wide-volume scan. *Coronary Artery Disease* **33**, 531–539 (2022).

[158] Kim, W.-K., Renker, M., Rolf, A., Liebetrau, C., Van Linden, A., Arsalan, M., Doss, M., Rieck, J., Opolski, M. P., Möllmann, H., , Walther, T. & Hamm, C. W. Accuracy of device landing zone calcium volume measurement with contrast-enhanced multidetector computed tomography. *International Journal of Cardiology* **263**, 171–176 (2018).

[159] Eberhard, M., Mastalerz, M., Frauenfelder, T., Tanner, F. C., Maisano, F., Nietlispach, F., Seifert, B., Alkadhi, H. & Nguyen-Kim, T. D. L. Quantification of aortic valve calcification on contrast-enhanced CT of patients prior to transcatheter aortic valve implantation. *Eurointervention* **13**, 921–927 (2017).

[160] Leber, A. W., Kasel, M., Ischinger, T., Ebersberger, U. H., Antoni, D., Schmidt, M., Riess, G., Renz, V., Huber, A., Helmberger, T. & Hoffmann, E. Aortic valve calcium score as a predictor for outcome after TAVI using the CoreValve revalving system. *International Journal of Cardiology* **166**, 652–657 (2013).

[161] Hong, C., Bae, K. T. & Pilgram, T. K. Coronary Artery Calcium: Accuracy and Reproducibility of Measurements with Multi–Detector Row CT—Assessment of Effects of Different Thresholds and Quantification Methods. *Radiology* **227**, 795–801 (2003).

[162] Hong, C., Becker, C. R., Schoepf, U. J., Ohnesorge, B., Bruening, R. & Reiser, M. F. Coronary Artery Calcium: Absolute Quantification in Nonenhanced and Contrast-enhanced Multi–Detector Row CT Studies. *Radiology* **223**, 474–480 (2002).

[163] Ohnesorge, B., Flohr, T., Fischbach, R., Kopp, A., Knez, A., Schröder, S., Schöpf, U., Crispin, A., Klotz, E., Reiser, M. & Becker, C. Reproducibility of coronary calcium quantification in repeat examinations with retrospectively ECG-gated multisection spiral CT. *European Radiology* **12**, 1532–1540 (2002).

[164] Santos, R. D., Rumberger, J. A., Budoff, M. J., Shaw, L. J., Orakzai, S. H., Berman, D., Raggi, P., Blumenthal, R. S. & Nasir, K. Thoracic aorta calcification detected by electron beam tomography predicts all-cause mortality. *Atherosclerosis* **209**, 131–135 (2010).

[165] Vahanian, A., Beyersdorf, F., Praz, F., Milojevic, M., Baldus, S., Bauersachs, J., Capodanno, D., Conradi, L., De Bonis, M., De Paulis, R., Delgado, V., Freemantle, N., Gilard, M., Haugaa, K. H., Jeppsson, A., Jüni, P., Pierard, L., Prendergast, B. D., Sádaba, J. R., Tribouilloy, C., Wojakowski, W., ESC/EACTS Scientific Document Group & ESC National Cardiac Societies. 2021 ESC/EACTS Guidelines for the management of valvular heart disease: Developed by the Task Force for the management of valvular heart disease of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS). *European Heart Journal* **43**, 561–632 (2022).

[166] van der Bijl, N., Joemai, R. M., Geleijns, J., Bax, J. J., Schuijf, J. D., de Roos, A. & Kroft, L. J. Assessment of Agatston Coronary Artery Calcium Score Using Contrast-Enhanced CT Coronary Angiography. *American Journal of Roentgenology* **195**, 1299–1305 (2010).

[167] Guilenea, F. N., Casciaro, M. E., Pascaner, A. F., Soulat, G., Mousseaux, E. & Craiem, D. Thoracic Aorta Calcium Detection and Quantification Using Convolutional Neural Networks in a Large Cohort of Intermediate-Risk Patients. *Tomography* **7**, 636–649 (2021).

[168] Graffy, P. M., Liu, J., O'Connor, S., Summers, R. M. & Pickhardt, P. J. Automated segmentation and quantification of aortic calcification at abdominal CT: application of a deep learning-based algorithm to a longitudinal screening cohort. *Abdominal Radiology* **44**, 2921–2928 (2019).

[169] Alqahtani, A. M., Boczar, K. E., Kansal, V., Chan, K., Dwivedi, G. & Chow, B. J. Quantifying Aortic Valve Calcification using Coronary Computed Tomography Angiography. *Journal of Cardiovascular Computed Tomography* **11**, 99–104 (2017).

[170] Yoon, H.-C., Greaser III, L. E., Mather, R., Sinha, S., McNitt-Gray, M. F. & Goldin, J. G. Coronary artery calcium: Alternate methods for accurate and reproducible quantitation. *Academic Radiology* **4**, 666–673 (1997).

[171] Bettinger, N., Khalique, O. K., Krepp, J. M., Hamid, N. B., Bae, D. J., Pulerwitz, T. C., Liao, M., Hahn, R. T., Vahl, T. P., Nazif, T. M., George, I., Leon, M. B., Einstein, A. J. & Kodali, S. K. Practical determination of aortic valve calcium volume score on contrast-enhanced computed tomography prior to transcatheter aortic valve replacement and impact on paravalvular regurgitation: Elucidating optimal threshold cutoffs. *Journal of Cardiovascular Computed Tomography* **11**, 302–308 (2017).

[172] Kofler, M., Meyer, A., Schwartz, J., Sündermann, S., Penkalla, A., Solowjowa, N., Klein, C., Unbehaun, A., Falk, V. & Kempfert, J. A new calcium score to predict paravalvular leak in transcatheter aortic valve implantation. *European Journal of Cardio-Thoracic Surgery* **59**, 894–900 (2021).

[173] Shahoveisi, F., Taheri Gorji, H., Shahabi, S., Hosseinirad, S., Markell, S. & Vasefi, F. Application of image processing and transfer learning for the detection of rust disease. *Scientific Reports* **13**, 5133 (2023).

[174] Hu, J., Shen, L. & Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 7132–7141 (2018).

[175] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. Attention Is All You Need (2023).

[176] Hendrycks, D. & Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv preprint:1606.08415* (2016).

He dicho. Caso cerrado.

– Dra. Ana María Polo

Medical imaging is crucial for non-invasive diagnosis, treatment planning, and image-guided interventions, yet accurate analysis requires advanced processing techniques. This thesis focuses on automating segmentation in computed tomography (CT), specifically for coronary geometries and aortic calcifications. Automation enhances consistency, accelerates diagnosis, and enables scalable, reproducible analysis, facilitating data-driven and personalized clinical decision-making.
By leveraging artificial intelligence, this work improves segmentation accuracy and addresses challenges such as CT artifacts. The integration of automated processes into clinical workflows optimizes operations, minimizes manual intervention, and enhances the reliability of medical imaging analysis in real-world applications.