IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 December 2023, accepted 4 January 2024, date of publication 11 January 2024, date of current version 20 February 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3353142*

RESEARCH ARTICLE

# No More Training: SAM's Zero-Shot Transfer Capabilities for Cost-Efficient Medical Image Segmentation

**JUAN D. GUTIÉRREZ[1], ROBERTO RODRIGUEZ-ECHEVERRIA[2], EMILIO DELGADO[2], MIGUEL ÁNGEL SUERO RODRIGO[3], AND FERNANDO SÁNCHEZ-FIGUEROA[2]**

[1]Department of Electronics and Computer Science, Universidad de Santiago de Compostela, 27002 Lugo, Spain
[2]i3 Laboratory @ Quercus Research Group, Department of Computer Systems Engineering and Telematics, Universidad de Extremadura, 10003 Cáceres, Spain
[3]Servicio Extremeño de Salud, Hospital Universitario de Cáceres, 10004 Cáceres, Spain

Corresponding author: Juan D. Gutiérrez (juandiego.gutierrez@usc.es)

This work was supported in part by MCIN/AEI/10.13039/501100011033 under Grant CPP2021-008491, and in part by the European Union NextGenerationEU/PRTR.

**ABSTRACT** Semantic segmentation of medical images presents an enormous potential for diagnosis and surgery. However, achieving precise results involves designing and training complex Deep Learning (DL) models specifically for this task, which is only available to some. SAM is a model developed by Meta capable of segmenting objects present in virtually any type of image. This paper showcases SAM's robustness and exceptional performance in medical image segmentation, even in the absence of direct training on these image types (lung Computed Tomographies (CTs) and chest X-rays, in particular). Additionally, it achieves this impressive outcome while requiring minimal user intervention. Although the dataset used to train SAM does not contain a single sample of both medical image types, processing a popular dataset comprised of 20 volumes with a total of 3520 slices using the ViT-L version of the model yields an average Jaccard index of 91.45 % and an average Dice score of 94.95 %. The same version of the model achieves a 93.19 % Dice score and a 87.45 % Jaccard index when segmenting a frequently-used chest X-ray dataset. The values obtained are above the 70 % mark recommended in the literature, and close to state-of-the art models developed specifically for medical segmentation. These results are achieved without user interaction by providing the model with positive prompts based on the masks of the dataset used and a negative prompt located in the center of bounding box that contains the masks.

**INDEX TERMS** Image segmentation, deep learning, zero-shot learning, medical imaging, semantic segmentation.

## I. INTRODUCTION

Meta[1] has recently introduced SAM, a model capable of segmenting images [1]. Described by its authors as a foundation model [2], SAM has been trained with a dataset of approximately 11 million images, with more than 1 billion segmentation masks. A foundation model is trained with a massive amount of data and can be used as part of a process in tasks of different natures. Moreover, it is a zero-shot transfer model [3], which can segment images in domains not trained for. The power of these models lies in their ability to perform specific tasks for which they have not been originally trained, thanks to having been trained with vast amounts of starting data and being able to generalize the problem. At its core, zero-shot generalization relies on leveraging latent representations and learning abstract features during training. These learned representations encode high-level information about various concepts and allow the model to make inferences and predictions in unseen scenarios. This model has various potential applications. For instance, it can take input prompts from different systems. In the future,

The associate editor coordinating the review of this manuscript and approving it for publication was Essam A. Rashed.

[1]https://ai.facebook.com/

it could utilize a user's gaze from an AR/VR headset to select an object. It can also enable text-to-object segmentation using bounding box prompts from an object detector. Additionally, it can generate output masks that other AI systems can use as inputs. However, one of the most valuable applications could be democratizing the segmentation of medical images.

Semantic segmentation of medical images consists of clearly identifying a desired organ or tissue within a given medical image. The potential of this operation is enormous, and it can be used for diagnosing [4] or performing more precise and less painful surgeries for the patient [5]. There is currently a significant movement in the application of techniques in this field of research [6], [7].

The SA-1B dataset used to train SAM does not contain a single medical image. Given that its authors describe it as a foundation model with zero-shot transfer capabilities, the question of its capability to perform high-quality medical image segmentation with minimum interaction from the user arises. If so, this would lead to the development of cheap interactive segmentation tools everyone could use in desktop and mobile environments.

Segmentation quality is measured using similarity values such as the Jaccard index or the Dice score, which will be described later. Values above the 70 percent threshold [8] are considered acceptable for these metrics. Therefore, the performance of SAM will be acceptable if the values obtained exceed that mark.

The effectiveness of SAM in performing the segmentation of a specific organ, namely the lung, in a CT or an X-ray is of particular interest. The decision to focus on lung medical imagery comes from the enormous interest this organ has aroused in the world of (AI)-aided segmentation recently, given that it holds the key to rapid detection of COVID-19, as well as determining the damage caused by it [9]. It is essential to consider the comparison of its performance with existing models in the medical imaging field.

In some of these works, a preprocessing phase highlights the lungs from the rest in the image. Traditional image processing techniques [10] or models trained for that specific purpose [11] can be used for this purpose. Using traditional image processing techniques only sometimes yields the desired results, and iterative adjustments could be necessary. In this process, the specialist's experience in performing the segmentation is vital to obtain satisfactory results. On the other hand, training models specifically for medical image segmentation requires extensive technical knowledge, large datasets, and high processing capacity, which is only available to some.

In summary, employing SAM for lung segmentation presents distinct advantages over conventional methodologies. It could be a compelling substitution by notably reducing the demanding nature of training models specifically crafted for medical image segmentation. Including SAM in the pipeline could circumvent the necessity for extensive technical expertise, expansive datasets, and substantial computational power, all resources accessible to a limited cohort. Moreover, SAM's reduced latency introduces the prospect of crafting interactive utilities for radiological specialists, augmenting their capabilities in the segmentation procedure.

Given the challenges posed by traditional techniques and the resource-intensive nature of training models for medical image segmentation, this work aims to compare how SAM performs and identify the factors influencing its efficacy for this critical task.

The remainder of the paper is organized as follows to address the questions above and to stimulate the discussion on whether the effort should be focused on generating foundation models or creating specific ones. Section II analyzes the results other works achieve when faced with the same problem. Section III describes the grounds on which the tests performed are based: which datasets have been used, how segmentation works with SAM, what previous preprocessing is performed with the images, what data is necessary to provide to SAM in order to perform the segmentation, i.e., the prompts employed, and the metrics used to measure the segmentation quality. In Section IV, the results obtained in the experimental tests are compared with those achieved by systems explicitly developed for lung segmentation. These results are further commented in Section V, where some specificities are discussed. Finally, in Section VI, these results are analyzed, and future lines of research are proposed.

## II. RELATED WORKS

Medical image segmentation remains a critical area in healthcare, demanding precise and adaptable models for accurate delineation of anatomical structures and pathological regions. Within this domain, SAM has already drawn attention due to its potential utility across diverse medical imaging datasets. Researchers have conducted extensive evaluations to measure SAM's performance, revealing varying degrees of success and challenges across different imaging modalities. Understanding the efficacy of SAM in this context is crucial, given its implications for enhancing medical image analysis and diagnosis. This section presents a comprehensive review of pertinent studies that assess SAM's applicability, performance, and potential in medical image segmentation.

First, Mazurowski et al. [12] evaluate the performance of SAM on a collection of 11 medical imaging sets encompassing diverse modalities and anatomical structures. They conclude that SAM's performance dramatically varies depending on the nature of the images, ranging from impressive (above 80 % when segmenting the ilium in X-ray hip images) to poor (around 10 % when segmenting gray matter in Magnetic Resonance Imaging (MRI) spine images). For example, in their test, SAM reaches a Jaccard index of around 50 % with the Montgomery chest X-ray dataset, while our method reaches 87.45 % using the same dataset.

After measuring the performance of SAM on 12 public medical image segmentation datasets, He et al. [13]

concludes that this model does not perform as accurately as models created specifically for this task. In their tests on the Montgomery chest X-ray dataset, SAM performance reached a 60.52 % Dice score. In contrast, the approach presented in this work reaches a 93.19 % in the same metric. The datasets and image segmentation architectures used for this study are very complete. However, there is room for improvement in how SAM has been used.

In Shi et al. [14], the authors investigate nine medical image segmentation benchmarks encompassing diverse imaging modalities (e.g., Optical Coherence Tomography (OCT), MRI, CT) and applications in dermatology, ophthalmology, and radiology. The authors of this work reach conflicting results. On the one hand, SAM fails when segmenting specific structured targets (e.g., blood vessels). In particular, when segmenting chest X-rays, the authors of this work achieved a 65.09 % Dice score and a 51.36 % Jaccard index. As a comparison, our system achieves a Dice score of 93.19 % and a Jaccard index of 87.45 % when segmenting the same kind of images. On the other, they think that SAM shows the potential to achieve the desired performance through a fine-tuning process.

In Ma and Wang [15], the authors take a different approach to the problem, developing a method to fine-tune SAM specifically for medical imaging. The result is MedSAM, which surprisingly only gets a 17.6 % Dice score. In our work, using SAM as Meta originally published it, we obtained Dice scores of 94.95 % and 93.19 % when segmenting lungs in CT and X-ray images, respectively. Also, they curate a dataset with around 200 000 masks over 11 types of medical images from different human organs and pathologies.

SAM capabilities to perform tumor segmentation, non-tumor tissue segmentation, and cell nuclei segmentation are tested in Deng et al. [16]. In it, the authors use the prompt types SAM accepts (one or more points, both positive and negative, and bounding boxes) in different combinations. Instead of a systematic approach to prompt selection like ours, prompts were randomly selected from manual annotations. Comparing SAM zero-shot capabilities with state-of-the-art domain-specific models suggests that the model achieves remarkable segmentation performance for large connected objects. However, it only sometimes achieves satisfying performance for dense instance object segmentation. These results confirm the findings presented in this work.

Lastly, Roy et al. [17] compare SAM with nnUNet capabilities to segment a comprehensive set of organs. Using random points and jittered box prompts, the results obtained fall behind those achieved by nnUNet in every organ type. Only when using a bounding box prompt with a moderate jitter SAM gets close to nnUNet. We strongly believe that a systematic approach to prompt selection, as presented in this work, could improve the results obtained while also democratizing the use of the Deep Learning models in medical image segmentation pipelines.

## III. METHODS

The preliminary steps to evaluate SAM's performance are described in the following subsections. The source code is available.[2]

### A. DATASETS

To evaluate the performance of SAM when segmenting medical images, two distinct popular public lung datasets have been selected, one with axial lung scans, and the other with frontal chest X-rays, so that the two imagery types have representation in the experiments. These datasets have been selected because of their availability, publication, and extensive use in other scientific work. Their recurrent use in comparative studies validates their relevance. Also, they are suitable for testing SAM's zero-shot capabilities because the dataset it was trained with did not include any medical image. Both datasets include masks of each lung, verified by radiology specialists, so properly labeled data is available for research purposes.

The first one, with 20 axial lung scans [18], contains two distinct subsets, one labeled ''coronacases'' (see Fig. 1a) and the other ''radiopaedia'' (see Fig. 1b). The quality of the CT slices in the first subset is better than in the second. The latter are cone beam CTs with lower resolution, hence the quality difference. Also, while ''coronacases'' subset only contains 512 px × 512 px slices, ''radiopaedia'' subset is more heterogeneous, containing nine 630 px × 630 px slices and one 630 px × 401 px. Each CT scan has an average of 176 slices. The selected dataset lets us evaluate the performance of SAM over around 3520 images. Owais et al. DMDF-Net's performance [11] is evaluated using this dataset.

The second one contains 138 frontal chest X-rays [19] in Portable Network Graphics (PNG) format (Fig. 1c), with corresponding left and right lung masks, also in PNG format. It will be referred to as ''montgomery'', as the X-ray set was collected in Montgomery County, Maryland, USA. This dataset is used in [20] to evaluate Chen et al. TransAttUnet's performance.

Following the idea presented in the introduction, both datasets contain disease traces. The first one, COVID-19 infection; the second one is tuberculosis. As such, the images present symptoms of infection that could negatively influence the segmentation process. One example of this can be seen in Fig. 1b, where the lower part of the left lung contains an infection (marked with a red arrow) that the specialists include inside the lung, but SAM considers it to be outside. This example stresses the necessity of providing the specialist with an interactive tool to deal with these segmentation issues easily.

### B. SAM

SAM is a model developed by Meta AI Research. It is capable of segmenting all the different objects present in an image with no interaction whatsoever from the user. Both the

[2]https://i3lab.unex.es/project/sam-mis/

model and the dataset are publicly available [1]. No medical image can be found in the dataset, which makes it ideal for evaluating its zero-shot transfer capabilities with the abovementioned lung imagery datasets.

Given an image, SAM can mark the different elements in it by itself. The result will be a series of masks for each possible object. These masks could overlap when there is more than one potential area to mark. Alternatively, the user can provide *flexible prompts* to narrow down the possible segmentation results. These prompts can be point coordinates or bounding boxes where the target object can be found (positive prompts or foreground) or points where the target is not (negative prompts or background). The bounding box does not crop the input image, it only instructs SAM where to focus its attention. SAM also accepts masks, but using them would defeat the research hypothesis presented in this work: that SAM can perform high-quality image segmentation with minimum interaction from the user.

In this work, the selected prompt combines minimal point coordinates within a bounding box. The reasoning behind this choice is to simulate an interactive tool in which the specialist could start by drawing a bounding box around the lungs, then touching the areas to be segmented (i.e., the lungs), and finish indicating what is part of the background in order to ignore it. Fig. 1 shows two examples of using these prompts with different slices of a CT, one for each subset of the chosen dataset (labeled "coronacases" and "radiopaedia"), and a chest X-ray. The orange box contains all the masks for the current slice. The green circles mark a point inside each lung, as centered as possible. The red circle marks the image's background, a point that should not be considered part of the lungs.

### C. IMAGE PROCESSING

Each of the datasets used in this work contains images in different formats. Even when they share a format, the characteristics of each sample are different. This section analyzes the peculiarities of each of them and the necessary steps to convert them to a common format.

Both the scans and the masks contained in the CT dataset are in Neuroimaging Informatics Technology Initiative (NIfTI) format.[3] Before we start processing them, they are converted to a more adequate format for the purpose of this work, as SAM is not capable of working directly with it.

Each CT scan comprises several slices corresponding to the axial samples taken from the patient. Each slice contains values distributed in a two-dimensional matrix in columns and rows. These values, known as Hounsfield Units (HUs), represent the X-rays' attenuation when passing through the material, using distilled water as a reference [21]. They are in the range $[-1000, +1000]$.

Depending on the part of the body to be studied and taking into account the nature of the tissues in question, it may be helpful to apply a contrast enhancement technique known as *windowing* [22]. For this, it is necessary to know the width and level of the window to be used. The window width determines which attenuation values we want to focus our attention on. The window level indicates where the center of the window is located. The result is that, with a window centered at its level, all values below and above the lead width are ignored.

In the dataset considered, the "coronacases" CT volumes are not windowed, while the radiopaedia ones are already windowed. Therefore, the former are properly windowed to get all in the same condition.

Finally, all CT slices are converted to grayscale, where each pixel can take values between 0 and 255.

X-rays themselves do not use HU but produce images where the brightness or darkness of the pixels represents the degree of X-ray absorption by different tissues. Consequently, no preprocessing steps are needed for the chest X-ray dataset.

### D. PROMPT SELECTION

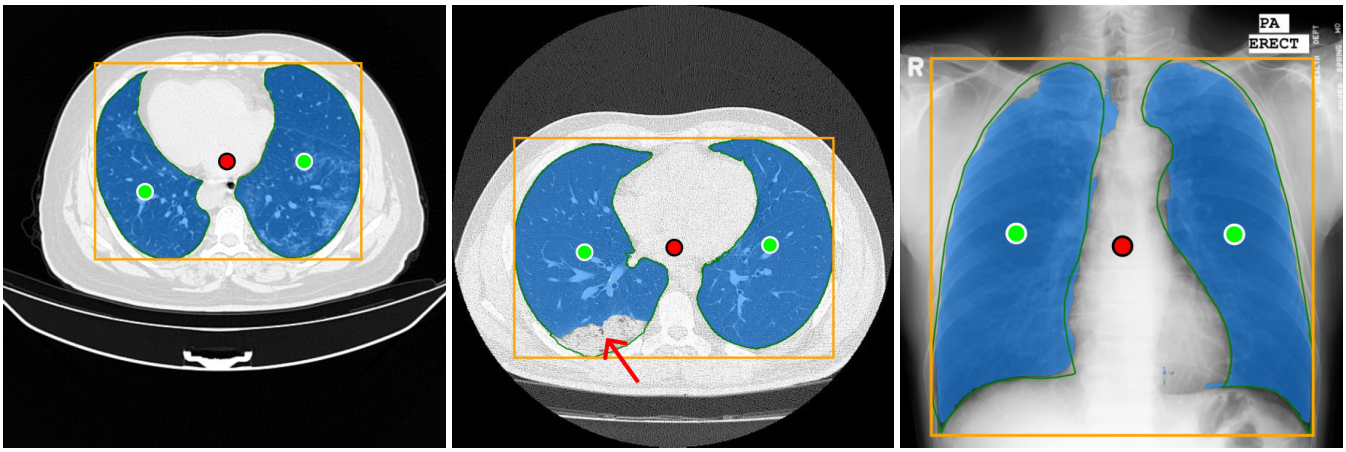
As explained at the end of Section III-B, point coordinates combined with a bounding box will be the prompts provided to SAM. To simulate the interactive selection of the points and the bounding box by a specialist, in this work we have defined the process described in Algorithm 1.

**Algorithm 1** Prompts Selection Algorithm

**Require:** $masks > 0$
  $points \leftarrow \varnothing$
  $labels \leftarrow \varnothing$
  $bounding\ box \leftarrow box(masks)$
  **for all** $mask \in masks$ **do**
    get *mask centroid*
    **if** $centroid \notin mask$ **then**
      move *centroid* inside *mask*
    **end if**
    add *centroid* to *points*
    add *positive* to *labels*
  **end for**
  add *bounding box center* to *points*
  add *negative* to *labels*

The selection process of the points used as prompts starts from the masks included in the datasets. First, for each sample, the bounding box containing all the masks is obtained. Then, each mask is processed, looking for its centroid. However, in some cases, given the lungs' shape, the centroid could be outside the mask. Fig. 2 shows how the algorithm operates in these cases. For the sake of simplicity, only one of the two centroids is shown. The centroid in Fig. 2a is the first attempt at getting a positive prompt. As Fig. 2b, it is outside the lung mask. This could be counterproductive because the prompt purpose is to guide SAM to be able to find the lungs. Our solution takes the centroid as a seed to find the intersections on the X and

[3] https://nifti.nimh.nih.gov/




**(a) Coronacases subset.** **(b) Radiopaedia subset.** **(c) Montgomery dataset.**

**FIGURE 1.** Prediction examples.

Y axes with the current mask, moving the centroid along the axis that contains the most points to the midpoint of that cut, thus placing the centroid inside the mask. The intersections shown in Fig. 2c are marked as 1, 2, and 3. Intersection 1 contains 34 pixels, intersection 2 contains 44, and intersection 3 contains 59 pixels. The centroid is then moved to the middle point of intersection 3, as shown in Fig. 2d. This change in the centroids is performed even when they are inside the masks, so they are better centered inside their corresponding masks.

To keep the proposed algorithm as simple as possible while still achieving its intended purpose, it might leave behind potential edge cases where the centroid may still fall outside the mask.

### E. METRICS

The quality of SAM's segmentation results will be measured using the Jaccard index *JCI* (1) and the Dice score *DSC* (2). Both metrics are commonly applied to compare how different two images $A$ and $B$ are.

$$JCI\ (A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

$$DSC\ (A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{2}$$

In Müller et al. [23], the authors argue that both the Jaccard index and the Dice score are presented as truthful and non-misleading metrics in evaluating the automatic segmentation of medical images. While Dice provides a balanced view between recall and accuracy, Jaccard provides a more severe penalty for under and over-segmentation, especially relevant when requiring a high degree of segmentation accuracy.

An example of the difference between the Jaccard index and the Dice score, and how the former is more severe than the latter, can be seen in Fig. 3. It shows two square areas. The side of the ground truth is 100 px long, while the prediction is 70 px. Their location relative to each other is shown next to both. With these conditions, the Jaccard index is 49.00 %, and the Dice score is 65.77 %.

In order to evaluate its performance, we will compare the values obtained by SAM in both metrics with the ones obtained in the literature, as there is no consensus on the threshold values that determine segmentation effectiveness [24].

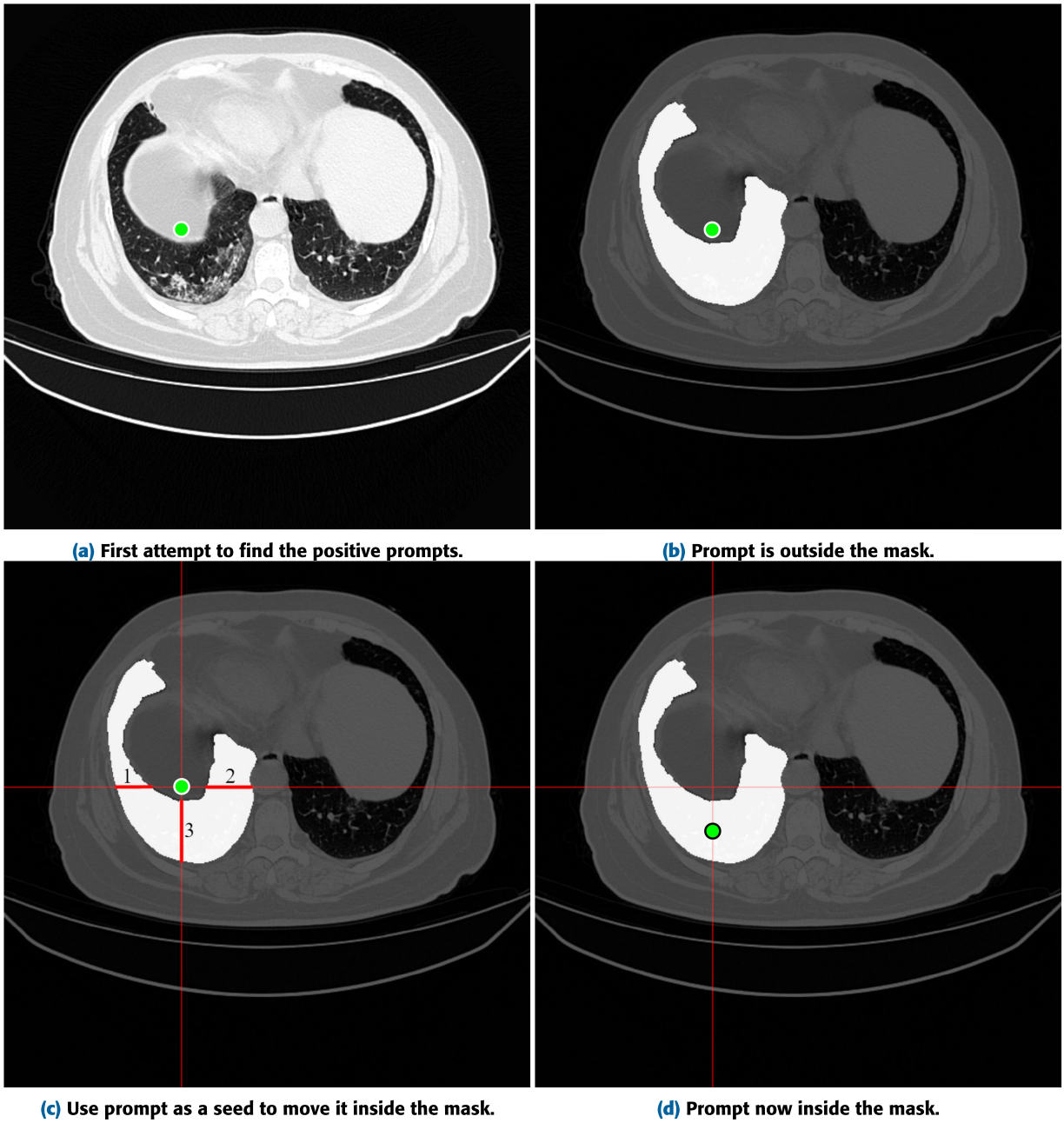
## IV. RESULTS

As Table 1's first two columns show, SAM obtains a better average Dice score than Jaccard index when segmenting lungs in CTs (93.69 % versus 89.97 %). The disparity of average values for the Dice score is also lower, as its average standard deviation is lower (9.69 % for Dice versus 12.76 % for Jaccard). Note that all the slices in the dataset have been considered despite they might contain a minimal lung area.

The results obtained when using SAM to segment chest X-rays are shown in Table 1's last two columns. Again, Dice score is better than Jaccard index (93.19 % versus 87.45 %), and its standard deviation is also lower (3.60 % versus 6.09 %). However, average values are still close to the 90 % mark.

Fig. 4 shows the quartile distribution of the segmentation results, divided by metric. Three subsets are presented: all volumes, only those labeled as ''coronacases'', and only those labeled as ''radiopaedia''. If only the ''coronacases'' subset were part of the experiment, the Jaccard index and Dice score results would clearly improve. Nevertheless, the global average is above the 90 % for both the Jaccard index and the Dice score. The Interquartile Range (IQR) of the Dice Score is above the 95 % in all the subsets, while the Jaccard index is above 92 % in the first two, but below for ''radiopaedia''.

The quartile distribution of SAM results segmenting Montgomery's chest X-rays is shown on Fig. 5. Although both Jaccard index and Dice score are close to the 90 % mark, SAM's performance is lower with X-rays than with CTs.




(a) First attempt to find the positive prompts.
(b) Prompt is outside the mask.
(c) Use prompt as a seed to move it inside the mask.
(d) Prompt now inside the mask.

**FIGURE 2.** Prompt location correction.

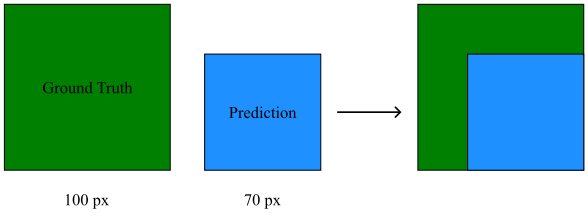



**FIGURE 3.** Ground truth versus prediction example.

The IQR for Dice scores is above the 90 % but starts below 85 % for Jaccard indexes. Also, minimum scores are lower for X-ray segmentation than for CT segmentation.

A rough generalization of the results, as shown in Fig. 4 and Fig. 5 is that the middle 50 % of all the data points are in the vicinity of the 90 % values. This result shows SAM's zero-shot transfer capabilities for medical image segmentation, as most results are positive.

A more detailed analysis of the results shows that Coronacases CT's results are best, followed by Radiopaedia's and X-rays. In the case of the CTs, the IQR is well above the 90 % for both the Dice score and the Jaccard index. However, there are more low results in the first quartile for the Radiopaedia subset than for the Coronacases one. X-ray segmentation results are not as good as CT ones,

**TABLE 1.** SAM's segmentation results.

| | Lung CT | | Chest X-ray | |
|---|---|---|---|---|
| | **Jaccard Index** | **Dice Score** | **Jaccard Index** | **Dice Score** |
| **Minimum** | 6.27 % | 11.79 % | 61.60 % | 76.24 % |
| **Maximum** | 98.84 % | 99.42 % | 95.94 % | 97.93 % |
| **Average** | 91.45 % | 94.95 % | 87.45 % | 93.19 % |
| **Standard Deviation** | 12.25 % | 9.30 % | 6.09 % | 3.60 % |

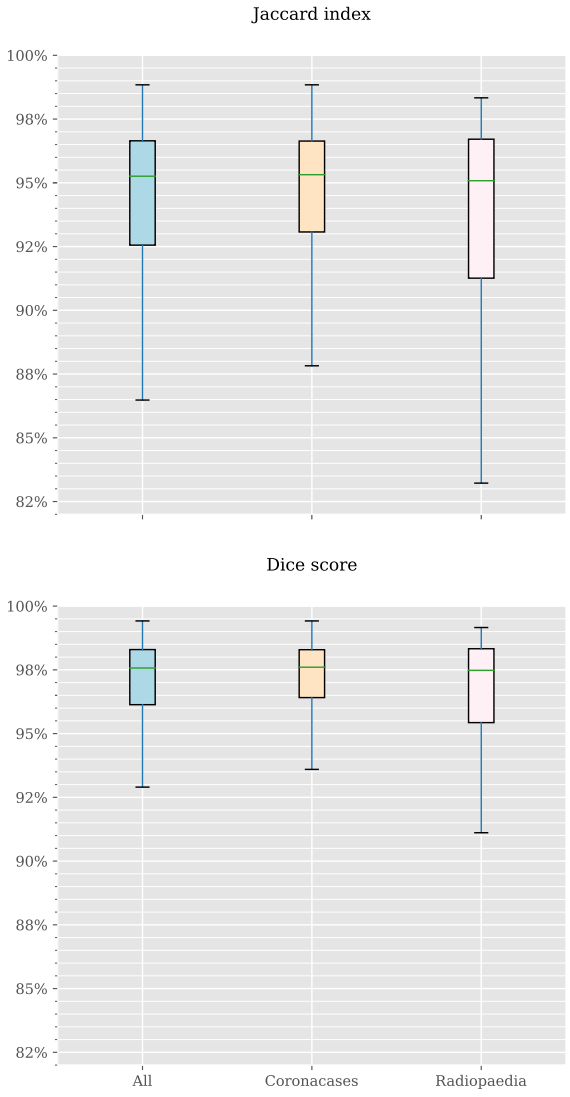



**FIGURE 4.** Lung CT segmentation results.

having the first quartile starting right above the 70 % mark for the Jaccard index and the 85 % mark for the Dice score. Nevertheless, the IQR is well positioned around the 90 % mark for both metrics. The samples with results in the first quartile shown in Fig. 4 and Fig. 5 could be used to create a subset suitable for improvement with prompt modifications, either by modifying the ones obtained with Algorithm 1 or by adding new ones.

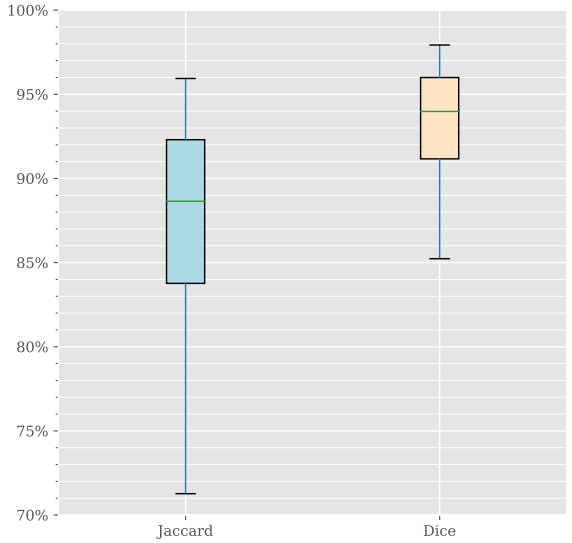



**FIGURE 5.** Chest X-ray segmentation results.

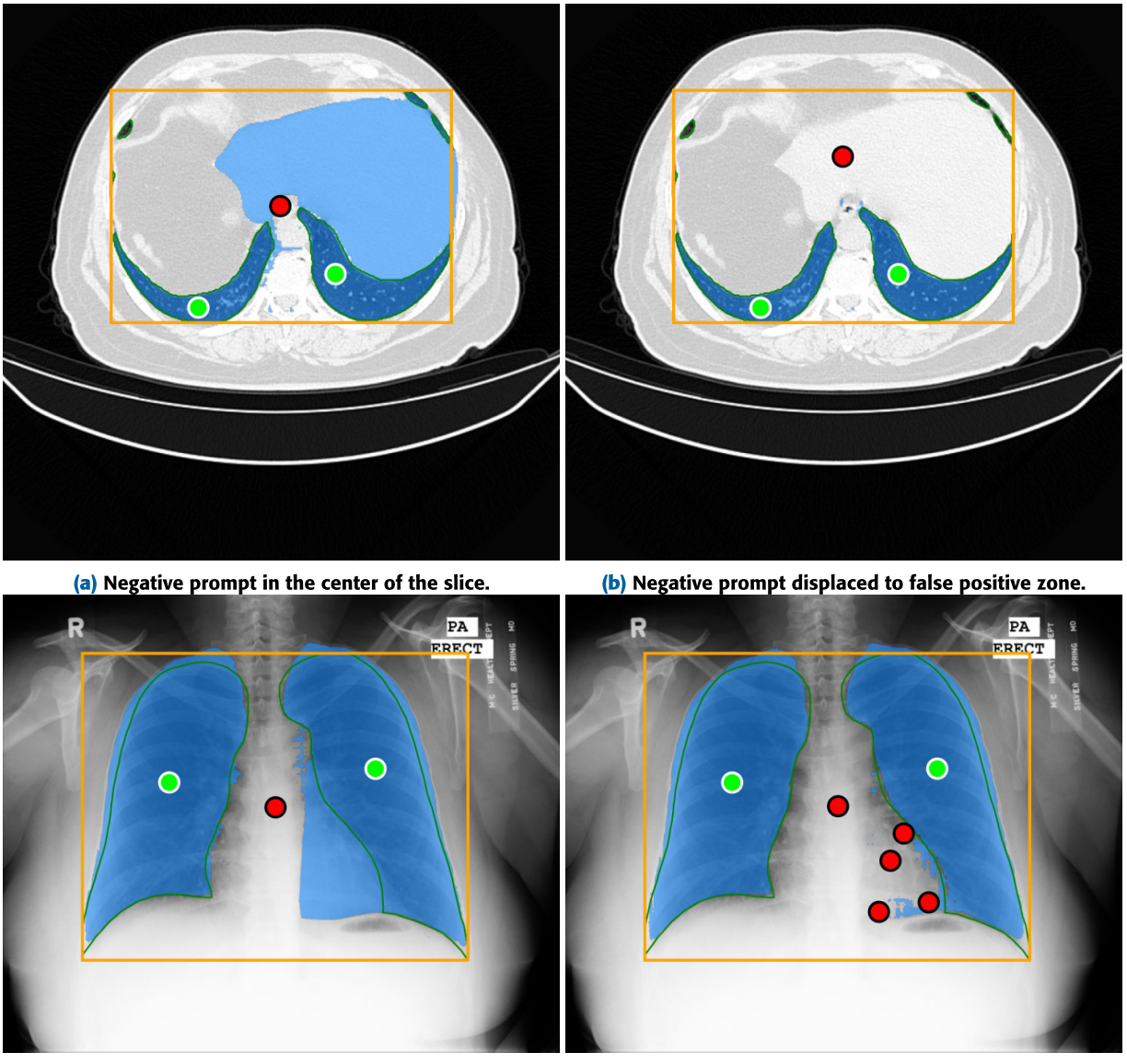
The median values shown in Fig. 4 are around 95 % for the Jaccard index and 98 % for the Dice score for both subsets. This statistical value is less affected by outliers that might skew the average values in a sequence. That is why median values are better in this particular case than average ones (around 91 % and 95 %, respectively).

In Table 2, the performance values obtained in the lung segmentation task by various state-of-the-art DL models can be compared with SAM's. Owais et al. [11] results are obtained with the same CT dataset used in this paper. While they reach a 97.38 % Jaccard index and a 98.66 % with their DMDF-Net, SAM obtains a 91.45 % Jaccard index and a 94.95 %. On the other hand, Chen et al. [20] employed the same X-ray dataset this work uses. In this case, they reach a 97.82 % Jaccard index and a 98.88 % using TransAttUnet, SAM obtains a 87.45 % Jaccard index and a 93.19 %. In both cases, custom solutions achieve better results than SAM. However, it is noteworthy to remember that although SAM has not been trained using medical images, its results are close to those obtained using custom-tailored solutions.

To close this section, it is worth noting the low processing latency of SAM (∼0.05 s per image on a web client [1]). Although this parameter highly depends on the execution

**TABLE 2.** Lung segmentation models' performance (SAM's performance delta).

| | Jaccard Index | Dice Score | Execution Time |
|---|---|---|---|
| **Owais et al.** [11] | 97.38 % (−5.93 %) | 98.66 % (−3.71 %) | 0.039 s (−0.011 s) |
| **Chen et al.** [20] | 97.82 % (−10.37 %) | 98.88 % (−5.69 %) | - |
| **Khanna et al.** [25] | 97.98 % | 98.98 % | - |
| **Hu et al.** [26] | - | 97.68 % | 11.2 s (11.15 s) |




**(a)** Negative prompt in the center of the slice.

**(b)** Negative prompt displaced to false positive zone.

**(c)** One single negative prompt.

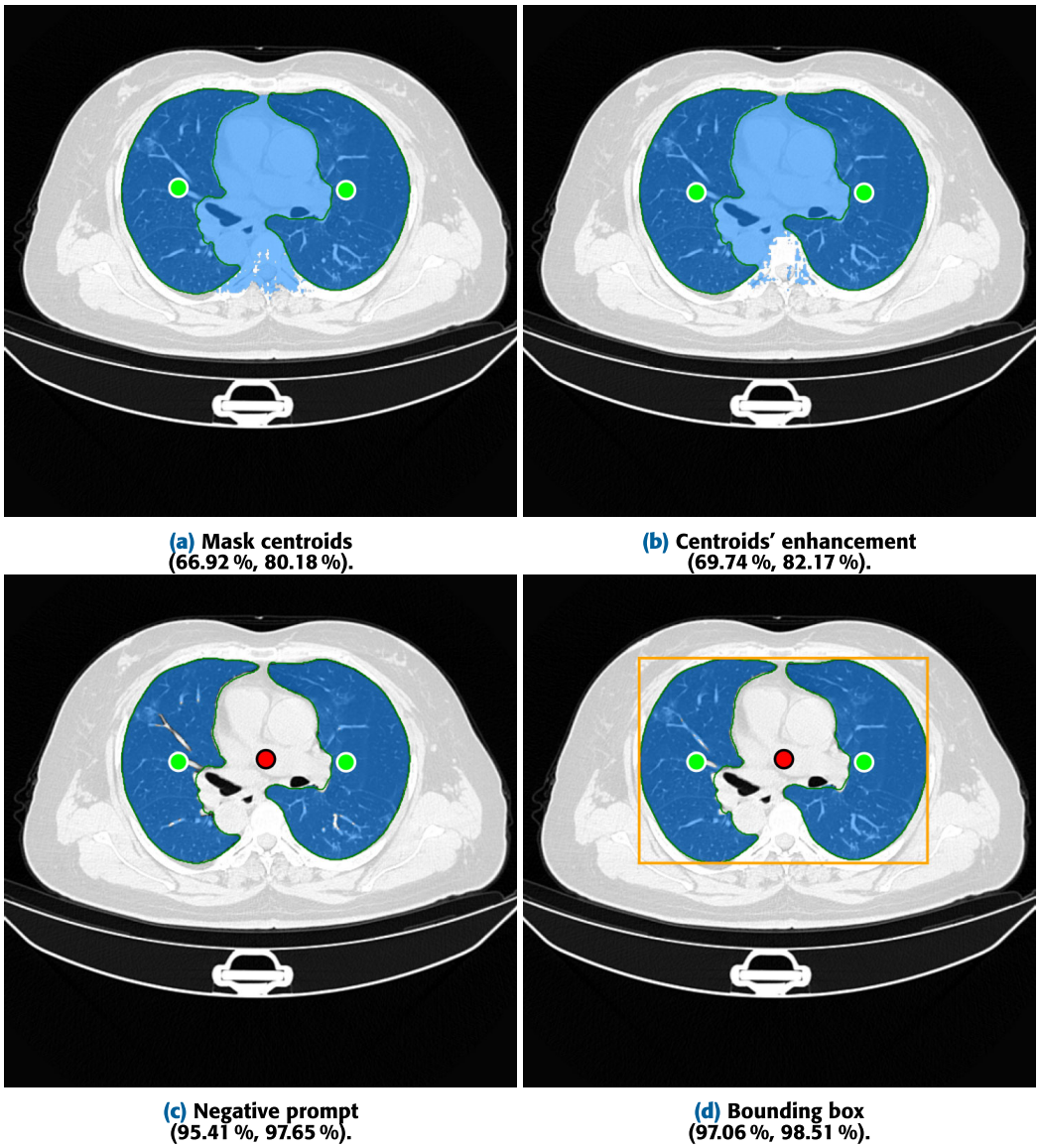
**(d)** Multiple negative prompts.

**FIGURE 6.** Improved results by changing negative prompt.

environment, there is data available from some of the works we are comparing ours with. Owais et al. [11] report that their system offers a throughput of 25.64 frames per second or, in other words, can process each image in approximately 0.039 s. Neither Chen et al. [20] nor Khanna et al. [25] mention the processing time of their systems. Finally, Hu et al. [26] report that the processing time per image is about 11.2 s.

## V. DISCUSSION

Given the results obtained in the experiments performed in this work, is SAM performance good enough for medical



**(a) Mask centroids (66.92 %, 80.18 %).** **(b) Centroids' enhancement (69.74 %, 82.17 %).**

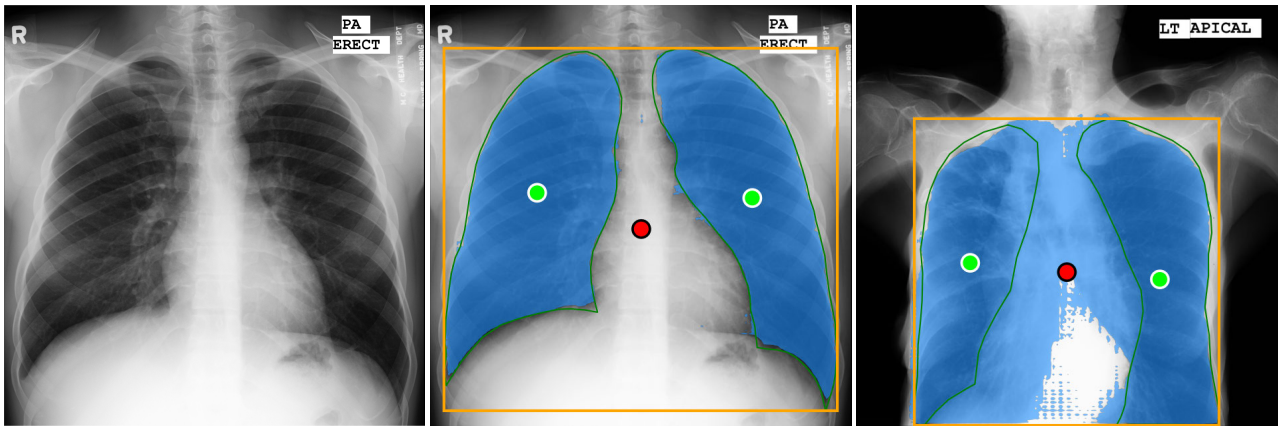
**(c) Negative prompt (95.41 %, 97.65 %).** **(d) Bounding box (97.06 %, 98.51 %).**

**FIGURE 7.** Segmentation results evolution (Jaccard index, Dice score) as the prompts provided change.

image segmentation, compared with models specifically created for said task shown in Table 2?

Firstly, there are some differences regarding the way data is processed. Although the goal of Owais et al. [11] is to segment infections caused by COVID-19, a preprocessing step consists of segmenting the lung itself is present in the pipeline. Using transfer learning with C-Blocks (5.85 M) on a network they call DMDF-Net-1, built on top of another Convolutional Neural Network (CNN) previously trained on a considerable collection of non-medical images. In [20], an attempt is made to build a general-purpose network for organ segmentation in medical imaging. Its results are obtained in sagittal slices rather than axial slices using the version of its architecture with residual connections. To achieve their results, Khanna et al. [25] has implemented a U-Net architecture with residual blocks to overcome a performance degradation problem. In addition, they have used different data augmentation techniques to improve the generalizability of their architecture.

The algorithm described in Algorithm 1 defines a narrow set of rules to perform the image segmentation. However, with minor modifications the results obtained could significantly improve. An example of correct segmentation can be seen in Fig. 1a, which shows slice 123 of the first volume of the CT dataset, with a Jaccard index of 98.37 % and a Dice score of 99.18 %. The green outline delimits the segmentation performed by the specialist (ground truth). In Fig. 6a can be seen the result obtained by SAM when working with slice 86 of the same volume, with a Jaccard index of 26.74 % and a Dice score of 42.20 %. This is the worst-performing slice of its volume, which shows why in some slices SAM performs worse than expected. In this particular case, the selected negative prompt is in the center of the bounding box containing the masks. However, this prompt falls outside




**(a) Unprocessed chest X-ray sample.** **(b) Best result (95.94 %, 97.93 %).** **(c) Worst result (61.60 %, 76.24 %).**

**FIGURE 8.** SAM's Montgomery dataset segmentation best and worst results (Jaccard index, Dice score).

of a potential shape that could be composed of both lungs. In Fig. 6b, it can be seen how it improves the result obtained in the same slice if the negative prompt were over the area incorrectly segmented as part of the lungs, breaking the potential shape, and obtaining now a Jaccard index of 90.57 % and a Dice score of 95.05 %.

For this experiment, prompts have been arbitrarily selected. The positive prompts should be inside the lungs, the negative at the center coordinates of the bounding box. If radiological specialists had an interactive tool, they could select these points so that the final result would be optimal, improving the results obtained for the Jaccard index and the Dice score. Fig. 6c presents the segmentation result of a sample from Montgomery's dataset, using one single negative prompt automatically selected using Algorithm 1. The segmentation result could be improved with additional negative prompts, as Fig. 6d demonstrates. In this particular example, using SAM with a single negative prompt reaches 78.85 % Jaccard index and 88.17 % Dice score. With multiple negative prompts, SAM performance improves, obtaining 88.30 % Jaccard index and 93.79 % Dice score.

The process described in Algorithm 1 is the result of the study of the influence of different prompts on the segmentation result. In Fig. 7, the results improve as the algorithm is modified. Fig. 7a only uses the centroids of the lung masks. These centroids are used as seeds, repositioned inside the mask if they were outside. In either case, whether inside or outside, they are centered inside the mask along the axis with the most points. As a result, in Fig. 7b, the centroids are a few rows down, which slightly improves the result. Using a negative prompt in Fig. 7c allows SAM to ignore the central part of the image and further improve the segmentation result. A final change, including a bounding box in the prompts, allows SAM to focus its attention on the lung part and provides a slight improvement, as can be seen in Fig. 7d.

Fig. 8a shows an unprocessed Montgomery dataset sample. The lung area is clearer in a CT than in an X-ray, leading to the conclusion that SAM's performance could be worse in the latter than in the former. Ribs are visible in the X-ray, and could interfere in SAM's results. As shown in Section IV, that is not the case. Although SAM segments CTs better than X-rays, in both cases the metrics values are close to the 90 % mark.

The best result obtained when working on the Montgomery dataset is shown in Fig. 8b, with 95.94 % Jaccard index and a 97.93 % Dice score. On the other hand, Fig. 8c shows the worst result, with 61.60 % Jaccard index and a 76.24 % Dice score.

## VI. CONCLUSION

The experimental results shown in this work demonstrate that it is possible to use SAM to perform medical image segmentation. With an average Jaccard index of 91.45 % and an average Dice score of 94.95 % when processing CTs, and an average 93.19 % Dice score and an average 87.45 % Jaccard index when segmenting chest X-rays, SAM is above the 70 % mark recommended in the literature [8] for this task. On top of that, the tests performed for this work SAM show results close to those of the best models trained specifically for lung segmentation, with an average Jaccard of 99.24 % and an average Dice score of 99.62 % for the CTs [25], and an average Jaccard of 97.82 % and an average Dice score of 98.88 % for the chest X-rays [20].

Currently, this model could be part of a medical image segmentation pipeline, helping specialists in their work. The selected datasets contain CTs of lungs and chest X-rays, together with the corresponding segmentation masks, which have been used as ground truth. It has only been necessary to provide SAM with the bounding box containing all the lung masks, the center points of each of these masks as a positive prompt (i.e., in this area is what you should find) and a point outside the lungs as a negative prompt (in this area is what you should not find).

As future work, given SAM's low processing latency (∼0.05 s per image in a web browser) compared to previous

works (11.2 s per image) [26], we plan to build an interactive tool to assess SAM performance with human feedback by means of a case study in a hospital. Moreover, additional experiments will be performed to evaluate the segmentation of more types of organs. We must be aware of certain potential limitations of using SAM for medical image segmentation tasks. When using SAM's zero-shot capabilities, it may not perform as well as a model trained on specific medical datasets, as demonstrated in the results section. Because SAM was not trained with images of that nature, it might lack a precise understanding of medical images, as they can be intricate, with subtle variations and nuances. Besides, SAM might struggle to generalize effectively across different types of medical data due to the complexity and variability within these images. Exploring more datasets of different natures will allow us to test SAM's generalizability on a larger scale. Furthermore, we also plan fine-tuning the model by expanding the original dataset with examples of the field to be applied [27]. Besides, data augmentation and its impact on SAM's performance while fine-tuning the model will also be explored.

## REFERENCES


[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023, *arXiv:2304.02643*.

[2] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[3] M.-W. Chang, L. Ratinov, D. Roth, and V. Srikumar, "Importance of semantic representation: Dataless classification," in *Proc. AAAI*, vol. 2, 2008, pp. 830–835.

[4] X. Fan and X. Feng, "SELDNet: Sequenced encoder and lightweight decoder network for COVID-19 infection region segmentation," *Displays*, vol. 77, Apr. 2023, Art. no. 102395.

[5] D. R. Bacon and D. J. Wilkinson, "Great moments in the history of anesthesiology," *Στο: Wylie & Churchill-Davidson's A Practice of Anesthesia*, T. E. J. Healy and C. P. J. Wylie, Eds., 7th ed. London, U.K.: Arnold, 2003, pp. 1–16.

[6] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102802.

[7] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, and D. Merhof, "Advances in medical image analysis with vision transformers: A comprehensive review," 2023, *arXiv:2301.03505*.

[8] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: Method and validation," *IEEE Trans. Med. Imag.*, vol. 13, no. 4, pp. 716–724, Dec. 1994.

[9] J. Prada, Y. Gala, and A. L. Sierra, "COVID-19 mortality risk prediction using X-ray images," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 6, no. 6, p. 7, 2021.

[10] F. Lu, C. Tang, T. Liu, Z. Zhang, and L. Li, "Multi-attention segmentation networks combined with the Sobel operator for medical images," *Sensors*, vol. 23, no. 5, p. 2546, Feb. 2023.

[11] M. Owais, N. R. Baek, and K. R. Park, "DMDF-Net: Dual multiscale dilated fusion network for accurate segmentation of lesions related to COVID-19 in lung radiographic scans," *Expert Syst. Appl.*, vol. 202, Sep. 2022, Art. no. 117360.

[12] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," *Med. Image Anal.*, vol. 89, Oct. 2023, Art. no. 102918.

[13] S. He, R. Bao, J. Li, J. Stout, A. Bjornerud, P. Ellen Grant, and Y. Ou, "Computer-vision benchmark segment-anything model (SAM) in medical images: Accuracy in 12 datasets," 2023, *arXiv:2304.09324*.

[14] P. Shi, J. Qiu, S. M. D. Abaxi, H. Wei, F. P.-W. Lo, and W. Yuan, "Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation," *Diagnostics*, vol. 13, no. 11, p. 1947, Jun. 2023.

[15] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Commun.*, vol. 15, no. 1, p. 654, Jan. 2024.

[16] R. Deng, C. Cui, Q. Liu, T. Yao, L. W. Remedios, S. Bao, B. A. Landman, L. E. Wheless, L. A. Coburn, K. T. Wilson, Y. Wang, S. Zhao, A. B. Fogo, H. Yang, Y. Tang, and Y. Huo, "Segment anything model (SAM) for digital pathology: Assess zero-shot segmentation on whole slide imaging," 2023, *arXiv:2304.04155*.

[17] S. Roy, T. Wald, G. Koehler, M. R. Rokuss, N. Disch, J. Holzschuh, D. Zimmerer, and K. H. Maier-Hein, "SAM.MD: Zero-shot medical image segmentation capabilities of the segment anything model," 2023, *arXiv:2304.05396*.

[18] M. Jun, G. Cheng, W. Yixin, A. Xingle, G. Jiantao, Y. Ziqi, Z. Minqing, L. Xin, D. Xueyuan, C. Shucheng, W. Hao, M. Sen, Y. Xiaoyu, N. Ziwei, Li Chen, T. Lu, Z. Yuntao, Z. Qiongjie, D. Guoqiang, and H. Jian, "COVID-19 CT lung and infection segmentation dataset," Dataset, 2020.

[19] S. Jaeger, S. Candemir, S. Antani, Y. X. Wang, P. X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant. Imag. Med. Surg.*, vol. 4, p. 475, Dec. 2014.

[20] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. Wai Kin Kong, "TransAttUnet: Multi-level attention-guided U-Net with transformer for medical image segmentation," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 1, pp. 55–68, Feb. 2024.

[21] D. R. Dance, S. Christofides, A. D. A. Maidment, I. D. McLean, and K. H. Ng, *Diagnostic Radiology Physics*. Vienna, Austria: Int. Atomic Energy Agency, 2014.

[22] A. S. Tidwell, "Advanced imaging concepts: A pictorial glossary of CT and MRI technology," *Clin. Techn. Small Animal Pract.*, vol. 14, no. 2, pp. 65–111, May 1999.

[23] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Res. Notes*, vol. 15, no. 1, p. 210, Dec. 2022.

[24] F. Kofler, I. Ezhov, F. Isensee, F. Balsiger, C. Berger, M. Koerner, B. Demiray, J. Rackerseder, J. Paetzold, H. Li, S. Shit, R. McKinley, M. Piraud, S. Bakas, C. Zimmer, N. Navab, J. Kirschke, B. Wiestler, and B. Menze, "Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient," 2021, *arXiv:2103.06205*.

[25] A. Khanna, N. D. Londhe, S. Gupta, and A. Semwal, "A deep residual U-Net convolutional neural network for automated lung segmentation in computed tomography images," *Biocybern. Biomed. Eng.*, vol. 40, no. 3, pp. 1314–1327, 2020, doi: 10.1016/j.bbe.2020.07.007.

[26] Q. Hu, L. F. de F. Souza, G. B. Holanda, S. S. A. Alves, F. H. dos S. Silva, T. Han, and P. P. Rebouças Filho, "An effective approach for CT lung segmentation using mask region-based convolutional neural networks," *Artif. Intell. Med.*, vol. 103, Mar. 2020, Art. no. 101792.

[27] R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, X. Ma, H. Dong, P. Gao, and H. Li, "Personalize segment anything model with one shot," 2023, *arXiv:2305.03048*.


**JUAN D. GUTIÉRREZ** is currently an Assistant Professor with Universidad de Santiago de Compostela (USC). With more than 20 years of experience in the computer world, he recently presented his Ph.D. research on the field of visible LED light-based indoor positioning systems (IPS). His current research interests include the application of artificial intelligence to different fields of knowledge. His training includes programming in different languages, system administration, application design, and databases and the internet. He has written more than 20 computer science books and translated another ten from English into Spanish. What began as a fun experience in the mid 90's has ended up being a real passion for him. He enjoys computing but, above all, learning new things.

**ROBERTO RODRIGUEZ-ECHEVERRIA** is currently a Professor in software architecture with the Computer Languages and Systems Department, Universidad de Extremadura (UEx), Spain. His research interests include software engineering, model-driven engineering, data-driven software development, machine learning, web engineering, and legacy software modernization. He is also the Head of the Applied Informatics Technology Institute. Moreover, he truly believes in local socioeconomic value generation through entrepreneurship, so he has recently created a new UEx spin-off company, named MetrikaMedia, which defines itself as a SaaS solution for multimedia content measurement.

**EMILIO DELGADO** is currently a Researcher with Universidad de Extremadura, whose primary focus is machine learning, particularly the study of deep learning. His current research interests include the intersection of artificial intelligence and healthcare, where he is applying deep learning techniques to solve medical problems. His work aims to use these algorithms to process and analyze large amounts of clinical and medical imaging data to improve people's standard of living. He is constantly looking for ways to improve and optimize deep learning algorithms for application in medicine, striving to ensure that they are accurate, efficient, and useful for healthcare professionals. He is exploring how machine learning can be used to improve medical diagnoses and treatments and investigating how these systems can be designed and trained to respect patient privacy and data security.

**MIGUEL ÁNGEL SUERO RODRIGO** is currently a Medical Physicist with Servicio Extremeño de Salud (SES), Hospital Universitario de Cáceres (HUC). His primary work areas include radiotherapy, radiodiagnosis, nuclear medicine, and radiological protection. Previously, he was the Head of the Medical Physics Department, Virgen del Puerto Hospital, and has been a member of several task groups dedicated to quality assurance and technology in healthcare. He has presented various communications related to his work. In recent years, his research interests have focused on leveraging artificial intelligence for applications in the medical physics field, including image anonymization, image segmentation, workflow optimization, and the advancement of cancer radiation therapy and diagnosis.

**FERNANDO SÁNCHEZ-FIGUEROA** was born in Trujillo, in 1968. He is currently a Full Professor with the Department of Computer and Telematics Systems Engineering, UEx, and accredited as a Professor. He is also the President and the Co-Founder of the spin-off Homeria Open Solutions with which he has participated in three research and development projects of the VII EU Framework Program. He is the author of more than 50 scientific-technical articles and has signed about 20 research and development contracts with both private and public entities.