

INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Daniel
Cores Costa

PhD Thesis

Spatio-temporal convolutional neural networks for video object detection

Santiago de Compostela, 2022

Doctoral Programme in Information Technology Research

TESE DE DOUTORAMENTO

SPATIO-TEMPORAL CONVOLUTIONAL NEURAL NETWORKS FOR VIDEO OBJECT DETECTION

Daniel Cores Costa

ESCOLA DE DOUTORAMENTO INTERNACIONAL DA UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

PROGRAMA DE DOUTORAMENTO EN INVESTIGACIÓN EN TECNOLOXÍAS DA INFORMACIÓN

SANTIAGO DE COMPOSTELA
2022

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Declaración do autor da tese

Don Daniel Cores Costa

Título da tese: Spatio-temporal convolutional neural networks for video object detection.

Presento a miña tese, seguindo o procedemento adecuado ao Regulamento, e declaro que:

1. *A tese abarca os resultados da elaboración do meu traballo.*
2. *De ser o caso, na tese faise referencia ás colaboracións que tivo este traballo.*
3. *Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.*
4. *A tese é a versión definitiva presentada para a súa defensa e coincide a versión impresa coa presentada en formato electrónico.*

E comprométome a presentar o Compromiso Documental de Supervisión no caso de que o orixinal non estea na Escola.

En Santiago de Compostela, 21 de setembro de 2022

Asdo. Daniel Cores Costa

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Autorización do director/titor da tese
Spatio-temporal convolutional neural networks for video object detection

Don Víctor M. Brea Sánchez, Profesor Titular da Área de Electrónica da Universidade de Santiago de Compostela

Don Manuel Mucientes Molina, Profesor Titular da Área de Ciencia da Computación e Intelixencia Artificial da Universidade de Santiago de Compostela

INFORMAN:

*Que a presente tese correspóndese co traballo realizado por **Don Daniel Cores Costa**, baixo a nosa dirección/titorización, e autorizamos a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como directores/titores desta non incorre nas causas de abstención establecidas na Lei 40/2015.*

*De acordo co indicado no Regulamento de Estudos de Doutoramento, declaramos tamén que a presente tese de doutoramento é idónea para ser defendida en base á modalidade de **Monográfica con reproducción de publicacións**, nas que a participación do doutorando foi decisiva para a súa elaboración e que as publicacións se axústan ó Plan de Investigación.*

En Santiago de Compostela, 21 de setembro de 2022

Asdo. Víctor M. Brea Sánchez
Director/a tese

Asdo. Manuel Mucientes Molina
Director/a tese

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Luck is what happens when preparation meets opportunity

Seneca

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AGRADECEMENTOS

En primeiro lugar, quero expresar o meu mais sincero agradecemento aos meus directores de tese Manuel Mucientes Molina e Víctor M. Brea Sánchez, pola confianza e os consellos que me brindaron e que fixeron que este traballo fose posible.

Quixera agradecer tamén ao Departamento de Electrónica e Computación, así como ao Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS), da Universidade de Santiago de Compostela, pola excelente contorna de traballo e os recursos que fixeron posible o desenvolvemento desta tese. Tamén me gustaría mencionar ao Dr. Lorezo Seidenari, ao Prof. Alberto del Bimbo e a todo o grupo do MICC (Media Integration and Communication Center) da Universidade de Florencia por acollerme durante a miña estadía.

Non podía faltar un agradecemento especial as amizades que fixen durante esta etapa no CiTIUS, especialmente a todos os compañeiros de laboratorio. Tamén agradecer o apoio fundamental da miña familia, aos meus pais e ao meu irmán. Un agradecemento especial para Jessica por estar sempre para todo, nos bos e nos malos momentos.

Doutra banda, gustaríame agradecer a todas as entidades que fixeron posible o desenvolvemento desta tese co seu financiamento. Comezando polo programa de Axudas de apoio a etapa predoutoral da Xunta de Galicia (ref. ED481A-2018/146), e a Consellería de Educación, Universidades e Formación Profesional da Xunta de Galicia (ref. ED431C 2018/29, ED431C 2017/69 e acreditacion 2016-2019, ED431G/08, ambas co-financiadas polo Fondo Europeo de Desenvolvemento Rexional do programa ERDF/FEDER). Tamen expresar o meu agradecemento ao Ministerio de Economía, Industria y Competitividad (ref. TIN2017-84796-C2-1-R, e TIN2017-90567-REDT) e ao Ministerio de Ciencia, Innovacion y Universidades (ref. RTI2018-097088-B-C32 e PID2020-112623GB-I00). Este proxecto tamén recibiu financiamento do programa de investigación e innovación Horizonte 2020 da Unión Europea baixo o acordo No. 101016734. Finalmente, quixera agradecer a cesión de GPUs por parte de NVIDIA.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Resumo

A detección de obxectos é un dos temas de investigación máis relevantes en visión por computador. A resolución deste problema comprende dúas tarefas principais: localización e clasificación. A localización busca definir unha caixa delimitadora que enmarque os obxectos de interese na imaxe de entrada, mentres que a clasificación calcula as categorías asociadas a ditos obxectos, obtendo unha puntuación de confianza para cada detección. As posibles aplicacións van dende a videovixilancia ata ser un compoñente básico en sistemas máis complexos como os vehículos autónomos. En xeral, é necesario un detector de obxectos como paso fundamental en calquera sistema no que sexa relevante a extracción de información dos obxectos presentes na escena. Isto inclúe aplicacións como sistemas de monitorización de tráfico, procesamento de imaxes médicas, recoñecemento facial ou reconto de multitudes. A precisión alcanzada na detección, sufriu un incremento substancial nos últimos anos, principalmente debido á adopción das Redes Neuronais Convolucionais profundas (CNN, *Convolutional Neural Networks*). Non obstante segue sendo un campo de investigación moi activo con grandes retos por resolver.

A maior parte das aplicacións mencionadas reciben un fluxo de fotogramas de entrada en lugar dunha única imaxe. Resolver o problema de detección de obxectos en vídeo entraña novos desafíos como, o desenfoque de obxectos por movemento, a aparición de determinados fotogramas mal enfocados ou unha maior probabilidade de oclusións. Ademais, os vídeos proporcionan información espazo-temporal que se pode explotar para abordar estes novos desafíos, mellorando en xeral a precisión da detección. Aproveitar esta información espazo-temporal require novas técnicas non consideradas no deseño de detectores de obxectos en imaxes. Polo tanto, aínda que a solución máis sinxela consiste en executar un detector deseñado para traballar con imaxes en cada un dos fotogramas de forma illada, son necesarios novos modelos deseñados especificamente para vídeos.

Os detectores de obxectos actuais precisan unha gran cantidade de datos etiquetados para o proceso de adestramento. Isto limita a aplicabilidade destes modelos a escenarios nos que se dispoñen de grandes cantidades de datos anotados. A anotación é un proceso manual que require unha considerable inversión en tempo aumentando os custos de implantación. Recentemente, estanse realizando avances no campo da detección de obxectos con poucos exemplos de adestramento mediante o deseño de novos modelos capaces de aprender a partir

dun número reducido de instancias anotadas por categoría. Non obstante, a aplicación destas técnicas á detección de obxectos en vídeo segue a ser un campo de investigación practicamente inexplorado.

Na detección de obxectos, como en moitos outros campos da visión por computador, como a clasificación de imaxes ou o seguimento de obxectos, a extracción de características xoga un papel esencial. Antes da introdución das CNN, os sistemas de detección de obxectos estaban baseados principalmente en características deseñadas manualmente para resolver problemas concretos. Un dos primeiros casos de éxito que segue esta tendencia é o detector Viola-Jones (VJ). VJ foi deseñado para a detección de caras mediante a clasificación de rexións xeradas desprazando unha xanela de tamaño fixo pola imaxe. O descritor baseado en histogramas de gradientes orientados (HOG, *Histogram of Oriented Gradients*) foi introducido máis tarde como un paso adiante na extracción de características. Os autores deste traballo propuxeron utilizar características HOG como entrada a un clasificador que consistía nunha Máquina de Soporte Vectorial (SVM, *Support Vector Machines*) para a detección de persoas. Aínda que orixinalmente foi proposto para a detección de persoas, o mesmo algoritmo pode empregarse para a detección de obxectos doutras categorías. Como solución máis avanzada DPM (*Deformable Part Model*) divide o problema da detección de obxectos complexos na detección de partes destes obxectos. Considérase o cumio desta fase inicial antes da irrupción dos métodos de aprendizaxe profunda.

A aplicación de *backpropagation* para aprender filtros convolucionais, foi proposto por primeira vez por Yann LeCun en 1989 para o recoñecemento de números escritos a man. Esta liña de investigación levou ao desenvolvemento de LeNet en 1998. Aínda que LeNet é unha das primeiras CNN, xa conta cos principais operadores que compoñen as CNN modernas. A introdución de AlexNet en 2012 finalmente representa un cambio de paradigma convertendo as CNN na aproximación dominante para extracción de características visuais. Os avances na computación de propósito xeral en unidades de procesamento gráfico (GPGPU, General-purpose Computing on Graphics Processing Units) foron chave para o rexurdir das CNN, mellorando drasticamente o seu rendemento. AlexNet propúxose como solución para resolver o problema da clasificación de imaxes, gañando o desafío de clasificación ImageNet por unha gran marxe. Este desafío propón un conxunto de datos que conta con imaxes de 1.000 categorías diferentes. O clasificador baseado en AlexNet acadou unha taxa de erros top-5 do 15,3% fronte ao 26,2% acadado polo segundo mellor modelo. O éxito de AlexNet foi o punto de partida dun intenso esforzo investigador no desenvolvemento de CNN máis profundas e complexas. Fitos importantes como VGG e ResNet trouxeron novas solucións que permitiron construír redes máis profundas, pasando das 5 capas convolucionais máis 3 capas totalmente conectadas de AlexNet ata 152 capas na variante de ResNet máis profunda. Este impresionante desenvolvemento reduciu drasticamente a taxa no erro de clasificación, impulsando tamén outros campos como a

detección de obxectos.

Os detectores de obxectos pódense clasificar en dous grandes grupos: dunha e dúas etapas. Os detectores de dúas etapas introducen un método de proposta de rexións como primeiro paso para identificar zonas da imaxe cunha alta probabilidade de conter un obxecto de interese. Nunha segunda etapa, estas rexións clasifícanse e modificase a caixa delimitadora para axustarse mellor os obxectos da imaxe. Os métodos que seguen unha aproximación dunha etapa realizan a regresión da caixa delimitadora e a clasificación de obxectos directamente sen un primeiro paso de xeración de propostas, simplificando a arquitectura do detector. En xeral, os detectores de obxectos de dúas etapas mostran unha maior precisión na detección mentres que os detectores dunha etapa son máis rápidos debido á súa sinxeleza.

Os detectores de dúas etapas popularizáronse coa introdución das redes neuronais convolucionais baseadas en rexións (R-CNN, Region-based Convolutional Neural Networks). Nas primeiras versións, cada unha das propostas calculadas mediante *Selective Search* transfórmanse nunha imaxe de tamaño fixo e propáganse a través dunha CNN que extrae un vector de características por cada proposta. A continuación, emprégase unha SVM para clasificar cada unha das propostas asignando unha confianza. Finalmente, as detección redundantes elimínanse mediante NMS (*Non-maximum Suppression*) e as caixas delimitadoras refínanse mediante unha SVM. Aínda que demostrou o gran potencial de empregar CNNs como extractores de características no contexto da detección de obxectos, este método presenta algunhas desvantaxes importantes. O principal problema radica na execución de todo o extractor de características para cada proposta, comprometendo o rendemento final do detector. SPP-Net resolve este problema executando o extractor de características sobre a imaxe completa e incluíndo un operador de mostraxe que extrae as características asociadas a cada proposta directamente no mapa de características calculado na última capa da rede. A saída deste operador de mostraxe ten unha dimensión fixa independentemente do tamaño e da relación de aspecto da rexión de entrada, abordando un dos principais problemas da arquitectura R-CNN. Esta aproximación permite, polo tanto, evitar a execución de capas convolucionais de forma redundante.

Ambos enfoques teñen unha estratexia de adestramento complexa con múltiples compoñentes que deben ser adestrados de forma independente. *Fast R-CNN* substitúe o clasificador SVM e o regresor da caixa delimitadora por capas totalmente conectadas que reciben como entrada a saída dun operador de mostraxe similar o proposto en SPP-Net. Os parámetros destas novas capas totalmente conectadas optimízanse simultaneamente co extractor de características sen necesidade de adestrar compoñentes externos. Non obstante, a etapa de xeración de propostas segue sendo unha etapa de preprocesamento, que ademais, é unha orde de magnitude máis lenta que o resto do proceso de detección. Tamén se debe ter en conta que esta etapa non aproveita as características extraídas pola CNN para a xeración de propostas. *Faster R-CNN* define unha rede de proposta de rexións (RPN, *Region Proposal Network*) que consisten unha rede totalmente

convolucional que recibe o mapa de características da imaxe. A RPN tamén recibe un conxunto de caixas predefinidas, dispostas de forma densa sobre a imaxe de entrada. Primeiro, a RPN calcula a probabilidade de que cada unha destas caixas conteña un obxecto e realiza un primeiro axuste da caixa delimitadora. A continuación, as caixas refinadas empréganse para extraer un mapa de características por cada proposta seguindo o mesmo esquema que Fast R-CNN. Tanto a RPN como as capas totalmente conectadas que realizan a clasificación e o axuste final, adéstranse simultaneamente coas demais capas do detector. *Faster R-CNN* segue sendo amplamente empregada como punto de partida para o desenvolvemento de novos detectores de obxectos, que xeralmente comparten os mesmos conceptos fundamentais presentes na proposta orixinal.

Unha das novidades máis relevantes sobre a arquitectura Faster R-CNN foi a introdución de FPN (Feature Pyramid Networks) que permite extraer características en diferentes niveis da pirámide calculada. As CNN teñen unha estrutura piramidal por deseño, na que canto máis profunda se extrae a información, menor será a resolución dos mapas de características. O modelo orixinal de Faster R-CNN toma a saída da capa convolucional máis profunda para alimentar a RPN e a cabeza de rede, xa que esta é a que contén información de máis alto nivel. Non obstante, tamén é a capa con menor resolución, o que dificulta a detección de obxectos pequenos. Para superar este problema, FPN define novas conexións que propagan información dende as capas máis profundas para calcular características semanticamente fortes a diferentes niveis de resolución. A detección lévase a cabo nos diferentes niveis da pirámide aumentando a robustez do sistema ante obxectos de tamaños moi diversos.

O primeiro detector de obxectos de aprendizaxe profunda que seguiu unha estratexia dunha etapa foi YOLO (*You Only Look Once*). No canto de incluír un mecanismo de proposta de obxectos, YOLO define unha cuadrícula que divide a imaxe de entrada en rexións. Partindo desta división, lévase a cabo a predición das caixas delimitadoras e a correspondente clasificación para cada rexión. Outro dos primeiros intentos de definir un detector de obxectos dunha etapa foi SSD (*Single Shot MultiBox Detector*). Este método introduce un detector multinivel, que busca detectar obxectos pequenos en capas menos profundas aproveitando a súa maior resolución. Non obstante, a información de baixo nivel dispoñible nestas primeiras capas prexudica a precisión da detección. Como alternativa máis recente, RetinaNet tamén propón empregar múltiples niveis de resolución, pero implementando unha arquitectura FPN que calcula mapas de características semanticamente fortes de alta resolución. Ademais, RetinaNet tamén inclúe unha nova función de coste que reduce as contribucións de exemplos sinxelos no adestramento. Esta función de perda mitiga o desequilibrio extremo entre rexións con obxectos de interese e fondo presente nos detectores dunha etapa. Este desequilibrio é debido á falta dun método de proposta de obxectos que filtre a maioría das rexións que so conteñen fondo. EfficientDet emprega EfficientNet como extractor de características e céntrase na eficiencia na execución. Está deseñado para cumprir

unha ampla gama de restricións de recursos mediante unha parametrización que pode aumentar a eficiencia sacrificando precisión.

A extracción de características está a piques de sufrir un novo cambio de paradigma, influenciado principalmente polos avances realizados no campo do procesamento da linguaxe natural. As arquitecturas baseadas en mecanismos de atención, especialmente os denominados *Transformers*, convertéronse recentemente no modelo dominante no campo do procesamento da linguaxe natural. A autoatención relaciona diferentes elementos dunha secuencia determinada para calcular unha representación desa secuencia. Polo tanto, pódese usar como modelo de extracción de características para secuencias de elementos en xeral. Esta idea aplicouse en combinación con CNNs para extraer relacións entre propostas no contexto dun detector de dúas etapas. DETR (*DEtection TRansformer*) describe unha arquitectura na que un *Transformer* recibe características extraídas previamente por unha CNN. Alternativamente, *Visual Transformer* (ViT) propón converter a imaxe de entrada nunha secuencia de rexións de tamaño predefinido para adaptar esta arquitectura o procesamento de imaxes. Isto posibilita a substitución por completo dos extractores de características convolucionais. Ao contrario das CNN, ViT non funciona xerarquicamente en diferentes niveis de resolución. Algunhas modificacións sobre a arquitectura ViT orixinal, como MViT e *Swin* propoñen calcular mapas de características de diferentes resolucións, implementando un extractor de múltiples escalas baseado en *Transformers*. Polo tanto, perseguen un obxectivo similar ao de FPN. ViTDet define un detector de obxectos de dúas etapas que xera mapas de características de varios niveis de resolución a partir da saída de ViT. Como resultado, VitDet consegue resultados moi competitivos na detección de obxectos en imaxes.

A métrica estándar para a avaliación da detección de obxectos é o AP (Average Precision). O AP calcúlase como a área baixo a curva de precisión fronte a *recall*. O AP final é a media do AP calculado de forma independente para cada categoría do conxunto de datos. Unha detección considérase un positivo verdadeiro se a intersección sobre unión (IoU) coa correspondente anotación é superior a un determinado limiar. Un mínimo de $IoU = 0,5$ emprégase habitualmente en moitos conxuntos de datos como valor de referencia para comparar diferentes algoritmos de detección. Dende a introdución do conxunto de datos Microsoft COCO, que se converteu no conxunto de datos estándar para a detección de obxectos en imaxe, a media do AP para diferentes valores de IoU fíxose máis popular. En concreto, COCO propón avaliar os detectores de obxectos calculando o AP medio variando o IoU mínimo entre 0,5 a 0,95 con incrementos de 0,05. Para o cálculo da curva, as detección ordénanse de acordo a confianza calculada pola rede. Avanzando na lista de detección, pódese calcular os puntos de precisión fronte a *recall*, tendo en conta todas as detección con maior confianza que a posición actual.

A detección de obxectos en vídeo empezou a cobrar máis protagonismo dende a publicación en 2015 do conxunto de datos de ImageNet para a detección de obxectos en vídeo (ImageNet

VID). O principal reto é como explotar a información espazo-temporal dispoñible nos vídeos para aumentar a precisión da detección. Este é un tema de investigación activo non só na detección de obxectos, senón tamén en campos relacionados como o recoñecemento de accións.

A maior parte dos detectores de obxectos en vídeo buscan enriquecer as características extraídas en un fotograma de referencia mediante métodos de agregación con características extraídas en un conxunto de fotogramas de soporte. A detección final realízase no fotograma de referencia empregando as características agregadas. O primeiro paso para agregar características ao longo do tempo debe ser establecer relacións entre ditas características nos diferentes fotogramas. En función de como se resolva este problema, as técnicas de agregación espazo-temporal pódense clasificar en dous grupos principais: a nivel de píxel ou a nivel de obxecto. Nos dous casos se inclúe un extractor de características que se executa en cada fotograma de forma independente. Os enfoques baseados en agregación a nivel de píxel establecen relacións entre os píxeles dos mapa de características dos diferentes fotogramas involucrados. Por outra banda, os métodos baseados en relacionar obxectos tamén realizan unha fase de proposta de obxecto por fotograma antes da agregación das características asociadas a cada proposta. O obxectivo destes métodos é vincular propostas de obxectos e agregar as características das propostas en lugar de traballar con un mapa de características global para toda a imaxe.

Existen diversas solucións para abordar o problema da asociación de características a nivel de píxel. O uso de fluxo óptico para enlazar píxeles en cadros consecutivos propúxose por primeira vez como parte de FGFA (*Flow-guided Feature Aggregation for Video Object Detection*). Aínda que esta técnica acadou resultados moi prometedores, converténdose nun fito importante no campo da detección de obxectos en vídeo, a sobrecarga asociada o cálculo do fluxo óptico levou á comunidade investigadora a avanzar cara alternativas máis eficientes. Este é o caso de STSN (*Spatio-temporal Sampling Network*), que implementa un mecanismo de mostraxe de características baseado en convolucións deformables. Os desprazamentos nas capas de convolución deformables apréndense para facer coincidir características do fotograma de referencia coas características correspondentes dos fotogramas de soporte. Como alternativa, PSLA (*Progressive Sparse Local Attention*) inclúe un módulo de atención para calcular pesos que establecen estas correspondencias.

Os métodos de agregación a nivel de obxecto asocian propostas calculadas pola RPN en cada fotograma, no canto de realizar unha agregación global sobre todo o mapa de características. Este enfoque xeralmente inclúe un método para asociar propostas en diferentes fotogramas, calculando traxectorias a curto prazo ou *tubelets*. Kang *et al.* propoñen unha rede de proposta de *tubelets* (TPN, *Tubelet Proposal Network*) que xeneralizar a fase de proposta de obxectos ao dominio espazo-temporal. Este método replica a mesma proposta ao longo dunha secuencia de fotogramas consecutivos. A continuación, unha fase de refinamento da caixa delimitadora modifica a proposta en cada fotograma para axustar mellor o obxecto en cada instante temporal.

Finalmente, unha LSTM (*Long Short Term Memory*) emprégase para clasificar os obxectos a partir das traxectorias calculadas. Como consecuencia do éxito acadado no campo do procesamento da linguaxe natural, os mecanismos de atención impuxéronse como método para a relación de propostas en fotogramas diferentes. En lugar de definir traxectorias de obxectos, esta nova tendencia propón calcular os pesos de atención entre cada par de propostas dun xeito denso. A agregación de características consiste nunha suma ponderada empregando estos pesos.

Como alternativa á agregación de características, propuxéronse técnicas para refinar o conxunto de detecciónss final. Estes métodos céntranse xeralmente en optimizar a confianza das detecciónss, mantendo a coherencia espazo-temporal. Feichtenhofer *et al.* propoñen un enfoque máis elaborado no que a detección e o seguimento de obxectos se levan a cabo conxuntamente. A información de seguimento úsase para calcular traxectorias ao longo do vídeo e optimizar as confianza das detecciónss asociadas a cada traxectoria. A hipótese de traballo destes métodos é que se unha detección pode asociarse con outras detecciónss cunha confianza alta, é máis probable que sexa unha detección correcta.

En xeral, as técnicas de aprendizaxe profunda, como as CNN, precisan dunha gran cantidade de exemplos etiquetados para aprender novas tarefas. Isto tamén ocorre no caso da detección de obxectos, no que se precisan varios centos de obxectos anotados por categoría para poder adaptar un detector a un novo dominio. Polo tanto, as limitacións na dispoñibilidade de datos anotados pode dificultar ou incluso imposibilitar a súa aplicación en determinados escenarios. A aprendizaxe con poucos exemplos, tamén coñecida como *few-shot*, ten como obxectivo definir estratexias para aprender novas tarefas a partir dun conxunto de datos reducido, que normalmente oscilan entre 1 e 30 exemplos por categoría. Nos últimos anos, acadáronse grandes avances en tarefas máis sinxelas como a clasificación de imaxes. Non obstante, a precisión acadada polos detectores de obxectos adestrados con poucos exemplos segue a ser moi inferior a acadada en escenarios tradicionais.

O adestramento dun detector de obxectos con poucos exemplos consta normalmente de dúas fases fundamentais. En primeiro lugar, o modelo adestrase nun conxunto de categorías base das que se conta con un gran número de anotacións, seguindo o mesmo esquema que un detector convencional. A continuación, ten lugar unha segunda fase de adaptación do modelo a un conxunto de categorías novas das que se dispón dun número reducido de anotacións. Existen dous enfoques principais na literatura para implementar un detector de obxectos de poucos exemplos: meta-aprendizaxe e a realización dun axuste fino. Os métodos de meta-aprendizaxe redefinen a tarefa de clasificación como a aprendizaxe dunha métrica de semellanza, na que a clasificación de obxectos se leva a cabo comparando cada detección cun prototipo asociado a cada categoría. Estes prototipos derívanse do conxunto de exemplos anotados dispoñibles para cada categoría nova. Por outra banda, os métodos de axuste fino seguen unha estratexia de adaptación máis convencional, desenvolvendo solucións para adaptar rapidamente o modelo

adestrado previamente nas categorías base para detectar obxectos pertencentes ás categorías novas.

Entre as técnicas que empregan solucións baseadas en meta-aprendizaxe destacan aquelas que estenden modelos tradicionais para adaptalos a este novo paradigma. FSRW aplica meta-aprendizaxe partindo da arquitectura YOLO mentres que Meta R-CNN propón un detector de dúas etapas baseado en Faster R-CNN. Un problema común en todos os modelos que aplican meta-aprendizaxe é calcular un prototipo por categoría que sexa o suficientemente xeral como para representala no seu conxunto sen provocar confusións con outras categorías. O método máis estendido para xerar estes prototipos consiste en calcular a media dos vectores de características dos exemplos anotados para cada categoría. DAnA (*Dual-awareness Attention*) introduce un bloque de atenuación para reducir a influencia das características asociadas ao fondo e un novo método para resumir a información de múltiples anotacións de cada categoría.

Os métodos baseados en axuste fino de modelos previamente adestrados foron popularizados por TFA. Este traballo consegue resultados competitivos partindo dun detector de obxectos de dúas etapas aplicando unha sinxela estratexia de axuste fino, na que só as últimas capas da rede se afinan nas categorías novas. DeFRCN afonda neste enfoque, identificando dous problemas principais na arquitectura orixinal de Faster R-CNN que dificultan a aprendizaxe con poucos exemplos: a función de perda multiobxectivo e multitarefa. A estratexia de aprendizaxe multiobxectivo consiste en optimizar simultáneamente a RPN e a cabeceira da rede. O obxectivo da RPN é ser un detector de obxectos agnóstico da categoría, mentres que a cabeceira da rede busca refinar e clasificar os obxectos. Esta discrepancia pode danar o proceso de aprendizaxe, especialmente con poucas instancias de adestramento. Ademais, as mesmas características compártense na cabeza da rede para tarefas de localización e clasificación. Polo tanto, DeFRCN inclúe unha capa de desacoplamento de gradiente para modular a influencia da RPN e da cabeza de rede no proceso de adestramento, e un bloque de calibración para desacoplar as tarefas de clasificación e localización.

Todos os métodos descritos anteriormente propoñen solucións para extraer coñecemento dun conxunto reducido de datos anotados, sen considerar a posibilidade de recuperar novas anotacións de imaxes sen etiquetar. Recentemente xurdiu unha liña de investigación que propón métodos para extraer automaticamente estas anotacións aumentando o conxunto de adestramento. Unha potencial fonte de novos datos son as propias imaxes que conteñen as anotacións das categorías base. Estas imaxes poden conter tamén obxectos das clases novas sen anotar. De feito, isto pode supoñer un problema xa que na primeira fase do adestramento, o sistema aprende a recoñecer estos obxectos como fondo. O número limitado de anotacións na fase de axuste fino pode non ser suficiente para aprender características discriminatorias que permitan diferenciar as clases novas do fondo. Este problema identifícase e resólvese en MINI (*Mining Implicit Novel Instances*) extraendo anotación das clases novas de forma automática das imaxes que

conteñen as clases base. Seguindo esta liña, traballos recendes propoñen recuperar anotacións de conxuntos de imaxes independentes non anotados. Defínese unha serie de etapas de verificación e corrección de detección calculadas por un detector adestrado con poucos exemplos. As detección filtradas e corrixidas inclúense despois no conxunto de adestramento do detector final. Acádanse resultados prometedores que demostran a utilidade de aumentar os conxuntos de adestramento sen incrementar os custos de anotación.

O campo da detección de obxectos con poucos exemplos de adestramento atópase nunha fase inicial de desenvolvemento. Como consecuencia, a aplicacións destas técnicas á detección de obxectos en vídeo segue practicamente sen explorar. O primeiro conxunto de datos para a detección de obxectos en vídeo con poucos exemplos de adestramento definiuse recentemente e recibe o nome de FSVOD500 (*Few-Shot Video Object Detection*). Este traballo tamén propón un método de meta-aprendizaxe que explota información espazo-temporal por primeira vez na detección de obxectos con poucos exemplos en vídeos.

O obxectivo principal desta tese de doutoramento é explorar como se pode explotar a información espazo-temporal para mellorar os resultados da detección de obxectos en vídeo. En concreto, analízase como aproveitar esta información nunha ampla gama de escenarios, incluíndo diferentes tamaños de obxectos, número de obxectos por fotograma, velocidade de movemento e conxuntos de adestramento de tamaño limitado. Para iso, abordaranse os seguintes obxectivos individuais:

1. **Explotar información dunha secuencia de fotogramas de entrada para mellorar a precisión da detección en cada fotograma.** Aplicar un detector de obxectos dunha soa imaxe a nivel de fotograma para resolver o problema da detección de obxectos en vídeo produciría resultados subóptimos, xa que non se considera información espazo-temporal. O primeiro obxectivo desta tese é deseñar novos métodos que aproveiten a información espazo-temporal considerando un grupo de fotogramas de entrada para calcular o conxunto de detección no fotograma de referencia.

2. **Incluír información espazo-temporal a longo prazo no proceso de agregación de características.** O obxectivo é deseñar un novo proceso de agregación que considere simultaneamente información de fotogramas próximos e afastados no tempo para enriquecer as características no fotograma de referencia. A información proporcionada relacionando propostas de obxectos a longo prazo é complementaria á información extraída en fotogramas próximos altamente redundantes. En lugar de incluír información a longo prazo como paso posterior a execución do detector, explorarase o uso de mecanismos de atención para realizar a agregación de características a longo prazo como paso fundamental do propio detector.

USC DE COMPOSTELA

3. **Detección de obxectos en vídeo con poucos exemplos de adestramento.** Búscase explotar a información espazo-temporal en escenarios nos que a dispoñibilidade de datos etiquetados é moi reducida. Baixo o paradigma de adestramento con poucos exemplos, non é factible ter vídeos etiquetados con anotacións en cada un dos fotogramas. Polo tanto, o detector de obxectos proposto deben aprender dun número moi limitado de instancias anotadas en imaxes illadas. Aínda que a información espazo-temporal non está dispoñible no momento do adestramento, si se pode empregar en inferencia. Esta circunstancia impón un gran desafío, xa que o modelo debe ser adestrado con imaxes illadas que non conteñen información espazo-temporal, mentres que a inferencia se realiza sobre secuencias de fotogramas.

Nesta tese de doutoramento descríbese o traballo realizado para a consecución dos obxectivos propostos. Inclúese unha descrición da metodoloxía seguida, dos resultados obtidos e as correspondentes conclusións extraídas en cada caso.

O capítulo 2 describe FANet, un novo detector de obxectos que aproveita información espazo-temporal para mellorar a precisión da detección en cada fotograma. Este modelo, combina un módulo de agregación de características a curto prazo cun método de optimización a longo prazo que se encarga de modificar as confianzas das deteccións. O módulo de agregación a curto prazo vincula características asociadas a propostas extraídas de un conxunto de fotogramas veciños. A continuación, introdúcese un operador de agregación que resume a información, calculando o mapa de características final asociado a cada proposta. Tamén se propón unha nova arquitectura para a cabeceira da rede, seguindo un enfoque de dúas ramas para procesar a información espacial do fotograma actual e a información espazo-temporal agregada a través de varios fotogramas. A optimización final a longo prazo corrixe a confianza das deteccións de obxectos, garantindo a coherencia espazo-temporal.

Capítulo 3 presenta SLTnet, unha versión mellorada de FANet que inclúe agregación de características a curto e longo prazo para calcular os mapas de características asociados a cada proposta. Este método inclúe unha agregación de características a curto prazo máis sólida baseada na semellanza entre propostas extraídas de diferentes fotogramas. A continuación, os mapas de características agregados vólvense a someter a un proceso de agregación a longo prazo. Esta segunda agregación lévase a cabo aplicando mecanismos de atención que identifican as relacións entre propostas calculadas en diferentes fotogramas. Este modulo ten en conta relacións entre características xeométricas a longo prazo por primeira vez na detección de obxectos en vídeo. En lugar de empregar a caixa delimitadora directamente como características xeométricas, este modulo calcula as traxectorias dos obxectos ao longo do vídeo para actualizar as coordenadas orixinais das caixas delimitadoras. Este capítulo tamén explora as capacidades de xeneralización dos detectores de obxectos aplicados a vídeo, proporcionando unha análise en profundidade en

cinco conxuntos de datos diferentes: ImageNet VID, UAVDT, Visdrone, USC-GRAD-STDdb e MOT20.

O capítulo 4 presenta FSVDet, un detector de obxectos en vídeo que se pode adestrar con un conxunto moi reducido de exemplos de adestramento. Proponse a xeración de fotogramas de soporte sintéticos a partir de imaxes para aumentar o conxunto de adestramento. A dobre cabeceira espazo-temporal esténdese para predicir o solape de cada detección cos obxectos reais no fotograma de referencia. Tamén se propón un método de optimización das confianzas das detecciónns baseado no cálculo de traxectorias. O proceso de creación destas traxectorias está guiado polas predicións de solape.

O capítulo 5 recolle as conclusións globais da investigación desenvolvida nesta tese e detalla o traballo futuro.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Contents

1 Introduction **1**
1.1 Motivation 1
1.2 Objectives 9
1.3 Contributions 10
1.4 Dissertation structure 14

2 Spatio-temporal Tubelet Feature Aggregation and Object Linking for Video Object Detection **17**
2.1 Introduction 18
2.2 Related work 19
2.3 FANet Architecture 21
2.4 Experiments 29
2.5 Conclusions 36

3 Short-Term Anchor Linking and Long-Term Self-Guided Attention for Video Object Detection **39**
3.1 Introduction 40
3.2 Related work 41
3.3 Method 43
3.4 Experimental Results 49
3.5 Conclusions 59

4 Few-shot Video Object Detection **61**
4.1 Introduction 61
4.2 Related Work 63
4.3 Proposed Method 65
4.4 Experiments 71
4.5 Conclusions 76

5 Conclusions **77**

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Bibliography **81**

List of Figures **91**

List of Tables **93**

A Reproduction Rights **95**

A.1 Journals 95

A.2 Conferences 96

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

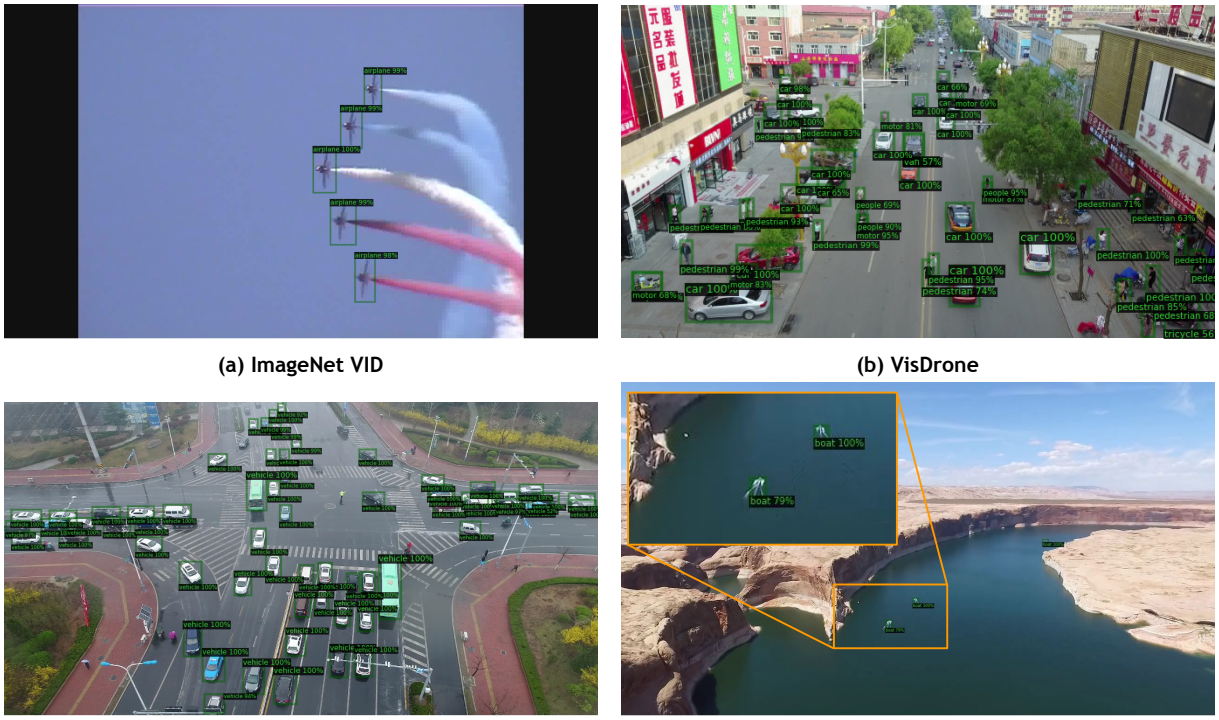
1 Introduction

1.1 Motivation

Object detection is one of the most relevant research topics in computer vision. The object detection problem comprises two main tasks: object localization and object classification. Object localization aims to define a bounding box framing the objects of interest in the image, while the classification task predicts object categories, yielding a confidence score for each detection. Figure 1.1 shows the output of an object detector in four different datasets used in this thesis to evaluate the models. Potential applications range from video surveillance to being a major component of more complex systems such as autonomous vehicles. In general, an object detector is required as a fundamental step in any system in which the extraction of information of the objects present in the scene is relevant. This includes applications such as traffic monitoring, medical imaging, face detection or crowd counting. The detection precision achieved by state-of-art object detectors has soared in recent years, mainly due to the adoption of deep convolutional neural networks (CNN). However, there are still challenges that remain unsolved, making it a very active research field.

Most of the aforementioned applications receive an input video stream instead of a single image. Solving the video object detection problem raises new challenges such as motion blur, frames out of focus or an increased likelihood of occlusions. Moreover, videos provide spatio-temporal information that can be exploited to overcome these new challenges, generally improving the detection precision. Leveraging this spatio-temporal information requires new techniques not considered in the design of single-image object detectors. Therefore, although running a single-image object detector at frame level is the most straightforward solution, new models specifically designed for videos are required.

Current state-of-the-art object detectors need a large amount of labeled data for the training process. This limits the applicability of these models to scenarios in which the availability of annotated data is not an issue. Furthermore, it also increases implementation costs, as the annotation process requires human intervention and is very time-consuming. Recently, some progress has been made in the field of few-shot object detection by designing new frameworks



(a) ImageNet VID (b) VisDrone

(c) UAVDT (d) USC-GRAD-STDdb

Figure 1.1: Object detection output in different datasets.

that only require few annotated instances per category. However, the few-shot video object detection problem remains almost unexplored.

This thesis focuses on both: traditional and few-shot video object detection with spatio-temporal convolutional neural networks. A video object detector can exploit information from many input frames to localize and classify the objects in each frame. Hence, the definition of techniques to gather this information is key in the development of the thesis. In addition, this thesis also discusses how these techniques can be transferred to data scarce scenarios.

1.1.1 Image Object Detection

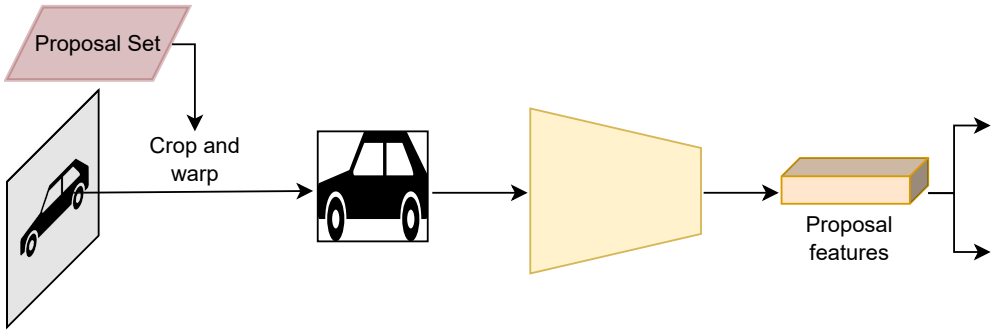
In object detection, as in many other computer vision fields such as image classification or object tracking, feature extraction plays a fundamental role. Before the introduction of CNNs, object detection systems were mainly based on handcrafted features [99]. An early success case following this trend is the Viola-Jones (VJ) detector [88], which uses the Adaboost algorithm [31] for the selection of Haar-like features. VJ was intended for human face detection based on the classification of regions of interest calculated through sliding window. Histogram of Oriented Gradients (HOG) descriptor [18] was later introduced as a step-forward in feature extraction. Authors proposed to feed HOG features to a Support Vector Machine (SVM) classifier for pedestrian detection. Although originally proposed for pedestrian detection, the same framework can be trained for other object categories. Deformable Parts Model (DPM) [30] breaks the

detection problem into detecting small parts. It is considered the peak of this early object detection phase before the irruption of deep learning methods [105].

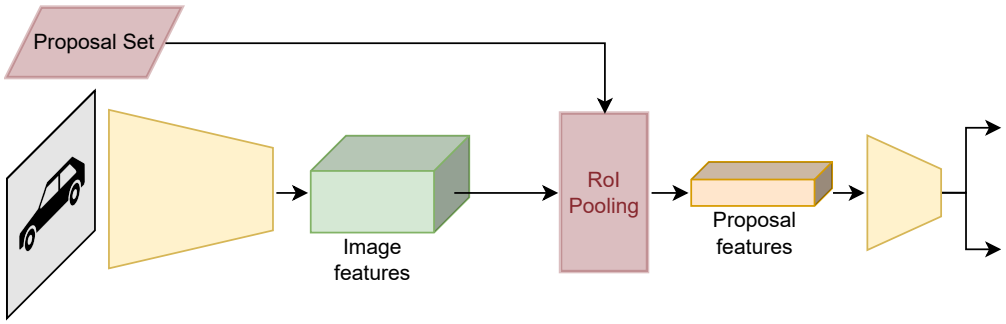
Applying backpropagation to learn convolutional kernels was first proposed in 1989 [55] for handwritten numbers recognition. This line of research led to the development of LeNet [56] in 1998. Although LeNet is an early CNN, it already has the main building blocks that compose modern CNNs. The introduction of AlexNet [53] in 2012 finally represents a paradigm shift from traditional handcrafted feature extractors to CNNs. Advances in *general purpose computing on graphics processing units* (GPGPUs) have been key to the resurgence of CNNs, dramatically improving their performance. AlexNet was first proposed to solve the image classification problem, winning the ImageNet classification challenge [74] by a large margin. This challenge proposes an image dataset with 1,000 different categories in which the AlexNet-based classifier achieved a top-5 error rate of 15.3% compared to 26.2% achieved by the second-best model. The success of AlexNet was the starting point for an intense research effort in the development of deeper and more complex CNNs. Important milestones such as VGG [78] and ResNet [44] brought new solutions that allowed building deeper networks, going from the 5 convolutional layers plus 3 fully connected layers of AlexNet up to 152 layers in the deepest ResNet variant. This impressive development has drastically reduced the classification error rate, boosting also other fields such as object detection.

Object detectors can be classified in two major approaches: one- and two-stage architectures. Two-stage detectors introduce a region proposal method as the first stage to generate candidate regions with a high probability of containing an object of interest. These regions are classified and further refined to better fit the object in a second stage. One-stage frameworks perform the bounding box regression and object classification directly without a proposal generator, simplifying the detection pipeline. In general, two-stage object detectors have shown higher detection precision while one-stage detectors are faster due to its simplicity.

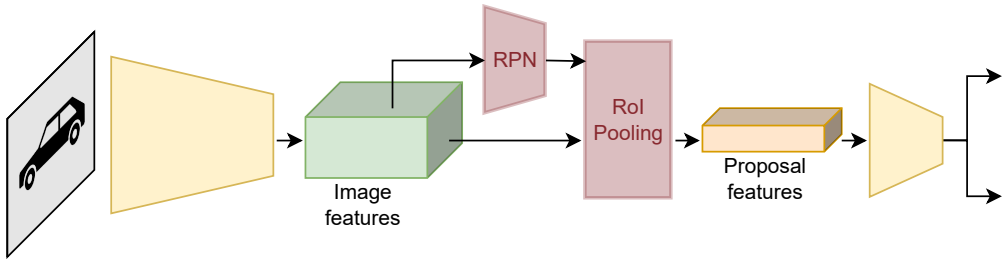
Two-stage CNN-based methods were first popularized by *Region-based Convolutional Neural Networks* (R-CNN) [34]. A set of object proposals calculated using Selective Search [86] are wrapped to a fixed shape and propagated through a CNN backbone that extracts a feature vector per proposal (Fig. 1.2a). Feature vectors are then fed to a per category SVM to predict confidence scores. Finally, redundant detections are removed by means of *Non-maximum Suppression* (NMS) and bounding boxes are refined by a bounding box regressor implemented as a per category SVM. Although it proved the great potential of CNNs as feature extractors for object detectors, this method still has some important disadvantages. It runs the whole feature extractor for each object proposal (Fig. 1.2a), compromising the detection performance. SPP-Net [43] solves this problem by including a spatial pyramid pooling layer after the last convolutional layer. The output of this pooling layer has a fixed shape independently of the input size and aspect ratio, addressing one of the main issues of the R-CNN architecture. This allows



(a) RCNN.



(b) Fast RCNN.



(c) Faster RCNN.

Figure 1.2: Different two-stage object detectors.

to run the CNN at image level and extract proposal features through the pooling layer in the last convolutional feature map, thus, avoiding redundant convolutional calculations.

Both approaches still have a cumbersome training strategy with multiple components that need to be trained independently. Fast R-CNN [33] (Fig. 1.2b) represents a major step forward towards an object detector that can be trained end-to-end. It replaces the SVM classifier and bounding box regressor by fully connected layers on top of the pooled feature maps. The Region of Interest (RoI) pooling operator used in Fast R-CNN is a simplification of the pyramid pooling layer implemented on SPP-Net with only one level. However, the proposal generation stage remains implemented as a preprocessing stage that is an order of magnitude slower than the detection stage. Moreover, it does not take advantage of CNN features for proposal generation. Faster R-CNN [72] defines a *Region Proposal Network* (RPN) implemented as a fully convolutional network on top of the CNN as shown in Fig. 1.2c. The RPN receives a set of predefined boxes, known as *anchor boxes*, arranged in a dense grid over the input image. Then,

it calculates the objectness score and refines each proposal box to better fit the objects on the image. Object proposal boxes are then fed to the pooling layer, as in the original Fast R-CNN model. Both the RPN and the network head, which includes the final classification and bounding box regression, are trained simultaneously, resulting in a fully end-to-end object detector. Faster R-CNN is still widely used as a baseline for the development of new two-stage object detectors, that generally share the same core concepts present in the original implementation.

One of the most relevant improvements over the Faster R-CNN architecture was the introduction of *Feature Pyramid Networks* (FPN) [60] to extract features at different pyramid levels. CNNs inherently have a pyramid structure, in which the deeper the information is extracted, the lower the resolution of the feature maps. The original Faster R-CNN model takes the deepest convolutional layer output to feed the RPN and the network head, as it is the layer with more semantic meaning. However, it is also the layer with the lowest resolution, hindering the detection of small objects. To overcome this issue, FPN defines top-down and lateral connections that propagate information from the deepest layers to calculate semantically strong features at different resolution levels. Then, objects of different sizes are detected from different level feature maps, improving the overall precision.

The first deep learning object detector to follow a one-stage strategy was *You Only Look Once* (YOLO) [70]. Instead of relying on object proposals, YOLO defines a grid dividing the input image into regions. Then, it simultaneously predicts bounding boxes and confidence scores for each region. Another early attempt to implement a one-stage object detector was *Single Shot MultiBox Detector* (SSD) [64]. This method introduces a multi-level detector, detecting small objects in earlier layers of the network, as they have higher resolution. However, the low-level information available in these first stages damages the detection precision. Lately, RetinaNet [61] also proposed using multiple levels in one-stage object detection, but implementing an FPN architecture that calculates high-resolution semantically strong feature maps for small object detection. Furthermore, RetinaNet [61] also includes a new loss function that reduces the contributions of easy examples. This loss function mitigates the extreme imbalance of foreground-background in one-stage object detectors due to the lack of an object proposal method that filters out most of the background regions. EfficientDet [82] is based on the EfficientNet [81] backbone, and focuses on execution efficiency. It is designed to meet a wide range of resource constraints by a parametrization that can reduce execution efficiency to gain detection precision.

Feature extraction is on the verge of a new paradigm shift, influenced by the advances made in the *natural language processing* (NLP) field. Architectures based on self-attention mechanisms, specially Transformers [87], have recently become the model of choice in NLP. Self-attention relates different elements of a given sequence to calculate a feature representation of that sequence. Hence, it can be used as a general feature extractor model for sequences of

elements. This idea was implemented in combination with CNNs to mine proposal relationships in the context of a two-stage detector [45]. *DEtection TRansformer* (DETR) [8] describes an end-to-end detection framework in which a Transformer is fed with the output of a convolutional backbone, also combining both models. Alternatively, Visual Transformer (ViT) [22] proposes to convert the input image into a sequence of patches to adapt this architecture to the image domain, completely replacing convolutional-based backbones. Contrary to CNNs, ViT does not work hierarchically at different resolution levels. Some modifications over the original ViT architecture such as MViT [24, 59] and Swin [65] calculate different resolution feature maps implementing a Transformer-based multiscale backbone. Hence, they pursue a similar goal than the original FPN [60] architecture, providing multiple pyramid levels. ViTDet [58] defines a two-stage object detector based on the original ViT backbone, generating multiple-level feature maps from the ViT output. Thus, it implements a multi-level object detector based on a plain non-hierarchical backbone. VitDet achieves very competitive results in image object detection with a ViT backbone pre-trained as a Masked Autoencoder (MAE) [41].

The standard metric for object detection evaluation is *Average Precision* (AP). AP is computed as the area under the precision-recall curve. The final AP is the mean of the independently computed AP for each category in the dataset. A detection is considered a true positive if the *intersection over union* (IoU) with the ground truth is higher than a given threshold. A minimum $IoU = 0.5$ is commonly used in many datasets as a reference value to compare different detection algorithms. Also, since the introduction of the Microsoft COCO dataset [62], which has become the standard dataset for single image object detection, averaging the AP for different IoU values has become more popular. Specifically, COCO proposes to evaluate object detectors averaging the AP with a minimum IoU ranging from 0.5 to 0.95 with increments of 0.05. For the computation of the curve, object detections are first ordered by the confidence score calculated by the network. Then, by advancing on the ordered detection list, precision and recall can be calculated, taking into account all the detections with higher confidence than the current list position.

1.1.2 Video Object Detection

The publication of the ImageNet challenge for video object detection (VID) [74] in 2015 has attracted the attention of the research community to the video object detection problem. The main challenge is how to exploit spatio-temporal information available in videos to boost the detection precision. This is an active research topic not only in video object detection, but also in related fields such as action recognition, in which video understanding is key.

The core idea behind most video object detectors is the aggregation of features individually extracted from a set of support frames to enhance features in a reference frame. Then, the final detection is carried out in the reference frame with improved features. The first step in

aggregating features over time must be establishing relationships among different frames linking those features. Based on how this issue is solved, spatio-temporal aggregation techniques can be classified into two main groups: pixel-level and object-level aggregation. Both cases include a per frame feature extractor implemented as a CNN. Then, pixel-level approaches establish relationships between feature map pixels in different frames to finally aggregate those features, calculating an enhanced feature map for the reference frame. On the other hand, object-based methods also perform a per-frame object proposal stage before the feature aggregation. The goal of these methods is to link object proposals and aggregate proposal features instead of working with the entire feature map.

There have been different proposals to address the pixel features association problem. The use of optical flow to link pixels in consecutive frames was first proposed in *Flow-guided Feature Aggregation for Video Object Detection* (FGFA) [104]. Although this technique achieved very promising results, being an important milestone in video object detection, the overhead of calculating the optical flow led the research community to move towards more efficient alternatives. *Spatio-temporal Sampling Network* (STSN) [1] implements a feature sampling mechanism based on deformable convolutions. Offsets in the deformable convolutions layers are learned to match features in the reference frame with features in the support frames. Alternatively, *Progressive Sparse Local Attention* (PSLA) [37] includes an attention module to compute weights to establish spatial correspondences between features.

Object-level aggregation methods rely on the RPN of a two-stage detection architecture to focus on proposal features rather than on the entire frame. This approach commonly includes a method to link object proposals in neighboring frames, creating object tubelets. Kang *et al.* [49] propose a *Tubelet Proposal Network* (TPN) to generalize the object proposal stage to the spatio-temporal domain. This method replicates the same object proposal throughout a sequence of consecutive frames. Then, a bounding box refinement stage modifies the proposal in each frame to fit the object in the corresponding frame. Finally, the authors propose a *Long Short-Term Memory* network (LSTM) to classify object tubelets. Inspired by the success of attention mechanisms in the natural language processing field, new techniques to mine proposal relationships between different frames have emerged [12,21]. Instead of defining object tubelets, this new trend proposes to calculate attention weights between each pair of proposals in a dense manner. The feature aggregation is implemented as a sum weighted by these attention weights.

Alternatively to feature aggregation, post-processing techniques have been proposed to improve the final detection set. These methods generally focus on optimizing the score of detections, keeping spatio-temporal consistency [39]. Feichtenhofer *et al.* [29] propose a more elaborated approach in which object detection and object tracking are jointly learned. Then, tracking information is used to build long tubes and optimize detections confidence scores throughout the tube. The working hypothesis of these methods is that if a detection can be

associated with a long tube, then it is more likely to be a correct detection.

1.1.3 Few-shot Object Detection

In general, deep learning techniques, such as deep CNNs, need thousands of labeled examples to learn new tasks. This is also the case for the object detection problem, in which an object detector is trained with at least several hundred annotated objects per category. This raises the annotation costs of adapting a state-of-the-art object detector to a new domain. Thus, it might not be even feasible due to unlabeled data availability limitations in some cases. Few-shot learning aims to define new strategies to learn new tasks from just a few examples, typically ranging from 1 to 30. There have been interesting developments in this field in recent years, achieving great results in simpler tasks like image classification. However, few-shot object detectors precision still lags far behind of state-of-the-art object detectors.

Few-shot object detection training typically consists of two phases. First, the model is pre-trained on a set of base categories with several annotated samples, as in the traditional object detection case. Then, a second phase of model adaptation to a set of novel categories with scarce annotations takes place. There are two main approaches in the literature to implement a few-shot object detector: meta-learning and fine-tuning. Meta-learning methods redefine the classification task as a similarity metric learning, in which object classification is implemented as a comparison between each detection and a per-category prototype. These prototypes are derived from the few annotated examples available for each novel category. On the other hand, fine-tuning-based methods follow a more traditional architecture, implementing solutions to rapidly adapt the pretrained model in base categories to novel categories by just fine-tuning the network with few annotations.

Meta-learning variations of traditional object detectors have been proposed to solve the few-shot object detection problem. FSRW [48] defines a one-stage few-shot detector based on YOLO [71] while Meta R-CNN [97] proposes a two-stage framework based on Faster R-CNN [72]. A common issue in all meta-learning approaches is calculating a category prototype that is general enough to represent each category. The most extended method to calculate the category prototype is by just averaging the feature vectors of annotated examples for each category. DAnA [11] introduces a background attenuation block to reduce the influence of background features in the final prototype and a new method to summarize the information, generating query position aware (QPA) support vectors.

Fine-tuning methods applied to few-shot object detection were popularized by TFA [90]. This work achieves competitive results based on a two-stage object detector following a simple fine-tuning strategy, in which just the last layers of a pre-trained model are fine-tuned on novel categories. DeFRCN [69] further develops this fine-tuning approach, identifying two main issues in the original Faster R-CNN framework that hinder the few-shot learning: the multi-objective

and multi-task learning. The multi-objective learning strategy simultaneously optimizes the RPN and the network head. The objective of the RPN is a class agnostic object detector, while the network head needs to refine and classify the objects. This discrepancy might damage the learning process, especially with few training instances. Moreover, the same features are shared in the network head for object localization and classification tasks. Therefore, this framework includes a gradient decoupled layer to modulate the influence of the RPN and network head in the training process, and a prototypical calibration block to decouple the classification and localization tasks.

Previous few-shot object detection frameworks focus on learning from few annotated instances, without considering any possibility of recovering new annotations from unlabeled images. However, a new line of research has recently emerged, proposing methods to automatically extract these annotations to increase the training set. *Mining Implicit Novel Instances* (MINI) [7] identifies a potential problem in the standard pre-train phase with base categories: images annotated with base categories are likely to contain also unlabeled instances from novel categories, making the model consider these objects as background in pre-training. The limited number of annotations in the fine-tuning phase might not be enough to learn discriminative features to differentiate instances of novel classes from background. MINI proposes to extract object annotations to mitigate the negative effect of unlabeled objects in training images. Another proposal is described in [52]. This work defines a label verification and correction pipeline that is fed with detections calculated by a few-shot object detector. The filtered and corrected detections are then included in the training set of the final detector. Both methods show promising results, proving that can boost the detection precision without increasing the annotation costs in applications in which unlabeled data availability is not an issue.

As the few-shot object detection field is still in an early stage, the few-shot video object detection problem remains almost unexplored. The first dataset for few-shot video object detection was recently defined in [25]. This work also proposes a meta-learning method that exploits spatio-temporal information for the first time in few-shot object detection. The proposed framework implements a tube proposal network and a tube-based matching network to compare features with category prototypes.

1.2 OBJECTIVES

The main goal of this PhD thesis is to explore how spatio-temporal information available in videos can be exploited to improve state-of-the-art results on video object detection. Specifically, we will leverage this spatio-temporal information in a wide range of scenarios including different object sizes, number of objects per frame, object movement velocity and data scarcity. To this end, the following individual objectives will be addressed:

USC DE COMPOSTELA

O1. Exploit information from a sequence of input frames to improve the per frame detection precision

Applying a single-image object detector at frame level to solve the video object detection problem would yield suboptimal results, as it does not consider spatio-temporal information. Therefore, frameworks specifically designed for this task are required in order to take advantage of all available information. The first goal of this thesis is to implement new methods that leverage spatio-temporal information by considering a group of consecutive input frames to calculate the final detection set for each reference frame.

O2. Include long-term spatio-temporal information in the feature aggregation process

The objective is to design a new aggregation process that simultaneously considers information from close and distant frames to enrich proposal features. The information provided by long-term object proposal relationships is complementary to the information extracted from highly redundant nearby frames. Thus, mining these proposal relationships is crucial to fully leverage short- and long-term spatio-temporal information. Instead of including long-term information as a post-procesing step, the use of attention mechanisms to perform long-term features aggregation will be explored.

O3. Few-shot video object detection

In this objective, we will exploit spatio-temporal information in scenarios in which labeled data is scarce. Under the few-shot paradigm, having labeled videos with annotations in every frame is not feasible. Thus, few-shot object detectors must learn from a very limited number of object instances annotated in single images. Although spatio-temporal information is not available at training time, it can be exploited in test videos. This data gap between the training and test sets imposes a major challenge for leveraging spatio-temporal information in few-shot scenarios.

1.3 CONTRIBUTIONS

The research conducted in this PhD thesis has resulted in the following contributions:

C1 The design o a new spatio-temporal object detection framework able to exploit both short- and long-term spatio-temporal information.

C1.1 The short-term feature aggregation method has two main components: (i) a tubelet proposal generator that links object instances in neighboring frames and (ii) a feature aggregation method that applies a temporal pooling operator to

propagate spatio-temporal information throughout the tubelets. Object tubelets are initialized as anchor cuboids, a spatio-temporal generalization of anchor boxes. Then, a per-frame RPN performs an independent bounding box regression to better fit the object in each frame. Finally, the temporal pooling operator summarizes tubelet features, generating a per object enriched feature map used by the network head.

C1.2 A new spatio-temporal double head that exploits both spatial information from the current frame and spatio-temporal information from neighboring frames. This model consists of two branches that process spatio-temporal aggregated features from neighboring frames and spatial information from the current frame independently. The hypothesis is that the most relevant information to localize the object is in the current frame, while both spatial information from the current frame and spatio-temporal information are valuable for the classification task.

C1.3 The final stage of the proposed framework is a long-term optimization method that updates detection confidences to ensure long-term spatio-temporal consistency. It calculates long object tubes linking object detections throughout the video and rescores all the detections that belong to the same object tube. In addition, it includes strategies to overcome missing detections in some frames that would otherwise hinder the long-term linking process.

C2 SLTnet, a new spatio-temporal object detector with enhanced short- and long-term feature aggregation.

C2.1 A more robust short-term aggregation method that considers deep feature similarities in the aggregation process. This aggregation process is reimplemented as a weighted average in which the features of each frame contribute differently to the final feature map. Therefore, errors associating features among different frames have a lower impact on the final result.

C2.2 A new component that also aggregates deep features in the long-term. Short-term aggregated feature maps are further enhanced by a multi-stage attention module that leverages proposal relationships between distant frames. This module utilizes visual features and object trajectories to establish these relationships and perform the final aggregation.

C3 FSVDet, a few-shot video object detection framework able to exploit spatio-temporal information available in videos, learning from a limited number of annotations.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C3.1 A new training strategy for the multi-stage attention module, including a method for synthetic support frame generation to augment each training image. This

allows finding object relationships between proposals in different frames in both inference and training, rather than relying on single images in the training phase. Synthetic frames are generated as copies of the input image, inserting the annotated objects in different positions. Random data augmentation is applied to the inserted objects to increase variability among different object instances.

C3.2 An extended spatio-temporal double head. The spatio-temporal branch includes a class-agnostic confidence score that predicts the overlap of the object proposal with the actual object. Therefore, the spatial branch provides spatial classification and object localization, while the spatio-temporal branch performs spatio-temporal classification and overlap prediction.

C3.3 A classification confidence score optimization based on the overlap prediction. This method associates detections in consecutive frames creating object trajectories. Then, it increases the classification score of detections that can be linked to an object trajectory that contains high-confidence detections.

Publications

The following publications constitute the scientific production of the research carried out in this thesis.

Journal Publications Publications directly related to the aforementioned contributions.

- **Daniel Cores**, Víctor M. Brea, Manuel Mucientes. Short-term anchor linking and long-term self-guided attention for video object detection. Image and Vision Computing, Vol. 110, 2021.
  Impact factor: 3.860 (JCR 2021).
  Journal ranked Q1 in JCR 2021.
  Category: COMPUTER SCIENCE, SOFTWARE ENGINEERING (22/110)
  Thesis contribution: C2.

- **Daniel Cores**, Víctor M. Brea, Manuel Mucientes. Spatiotemporal tubelet feature aggregation and object linking for small object detection in videos. Applied Intelligence, 2022.
  Impact factor: 5.019 (JCR 2021).
  Journal ranked Q2 in JCR 2021.
  Category: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (46/144)
  Thesis contribution: C1.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Conferences

- **Daniel Cores**, Manuel Mucientes, Víctor M. Brea. RoI feature propagation for video object detection. *In Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, Santiago de Compostela (Spain), 2020.
  Conference Ranking (GGS Rating): A, Class 2.
  Thesis contribution: C1.

- **Daniel Cores**, Víctor M. Brea, Manuel Mucientes. Spatio-Temporal Object Detection from UAV On-Board Cameras. *In Proceedings of the 19th International Conference on Computer Analysis of Images and Patterns (CAIP)*, Nicosia (Cyprus), 2021.
  Conference Ranking (GGS Rating): B, Class 3.
  Thesis contribution: C2.

Other Journal Publications List of publications not directly related to the main contributions of this thesis.

- Adrián González-Sieira, **Daniel Cores**, Manuel Mucientes, Alberto Bugarín. Autonomous navigation for UAVs managing motion and sensing uncertainty. Robotics and Autonomous Systems, Vol. 126, 2020.
  Impact factor: 3.120 (JCR 2020).
  Journal ranked Q2 in JCR 2020.
  Category: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (59/139)

- Brais Bosquet, **Daniel Cores**, Lorenzo Seidenari, Víctor M. Brea, Manuel Mucientes, Alberto Del Bimbo. A Full Data Augmentation Pipeline for Small Object Detection based on Generative Adversarial Networks. Pattern Recognition, Vol. 133, 2022.
  Impact factor: 8.518 (JCR 2021).
  Journal ranked Q1 in JCR 2021.
  Category: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (22/144)

- **Daniel Cores**, Nicolás Vila-Blanco, María Pérez-Alarcón, Anxo Martínez-de-Alegría, Manuel Mucientes, María J. Carreira. A few-shot approach for COVID-19 screening in standard and portable chest X-Ray images. Article **submitted** to Scientific Reports in August 2022 and currently under review.
  Impact factor: 4.996 (JCR 2021).
  Journal ranked Q2 in JCR 2021.
  Category: MULTIDISCIPLINARY SCIENCES (19/73)

USC DE COMPOSTELA

1.4 DISSERTATION STRUCTURE

This dissertation is organized into five chapters. Chapters 2, 3 and 4 provide a description of the contributions and experimental results. Chapter 5 presents the conclusions drawn from the research conducted in this thesis. Finally, Appendix A includes licensing information of the publications from which the content of this thesis is based on. More precisely, the structure of the document is as follows:

- Chapter 2 describes FANet, a new object detection framework that exploits spatio-temporal information to improve detection precision. It combines a short-term feature aggregation module fed with object proposals and a long-term optimization module that propagates the scores of the final detection set to ensure spatio-temporal consistency. The short-term aggregation module links features through neighboring frames by calculating object tubelets initialized as anchor cuboids. This method relies on the RPN receptive field to independently adapt anchor cuboids in each frame to better fit the object in the corresponding frame. Then, a new temporal pooling operator aggregates and summarizes the information, calculating the final proposal feature maps. A new network head architecture is also proposed, following a two branch approach to process spatial information from the current frame and spatio-temporal aggregated information from multiple frames. The final long-term optimization corrects the confidence of object detections, ensuring spatio-temporal consistency.

- Chapter 3 presents SLTnet, an improved version of FANet that includes short- and long-term feature aggregation to calculate the final proposal feature maps. This proposal includes a more robust short-term feature aggregation based on similarity weights among different proposal features in different frames. Short-term aggregated feature maps are further enhanced by including long-term information. An attention module mines proposal relationships in the long-term taking into account for the first time in video object detection appearance features and geometry features. Instead of bounding box geometry features, object trajectories are used in this module to compare proposal positions on different frames. This chapter also explores the generalization capabilities of state-of-the-art video object detectors, providing an extensive analysis in five different video datasets: ImageNet VID, UAVDT, Visdrone, USC-GRAD-STDdb and MOT20.

- Chapter 4 introduces FSVDet, a few-shot video object detection framework. This chapter describes the main challenges of leveraging spatio-temporal information in a data scarce scenario in which training data is limited. The generation of synthetic support frames from single images is proposed to augment the training set in each training iteration. The spatio-temporal double head is extended to also predict the overlap of each detection with

the actual objects in the current frame. This framework also includes a confidence score optimization method throughout long object tubes. The object linking process is guided by the predicted overlap score to increase the classification score throughout the tubes.

- Chapter 5 summarizes the conclusions of the research developed in this thesis. This chapter also discusses future work based on these findings.
- Appendix A contains license information and reproduction rights of the work published throughout the development of this PhD thesis. The content of this dissertation is partially based on these publications.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

2 Spatio-temporal Tubelet Feature Aggregation and Object Linking for Video Object Detection

This chapter describes a new video object detection framework that leverages spatio-temporal information to improve object detection precision in videos. To exploit all the information contained in videos for object detection, it is necessary to design specific detection frameworks, rather than directly applying single-image object detectors at frame level. The proposed solution is based on a two-stage object detector. It focuses on improving proposal feature maps through short-term aggregation and long-term confidence optimization over the final detection set. A temporal pooling operator takes short object tubelets that link proposals in neighboring frames and calculates aggregated feature maps. We also propose a novel network head architecture based on two branches, a spatial branch that localizes the object based on the information provided by the current frame and performs the classification task based on both: spatio-temporal information from neighboring frames and spatial information from the current frame. Finally, a long-term object linking optimizes detections' scores, ensuring spatio-temporal consistency. This component incorporates mechanisms to overcome detection errors such as missing detections that might hinder the long-term tube creation. The whole system has been evaluated in the widely used ImageNet VID dataset and in the USC-GRAD-STDdb dataset, specifically designed to evaluate the small object detection performance.

The contents of this chapter are extracted from the following publications:

> Daniel Cores*, Víctor M. Brea*, Manuel Mucientes*. Spatiotemporal tubelet feature aggregation and object linking for small object detection in videos. Applied Intelligence, 2022. DOI: 10.1007/s10489-022-03529-w

> Daniel Cores *, Manuel Mucientes *, Víctor M. Brea *. RoI feature propagation for video object detection. *In Proceedings of the 24th European Conference on Artificial Intelligence*

USC
DE SANTIAGO
DE COMPOSTELA

*Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) and Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.

(ECAI), pp. 2680-2687, Santiago de Compostela (Spain), 2020. DOI: 10.3233/FAIA200406

2.1 INTRODUCTION

Object detection has been one of the most active research topics in computer vision for the past years. However, the use of temporal information in videos to boost the detection precision is still an open problem. Although a traditional object detection framework can be executed at frame level, it does not exploit temporal information available on videos that can be crucial to address challenges such as motion blur, occlusions or changes in objects appearance at certain frames.

In general, object detection frameworks implement two main tasks: bounding box regression and object classification. The bounding box regression is responsible for the spatial object localization. Therefore, in video processing, the current frame at each time instant provides the most valuable information to perform this task. In contrast, we hypothesize that extracting spatio-temporal information from the appearance of the object in previous frames can improve significantly the classification task accuracy. This raises the issue of linking and aggregating spatio-temporal features throughout time.

State-of-the-art object detection frameworks are based on two major approaches: one stage and two stage architectures. In one stage methods [64], [70], the network head has to process a dense set of candidate object locations, with a high imbalance between objects of interest and background sample. Two stage frameworks [33], [34], [72] address this issue adding an object proposal method as the first stage that filters out most of the candidates associated with background. Then, the network head refines the proposal set. We develop a two stage spatio-temporal architecture in which the proposals generated by the first stage are also used to propagate information from previous frames.

The proposed framework is an extension of the Faster R-CNN model to use both temporal and spatial information to enhance the network classification accuracy. As the original Faster R-CNN, our deep convolutional network can be trained end-to-end without any precomputed proposals. The main novelties of our proposal are:

- A tubelet proposal method that works with Feature Pyramid Network (FPN) models dealing with multiple Region Proposal Networks (RPN) and extracting RoI features at different pyramid levels. As far as we know, this is the first spatio-temporal framework with a multiple-level backbone such as FPN.
- A temporal pooling method able to summarize information from the previous N frames producing a feature map with the same size as if the network was working with a single

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

frame. Thus, it works with a fully connected network head with the same number of parameters and a constant execution time independently of N.

- A spatio-temporal double head. One branch takes spatial information from the current frame to perform the bounding box regression and the spatial classification, and another branch complements the classification output using information from the previous N frames.

- A long-term optimization method creates long tubes associating object instances throughout the video, rescoring all detections belonging to the same tube. This method reuses short-term information to improve the long tubes creation process, overcoming network errors at certain frames such as missing detections that could otherwise break the tubes.

A preliminary version of this spatio-temporal framework was described in [16]. The main differences in the CNN come from the new double head architecture capable of better combining spatio-temporal information with the spatial information extracted from the current frame. Moreover, we extend the experimentation evaluating new training and testing strategies. All these changes lead to an improvement of 2.7% mAP with respect to [16] in the widely used ImageNet VID dataset. We also add a small object detection dataset (USC-GRAD-STDdb) [4] to assess the performance of our network by providing new scenarios that are not covered by Imagenet VID.

2.2 Related work

The object detection problem was first defined in the single image domain following two main approaches: two stage and one stage architectures. Lately, spatio-temporal frameworks were proposed based on these methods but taking into account the temporal information available in videos to improve the detection precision.

Two stage architectures were first popularized by R-CNN [34]. R-CNN needs to apply a Convolutional Neural Network (CNN) on each region of a precalculated proposal set, resulting in a very slow approach. This issue is addressed in Fast R-CNN [33], adding an RoI pooling layer on top of the CNN. Instead of executing the CNN over each proposal, it takes the whole image as input, generating a global deep feature map. Then, the RoI pooling generates a per proposal feature map extracting the corresponding features. This way, it improves dramatically both the training and test times by sharing all the CNN backbone calculations.

All these methods rely on an external region proposal method. The Faster R-CNN framework [72] defines an RPN to generate the proposal set in a fully convolutional fashion reusing the backbone calculations. This makes it possible to perform an end-to-end training of the whole system without any precomputed information. The Feature Pyramid Network (FPN) [60]

proposes a change in the definition of the CNN backbone extracting feature maps at different depth levels instead of taking just the deepest one. Therefore, the RPN and the network head have to calculate object proposals and the final detections at different feature map levels. FPN implements a top-down pathway and lateral connections to combine low-resolution, semantically strong features with high-resolution, semantically weak features to be able to work at different levels without loosing semantic meaning.

The R-FCN object detector [17] re-implements the network header avoiding the fully connected layers used by previous work. Instead, it follows a fully convolutional approach changing the RoI pooling by a position sensitive RoI pooling. Many other works propose new header implementations based on the original Faster R-CNN framework such as the Cascade R-CNN [6]. This method defines a multi-stage head that refines the object proposal set iteratively. Wu *et al.* [93] studied how the head architecture affects the bounding box regression and the classification tasks proposing a double head approach to take advantage of the benefits of each architecture in each task.

One stage detectors such as SSD [64] and YOLO [70] do not refine a previously calculated proposal set. Instead, they generate a dense grid of bounding boxes and directly calculate the final detection set. Recent works [61] try to overcome the object/background imbalance problem modifying the cost function to prevent easy examples from having a huge impact on the network training process.

Since the introduction of the ImageNet object detection from videos (VID) challenge, the video object detection problem has drawn the attention of the research community. Even so, how temporal information available in videos should be used to improve the detection performance remains an open problem. In fact, the same issue also remains unsolved in the action recognition field.

Two-stream networks such as [9], [27], [28] or [77] have become the standard approach in action recognition. One of the branches processes video frames, while the other one takes precomputed dense optical flow frames as input. Although action recognition is a related problem, the benefits of adding optical flow information to spatio-temporal object detectors might not be so straightforward. To be able to distinguish some action classes such as *"sitting down"* and *"getting up"*, motion information given by optical flow might be crucial. This is not so evident in object detection. For instance, authors in [47] conclude that traditional methods for mixing movement and appearance information work for actions but not for objects using the same network architecture.

Still, optical flow information has proven successful in [104] for video object detection. This work proposes to aggregate deep spatial features throughout time to improve the per-frame feature maps. To do so, it utilizes the movement information given by the optical flow to find the correspondences between the current features and the nearby feature maps. We propose

a novel method to find these correspondences working with two stage frameworks by linking object proposals throughout time.

Several approaches have been proposed to link object detections throughout neighbouring frames making up short object tubelets. Object tracking has been used to link detections generated by a frame level object detector in [51]. T-CNN [50] works on two single frame detectors and also applies tracking to link these detections over time. Kang *et al.* [49] define a Tubelet Proposal Network (TPN) with two main components. First, it propagates static proposals at frame level across time. Then, the second network component calculates the bounding box displacement in each frame to build the tubelet proposal. Although this second component works with pooled features extracted using the same bounding box across time, the network can handle moving objects due to the generally large receptive field of CNNs.

Alternatively, the idea of *anchor box* defined in single frame object detectors can be extended to the spatio-temporal domain. The ACtion Tubelet detector (ACT-detector) [46] utilizes *anchor cuboids* to initialize the action tubelets. Tang *et al.* [83] propose a Cuboid Proposal Network (CPN) as the first step for short object tubelet detection. Unlike these previous methods, in our implementation, each *anchor box* in the *anchor cuboid* is regressed independently by the corresponding RPN with information from the corresponding frame, allowing us to reuse part of the computations from previous iterations while preserving the proposals linked throughout consecutive frames.

The aim of the described methods is to link objects in the short-term. Therefore, they only take into account the nearby frames wasting long-term information. To overcome this, Feichtenhofer *et al.* [29] use tracking information as input to an object linking algorithm to build long tubes, and aggregate detection scores throughout the tube. To do that, tracking and object detection are performed and learned simultaneously using a multi-task objective training.

Tang *et al.* [83] also perform a long-term object linking concatenating small tubelets. First, they calculate short temporally overlapping tubelets so one single frame could have detections associated with more than one tubelet. Then, tubelets are joined by analyzing the overlap in the shared frames.

All these methods only consider the final detections set to perform the long-term object linking. In this work, we propose to include short-term information calculated by the RPN to overcome missed detections, being able to build larger tubes. In doing so, we only use object proposals and detections given by the network without any external tracking method to aid the object linking process.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

2.3 FANet Architecture

The proposed framework (Fig.2.1) generalizes a single frame two stage object detector to take advantage of working with a sequence of N frames $f_{t-N-1}, ..., f_{t-1}, f_t$ to improve the accuracy

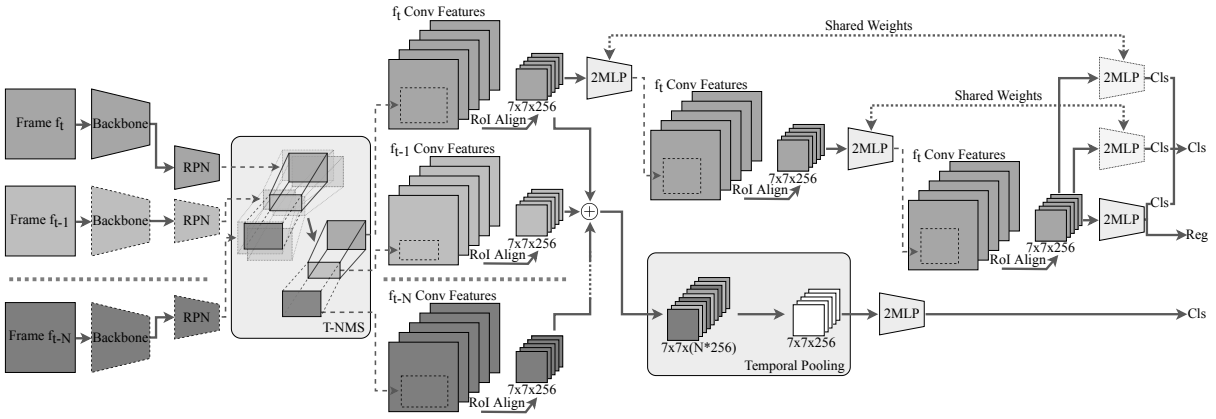


Figure 2.1: FANet architecture with a single level backbone. Dotted backbone and RPN boxes represent components that are reused without new computations. Blocks labeled as 2MLP in the network head implement a multilayer perceptron with two fully connected layers.

in each frame f_t. Even though we build our system on Faster R-CNN [72] with a Feature Pyramid Network (FPN) backbone [60] to illustrate the network architecture, the same ideas could be applied to other models. Indeed, since the spatio-temporal tubelet proposal is a core concept in our architecture, the multi-scale level approach imposes higher complexity than single scale models due to the multiple Region Proposal Networks (RPN). These object tubelets link proposals throughout time allowing the network to combine object information from multiple temporal instants.

We initialize object tubelets as a sequence of N *anchor boxes*, one per frame, in the same position, with the same area and aspect ratio. Then, the RPN modifies each *anchor box* independently in the corresponding frame using features calculated by the corresponding backbone. This, together with the fact that both the network backbone and the RPN share the convolutional weights among all input frames, allows us to reuse the backbone and RPN computations to reduce the overhead associated with our spatio-temporal approach with respect to the single frame method. This way, feature maps and object proposals associated with frame f_{t-1} at time t become feature maps and proposals associated with frame f_{t-2} at time $t+1$. As a result, this process outputs a set of tubelet proposals that must be filtered to remove the spatially redundant ones. To do that, we add a Tubelet Non-Maximum Suppression (T-NMS) [83] algorithm on top of the RPNs. The tubelet generation is further addressed in Sec. 2.3.1.

Having the tubelet proposals and the backbone features, the RoI Align method [42] extracts a per frame and per tubelet feature map of fixed size centered on the object. Therefore, features from these feature maps can be directly associated by position and propagated throughout previous frames. Fig. 2.1 shows the simplest case in which we have a single level backbone rather than the more complex FPN. After the RoI Align, we concatenate all feature maps associated with proposals belonging to the same tubelet. Then, we shuffled the channels, so channels in the same position in the original feature maps are consecutive in the concatenated

one (Fig.2.1). The resulting feature map has a joined dimension N times the original RoI Align size, making it dependent on the number of input frames. The Temporal Pooling method reduces this dimension to a fixed size independently of N. Sec. 2.3.2 describes the joining and pooling processes in detail.

We introduce a double head approach to process both the current frame and the spatio-temporal information (Sec. 2.3.3). The spatial branch is fed by the RoI Align output in the current frame, while the spatio-temporal branch takes the Temporal Pooling output. Since the spatial branch is mainly intended to localize the object in current frame, we follow a multi-stage object architecture [6] in which each stage output is the input to the next. This way, it refines object proposals gradually to maximize the overlap with the actual object. The complete proposed network can be trained end-to-end without heavily engineered proposals.

Lastly, per frame object detections are linked making up long tubes. Short-term object tubelets provide helpful information about whether or not two detections in different neighbouring frames are the same object. The proposed long-term object linking algorithm takes this information as input to grow the final tubes. Then, classification score is propagated throughout each object tube (Sec. 2.3.4).

2.3.1 Tubelet proposals

In general, the starting point for most object detectors is a set of predefined *anchor boxes*. Then, they adjust these *anchor boxes* to better fit the object and assign an object category removing those classified as background. Similarly, tubelet proposals are initialized as a sequence of N *anchor boxes*, generating *anchor cuboids* [46] in our spatio-temporal framework. Each *anchor box* in the *anchor cuboid* has exactly the same size, aspect ratio, and is in the same position for all input frames. Therefore, the number of *anchor cuboids* remains the same as the number of *anchor boxes* in the single frame approach, distributing k *anchor boxes* at each sliding position $W \times H$, generating $W \times H \times k$ *anchor cuboids* in total.

The FPN (Feature Pyramid Network) backbone [60] distributes the *anchor boxes* among the RPNs by area. In our implementation, every single *anchor box* b^i in the *anchor cuboid* sequence $(b^{t-N-1}, ..., b^{t-1}, b^t)$ is regressed independently in its corresponding frame at the appropriate level. This means that, the RPN can reuse the bounding box regression calculations for $b^i \in (b^{t-N-1}, ..., b^{t-1})$, and only needs to calculate b^t at the current instant. Adapting this strategy to single level models just implies that *anchor boxes* belonging to all *anchor cuboids* are being mapped to the same level.

Regressing every *anchor cuboid* leads to spatially redundant tubelets in the proposal set. In two stage single frame frameworks, this problem is generally solved executing the Non Maximum Suppression (NMS) method over the proposal set. In our case, doing a per frame NMS might break the tubelets, removing some of their bounding boxes. Instead, we apply an extension of

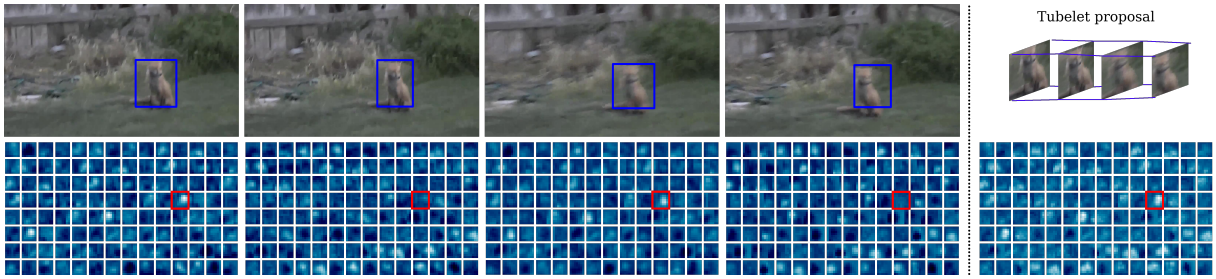


Figure 2.2: Temporal pooling example with a number of input frames $N = 4$. From left to right, there are the four input frames with an object proposal (the blue bounding box) in each one, belonging all of them to the same tubelet. Under each input frame, a subset of channels from the RoI Align output are represented. On the right, we show the tubelet proposal linking all box proposals. At the bottom right, we also show the aggregated feature map calculated by our Temporal Pooling method. We highlight one channel (surrounded in red) as an example of how the highest activations (lighter colors) in each frame contribute to the aggregated feature map.

the NMS algorithm called Tubelet Non-Maximum Suppression (T-NMS) based on [83], but with different metrics to calculate the tubelet score $ts(\tau_i)$ and the overlap between two tubelets τ_i and τ_j. This method aims to remove spatially redundant tubelets instead of single boxes.

The score of a given tubelet τ_i is calculated as:

$$ts(\tau_i) = mean(bs_i^{t-N-1}, bs_i^{t-1}, ..., bs_i^t). \tag{2.1}$$

being bs^t the score associated with the proposal b at frame t.

The overlap between a pair of tubelets τ_i and τ_j is defined as:

$$overlap(\tau_i, \tau_j) = \underset{k=t-N-1,...,t}{\text{mean}} IoU(b_i^k, b_j^k). \tag{2.2}$$

Unlike the original FPN strategy which performs a per level NMS, our T-NMS implementation removes globally the spatially redundant proposals, taking as input the whole set of tubelets. The resultant subset $\mathcal{T}$ represents the final collection of proposals.

2.3.2 Temporal Pooling

In Faster R-CNN based object detectors, an RoI feature pooling method takes the proposal set to produce a per proposal fixed size feature map. Working with FPN, object proposals are distributed among the pyramid levels according to their size to perform the RoI pooling over the corresponding feature map. We use the RoI Align [42] method to perform the feature pooling operation taking each bounding box b^j belonging to each tubelet $\tau_i = (b_i^{t-N-1}, ..., b_i^{t-1}, b_i^t)$ to extract features from the corresponding pyramid level in the frame f_j. Performing this mapping process independently in each frame rather than per tubelet allows to map each box b^j within the same tubelet τ_i to a different pyramid level making our system robust against scale variations in the tubelet sequence. As a result, we have a fixed size feature map (in our case of $7 \times 7 \times 256$, Fig. 2.1) associated with each box in the tubelet.

On top of that, we introduce an operator called *Temporal Pooling* that summarizes the whole tubelet information generating a feature map with just the same size as the RoI Align output for a single frame, independently of the number of input frames N. This method requires N to be small enough to allow the RPN to adapt the corresponding *anchor box* in the *anchor cuboid* sequence to fit the object in each frame. Working with a large N, the object movement might exceed the RPN receptive field, making impossible to adjust the same *anchor box* in every frame to the target object *. Since, all RoI Align outputs associated with the same tubelet have a fixed size and are centered in occurrences of the same object over time, we can associate features in the same position of the RoI pooled feature map in every frame.

Although errors in the RPN bounding box regression might cause feature mismatches, our method is robust against small variations since we use pooled features. These pooled features are a broad representation of features extracted from the backbone, so slight changes on the proposal bounding box do not have a significant effect in the RoI Align output.

Firstly, to perform the temporal aggregation, our method concatenates the N input feature maps with size $W \times H \times C$ into one feature map with size $W \times H \times N \cdot C$ (Fig. 2.1). Then, we reorder the resulting feature map channels, so that channels at the same position in all the input feature maps are concatenated consecutively (see the input to the Temporal Pooling operator in Fig. 2.1). Finally, the Temporal Pooling applies the following transformation to get each element of the resultant feature map X:

$$x_{ijk} = \max_{t=0,\ldots,N-1} (y_{ij(N(k-1)+t)}) \tag{2.3}$$

being x_{ijk} an element in the position $i \times j$ in channel k of the output feature map, and $y_{ij(N(k-1)+t)}$ an element in the position $i \times j$ in channel $(N(k-1)+t)$ from the concatenated feature map. As Fig. 2.2 shows, this propagates the highest activation values throughout the tubelet.

2.3.3 Spatio-temporal double head

The network head follows a double head approach based on the idea that aggregated spatio-temporal information is valuable to improve object classification while spatial information extracted from the current frame is crucial for bounding box regression. Consequently, we design each head branch to perform better in their respective task taking into account the input data.

The spatial head (Fig. 2.1 at the upper right) takes as input the RoI Align output in the current frame f_t to calculate a spatial object classification and a class agnostic bounding box regression. Since the main goal of this branch is the object localization, we implement a cascade of object

*As Section 2.4.3 shows, the network achieves the best result with $N = 6$.

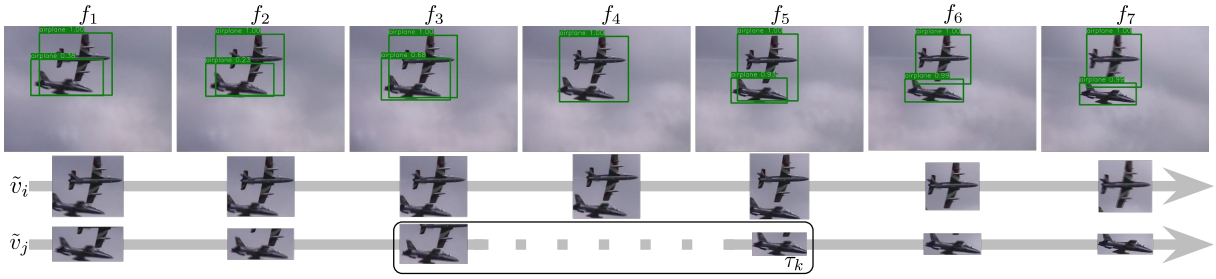


Figure 2.3: Long-term object linking. Green boxes are actual network detections. The network does not detect the two objects in f_4 breaking $\tilde{v}_j$ in two fragments. The last detection of the first fragment and the first detection of the second fragment belong to the same RPN tubelet (τ_k).

detectors [6] to iteratively refine the object proposal set until the final bounding box regression is done.

Following the Cascade R-CNN training strategy, we perform a proposal resampling after each stage applying an increasing IoU threshold to assign each proposal to a ground truth object. This way, the requirements to consider one proposal as positive example to train the corresponding stage are harder as we advance in the cascade. In general, an IoU too high might assign all proposals to background vanishing the positive examples. Nevertheless, as each stage takes as input the refined proposal set by the previous stage, we can increase the IoU achieving more accurate boxes in each stage. On test time, we use the average of the classification scores calculated by every stage detector over the final proposal set [6].

The spatio-temporal head (Fig. 2.1 at the bottom right) classifies the object based on the output of the Temporal Pooling operator. Therefore, this branch takes into account the appearance of the object in the previous N frames.

As a result, this strategy produces a bounding box regression and two object classification vectors, one calculated using features in the current frame and another calculated using aggregated features through the previous N frames. The final classification is calculated as follows [93]:

$$p = p_{tmp} + p_{spt}(1 - p_{tmp}) \tag{2.4}$$

where p_{spt} and p_{tmp} are the score vectors from the spatial and temporal heads, respectively. As a result, we are considering spatial classification in the current frame and spatio-temporal classification at the same level. That makes the current frame f_t to have a greater influence in the final classification result than any previous frame f_i in $f_{t-N-1}, ..., f_{t-1}$.

2.3.4 Long-term object linking

We propose a two step long-term object linking algorithm that takes network detections and produces long object tubes. Linking network detections over time to identify single action/object instances has become a standard approach in both action recognition [35], [75], and recently in

object detection [29], [50], [83]. These methods try to join single boxes or small tubelets into larger tubes in order to propagate the classification score throughout the video.

In the first stage of our method, the frame to frame linking problem is defined as an optimization problem that maximizes the accumulated linking score in each tubelet. Network errors such as false negatives or misclassified detections might brake large tubes since it is not possible to find a detection to link in some frames. The second step of the long term linking method utilizes the short-term tubelet information to overcome some of these problems allowing the algorithm to produce larger tubes.

In this implementation, each detection $d_t^i = \{x_t^i, y_t^i, w_t^i, h_t^i, p_t^i\}$ in the set D_t indexed by i in the frame t, is composed of a box centered at (x_t^i, y_t^i) with width w_t^i and height h_t^i, and an associated classification confidence p_t^i for the object class. A lower threshold β is applied over the detection set D_t before the object linking algorithm to prevent that low confidence detections adversely affect the tube creation. The linking score $ls(d^i, d^j)$ between two detections d^i and d^j at different frames t and t' is defined as:

$$ls(d_t^i, d_{t'}^j) = p_t^i + p_{t'}^j + IoU(d_t^i, d_{t'}^j). \tag{2.5}$$

Then, the optimal tube $\hat{v}$ can be found by solving the following optimization problem applying the Viterbi algorithm per object category:

$$\hat{v} = \arg\max_{\mathcal{V}} \sum_{t=2}^{T} ls(D_{t-1}, D_t) \tag{2.6}$$

where $\mathcal{V}$ is the set of all possible tubes.

Algorithm 1 describes how we create long-term object tubes in detail. This method solves Equation 2.6 (Algorithm 1:4) iteratively to find all tubes ending in frame i, starting with tubes ending in the last frame $i = T$. In each iteration, all detections belonging to the best tube $\hat{v}$ are removed from the candidate set $\mathcal{D}$ (Algorithm 1:5) and the new tube is added to the result set (Algorithm 1:6). When there are no more candidate detections in the current frame (Algorithm 1:3), the same process is applied to find all tubes ending in the previous frame (Algorithm 1:2). The process ends when there are no more tubes ending in the second frame of the video. This is due to the fact that we directly create a tube of length $|\hat{v}| = 1$ for each detection in the first frame that does not belong to any tube computed in previous iterations.

Previous works defined action/objects tubes as sequences of consecutive detection boxes, without interruptions. In fact, the output of the method described above follows this definition of object tube. Nevertheless, object occlusions or even network errors such as false negatives or misclassified examples might artificially break large tubes into smaller chunks.

Information given by RPN tubelets can be used to link small tubes to provide larger ones, making up the second step of the long-term object linking. In Fig. 2.3, both the last detection of

Algorithm 1: Long-term tubes creation

Input : Per frame detection set $\mathcal{D} = \{D_t = (d_t^1, ..., d_t^{n_t})\}_{t=1}^{T}$
Input : All possible tubes: $\mathcal{V}$
Output Object tubes $\hat{\mathcal{V}}$
:
$\hat{\mathcal{V}} \leftarrow \emptyset$
for i **in** $T, ..., 2$ **do**
  while $D_i \neq \emptyset$ **do**
    $\hat{v} \leftarrow \arg\max_{\mathcal{V}} \sum_{t=2}^{i} ls(D_{t-1}, D_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \setminus \hat{v}$
    $\hat{\mathcal{V}} \leftarrow \hat{\mathcal{V}} \cup \hat{v}$

Algorithm 2: Long-term object tube linking

Input : Per frame detection set $\mathcal{D} = \{D_t = (d_t^1, ..., d_t^{n_t})\}_{t=1}^{T}$
Input : Tubelet set $\mathcal{T} = \{\tau_i = (b_i^1, ..., b_i^N)\}_{i=1}^{\theta}$
Input : Object tubes $\hat{\mathcal{V}} = \{\hat{v}_i = (d^{i,1}, ..., d^{i,m_i}\}_{i=1}^{\delta}$
Output Joined object tubes $\tilde{\mathcal{V}}$
:
$\tilde{\mathcal{V}} \leftarrow \hat{\mathcal{V}}$
for $\hat{v}_i$ **in** $\hat{\mathcal{V}}$ **do**
  for $\hat{v}_j$ **in** $\hat{\mathcal{V}}$ **do**
    $ts_{max} = 0$
    for τ_l **in** $\mathcal{T}$ **do**
      if $\exists b_l^k \in \tau_l \mid \gamma(b_l^k, d^{i,m_i})$ **and** $\exists b_l^{k'} \in \tau_l \mid \gamma(b_l^{k'}, d^{j,1})$ **and** $time(d^{i,m_i}) > time(d^{j,1})$ **then**
        if $ts(\tau_l) > ts_{max}$ **then**
          $ts_{max} = ts(\tau_l)$
    $C_{ij} = ts_{max}$
$\mathcal{H} \leftarrow Hungarian(C)$
for h_i **in** $\mathcal{H}$ **do**
  $\tilde{\mathcal{V}} \leftarrow \tilde{\mathcal{V}} \setminus \hat{v}_{h_i}$
  $\tilde{v}_i \leftarrow \tilde{v}_i \cup \hat{v}_{h_i}$
for $\tilde{v}_i$ **in** $\tilde{\mathcal{V}}$ **do**
  $updateScores(\tilde{v}_i)$

the first fragment and the first detection of the second fragment of $\tilde{v}_j$ belong to the same RPN tubelet. Having this information into account, our algorithm links the two fragments making one larger tube.

Algorithm 2 describes the tube linking method in detail. Formally, giving two tubes $\hat{v}_i = (d^{i,1}, ..., d^{i,m_i})$ and $\hat{v}_j = (d^{j,1}, ..., d^{j,m_j})$ with size m_i and m_j respectively, both will be joined in

a tube of size $m_i + m_j$ if $d^{j,1}$ follows d^{i,m_i} in time, and both detections belong to the same RPN tubelet (Algorithm 2:6). Thus, the tubelet allows to link both tubes as it contains detections from both of them, although the tubes do not have temporal overlap.

The detection set D_t is the output of Non-Maximum Suppression (NMS) followed by a Bounding Box Voting transformation [32] to the actual network output. This method takes the highest score detection d and removes all other detections with an overlap with d higher than a given threshold. Therefore, the final detection d has many associated network outputs. Since each network detection came from one RPN tubelet, d has also many RPN tubelets associated, one per each suppressed detection. Therefore, there can be many tubelets τ that contain the first or the last detection box of a specific tube. The function $\gamma(b_l^k, d^{i,m_i})$ returns True if the detection d^{i,m_i} is associated with the box proposal b_l^k in the tubelet τ_l (Algorithm 2:6).

For a given tube $\hat{v}_i = (d^{i,1}, ..., d^{i,m_i})$ we might have more than one candidate tube $\hat{v}_j$ to join with. We use the RPN tubelet score defined by Equation 2.1 to choose the best candidate to link for $\hat{v}_i$. We choose the highest score of all tubelets associated with d^{i,m_i} and $d^{j,1}$ (Algorithm 2:7-8). Then, a cost matrix C can be constructed with as many rows as ending candidate tubes and as many columns as starting tubes. Each element c_{ij} is the maximum score for tubelets containing d^{i,m_i} and $d^{j,1}$. This becomes an assignment problem that can be easily solved with the *Hungarian Method* (Algorithm 2:10-13). For each tube assignment (i, j), we remove the second tube from the output set (Algorithm 2:12) and build a new tube joining the two original fragments (Algorithm 2:13). Once this process has finished, for a giving tube $\tilde{v}_i$ the score of all detections $d^{i,j} \in \tilde{v}_i$ is set to the mean of the $\alpha = 10\%$ highest scores in that tube (Algorithm 2:15).

2.4 Experiments

2.4.1 Datasets

We use the ImageNet VID dataset [74] to evaluate the proposed framework and how each component contributes to the final result. This dataset contains 30 different objects categories annotated in 3,862 training and 555 validation videos. As the test subset annotations are not publicly available and the challenge evaluation server is closed, we use the Average Precision (AP) and Mean AP over the validation set as the main evaluation metrics following the standard approach established by previous works [29, 51, 83, 104].

To train the single frame baseline we select 20 uniformly distributed frames over time from each video in the VID training set. Nevertheless, the spatio-temporal network needs N consecutive input frames in each iteration instead of just one. To do that, we select 2 groups of input frames with length $N \times 15$. We have trained our spatial baseline also with this sampling

strategy to assess that it does not bias the results analysis. We have not seen any significant differences in the precision values of the test for the two strategies in the single frame case.

Following the training procedure proposed by [29], we also use data from the ImageNet DET dataset to enhance the training set. ImageNet DET consists of 456,567 training and 20,121 validation images of 200 different categories that include the 30 object classes used in VID. We add to the training set at most 2,000 images per VID object class from ImageNet DET. This upper bound prevents the bias of the training set for large object categories in DET. In the spatio-temporal case, we transform these images into videos of N repeated frames.

In addition to the ImageNet VID dataset, we also test our proposal with the Small Target Detection database (USC-GRAD-STDdb) [4] to evaluate the proposed framework in a completely different scenario. This dataset contains 115 videos with over 56,000 annotated small objects of 5 different categories. Small objects are defined in USC-GRAD-STDdb as those that fit in a bounding box with an area range between 16 ($\approx 4\times4$) to 256 ($\approx 16\times16$) pixels. This makes a huge difference between both datasets since all anotations in ImageNet VID have an area greater than 256 pixels.

2.4.2 Implementation details

The chosen backbone is a Feature Pyramid Network (FPN) based on ResNeXt-101 [96] pre-trained on the ImageNet classification dataset. We add to both the single frame and the spatio-temporal approach a 3 stage cascade of detectors [6] as described in Fig. 2.1. To train our spatio-temporal framework, we first train the single frame model and then we initialize the spatio-temporal network with the same weights keeping all learned layers frozen. Thus, we only have to train the spatio-temporal head if we have the equivalent single frame model already trained.

All input images are scaled so that the smallest dimension is 720 at most. If the highest dimension remains above 1,280 pixels, the image is scaled down to prevent that. In any case, the image scaling keeps the original aspect ratio.

For the spatial baseline, we use the SGD learning algorithm with a learning rate set to 2.5×10^{-4} for the first 240K iterations, 2.5×10^{-5} for the next 80K iterations, and 2.5×10^{-6} for the last 40K iterations. RPN redundant proposals are suppressed by NMS with a threshold of 0.7, while the final detection set is filtered by means of NMS with an IoU threshold equal to 0.5.

To train the spatio-temporal head we set the learning rate to 1.25×10^{-3} for the first 180K iterations, 1.25×10^{-4} for the next 60K iterations, and 1.25×10^{-5} for the last 30K iterations. The network needs N input images for each example: the current frame and the $N-1$ previous ones. For this reason, we replicate the first frame $N-1$ times to be able to process the first $N-1$ frames of each video.

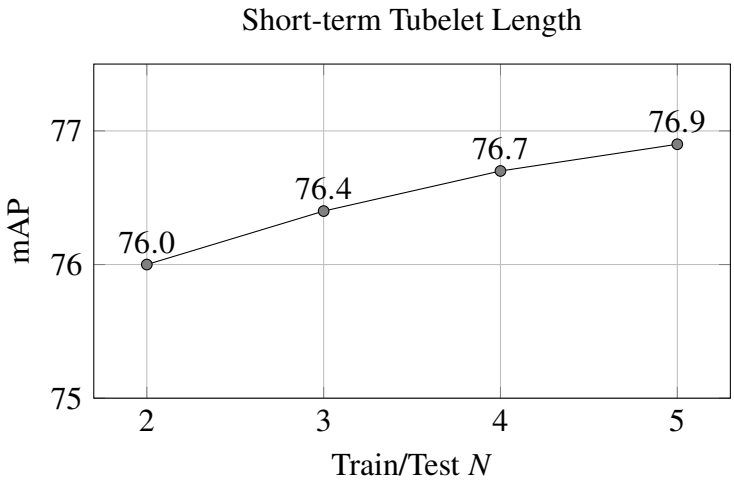


Figure 2.4: Detection mAP when training and testing with the same tubelet lengths (N) on the ImageNet VID validation set without long-term information.

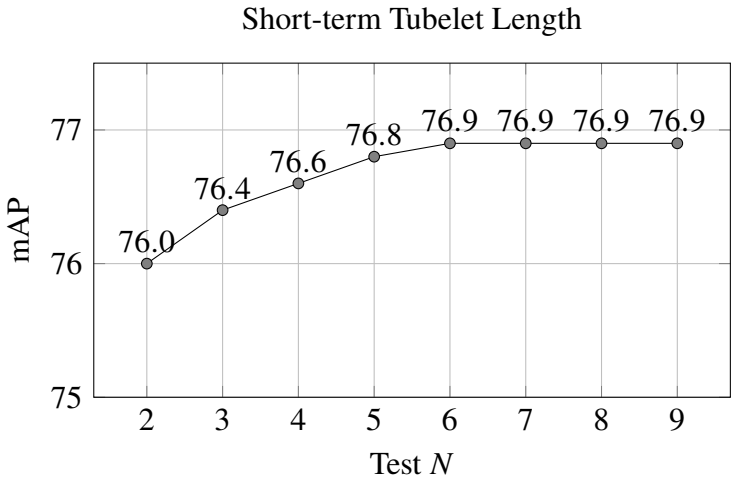


Figure 2.5: Detection mAP when training with $N = 2$ and testing with different tubelet lengths (N) on the ImageNet VID validation set without long-term information.

Finally, we apply a Bounding Box Voting transformation [32] and a confidence threshold $\beta = 0.05$ over the output detection set. This filtering prevents detections with lower scores from degrading the long-term object linking output.

2.4.3 Ablation studies

We conducted a series of experiments to asses how the number of input frames affects the network precision. Moreover, we also perform a collection of ablation studies in order to better understand how each network component contributes to the final result.

Fig. 2.4 shows how the number of input frames N influences the network precision on the ImageNet VID validation set. Our current training implementation keeps all backbone

Table 2.1: Influence of each component on the framework precision on ImageNet VID dataset.

Spatial head Cls	Spatio-temporal head Cls	Cascade head	Long-term object linking	Mean AP
✓				74.4
	✓			73.7
✓	✓			76.0
✓		✓		75.3
	✓	✓		73.9
✓	✓	✓		76.9
✓	✓		✓	78.5
✓	✓	✓	✓	78.6

computations for every input frame in GPU memory, making unfeasible to train with more than $N = 5$ in a single Nvidia Tesla P40 24GB. However, since the number of parameters in our framework is independent of the number of input frames, we can tests the network with a higher N independently of the training configuration. Fig. 2.5 shows the network precision training with $N = 2$ input frames and testing up to a maximum of 9 frames. By comparing Fig. 2.4 and Fig. 2.5, we can conclude that training and testing with a different number of input frames barely affects the network mAP, compared with training and testing with the same N, showing a maximum difference of 0.1% mAP for $N = 4$ and $N = 5$.

Fig. 2.5 shows how increasing the number of input frames up to 6 leads to an improvement in the network precision, proving that our approach exploits spatio-temporal information in the short-term. Fig. 2.5 also shows that increasing N above 6 has no impact in mAP. This might be because of the tubelet initialization based on *anchor cuboids*: it expects that the same object can be associated with the same *anchor box* in the same position in every input frame. If the object movement exceeds the network receptive field this assumption is not true, which is more likely for large N values.

Table 2.1 details how each network component contributes to the final mean average precision (mAP) testing on the ImageNet VID validation set. We use in all the experiments the bounding boxes calculated by the spatial header, choosing the classification scores from the spatial branch, the spatio-temporal branch or the combination of both (Eq. 2.4). Regarding the bounding boxes, we also compare the network precision with and without refining the object proposals using a cascade of detectors. In these experiments, using only the classification score calculated by the spatial head is different from just running the spatial network baseline. This is due to the T-NMS method that filters the proposal set in a different fashion than the conventional NMS.

Table 2.1 shows how the combination of the spatial and the spatio-temporal data outperforms the execution of each of them independently. It happens with and without applying a cascade refinement, proving that they provide complementary information in both cases. The cascade

Table 2.2: VID validation set results. We use a number of input frames $N = 6$ to test our framework.

	airplane	antelope	bear	bicycle	bird	bus	car	cattle	dog	domestic cat	elephant
Kang *et al.* [49]	84.6	78.1	72.0	67.2	68.0	80.1	54.7	61.2	61.6	78.9	71.6
Kang *et al.* [50]	83.7	85.7	84.4	74.5	73.8	75.7	57.1	58.7	72.3	69.2	80.2
Lee *et al.* [57]	86.3	83.4	88.2	78.9	65.9	90.6	66.3	81.5	72.1	76.8	82.4
Feichtenhofer *et al.* [29]	90.2	82.3	87.9	70.1	73.2	87.7	57.0	80.6	77.3	82.6	83.0
Wang *et al.* [89]	88.7	88.4	86.9	71.4	73.0	78.9	59.3	78.5	77.8	90.6	79.1
Tang *et al.* [83]	90.5	80.1	89.0	75.7	75.5	83.5	64.0	71.4	81.3	92.3	80.0
FPN-X101 baseline	91.7	82.3	81.3	68.8	74.8	80.0	62.8	57.4	72.0	80.5	75.9
FPN-X101 Cascade baseline	93.1	83.3	82.2	72.0	75.0	82.1	63.7	64.3	71.9	77.1	77.4
FANet (Short-Term)	93.0	84.5	82.8	72.5	75.3	83.0	63.4	64.5	73.8	79.3	77.8
FANet (Short&Long-Term)	90.2	85.0	85.1	72.2	76.0	84.7	62.3	68.2	79.4	86.2	77.5
FANet* (Short-Term)	93.7	85.2	86.1	78.1	75.0	84.6	66.8	71.0	76.1	86.2	79.2
FANet* (Short&Long-Term)	90.4	85.4	88.2	77.0	76.5	87.3	65.5	76.6	81.5	92.5	79.3

	fox	giant panda	hamster	horse	lion	lizard	monkey	motorcycle	rabbit	red panda	sheep
Kang *et al.* [49]	83.2	78.1	91.5	66.8	21.6	74.4	36.6	76.3	51.4	70.6	64.2
Kang *et al.* [50]	83.4	80.5	93.1	84.2	67.8	80.3	54.8	80.6	63.7	85.7	60.5
Lee *et al.* [57]	88.9	91.3	89.3	66.5	38.0	77.1	57.3	88.8	78.2	77.7	40.6
Feichtenhofer *et al.* [29]	97.8	85.8	96.6	82.1	66.7	83.4	57.6	86.7	74.2	91.6	59.7
Wang *et al.* [89]	96.3	84.8	98.5	77.4	75.5	84.8	55.1	85.8	76.7	95.3	76.2
Tang *et al.* [83]	96.1	87.6	97.8	77.5	73.1	81.5	56.0	85.7	79.9	87.0	68.8
FPN-X101 baseline	88.8	88.7	89.9	75.0	54.8	78.6	55.2	85.8	66.7	68.6	60.1
FPN-X101 Cascade baseline	86.2	87.5	85.9	74.5	58.3	78.0	55.6	84.7	65.8	81.5	62.1
FANet (Short-Term)	87.1	87.6	89.0	74.3	62.5	78.8	57.3	85.2	63.1	82.8	62.3
FANet (Short&Long-Term)	88.3	86.4	94.8	76.6	70.1	80.7	57.5	88.7	71.2	89.3	62.9
FANet* (Short-Term)	88.0	88.3	96.4	74.2	66.3	80.4	58.3	86.2	71.1	85.1	59.0
FANet* (Short&Long-Term)	89.6	87.8	98.8	77.2	74.3	82.7	58.2	89.7	78.2	91.8	62.0

Continued on next page

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Table 2.2: Continued from previous page.

	snake	squirrel	tiger	train	turtle	watercraft	whale	zebra	Mean AP
Kang *et al.* [49]	61.2	42.3	84.8	78.1	77.2	61.5	66.9	88.5	68.4
Kang *et al.* [50]	72.9	52.7	89.7	81.3	73.7	69.5	33.5	90.2	73.8
Lee *et al.* [57]	50.3	44.3	91.8	78.2	75.1	81.7	63.1	85.2	74.5
Yang *et al.* [98]	-	-	-	-	-	-	-	-	76.2
Feichtenhofer *et al.* [29]	76.4	68.4	92.6	86.1	84.3	69.7	66.3	95.2	79.8
Zhu *et al.* [104]	-	-	-	-	-	-	-	-	80.1
Wang *et al.* [89]	75.7	59.0	91.5	81.7	84.2	69.1	72.9	94.6	80.3
Bertasius *et al.* [1]	-	-	-	-	-	-	-	-	80.4
Xiao and Lee [94]	-	-	-	-	-	-	-	-	80.5
Tang *et al.* [83]	80.7	61.6	91.6	85.5	81.3	73.6	77.4	91.9	80.6
FPN-X101 baseline	59.2	53.8	89.6	84.2	77.2	72.0	75.2	90.5	74.7
FPN-X101 Cascade baseline	64.7	57.3	89.9	87.1	76.0	75.0	76.7	89.3	75.9
FANet (Short-Term)	68.2	59.5	90.5	87.2	76.8	76.2	76.8	90.4	76.9
FANet (Short&Long-Term)	69.3	62.4	90.6	85.0	77.6	76.9	72.4	90.6	78.6
FANet* (Short-Term)	75.7	60.8	92.0	87.8	78.9	76.6	77.0	90.7	79.2
FANet* (Short&Long-Term)	76.1	64.3	91.5	84.6	80.0	77.0	72.2	90.7	80.9

bounding box refinement improves the mAP in every case. As expected, this precision gain has less influence when working only with spatio-temporal classification, as it only has an effect over the bounding box regression. In the other cases, the benefits of having this cascade also influence the spatial classification, resulting in a higher mAP improvement.

Finally, Table 2.1 also shows how the long-term object linking highly improves the final mAP. This is consistent with previous works that implement this kind of post-processing methods, and reveals that the RPN tubelet information can be valuable to link object instances throughout the video. Nevertheless, applying long-term object linking reduces the improvement of the cascade head. This might be due to the fact that improving the detection set calculated without cascade is more straightforward, while the version with cascade produces more accurate detections and are, therefore, harder to improve.

2.4.4 Results

2.4.4.1 ImageNet VID

We compare our framework with state-of-the-art spatio-temporal object detectors in Table 2.2. T-CNN [50] is the winner of the ILSVRC2015 with 73.8% mAP on the validation set. This work uses two single frame object detector frameworks to calculate the initial detection set and then applies context suppression and detection propagation using optical flow. Then, it utilizes visual tracking to make up the object tubelets. Visual tracking techniques are also used in Multi-Class

Multi-Object Tracking (MCMOT) [57], the ILSVRC2016 winner [98] (they are able to boost up the system precision from 76.2% to 81.2% mAP with model ensembles) and in Detect to Track and Track to Detect [29]. Our system relies on the definition of *anchor cuboids* to build tubelets, instead of the more complex visual tracking algorithms, outperforming the best tracking-based method [29] by 1.1%.

Flow-Guided Feature Aggregation (FGFA) [104] is a core component of the ILSVRC2017 winner, achieving 80.1% mAP without bells-and-whistles on the validation set. Instead of building object tubelets, they find feature correspondences based on optical flow information. Finding correspondences based on optical flow is also present in the Motion-Aware Network (MANet) proposed by Wang *et al.* [89]. Alternatively, Bertasius *et al.* [1] define a spatio-temporal Sampling Network (STSN) based on deformable convolutions to establish feature correspondences across the video getting 80.4% mAP. We propose to find those correspondences by associating RoI pooled features through object tubelets, avoiding to calculate the optical flow, and getting 80.9% mAP.

Xiao and Lee [94] propose a Recurrent Neural Network (RNN) based on a Spatial-Temporal Memory Module (STMM) that aggregates appearance information throughout time. The STMM modifies the spatial-temporal memory taking as input deep feature maps from each frame and the previous memory state, which carries information of all previous frames. This memory contains similar information to the temporal pooling output in our implementation, aggregating spatial features over time. The extra complexity of adding a RNN over the CNN backbone does not represent a clear benefit, getting slightly worse results than our method.

Tang *et al.* [83] implement a short tubelet detection framework to identify tubelets with temporal overlapping. Then, given two tubelets, they join those tubelets analyzing the spatial overlap between bounding boxes belonging to each tubelet in the common frame. They achieve 80.6% mAP.

Our framework outperforms all methods in Table 2.2 achieving 80.9% mAP with a precision gain of 5.0% mAP compared to our single frame baseline (*FPN-X101 Cascade* in Table 2.2). We report our results with multi-scale testing —marked with * in Table 2.2— following [83], and without testing augmentation.

2.4.4.2 USC-GRAD-STDdb

We show the resuts on the USC-GRAD-STDdb dataset in Table 2.3. Our method outperforms the single frame baseline in every object category just using short-term information. Unlike [4] and [5] that give a class agnostic mAP, we report the results taking into account the object categories. Adding both long-term and short term-information leads to 49.9% mAP with an IoU threshold of 0.5. Our system improves the single frame baseline (FPN-X101 with cascade

Table 2.3: USC-GRAD-STDdb results

	Drone	Boat	Vehicle	Person	Bird	Mean $AP_{0.5}$	Mean $AP_{0.5-0.95}$
FPN-X101 baseline	65.2	44.9	36.1	62.8	13.4	44.5	16.3
FANet (Short-Term)	67.9	49.5	38.9	66.2	20.3	48.5	17.6
FANet (Short&Long-Term)	66.8	53.3	35.9	68.4	25.1	49.9	18.3

head) by 5.4% in this metric. We also compare the average mAP changing the IoU threshold from 0.5 to 0.95 with a step of 0.05, improving the single frame baseline by 2.0%.

In general, working with objects smaller than 16×16 pixels, represents a challenge for our spatio-temporal approach due to the relative movement of the objects with respect to their size: a small displacement of a few pixels might cause the object overlap among neighbouring frames to decrease dramatically, making harder to build accurate tubelet proposals from *anchor cuboids*. Therefore, we reduce the tubelet length, working with N=4 for all experiments on the USC-GRAD-STDdb dataset. In this line, the most challenging class of the dataset is *Bird*, as objects in this category are generally fast moving and with highly irregular trajectories. Nevertheless, analyzing the per category results, the improvement in the *Bird* class boosting the AP from 13.4% to 25.1% is very promising and shows the benefits of adding spatio-temporal information to the small object detection task.

2.5 CONCLUSIONS

We have described a novel framework that extends two stage object detectors to exploit spatio-temporal information to improve object detection in videos. We have proposed a feature aggregation method throughout short tubelet proposals without any other external information, such as associations given by tracking algorithms or optical flow. We have redesigned the network head, to take advantage of this aggregated spatio-temporal data, with a double head implementation specialized in both spatial and spatio-temporal information. Our experimentation has proved that this short-term information is complementary to the long-term information calculated by our linking method. The overall framework achieves competitive results in the widely used ImageNet VID dataset with 80.9% mAP. Also, it significantly outperforms the single frame baseline in the USC-GRAD-STDdb dataset.

Although our tubelet initialization based on anchor cuboids provides a light computational method to link objects through neighbour frames, it imposes a limitation in the maximum tubelet

USC DE SANTIAGO DE COMPOSTELA

length. In the future, we will further develop this component making it more flexible in order to avoid this limitation.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3 Short-Term Anchor Linking and Long-Term Self-Guided Attention for Video Object Detection

The proposed solution is a step forward compared to the version presented in Chapter 2. Instead of working with object tubelets, proposals are processed independently in the corresponding frame. Then, proposals that came from the same anchor box are linked after an independent NMS in each reference frame. Finally, short-term information is aggregated, taking into account features similarity, instead of the max temporal pooling described in Chapter 2. Besides short-term aggregation, long-term aggregation is also explored to enhance proposal feature maps. Long-term information is incorporated through a self-guided self-attention module that mines proposal relationships among different frames. This module takes into account object appearances and trajectories to establish relationships among proposals from distant frames. Trajectories calculation is guided by the similarity weights calculated by the attention module in previous frames. Thus, it includes both short- and long-term information in the detection process rather than only exploiting long-term information in a post-processing step.

We evaluated our proposal on five datasets with different characteristics, varying the number of objects per frame, object sizes and relative movement between consecutive frames. The code is available at `https://github.com/daniel-cores/SLTnet`.

The contents of this chapter are extracted from the following publications:

> Daniel Cores*, Víctor M. Brea*, Manuel Mucientes*. Short-term anchor linking and long-term self-guided attention for video object detection. Image and Vision Computing, Vol. 110, 2021. DOI: 10.1016/j.imavis.2021.104179

> Daniel Cores*, Víctor M. Brea*, Manuel Mucientes*. Spatio-Temporal Object Detection from UAV On-Board Cameras. *In Proceedings of the 19th International Conference on Computer Analysis of Images and Patterns (CAIP)*, pp. 143-152, Nicosia (Cyprus), 2021. DOI: 10.1007/978-3-030-89131-2_13

USC DE SANTIAGO DE COMPOSTELA

*Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) and Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.

3.1 INTRODUCTION

The advances in Convolutional Neural Networks (CNNs) have dramatically boosted the precision of single image object detectors. Nevertheless, applying single image methods directly on isolated video frames might produce unsatisfactory results due to challenges such as motion blur, out of focus or occlusions in some specific frames. Videos contain spatio-temporal information that single image object detectors do not exploit, and that can be very valuable to address these issues. Thus, to calculate the detection set for a given frame, spatio-temporal object detectors enrich features in the reference frame by analyzing a set of supporting frames that provide spatio-temporal context.

Spatio-temporal feature aggregation is a fundamental part in almost every state-of-the-art video object detector. Features from a set of supporting frames are aggregated to features in the reference frame, achieving more robust feature maps. Regarding the aggregation strategy, there are two main categories of spatio-temporal object detectors: pixel level aggregation methods [1,37,66,89,94,104] and object level aggregation methods [12,16,20,21,50,51,76,83]. Pixel level based methods aggregate information using per frame full size feature maps, while object level approaches focus on aggregating box features throughout time. Thus, the main goal of object level methods is to improve per proposal feature maps instead of improving the whole frame feature maps, concentrating on regions with high probability of containing an object.

We propose a new box level spatio-temporal object detection framework that exploits both short and long-term spatio-temporal information. First, we aggregate box features throughout nearby frames by applying a new proposal linking algorithm based on anchor boxes. We avoid short object tubelets used in previous work [16, 49–51, 83] to establish short-term relationships, providing a simpler and more efficient yet effective method. Also, we define a self-guided multi stage attention module that takes the short-term enhanced box features to establish long-term relationships. For the best of our knowledge, this is the first implementation that can handle both geometry and appearance features in long-term attention modules, since previous work such as [21] only exploits a short-term temporal context, and [12] only takes into account appearance features when it comes to long-term. Our approach shows state-of-the-art results in a wide variety of video object detection datasets.

The main contributions of this work are:

- A new short-term linking method throughout neighboring frames to associate object proposals. This method links each proposal in the reference frame with proposals that come from the same anchor in the supporting frames, relying on the Region Proposal Network (RPN) to adjust the anchor to the object in the corresponding frame.
- A new self-guided multi-stage attention module that can handle both appearance and geometry features in the long-term. Object position becomes meaningless in the long-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

term when it comes to compare two bounding boxes. To solve this issue, we keep track of the bounding box center, updating the bounding box position. We call this method self-guided because it reuses the attention weights from previous frames to guide the proposal tracking.

- An in-depth experimental study in five object detection video datasets with different characteristics regarding the number of objects per frame, the size of the objects, and the speed at which the position of the object changes between consecutive frames. We have compared our proposal with the state-of-the-art approaches, and we have applied a non-parametric statistical test, which shows that our method ranks first, and that the differences with the other approaches are statistically significant.

3.2 Related work

Image object detection

State-of-the-art single image object detectors follow two main approaches: two stage and one stage architectures.

Two stage frameworks were first popularized by R-CNN [34]. Then, Fast R-CNN [33], introduced an RoI pooling layer that allows the network to use a per image feature map instead of one for each object proposal. The generation of the object proposals was first integrated in the network by Faster R-CNN [72] defining a Region Proposal Network (RPN). Feature Pyramid Network (FPN) [60] produces feature maps of different resolutions with high-level semantics throughout by adding a top-down architecture with skip connections to Faster R-CNN. That idea has also been improved in current state-of-the-art networks like PANet [63] and EfficientDet [82].

One stage object detectors such as SSD [64] and YOLO [70] directly calculate the final detection set taking a dense grid of bounding boxes as input, instead of proposals targeting objects of interest. Therefore, these architectures must deal with a high imbalance between objects of interest and background examples in the network head. Authors in [61] propose a new cost function to deal with this issue. A more recent work [84] proposes an anchor free approach, avoiding the complicated computation related to anchor boxes.

All previous work considers object instances individually, without exploiting any relationship between them. Attention modules were first introduced in the object detection domain by [45] to model these relations. This work was motivated by the success of attention modules in natural language processing (NLP), modeling dependencies between different elements [87].

USC DE COMPOSTELA

Video object detection

The main idea behind most of the state-of-the-art spatio-temporal object detection frameworks is to include feature aggregation throughout several input frames to enhance the per-frame features. Some works such as [89, 104] use optical flow information to find correspondences between features in the reference frame and features in the supporting frames. Recent methods try to avoid the optical flow calculation time, for instance by learning these correspondences based just on deformable convolutions [1]. Alternatively, [94] proposes a Recurrent Neural Network (RNN), defining a module called Spatial-Temporal Memory Module (STMM) that aggregates spatial information throughout time. A more recent work [37] proposes a new module, Progressive Sparse Local Attention (PSLA), based on attention mechanisms but working in a local fashion. All these methods aim to find correspondences and aggregate features at pixel level.

Several approaches have proposed to work at object level instead of pixel level [12, 16, 20, 21, 50, 51, 76, 83], linking objects throughout time. Object level methods aggregate only useful information in areas with high probability of containing an object. We follow this object oriented approach in our architecture.

Object tracking techniques have been applied to link detections calculated at frame level in [50, 51]. As an alternative to object tracking, a Tubelet Proposal Network (TPN) was first introduced in [49]. This network has two main steps: propagation of static proposals across time and calculation of the corresponding displacement in each frame. It takes advantage of the generally large receptive field of CNNs and the spatio-temporal redundancy between consecutive frames to be able to handle moving objects using static proposals throughout neighboring frames. A similar idea is present in [83] with the definition of a Cuboid Proposal Network (CPN) as the first step for short object tubelet detection. This CPN works with *anchor cuboids*, a spatio-temporal generalization of *anchor boxes* from the single image domain, to generate short tubelets that link objects in the short-term. Using *anchor cuboids* to link object proposals is also included in [16], and was also proposed to solve the action recognition problem in [46]. In our implementation, we avoid tubelet proposals, working directly with box proposals. This reduces the overhead of adding spatio-temporal context to single image object detectors.

As in pixel level methods, attention mechanisms have also grown in popularity among object level methods [12, 20, 21, 76]. The method described in [45] for single image object detection was extended to videos in [21], modeling relationships between object proposals across nearby frames with a multi stage attention module. This module takes into account both appearance (RoI pooling output) and geometry features (bounding box definition) to establish the relation weights. These relation weights among proposals of different frames are used to enhance box features in the current frame. The solution in [76] also searches for similar proposals in the supporting frames, focusing on long term relationships.

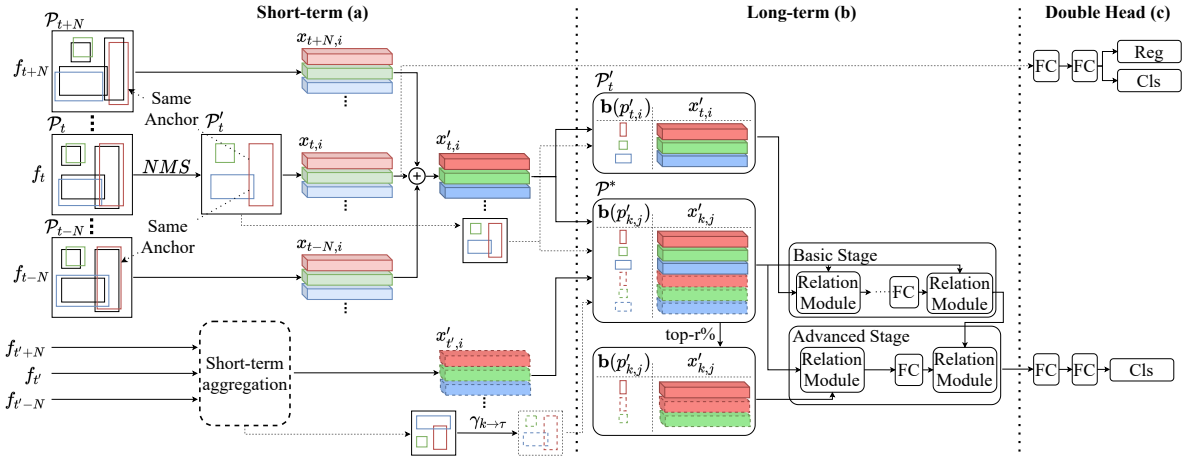


Figure 3.1: Our approach has three main components: (a) short-term object linking based on anchor boxes, and box aggregation features throughout the nearby frames; (b) long-term self-guided attention module that enhances short-term aggregated features with key frame information ($k\,f_{t'}$); (c) spatio-temporal double head.

Previous work only considers short- or long-term information to implement attention mechanisms, but not the combination of both to take full advantage of the whole spatio-temporal context. This issue is tackled in [12] by integrating both information from nearby frames and randomly selected key frames from the entire video. Nevertheless, since original bounding box positions are not meaningful to compare proposals in distant frames, they just get rid of geometric features, proposing a location free implementation. As a step forward, we propose a new method to integrate these geometric features in the long-term aggregation method. Our approach updates bounding box positions throughout time, making possible to use previous locations to establish proposal relationships. This way, object trajectories guide the attention process, associating proposals corresponding to the same object in the past.

3.3 Method

We propose a new spatio-temporal framework able to improve object detection precision in videos by exploiting short- and long-term temporal context. Although our implementation is based on the two-stage object detector Faster R-CNN [72] with a Feature Pyramid Network (FPN) structure [60], the core components of our approach can be directly applied to any two-stage object detector architecture.

Two-stage single image object detector architectures take a predefined set of *anchor boxes* to initialize object proposals. Then, a Region Proposal Network (RPN) calculates the final object proposals set by modifying these *anchor boxes* to better fit the objects in the image. In addition, the RPN also gives the probability of containing an object of interest for each proposal box. Finally, spatially redundant proposals with lower confidence are removed, typically applying Non-Maximum Suppression (NMS). Our spatio-temporal framework keeps this same pipeline to initialize the per frame object proposal set. Thus, we do not add an extra overhead in the

proposal generation in comparison with the single image counterpart.

Once the per-frame object proposals are calculated, they are linked throughout the nearby frames to exploit short-term information. We define two modes of operation: (i) an approach working with N input previous frames —$f_{t-N-1}, ..., f_{t-1}, f_t$— for each reference frame f_t; (ii) a symmetric approach using frames in advance taking into account $f_{t-N}, ..., f_{t-1}, f_t, f_{t+1}, ..., f_{t+N}$ for each reference frame. For the sake of simplicity we only consider the symmetric approach in further explanations. We report the precision for both approaches in the results section.

Box features in the reference frame f_t are enhanced with features in the nearby supporting frames, performing an adaptive weight feature aggregation (Fig. 3.1a). More details about both the linking and aggregation process are given in Section 3.3.1.

We also exploit long-term relations among proposals to further enrich box features. Our long-term method works with short-term aggregated features and attention mechanisms to define the long-term relationships (Fig. 3.1b). Section 3.3.2 describes this component.

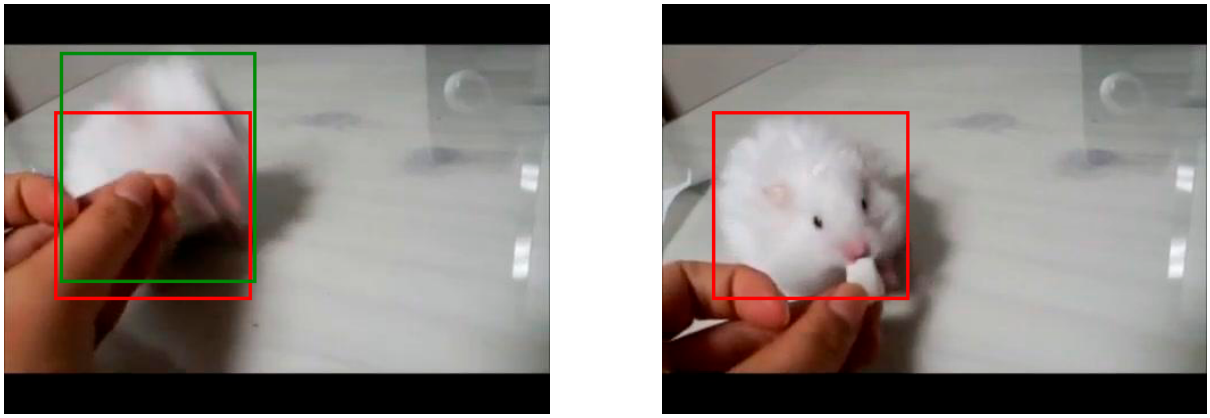
Finally, the spatio-temporal double head classifies the objects of interest using both the enhanced box features and spatial features, while the bounding box regression is performed by just taking features from the reference frame (Fig. 3.1c). This differs from the extended trend in the state-of-the-art of applying both the bounding box regression and object classification heads over the spatio-temporal aggregated features. Even if objects in the current frame are not well defined, we argue that the most relevant information to localize the object must come from this frame. For instance, although Fig. 3.2a suffers from motion blur, making the classification task really challenging, the localization task can still be done. In fact, even though Fig. 3.2b gives relevant information to address the classification issue, the most relevant information to localize the object regarding its position and shape is still in Fig. 3.2a. The final classification score is calculated as:

$$p = p_{tmp} + p_{spt}(1 - p_{tmp}) \tag{3.1}$$

being p_{tmp} the classification score calculated with spatio-temporal features and p_{spt} the score of the classification in the reference frame without spatio-temporal information.

3.3.1 Short-term anchor linking and aggregation

Our short-term module first links the proposal boxes in the nearby frames and, then, aggregates the corresponding box features. Each object proposal p_i has an associated confidence score $\mathbf{s}(p_i)$ and a bounding box $\mathbf{b}(p_i)$ used to extract box features with an RoI pooling layer —RoI Align in our implementation. These box features have the same shape regardless of the object size. Thus, the output of this short-term module keeps the same dimension as the original RoI pooling layer, independently of the number of input frames.



(a) Reference frame.

(b) Supporting frame.

Figure 3.2: Images from ImageNet VID validation set. Red boxes represent the bounding box in the supporting frame. The green box represents the object location in the reference frame. Location information from the supporting frame (red box) is not accurate in the reference frame (green box).

To link proposals in the short-term, we exploit the spatial redundancy between close frames. For every object in the image, it is very likely that the same object appears in a similar position in the nearby frames. This is a core concept in our implementation: we link proposals that come from the same *anchor box* for every frame in $\{f_{t-N}, ..., f_{t-1}, f_t, f_{t+1}, ..., f_{t+N}\}$. This method also relies on the generally large field of view of deep CNNs that allows the RPN to fit each anchor to the actual object even when the object is not very close to the predefined *anchor box*.

For every frame f_t in the input video, the RPN generates a proposal set $\mathcal{P}_t = \{p_{t,i}\}_{i=1}^{A}$, being A the total number of *anchor boxes*, calculated by multiplying the number of *anchor boxes* per position by the number of grid positions in one image. Hence, A is also the initial number of proposals per image. The final proposal set used by the network head is calculated as $\mathcal{P}'_t = top_s(NMS(\mathcal{P}_t))$. Thus, $\mathcal{P}'_t$ only contains the top s proposals ordered by confidence after removing the spatially redundant ones by means of Non-Maximum Suppression, resulting in $|\mathcal{P}'_t| \leq |\mathcal{P}_t|$. Therefore, it is very likely that there are no proposals associated with the same *anchor box* for every input frame f_j in $\{f_{t-N}, ..., f_{t-1}, f_t, f_{t+1}, ..., f_{t+N}\}$, making impossible to link proposals directly using the $\mathcal{P}'_j$ proposal sets. Instead, we keep the original proposal set $\mathcal{P}_j$ for every supporting frame $\{f_{t-N}, ..., f_{t-1}\} \cup \{f_{t+1}, ..., f_{t+N}\}$, and we link each proposal in $\mathcal{P}'_t$ with proposals from $\{\mathcal{P}_{t-N}, ..., \mathcal{P}_{t-1}\} \cup \{\mathcal{P}_{t+1}, ..., \mathcal{P}_{t+N}\}$ that came from the same *anchor box*. This process is shown in Fig. 3.1a: boxes with the same colors for nearby frames come from the same *anchor box*.

Then, we aggregate the box features extracted with the RoI Align method at the corresponding bounding box of each frame:

$$x'_{t,i} = \sum_{l=-N}^{N} \omega^{s}_{t+l,i} x_{t+l,i} \tag{3.2}$$

being $x'_{t,i}$ the aggregated feature map, $\omega^{s}_{t+l,i}$ the short-term weight for the feature map that came

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

from the proposal associated with anchor i in frame f_{t+l}, and $x_{t+l,i}$ the feature map associated with proposal i in frame f_{t+l}. The short-term weight is based on the cosine similarity metric between supporting proposals and proposals in the reference frame:

$$\omega^s_{t+l,i} = exp\left(\frac{x_{t,i}\, x_{t+l,i}}{|x_{t,i}||x_{t+l,i}|}\right) \tag{3.3}$$

Then, the weights are normalized using a Softmax function to ensure that $\sum_{l=-N}^{N} \omega^s_{t+l,i} = 1$.

This goes beyond the highly effective pixel level method reported in [104] and [1], by designing a new box level method. In contrast with previous methods that use a subnetwork to calculate an intermediate feature representation of the full frame feature maps, we work directly with the box features calculated by the RoI Align method. Hence, we focus just on promising regions instead of aggregating the complete frame information, which simplifies the process.

3.3.2 Long-term self-guided attention module

In the long-term scenario, we cannot rely on spatial redundancy to link the proposals as in the short-term case. Therefore, we follow a more flexible approach in which every reference proposal is compared to every supporting proposal. This idea is based on attention methods proposed in [87] applied to NLP, and lately in [45] applied to the single image object detection problem. In spatio-temporal long-term aggregation, this technique allows to establish the relationship between each proposal in the reference frame and every proposal in a set of supporting key frames. In this case, supporting key frames are selected at a fixed interval I rather than consecutively as in the short-term phase. We are also considering the reference frame as a supporting frame, so that we can use proposals from the reference frame in the aggregation process.

Formally, given a proposal in the reference frame $p'_{t,i}$ and a set of supporting proposals $\mathcal{P}^{KF}$, the goal of the relation module is to calculate M relation features $\mathbf{f}^m_R$:

$$\mathbf{f}^m_R(p'_{t,i}, \mathcal{P}^{KF}) = \sum_{k=1}^{K} \sum_{j=1}^{|\mathcal{P}'_k|} w^m_{t(i),k(j)} \, (W_V \, x'_{k,j}), \quad m = 1, ..., M \tag{3.4}$$

where $\mathcal{P}^{KF} = \{\mathcal{P}'_1, \mathcal{P}'_2, ..., \mathcal{P}'_K\}$, being K the number of long-term supporting key frames, and where each proposal $p'_{t,i}$ is defined by its appearance features $x'_{t,i}$ and geometry features $\mathbf{b}(p'_{t,i})$. The linear transformation matrix W_V is optimised through backpropagation in an end-to-end fashion. Previous attempts to adapt the relational module to the spatio-temporal domain [12,21] use as appearance features the RoI pooled object proposals directly. In contrast, the appearance features that feed the relational module in our implementation are the output of the short-term aggregation process $x'_{k,j}$. This way, we can work with a more robust representation of the object.

The relational weight $w^m_{t(i),k(j)}$ is calculated as:

Algorithm 3: $\gamma_{k\to\tau}$

Input : Previous frame proposals: $\mathcal{P'}_\tau = \{p'_{\tau,i}\}_{i=1}^{\eta}$
Input : Key frame proposals: $\mathcal{P'}_k = \{p'_{k,j}\}_{j=1}^{\eta}$
Input : Attention weights: $\{w^m_{\tau(i),k(j)}\} \,|\, \forall p'_{\tau,i} \in \mathcal{P'}_\tau, \forall p'_{k,j} \in \mathcal{P'}_k$

$\overline{w}_{\tau(i),k(j)} = mean_m(w^m_{\tau(i),k(j)})$
for k **in** $1, ..., K$ **do**
  $\mathcal{S}_{i,j} \leftarrow \mathbf{s}(p'_{\tau,i})\, exp(\overline{w}_{\tau(i),k(j)}) \,|\, \forall p'_{\tau,i} \in \mathcal{P'}_\tau, \forall p'_{k,j} \in \mathcal{P'}_k$
  $\mathcal{H} \leftarrow Hungarian(\mathcal{S})$
  for (i, j) **in** $\mathcal{H}$ **do**
    $\mathbf{b}_{xy}(p'_{k,j}) \leftarrow \mathbf{b}_{xy}(p'_{\tau,i}) | p'_{\tau,i} \in \mathcal{P'}_\tau, p'_{k,j} \in \mathcal{P'}_k$
return updated $\mathcal{P'}_k$

$$w^m_{t(i),k(j)} = \frac{g^m_{t(i),k(j)}\, exp(a^m_{t(i),k(j)})}{\sum_q g^m_{t(i),q}\, exp(a^m_{t(i),q})} \tag{3.5}$$

being $a^m_{t(i),k(j)}$ the appearance weight and $g^m_{t(i),k(j)}$ the geometry weight. The appearance weight is calculated as a normalized dot product:

$$a^m_{t(i),k(j)} = \frac{\langle W_H\, x'_{t,i}, W_Q\, x'_{k,j} \rangle}{\sqrt{d_h}} \tag{3.6}$$

where W_H and W_Q of Eq. 3.6, as well as W_G in Eq. 3.7, are also learnt in the training process as W_V (Eq. 3.4). W_H and W_Q project the appearance features in the reference frame and supporting key frames, respectively, being d_H the projected dimension.

Geometry weights are computed as:

$$g^m_{t(i),k(j)} = max\{0, W_G\, \mathcal{E}(\mathbf{b}(p'_{t,i}), \gamma_{k\to\tau}(\mathbf{b}(p'_{k,j})))\} \tag{3.7}$$

where each geometric feature $\mathbf{b}(p')$ is a 4-d vector representing the bounding box parameters (x, y, w, h) associated with proposal p'. Function $\mathcal{E}$ embeds the vector $\left(log\left(\frac{|x_i - x_j|}{w_i}\right), log\left(\frac{|y_i - y_j|}{h_i}\right), log\left(\frac{w_j}{w_i}\right), log\left(\frac{h_j}{h_i}\right)\right)$ in a high dimensional representation following the method outlined in [87]. We introduce a new function $\gamma_{k\to\tau}$ to transform geometric features from supporting key frames, so that they can be compared with geometric features from proposals in the reference frame. Otherwise, comparing box positions in distant frames would not be meaningful to calculate strong attention weights. This new method allows to exploit geometric features in long-term attention mechanisms for the first time in video object detection.

The core idea behind $\gamma_{k\to\tau}$ is to update proposal box positions using the attention weights to predict the object movement throughout the video by matching object proposals (Alg. 3). In doing that, this function considers the previous frame proposal set $\mathcal{P'}_\tau$ and the supporting key frame proposal set $\mathcal{P}_k$. At a certain reference frame f_t we have already calculated the relation

weights that associate every proposal in $\mathcal{P}'_\tau$ with proposals in $\mathcal{P}'_k$: $w^m_{\tau(i),k(j)}$ (Eq. 3.5). As there are M relation weights for every pair of proposals, we aggregate them calculating the average relation weight $\overline{w}_{\tau(i),k(j)}$ (Alg. 3:1). The output of $\gamma_{k\to\tau}$ are the per key frame proposal sets $\mathcal{P}'_k$ with the proposal positions updated to the predicted position of the objects in the previous frame f_τ.

Then, to link proposals in key frames with proposals in f_τ, a score matrix $\mathcal{S}$ is populated (Alg. 3:3), taking into account proposals score $\mathbf{s}(p'_{\tau,i})$ and relational weights $\overline{w}_{\tau(i),k(j)}$. By considering the RPN confidence in f_τ, we avoid to link with low confidence proposals in the previous frame. The association problem can be solved with the Hungarian method (Alg. 3:4) [54]. Then, each bounding box position $\mathbf{b}_{xy}(p'_{k,j})$ in the supporting key frame k is updated to the corresponding position in f_τ (Alg. 3:6). $\mathbf{b}_{xy}(p')$ represents the bounding box center coordinates of proposal p'. We experimentally found that it is better to keep the original bounding box size $(w;h)$, and just updating the center coordinates. Now, Eq. 3.7 compares bounding box positions in consecutive frames rather than in arbitrary distant frames.

The final feature map for each proposal used by the network head is calculated as:

$$\mathbf{f}_R(p'_{t,i},\mathcal{P}^{KF}) = \mathbf{f}_R(p'_{t,i},\mathcal{P}^{KF}) + concat[\{\mathbf{f}^m_R(p'_{t,i},\mathcal{P}^{KF})\}^M_{m=1}] \tag{3.8}$$

This is the concatenation of the M relational features $\mathbf{f}^m_R(p'_{t,i},\mathcal{P}^{KF})$ (Eq. 3.4), and adding the result to the original proposal appearance feature $\mathbf{f}_R(p'_{t,i},\mathcal{P}^{KF})$.

We follow a multi stage implementation similar to [21] with a set of stacked relation modules. The aim of this architecture is to iteratively refine object proposals defining two main stages, a basic stage and an advanced stage (Fig. 3.1b).The basic stage inputs are the top λ proposals of every key frame and the reference frame proposals $\mathcal{P}'_t$. The advanced stage has two steps. First, the top r% proposals in $\mathcal{P}^{KF}$ are enhanced by an attention module with $\mathcal{P}^{KF}$ as supporting proposals —first relation module in the advanced stage in Fig. 3.1b. Then, these enhanced proposals are used as supporting proposals to further improve proposal features calculated in the basic stage —second relation module in the advanced stage in Fig. 3.1b.

3.3.3 Training and inference

Both nearby frame and long-term key frame selections are implemented in a different way in the training and inference phases. This is mainly because of the ground truth availability in the training stage and the lack of constraints on which frames can be used in each moment.

As explained in Sec. 3.3.1, we resort to a set of neighboring frames $\{f_{t-N},...,f_{t-1},f_t,f_{t+1},...,f_{t+N}\}$ to enhance the reference frame box features. In the inference stage, all video frames are sequentially processed by the network. Therefore, we can reuse all the backbone and RPN calculations from the nearby frames, drastically reducing the impact of

enlarging the reference frame neighborhood. In contrast, in the training stage, instead of all video frames, we select a fixed size subsample of evenly spaced frames. This way, we prevent from large videos to bias the training process. In consequence, the idea of reusing computations in training cannot be applied since close frames are not selected as reference frames. Thus, the training approach is slightly different, taking just three input frames: the reference frame and two supporting frames. The two supporting frames f_{s1} and f_{s2} are randomly selected from $\{f_{t-N}, ..., f_{t-1}\}$ and $\{f_{t+1}, ..., f_{t+N}\}$ respectively.

In the long-term method we also face the same issue. Instead of several key frames, we randomly select two frames from the whole video for each reference frame. In both long- and short-term cases the number of input frames does not change the number of parameters in the network, so training and testing with different number of input frames does not need any modification in the architecture.

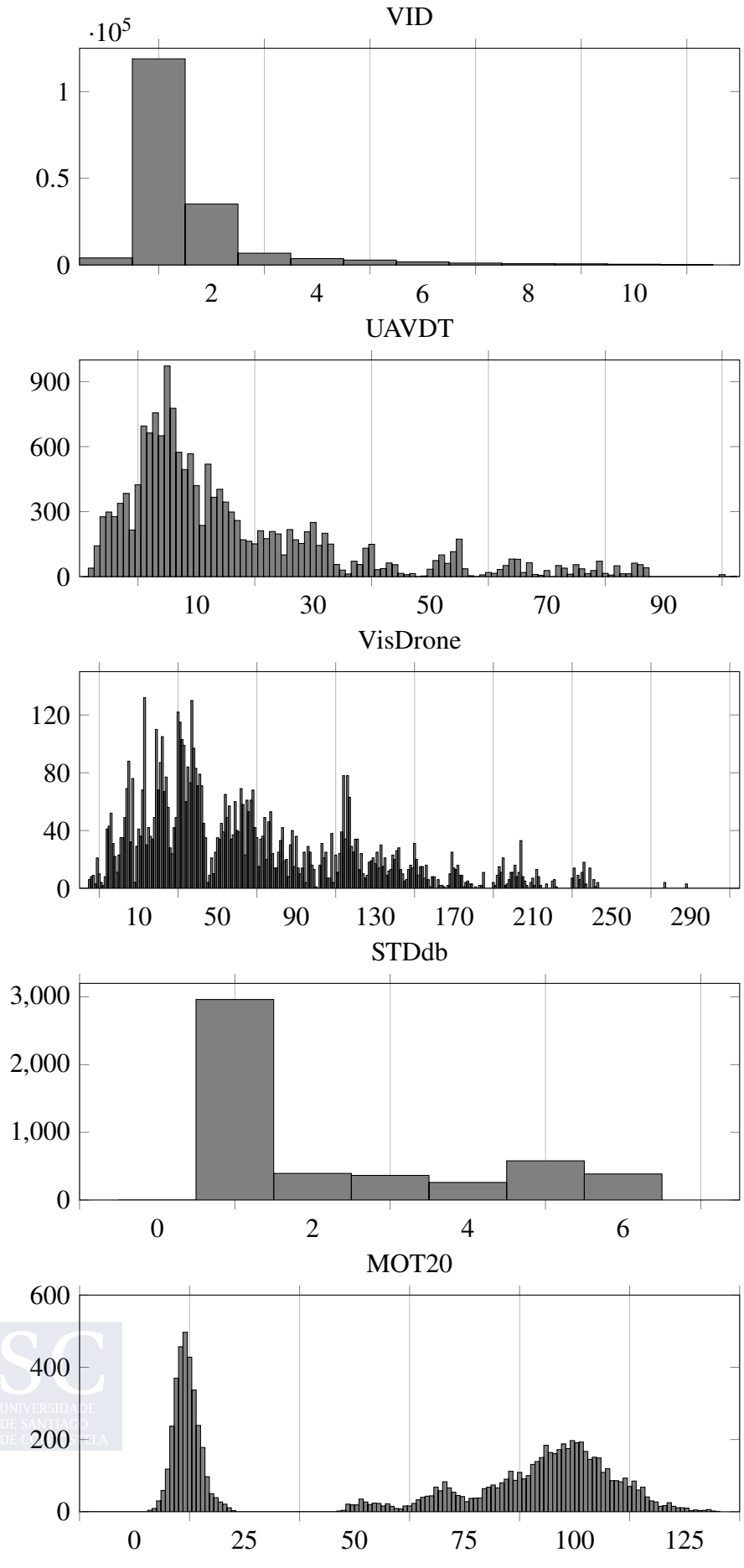
Moreover, the implementation in the training stage of $\gamma_{k\to\tau}$ (used in Eq. 3.7) also differs from the inference version described in Sec 3.3.2. As the network does not process frames sequentially, relational weights for previous frames are not available. However, most video object detection datasets include object identity annotations that link appearances of the same object throughout the whole video. We exploit these annotations during training to update the position of proposals in the key frame following the actual object movement. First, each object proposal is linked to the ground truth box with higher Intersection-Over-Union (IoU). Then, we apply the ground truth object translation to proposal boxes. These updates allow to use Eq. 3.7 during training.

3.4 Experimental Results

3.4.1 Datasets

ImageNet VID dataset [74] has become the standard benchmark to evaluate spatio-temporal object detection frameworks. In fact, most recent solutions report their metrics only on it [1, 12, 21, 29, 37, 76, 83, 89, 94]. Nevertheless, we believe that a complete and reliable evaluation requires tests in several datasets with different characteristics to assess the quality of the detector in a wide number of scenarios. In so doing, we have selected 5 different video datasets to evaluate the performance of both our proposal and state-of-the-art approaches: ImageNet VID [74], UAVDT [23], VisDrone [102], USC-GRAD-STDdb [4] and MOTChallenge [19]. As we will show in the results section (Sec. 3.4.4), the performance of some of these approaches highly changes with the dataset in comparison with the baseline.

There are many characteristics of the datasets that influence the detection precision. In this paper, we focus the analysis on three of them:



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Figure 3.3: Average number of objects per frame for each test set.

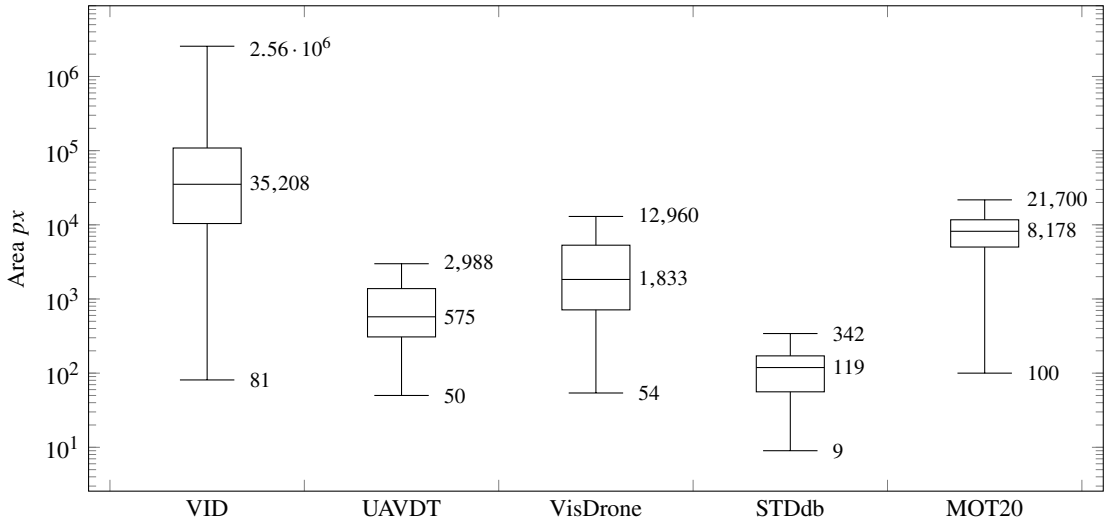


Figure 3.4: Box plot for the object size of each test set.

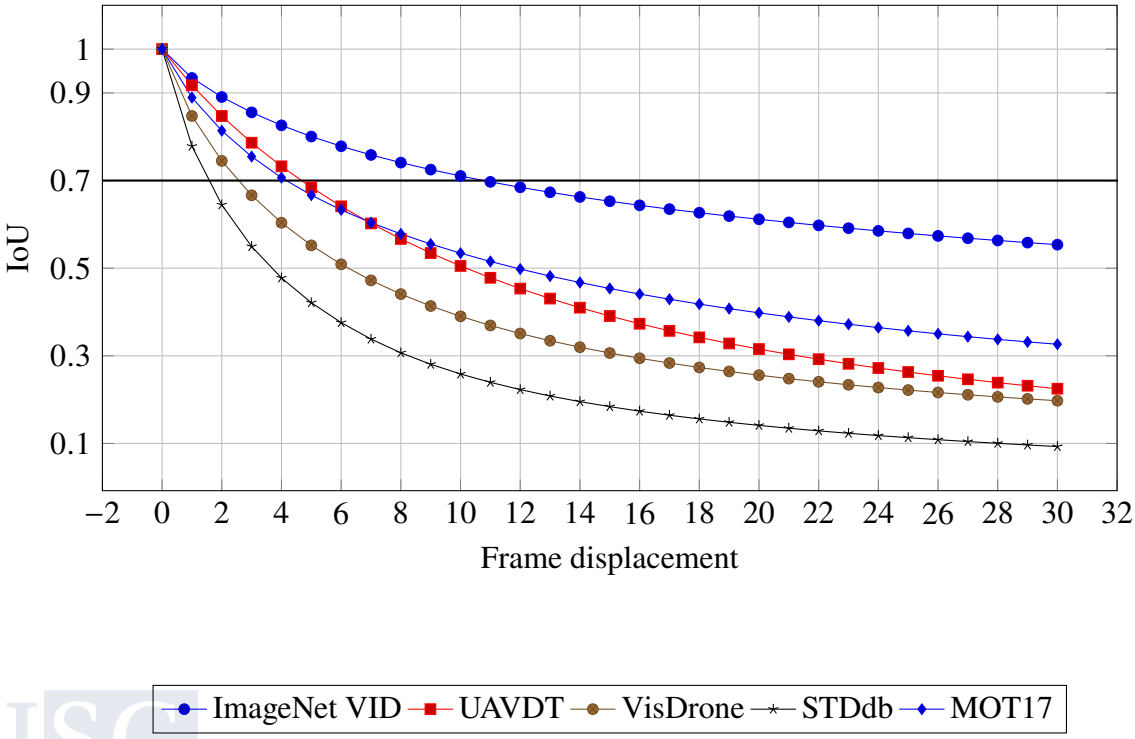
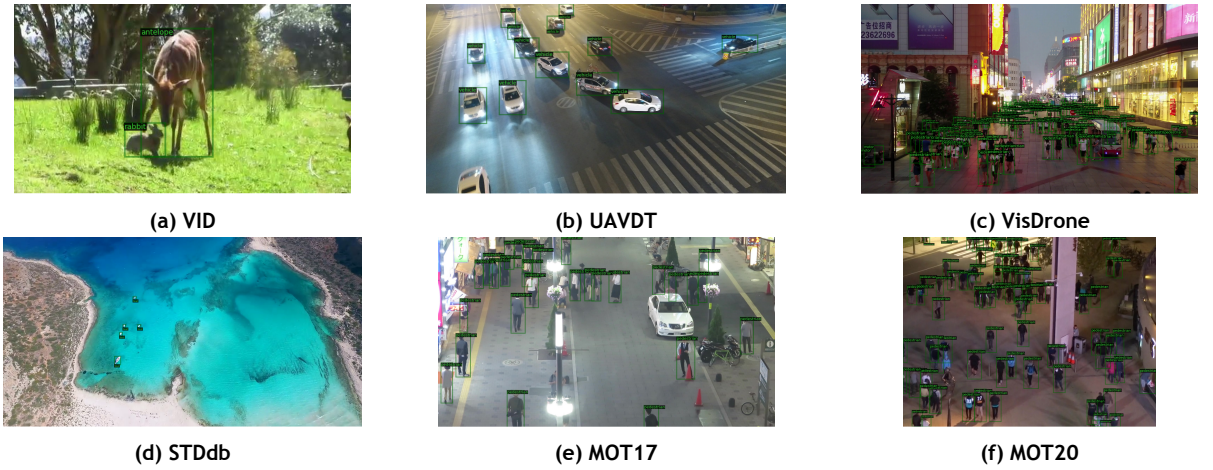


Figure 3.5: Mean IoU for the same object bounding boxes separated by different number of frames, and for all the training sets.

USC DE SANTIAGO DE COMPOSTELA



(a) VID (b) UAVDT (c) VisDrone

(d) STDdb (e) MOT17 (f) MOT20

Figure 3.6: Examples from the different video datasets evaluated with state-of-the-art solutions and our approach.

- Number of objects per frame (Fig. 3.3), which influences both spatial and spatio-temporal object detectors. Most spatio-temporal detectors exploit object relations between different frames and, therefore, a greater number of objects per frame would be preferable to assess how these methods can establish robust relationships working with many objects simultaneously.

- Size of the objects (Fig. 3.4), which affects both spatial and spatio-temporal detectors. In fact, small object detection is a challenge itself [5], specially for objects with areas smaller than 256 pixels ($\approx 16 \times 16$).

- The speed at which the position of the objects changes due to the own objects motion or the camera motion. This influences the performance of the spatio-temporal detectors, as most of them make feature aggregation throughout nearby frames. We measure this speed with the Intersection over Union (IoU) of the bounding boxes of the same object in two frames (Fig. 3.5).

Fig. 3.6 shows some examples of frames extracted from different datasets. The datasets have been selected to cover a wide variety of scenarios, including low/high number of objects per frame, small/large objects, and low/fast position changing

ImageNet VID dataset [74] contains 30 different objects categories in 3,862 training and 555 validation videos. Following the training procedure proposed by [29], we also include data from ImageNet DET, a single image object detection dataset. This dataset contains 456,567 training and 20,121 validation images with annotated objects of 200 different categories that include the 30 classes considered in ImageNet VID. We select at most 2,000 images per VID object class from ImageNet DET to prevent from biasing the training set by including categories with a large number of images in ImageNet DET. To be able to train our spatio-temporal framework with still images, each image is repeated to create short input videos. The dataset has a very low

number of objects per frame, the objects are large, and the position of the objects changes very slowly —IoUs of 0.7 on average for an object 10 frames apart.

Unmanned Aerial Vehicle Benchmark (UAVDT) [23] is focused on videos recorded by cameras mounted on Unmanned Aerial Vehicles (UAVs) with about 40,000 annotated frames belonging to 30 training videos and 20 testing videos, with just one category. The number of objects per image is higher than Imagenet VID, the size of the objects is medium/small, and the position of the objects changes slowly.

VisDrone dataset [102] is also focused on UAV recorded images, with 56 training videos and 17 videos for testing, containing 11 object categories. Nevertheless, it has much more objects per frame than UAVDT, and the size of the objects is medium/large. Also, the position of the objects changes fast.

USC-GRAD-STDdb dataset is specifically designed for small object detection. It is composed of 92 training videos and 23 testing videos with over 56,000 annotated small objects of 5 different categories. The number of objects per frame is very low, and their size ranges 256 ($\approx 16\times16$) to as small as 16 ($\approx 4\times4$) pixels. Moreover, the object position changes very fast —average IoUs below 0.7 for an object in two consecutive frames— due to the small object sizes and the camera movement.

MOTChallenge [19] proposes pedestrian focused annotated video sequences. In this paper we train all the object detectors with the 7 training sequences from MOT17 and evaluate with the 4 sequences in the MOT20 training set. The number of objects per frame in MOT20 is considerably higher than in MOT17, changing the training and testing conditions. Following the same strategy that we use with ImageNet VID and ImageNet DET, we also add single images from CUHK-SYSU dataset [95] to the training set.

3.4.2 Implementation details

Our proposal has as per frame feature extractor a Feature Pyramid Network (FPN) [60] with ResNeXt-101 backbone and deformable convolutions [103] on *conv3*, *conv4* and *conv5*. We initialize the backbone with pre-train ImageNet classification weights to train the single frame baseline. To train the spatio-temporal network, we reuse the baseline weights keeping them frozen. This way, we only have to train the attention module and the temporal classification head if we have the single image counterpart trained, speeding up the training process dramatically. In the ablation studies we replace the ResNeXt-101 backbone by the smaller ResNet-50 due to the high number of different experiments needed.

The results of the state-of-the-art proposals in the different datasets have been obtained from: (i) the results reported by authors in their original work —ImageNet VID dataset—; and (ii) training and testing them with the implementations provided in [12] —UAVDT, VisDrone, STDdb, MOT.

Table 3.1: Long-term strategy, without short-term information.

Method	Appearance	Geometry	Mean AP
location free	✓		77.1
self-guided	✓	✓	77.3

For experimentation on ImageNet VID, images are scaled, so the smallest size is at most 600px. To train the single image baseline we set the base learning rate to 2.5×10^{-4} for 360K iterations, reducing by $*0.1$ at 280K and 250K iterations. For the spatio-temporal network the initial learning rate is set to 1.25×10^{-3} for 270K iterations with learning rate reductions at 210K and 250K iterations. For UAVDT, VisDrone, SDTdb and MOT we run 45K training iterations for the baseline network with an initial learning rate of 1.25×10^{-3} and learning rate reduction steps at 30K and 45K iterations. For the spatio-temporal network the number of iterations is set to 15K with a learning rate of 1.25×10^{-3} and just one reduction step at 12K iterations. We set the shortest image dimension to 720px in VisDrone, STDdb and MOT, and 540px for UAVDT, keeping the largest dimension below 1280px and 1024px, respectively.

We define a heuristic rule to set the number of short-term supporting frames N in the different datasets. The rule takes into account the object movement average in the training set to determine this hyperparameter. Based on data from Fig. 3.5, we select the value of N that keeps the IoU for the same object higher than 0.7 for every displacement lower than N frames. Therefore, N is set to 10 for ImageNet VID, 4 for UAVDT and MOT, 2 for VisDrone, and 1 for STDdb. We keep all the other parameters unchanged, irrespective of the dataset.

3.4.3 Ablation studies

We conducted a series of ablation studies to prove that each component in our architecture is contributing to the network precision. For the sake of simplicity, we use ImageNet VID as the reference dataset for these experiments. Fig 3.7 shows how the number of short-term input frames affects the mAP in both sequential and symmetric modes. Moreover, Fig 3.7 shows how the symmetric setting yields a higher mAP than the sequential approach. This was expected since the network can aggregate more information representing the same object in close frames than in the sequential approach. In both cases, our spatio-temporal approach shows a significant improvement over the single frame baseline: 0.6 by just using one supporting frame in the sequential case, and 1.4 with two supporting frames in the symmetric case, which is the minimum possible value in this approach.

Regarding the long-term strategy, Table 3.1 proves the effectiveness of considering geometry features in the long-term aggregation process. Our approach outperforms the location free version —first proposed in [12]— in which geometry features are just not taken into account.

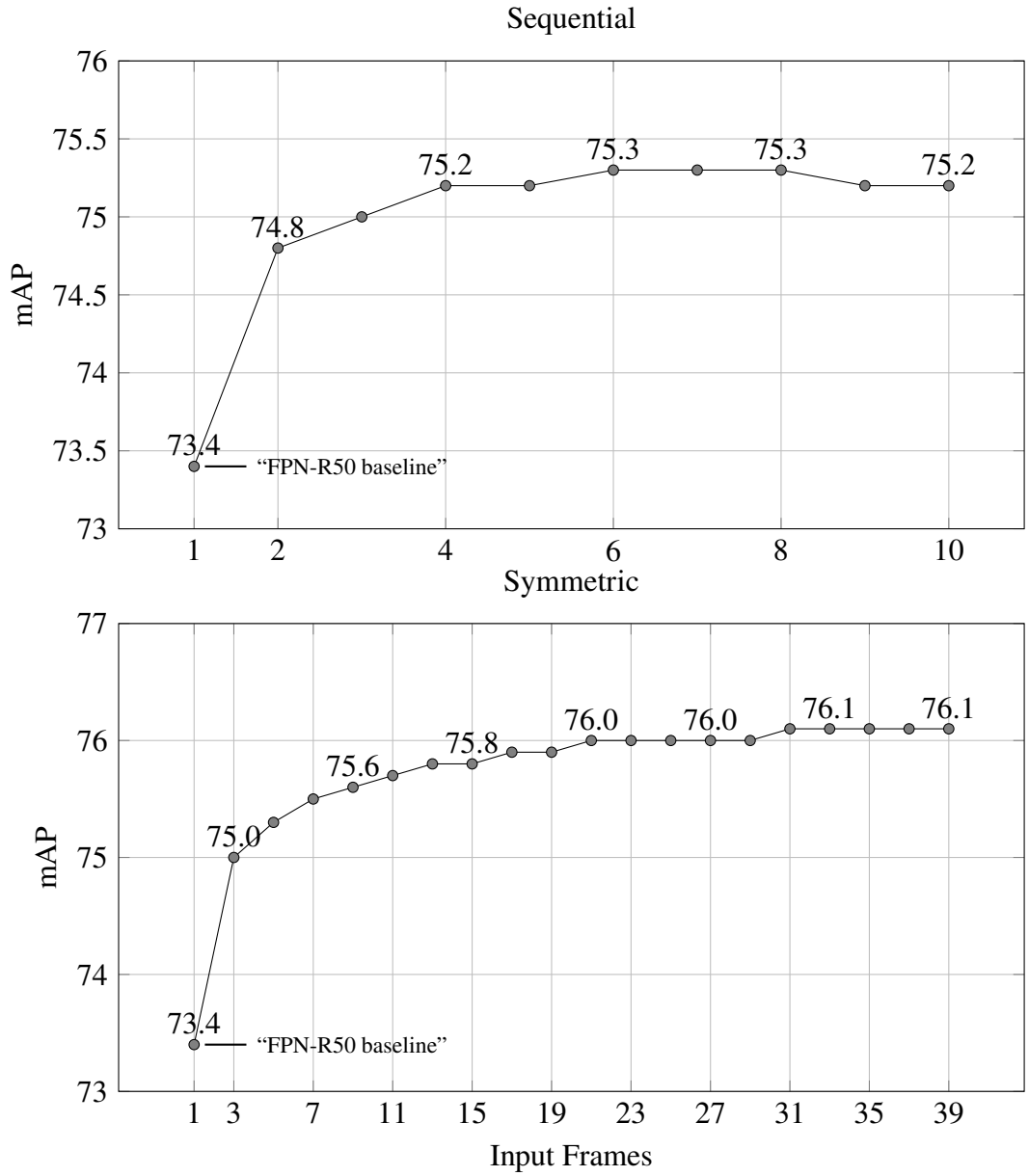


Figure 3.7: mAP varing the number of short-term input frames without considering long-term information.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Table 3.2: Influence of each component on the framework precision on ImageNet VID dataset.

Short-term	Long-term	Mean AP
		73.4
✓		76.1
	✓	77.3
✓	✓	77.8

In order to evaluate the long-term method isolated, we do not include short-term aggregation in these experiments.

Finally, Table 3.2 shows how each component contributes to the final mAP. The results prove that short- and long-term spatio-temporal information are complementary and, thus, both are valuable to boost the object detection precision.

3.4.4 Results

In this section we compare our framework with the state-of-the-art spatio-temporal object detectors in the 5 selected datasets. Table 3.3 shows the results for the ImageNet VID dataset. We compare our model in both sequential and symmetric configurations, as well as with box level post-processing techniques. In our case, the post-processing is just box rescoring over long tubes calculated with the Viterbi algorithm over the per frame final detection set. Post-procesing methods are a special case of a symmetric approach, as they use frames in advance. However, these methods need the detection set for every frame in the video in order to be executed. Our method ranks second in the sequential mode and third in both symmetric and post-processing modes, with mAPs close to MEGA. It is important to notice that in the symmetric mode MEGA selects random frames from the whole video, while the other methods —including our proposal— just use a small number of consecutive frames in advance.

For the rest of datasets, the comparison has been done with those proposals provided in [12] —FGFA [104], RDN [21], MEGA [12]; all of them in the symmetric mode, which is the best available one—, and the baseline FPN-X101. We report $mAP_{@0.5}$ and, also, $mAP_{@0.5–0.95}$, which is much more exhaustive.

Table 3.4 shows the results in the UAVDT dataset. Our method outperforms the best spatio-temporal object detector by 1.4 points in $mAP_{@0.5}$ and 2.5 points in $mAP_{@0.5–0.95}$. For the Visdrone dataset (Table 3.5) our method also achieves the best results with higher difference in $mAP_{@0.5–0.95}$. In the challenging small object detection problem (Table 3.6), the other spatio-temporal object detectors degrade their performance with results below our single frame baseline. Our approach achieves again the best results, overcoming the single frame baseline by 2.6 $mAP_{@0.5}$, and the best spatio-temporal framework by 4.8 $mAP_{@0.5}$. Finally, Table 3.7

Table 3.3: Results on ImageNet VID dataset.

Method	Mode	$mAP_{@0.5}$
FPN-X101 baseline	*Sequential*	*78.6*
D&T [29]	Sequential	78.7
PSLA [37]	Sequential	80.0
OGEMN [20]	Sequential	80.0
MEGA [12]	Sequential	81.9
ours	Sequential	**81.3**
FGFA [104]	Symmetric	77.8
STSN [1]	Symmetric	78.9
MANet [89]	Symmetric	78.1
SELSA [91]	Symmetric	80.3
RDN [21]	Symmetric	83.2
MEGA [12]	*Symmetric*	*84.1**
ours	Symmetric	**81.9**
D&T [29]	Post-processing	79.8
FGFA [104]	Post-processing	80.1
STSN [1]	Post-processing	80.4
MANet [89]	Post-processing	80.3
STMN [94]	Post-processing	80.5
SELSA [91]	Post-processing	80.5
PSLA [37]	Post-processing	81.4
OGEMN [20]	Post-processing	81.6
RDN [21]	Post-processing	84.7
MEGA [12]	Post-processing	85.4
ours	Post-processing	**82.4**

Table 3.4: Results on UAVDT dataset.

Method	$mAP_{@0.5}$	$mAP_{@0.5-0.95}$
FGFA [104]	57.6	28.9
RDN [21]	60.4	32.5
MEGA [12]	59.4	31.7
FPN-X101	59.6	33.6
ours	61.8	35.0

shows the results for the MOT dataset, where our method ranks second, 0.9 $mAP_{@0.5}$ below FGFA, which is the best.

In summary, our method achieves the best results in UAVDT, Visdrone and STDdb, while MEGA [12] provides the best results in ImageNet VID —our method ranks third—, and FGFA [104] in the MOT20 dataset —our method ranks second. All in all, it can be stated that our method ranks better than the state-of-the-art in the collection of tested datasets, showing an

Table 3.5: Results on Visdrone dataset.

Method	$mAP_{@0.5}$	$mAP_{@0.5-0.95}$
FGFA [104]	30.7	14.1
RDN [21]	31.5	14.4
MEGA [12]	31.8	14.5
FPN-X101	31.4	15.2
ours	31.9	15.3

Table 3.6: Results on STDdb dataset.

Method	$mAP_{@0.5}$	$mAP_{@0.5-0.95}$
FGFA [104]	29.3	9.0
RDN [21]	40.1	13.2
MEGA [12]	37.5	12.6
FPN-X101	42.3	15.6
ours	44.9	16.6

Table 3.7: Results on MOT20 dataset.

Method	$mAP_{@0.5}$	$mAP_{@0.5-0.95}$
FGFA [104]	67.0	29.4
RDN [21]	66.1	26.6
MEGA [12]	46.5	17.4
FPN-X101	65.6	28.2
ours	66.1	28.4

excellent performance under very different conditions of number of objects per frame, size of the objects, and the speed the objects move in consecutive frames. In order to assess if the differences between our method and the state-of-the-art are statistically significant, we conducted a series of non-parametric tests with the STAC platform [73]. As data is non symmetric and paired, we run the Binomial Sign test comparing our proposal with each of the state-of-the-art approaches (Table 3.8). In this comparison we use the more robust $mAP_{@0.5-0.95}$ for every dataset except for ImageNet VID, in which we use the $mAP_{@0.5}$ originally reported by the authors. The test shows that the probability of having statistically significant differences with FGFA, RDN and MEGA is 93.75%, and of 100% with our single frame baseline. Therefore, we can conclude that, overall, our proposal outperforms the state-of-the-art, and that the differences

Table 3.8: Binomial Sign test. p is the calculated p-value comparing each method with our proposal.

	FPN	FGFA	RDN	MEGA
$1-p$	1.000	0.9375	0.9375	0.9375

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

are statistically significant.

3.5 Conclusions

We have proposed a new framework for spatio-temporal object detection that takes into account both short- and long-term information. First, short-term information is linked and aggregated based on anchor association. Then, long-term information is taken into account by means of our self-guided attention module. This component allows to consider geometrical features in the long-term for the first time in the video object detection domain.

We have tested our proposal with 5 video object detection datasets, in order to analyze the performance in very different scenarios. Moreover, we have compared our approach with the state-of-the-art. Results show that our proposal ranks first on average in the collection of datasets, and non-parametric statistical tests indicate that the differences are statistically significant.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

4 Few-shot Video Object Detection

This chapter presents FSVDet, a video object detector specifically designed to work with a limited training set. Few-shot object detection in general, and few-shot video object detection in particular, is an emerging field of research that is still at an early stage. A few-shot detector must quickly adapt to a new domain in which the number of annotations per category is very limited. This chapter describes the main challenges of few-shot video object detection such as the limitations on the training set that impose the need for training the model with single images rather than complete videos. We propose to augment each training image with synthetic frames to train the relation module. Thus, considering spatio-temporal information from a set of input frames at inference. Spatio-temporal information is processed by a modified version of the spatio-temporal double head that includes a category-agnostic confidence predictor. Including this new component in the loss function reduces the overfitting, a key issue considering the limited training set. Moreover, the predicted score is the input to a long-term object linking method that calculates object tubes throughout the entire video. The classification score is optimized based on these tubes, ensuring spatio-temporal consistency.

The experimental results also show that the classification task is particularly difficult compared to the localization task. This highlights the potential of leveraging spatio-temporal information considering information from a set of input frames to classify the object in the current frame. The proposed solution addresses this issue by adding a spatio-temporal classification prediction to the spatial classifier and by the final score optimization. Final results prove that these modifications lead to a significant precision improvement.

The contents of this chapter will be submitted for publication to a peer reviewed journal.

4.1 Introduction

Object detection aims to localize and classify objects of interest in images. This field has witnessed a significant improvement, mostly due to the adoption of deep learning techniques. Nevertheless, training current state-of-the-art models requires large amounts of labeled data for each object category. This makes infeasible to apply these techniques in data scarce scenarios.

Also, the high annotation costs, even when sufficient unlabeled data is available might limit the usefulness of these methods in real-world applications. Therefore, the development of few-shot object detection models capable of learning from few annotated examples has become a very active research topic.

An image object detector can be directly generalized to solve the video object detection problem by simply processing video frames independently. However, detecting objects in videos presents new challenges such as motion blur, out of focus, occlusions or big changes in objects appearances that may hinder the detection process in some frames. Moreover, recent work [12,13,21,104] shows that spatio-temporal information available in videos can be exploited to boost the detection precision, overcoming these issues. Although the success reached in traditional video object detection suggests the need for few-shot algorithms specifically designed for videos, this field remains almost unexplored.

Attention mechanisms have proven to be highly effective modeling sequences in the Natural Language Processing (NLP) field. Since the introduction of the multi-relation head [87], this technique has become the standard approach not only in NLP, but also for modeling object proposal relationships in the traditional video object detection domain. However, training these models requires multiple video frames in each training iteration to optimize the relation module parameters. A few-shot video object detector should learn from a limited number of labeled instances per object category while exploiting spatio-temporal information at inference time. In a few-shot setting, it is not possible to have complete videos for training. Thus, the training set consists of few single images, generating a data gap between train and test sets that must be tackled.

In view of these considerations, we propose a few-shot object detector that exploits spatio-temporal information available in videos. The proposed framework includes a proposal relation module that applies attention mechanisms for spatio-temporal feature aggregation. It also includes a classification score optimization that ensures spatio-temporal consistency.

The main contributions of this work are:

- A proposal relation module trained with single images that performs feature aggregation throughout neighboring frames, bridging the data gap between test and train sets. As this module mines object proposal relationships among different frames, training it with single images degrades the inference performance. Therefore, the proposed training method augments each input image by generating synthetic frames replicating the input image and placing variations of its objects in different positions. These generated synthetic frames allow relating the added objects with proposals from the current image in the training phase.

- A new spatio-temporal double head that leverages both spatial and spatio-temporal infor-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

mation. This double head includes a spatial branch that localizes the objects in the current frame and performs a spatial classification. It also exploits spatio-temporal information in an independent branch to improve the classification and predict the overlap of detections with the ground truth as a detection precision metric.

- An object tubes generation strategy based on the predicted overlap to link detections in consecutive frames. This method benefits from the generally high spatio-temporal redundancy of videos to modify the classification confidence of object detections based on the calculated object tubes.

Our framework outperforms both single image and previous video object detectors in the challenging FSVOD500 dataset, which has been specifically designed for the evaluation of few-shot video object detectors.

4.2 Related Work

4.2.1 Object Detection

General object detectors based on deep neural networks fall mainly into two main categories: one- and two-stage detectors. Two-stage detectors [6, 17, 72] generate a set of object proposals with high probability of containing an object of interest, and then perform a bounding box refinement and object classification. In contrast, one-stage detectors [61,64,71] directly calculate the final detection set by processing a dense grid of unfiltered candidate regions.

New object detection frameworks specifically designed to work with videos [12, 13, 21, 104] were introduced to take advantage of spatio-temporal information, improving detection precision. The main idea behind these methods is to aggregate per-frame features throughout time, achieving more robust feature maps. This aggregation can be done at pixel- [1, 37, 104] or at object-level [12, 13, 21, 36, 83], linking object instances through neighboring frames. Besides feature aggregation among frames of the same video, considering information from different videos was also successfully explored in [38].

4.2.2 Few-Shot Object Detection

Learning new tasks from just a few training examples is very challenging for most machine learning algorithms, especially for deep neural networks. Few-shot learning techniques aim to extract general knowledge from base data and rapidly adapt to novel scarce data. Few-shot image classification has been widely studied [40, 79, 85] as the first attempt to apply few-shot

techniques in computer vision. Recently, the few-shot object detection problem has drawn significant attention in order to replicate the success achieved in the image classification field.

Meta-learning based methods is one of the research lines to address the few-shot object detection problem. FSRW [48] proposes a framework based on YOLOv2 [71] including a feature reweighting module and a meta-feature extractor. Alternatively, a two-stage approach based on Faster/Mask R-CNN is proposed on Meta R-CNN [97]. This work proposes a new network head that applies channel-wise soft attention between RoI features given by a Region Proposal Network (RPN), and category prototypes to calculate the final detection set. The authors in [26] also propose to modify the RPN of a two-stage architecture, including an attention RPN that takes advantage of the support information to increase the RPN recall. A key component in all meta-learning based object detectors is to compute category prototypes from a set of annotated objects that are general enough to represent each object category. DAnA [11] implements a background attenuation block to minimize the effect of background in the final category prototype and also proposes a new method to summarize the support information in query position aware (QPA) support vectors.

Alternatively to meta-learning based approaches, fine-tuning techniques were proposed. TFA [90] proves that a simple transfer-learning approach, in which only the last layers of existing detectors is fine-tuned with new scarce data, achieves results comparable to those of state-of-the-art meta-learning based detectors. The fine-tuning approach is further developed in DeFRCN [69], including a gradient decoupled layer (GDL) that modulates the influence of the RPN and the network classification and localization head in the training process and a prototypical calibration layer (PCB) to decouple the classification and localization tasks. Few-shot detectors that follow a fine-tuning approach suffer from a non-exhaustive training process, in which objects from novel categories in the base images are not annotated. Therefore, the model learns to treat novel categories as background in the pre-training stage. This issue is addressed in [7] by mining annotations from novel categories in base images. An automatic annotation framework was also proposed in [52] to expand the annotations available for the novel categories through an iterative process of object detection and refinement. This work focuses on problems in which the availability of unlabeled data is not an issue.

To the best of our knowledge, there is only one first attempt [25] to consider spatio-temporal information for few-shot object detection. [25] also proposes for the first time a dataset specifically designed for few-shot video object detection. The method is based on a meta-learning approach, including a tube proposal network to associate objects across frames and a Tube-based Matching Network to compare tube features with support prototypes. Alternatively, we propose a fine-tuning based video object detector that implements feature aggregation at object level to enhance per-frame proposal features.

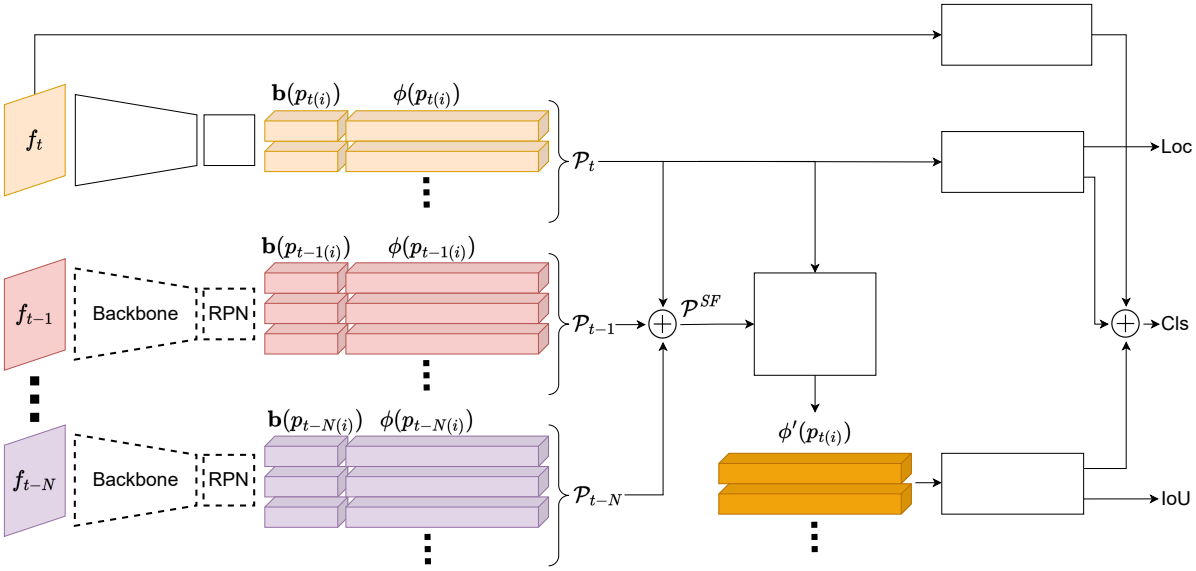


Figure 4.1: FSVDet architecture overview, without the CSO module.

4.3 Proposed Method

4.3.1 Problem Definition

We follow the standard few-shot object detection setting established in previous works [26, 48, 69, 90]. The whole dataset is divided into $\mathcal{D}_{base}$, with several annotated objects of the base classes $\mathcal{C}_{base}$, and $\mathcal{D}_{novel}$, with only a few annotations of the novel classes $\mathcal{C}_{novel}$ for training, with $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$. The detection problem is defined as K-shot, where K is the number of objects annotated for each novel category $\mathcal{C}_{novel}$. Typically, the number of annotations K used to evaluate few-shot detectors in the literature ranges from 1 to 30 examples. As the proposed framework follows a fine-tuning approach, the training process is organized in two general stages. First, the network is trained in $\mathcal{D}_{base}$. Then, it is fine-tuned in $\mathcal{D}_{novel}$.

Although we propose a video object detector, annotations are object instances in single video frames. Thus, $\mathcal{D}_{novel}$ follows the same definition as for single-image approaches: it contains K annotated object instances in isolated video frames. Annotating object trajectories throughout the video would increase the annotation effort and data availability requirements, setting a completely different scenario than the single image domain.

Video object detection aims to localize and classify objects of interest in every frame of an input video. Object detection at each time step t is performed by considering information from the reference frame f_t and N previous frames $f_{t-N}, ..., f_{t-1}$. In the proposed framework, we also include a long-term optimization that considers information from the complete video to optimize the confidence of the detections.

USC DE SANTIAGO DE COMPOSTELA

4.3.2 FSVDet: Few-shot video object detection

We propose FSVDet, a two-stage video object detection framework that can be trained with few labeled examples. The selected spatial baseline is DeFRCN [69], a modification of the original Faster R-CNN [72] able to perform a quick adaptation from a set of base categories C_{base} to a new domain C_{novel}.

Fig. 4.1 shows an overview of the proposed network architecture. First, per-frame image features and object proposals are independently calculated. Then, proposal boxes $\mathbf{b}(p_{t(i)})$ and the corresponding features $\phi(p_{t(i)})$, extracted through RoI Align for proposal $p_{t(i)}$ in the current frame f_t, are fed to the spatial branch of the double head. Following previous work [13], object localization is exclusively performed with information extracted from the current frame. On the other hand, the classification task combines both spatial information from the current frame and spatio-temporal information from a set of input frames to improve the single image classification accuracy.

Feature proposals from the current frame $\phi(p_{t(i)})$ are enhanced by a feature aggregation module that mines proposal relationships with proposals in $\mathcal{P}^{SF}$, extracted from $f_{t-N}, ..., f_{t-1}, f_t$. The aggregation process is described in Sec. 4.3.4. Enhanced features $\phi'(p_{t(i)})$ are the input to the spatio-temporal head that classifies the objects and predicts the overlap of the detection with the ground truth, calculating a class agnostic confidence score. Spatial classification and spatio-temporal classification scores for each category are combined as:

$$s = s_{tmp} + s_{spt}(1 - s_{tmp}) \tag{4.1}$$

being s_{spt} the spatial classification score and s_{tmp} the spatio-temporal score.

As our spatial baseline is DeFRCN [69], we also include the Prototypical Calibration Block (PCB) originally proposed in that work. This module applies a classification score refinement based on a similarity distance metric between network detections and category prototypes. For the category prototype calculation, annotated images from $\mathcal{D}_{novel}$ are fed to a CNN pre-trainend on ImageNet to extract deep image features. Then, a RoI Align layer extracts object features with ground truth boxes, calculating K feature maps per each category in C_{novel}. The final category prototype is calculated by an average pooling operator over the K feature maps of each category. The final classification score is calculated as follows:

$$s' = \beta \cdot s + (1 - \beta) \cdot s^{\cos} \tag{4.2}$$

being $s^{\cos}$ the cosine distance used as similarity metric between category prototypes and detection features extracted by the same CNN as the category prototypes. The hyperparameter β sets the tradeoff between the two confidence scores.

Finally, we define the Confidence Score Optimization method (CSO) that links object detections throughout the video and updates their classification scores. The linking method is based on the category agnostic scores and the overlap between detections in consecutive frames. The

Algorithm 4: Synthetic support frames generation

Input : Single image from the training set: I_t
Input : Object annotations: $\mathcal{A}_t = \{\alpha_{t(j)}\}_{j=1}^{\eta_t}$
Output : Generated synthetic frames: $\mathcal{F}$

1 $\mathcal{F} \leftarrow \emptyset$
2 **for** q **in** $1, ..., Q$ **do**
3 $f_q \leftarrow I_t$
4 **for** j **in** $1, ..., \eta_t$ **do**
5 $L_x \leftarrow \lceil width(I_t)/width(\alpha_{t(j)}\}) \rceil$
6 $L_y \leftarrow \lceil height(I_t)/height(\alpha_{t(j)}\}) \rceil$
7 $L \leftarrow min(L_x, L_y, \gamma)$
8 **for** l **in** $1, ..., L$ **do**
9 $\hat{\alpha}_l \leftarrow \mathcal{T}(\alpha_{t(j)})$
10 $(x, y) \leftarrow \mathcal{P}(f_q)$
11 $f_q \leftarrow C(f_q, \hat{\alpha}_l, (x, y))$
12 $\mathcal{F} \leftarrow \mathcal{F} \cup \{f_q\}$
13 **return** $\mathcal{F} = \{f_q\}_{q=1}^{Q}$

goal of this method is to modify the classification score of detections in each tube, ensuring spatio-temporal consistency. Further details of this method are given in Sec. 4.3.6.

4.3.3 Single image spatio-temporal training

Traditional video object detectors randomly select support frames in the input video sequence for each reference frame for training [12, 13, 21, 104]. However, few-shot object detectors are trained with a limited number of annotated instances per category. Hence, the training set is composed of single images rather than fully annotated videos. To overcome this issue, we propose to generate synthetic support frames for each training image.

Algorithm 4 describes the process of generating a set of synthetic support frames $\mathcal{F}$. The goal of this method is to generate Q synthetic support frames f_q for each training image I_t by inserting transformations of objects from I_t in different positions of f_q. First, f_q is initialized as I_t (Alg. 4:3). Then, each annotated object $\alpha_{t(j)}$ from I_t is randomly transformed and inserted L times in f_q. The number of insertions depends on the size of the object relative to the image size, with a maximum number of insertions of γ (Alg. 4:7). The random transformation operator applies random horizontal flipping over the original object (Alg. 4:9). A random position is also generated to insert the object in f_q (Alg. 4:10), ensuring that the new object fits inside the image boundaries. Directly inserting the cropped object in a different position generates undesirable artifacts that might damage image features, hindering the learning process. Therefore, each transformed object $\hat{\alpha}_l$ is inserted in the position (x, y) in f_q by a seamless cloning operator [68]

(Alg. 4:11). This method can insert an object from a source image into a target image, removing undesirable artifacts. It does not require a segmentation mask, which allows to define the objects boundaries in the source image based on the annotated bounding boxes.

Reusing the original image background to place new objects decreases the probability of inconsistencies between background and foreground features. Reducing these background mismatches is crucial as the detector learns not only the object features but also the context features [10]. Previous methods for position selection to insert new objects in video frames rely on spatio-temporal consistency and many annotated objects to select valid positions [3, 10]. However, the data availability restrictions of few-shot training makes the application of these methods infeasible.

4.3.4 Proposal feature aggregation

Relations among objects in the same image were successfully explored in [45], defining an object relation module based on the multi-head attention introduced in [87]. Similar approaches were also successfully applied in video object detection, mining relations between objects in different frames [13, 21, 45].

The goal of the object relation module is to compute M relation features $\mathbf{r}_R^m$ for each object proposal $p_{t(i)} \in \mathcal{P}_t$ in the reference frame f_t:

$$\mathbf{r}_R^m(p_{t(i)}, \mathcal{P}^{SF}) = \sum_{r=1}^{R} \sum_{j=1}^{|\mathcal{P}_r|} w_{t(i),r(j)}^m \, (W_V \, \phi(p_{r(j)})), \quad m = 1, ..., M \tag{4.3}$$

where $\mathcal{P}^{SF} = \{\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_R\}$ contains object proposals in the R support frames. In training, $\mathcal{P}_t$ contains the proposals extracted from the input image I_t while $\mathcal{P}^{SF}$ contains proposals from the synthetic frames set $\mathcal{F}$ generated through Algorithm 4. $w_{t(i),r(j)}^m$ is a pairwise relation weight between $p_{t(i)}$ and each proposal $p_{r(j)}$ in the support frame f_r based on appearance and geometry similarities. Each proposal is defined by its appearance features $\phi(p_{t(i)})$ and geometry features $\mathbf{b}(p_{t(i)})$. W_V is a learnable transformation matrix.

The definition of the pairwise relational weight $w_{t(i),r(j)}^m$ is:

$$w_{t(i),r(j)}^m = \frac{g_{t(i),r(j)}^m \, exp(a_{t(i),r(j)}^m)}{\sum_q g_{t(i),q}^m \, exp(a_{t(i),q}^m)} \tag{4.4}$$

being $a_{t(i),r(j)}^m$ the pairwise appearance similarity weight and $g_{t(i),r(j)}^m$ the geometry similarity weight.

The appearance similarity weight is as a normalized dot product:

$$a_{t(i),r(j)}^m = \frac{\langle W_H \, \phi(p_{t(i)}), W_Q \, \phi(p_{r(j)}) \rangle}{\sqrt{d_h}} \tag{4.5}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

where W_H and W_Q are learnable parameters that project original appearance features to a subspace to measure their similarity, being d_H the feature dimension after the projection.

The geometry similarity weight is calculated as:

$$g^m_{t(i),r(j)} = max\{0, W_G\,\mathcal{E}(\mathbf{b}(p_{t(i)}), \mathbf{b}(p_{r(j)}))\} \tag{4.6}$$

where W_H is also a learnable parameter. Geometry features containing the parameters of the bounding box (x, y, w, h) are encoded as $\left(log\left(\frac{|x_i - x_j|}{w_i}\right), log\left(\frac{|y_i - y_j|}{h_i}\right), log\left(\frac{w_j}{w_i}\right), log\left(\frac{h_j}{h_i}\right)\right)$. Function $\mathcal{E}$ represents an embedding calculation into a high-dimensional representation following [87].

The final feature map for proposal $p_{t(i)}$ is enhanced by adding the concatenation of the M relational features $\mathbf{r}^m_R(p_{t(i)}, \mathcal{P}^{SF})$ (Eq. 4.3):

$$\phi'(p_{t(i)}) = \phi(p_{t(i)}) + concat[\{\mathbf{r}^m_R(p_{t(i)}, \mathcal{P}^{SF})\}^M_{m=1}] \tag{4.7}$$

Following previous work [13, 21], we implement a multi-stage relation module with a basic stage and an advanced stage. The basic stage enhances proposal features in the current frame $\mathcal{P}_t$ by mining relationships with the proposals in the support frames $\mathcal{P}^{SF}$, generating $\mathcal{P}'_t$. In the advanced stage, the proposals in $\mathcal{P}^{SF}$ are first improved by aggregating them with the top-ϵ% scoring proposals in $\mathcal{P}^{SF}$. This enhanced support proposal features are finally used in this advanced stage to aggregate with proposals in $\mathcal{P}'_t$, generating the final proposal set.

4.3.5 Class-agnostic confidence score

We propose a double objective optimization loss for the training of the spatio-temporal branch of the network double head: a classification loss and an overlap prediction. On the one hand, the classification loss function is implemented as a cross-entropy loss, following the standard approach to train a multi-class classifier. On the other hand, the overlap prediction is based on estimating the overlap of each detection with the actual objects in the image. This loss function is implemented as a binary cross-entropy loss. Hence, the final loss for the spatio-temporal branch is defined as:

$$\mathcal{L} = \mathcal{L}_{CLS} + \mathcal{L}_{IoU} \tag{4.8}$$

In this implementation, the overlap is calculated as the intersection over union (IoU). The following equation describes the target generation process for training:

$$\psi(p_{i(j)}, I_t) = \max_{\forall \alpha_{i(k)} \in I_t} IoU(p_{i(k)}, \alpha_{i(j)}) \tag{4.9}$$

This equation calculates the maximum overlap between each object proposal $p_{i(k)}$ in the current training image I_t and every annotation $\alpha_{i(j)}$ in I_t.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Previous work on the use of the IoU to increase the correlation of the confidence score with the actual detection precision has shown the effectiveness of binary cross-entropy over other common regression loss function alternatives such as L2 loss [92]. Therefore, our $\mathcal{L}_{IoU}$ is defined as:

$$\begin{aligned}\mathcal{L}_{IoU} = \underset{n=1,\dots,B}{\text{mean}} [&\psi(p_{i(j)}, I_t) \cdot \log \hat{\psi}(p_{i(j)}) + \\ &(1 - \psi(p_{i(j)}, I_t)) \cdot \log(1 - \hat{\psi}(p_{i(j)}))]\end{aligned} \tag{4.10}$$

being $\hat{\psi}(p_{i(j)})$ the predicted overlap for a proposal $p_{i(j)}$ and $\psi(p_{i(j)}, I_t)$ the actual maximum overlap with the ground truth defined in Eq. 4.9.

Experimental results show that, by simply including the IoU-aware loss function, it increases the detection precision (Sec. 4.4.3).

4.3.6 Confidence Score Optimization (CSO)

Confidence prediction plays a fundamental role in object detection. The network head of an object detector calculates a per-proposal final detection with an associated confidence. Then, redundant detections are removed by means of Non-Maximum Suppression (NMS). A confidence threshold is also commonly applied to discard false positive detections. Thus, the confidence score strongly influences the final detection set. State-of-the-art object detectors directly use the classification score as a detection confidence.

Long tube generation has been successfully applied to optimize detection confidence leveraging spatio-temporal consistency for object detection [15, 21, 29] and action recognition [35, 75]. This idea is the basis of our Confidence Score Optimization (CSO) method, which first builds long tubes from detections of one object and, then, increases the classification scores based on the detections of the tube. We propose to calculate the link score *ls* between a detection $d_{t(i)}$ in frame f_t and a detection $d_{t'(j)}$ in $f_{t'}$ as follows:

$$ls(d_{t(i)}, d_{t'(j)}) = \hat{\psi}(d_{t(i)}) + \hat{\psi}(d_{t'(j)}) + 2 \cdot IoU(d_{t(i)}, d_{t'(j)}). \tag{4.11}$$

being $\hat{\psi}(d_{t(i)})$ the predicted IoU score described in Sec. 4.3.5. This increases the probability of linking detections with greater predicted overlap with the ground truth. As the output of $\hat{\psi}(d_{t(i)})$ is between 0 and 1 the factor 2 ensures that the network confidence and the overlap between $d_{t(i)}$ and $d_{t'(j)}$ have the same influence on the final link score.

Each object tube $\hat{v}$ from all possible tubes $\mathcal{V}$ is calculated by iteratively solving the following equation:

$$\hat{v} = \underset{\mathcal{V}}{\arg\max} \sum_{t=2}^{T} ls(D_{t-1}, D_t) \tag{4.12}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

where D_t is the detection set in frame f_t. Once an object tube $\hat{v}$ is calculated, all detections belonging to $\hat{v}$ are removed from the candidate detection set to build a new object tube. The output of *ls* is a matrix with the linking score between each detection in D_{t-1} and each detection in D_t. Eq. 4.12 can be solved applying the Viterbi algorithm. Then, following previous methods [12,13,21,29], detections belonging to each tube $\hat{v}$ are updated, setting their classification score to the mean classification score of the top-20% in $\hat{v}$.

4.4 Experiments

4.4.1 Experimental settings

We use the FSVOD-500 dataset [25] to evaluate our model. It contains 2,553 annotated videos with 320 different object categories for $\mathcal{D}_{base}$, and 949 videos with 100 object categories for $\mathcal{D}_{novel}$*. Annotations in FSVOD-500 are provided at 1fps, thus there are no annotations available for every frame in the video. Following the few-shot setting, object categories from these splits are completely different: base categories C_{base} mainly contain common categories, while novel categories C_{novel} are rare categories for which available data is scarce. This simulates a real world scenario in which there are few annotations for target categories, but public datasets can be used as $\mathcal{D}_{base}$.

Following [25], instead of defining a unique global training set for fine-tuning ($\mathcal{D}^{train}_{novel}$) and a testing set with the remaining videos ($\mathcal{D}^{test}_{novel}$), we select multiple fine-tuning sets, so that each video in $\mathcal{D}_{novel}$ is picked once for test. For this, in our implementation $\mathcal{D}_{novel}$ is randomly divided into two subsets, keeping the same distribution of videos per object category. Thus, one subset is used as $\mathcal{D}^{test}_{novel}$, while $\mathcal{D}^{train}_{novel}$ is selected from the other subset, picking randomly K object annotations per category. Then, the subsets are interchanged, so each video is in $\mathcal{D}^{test}_{novel}$ once. We repeat this whole process 5 times —with different random splits—, and the reported results include the mean and standard deviation of these 5 executions The same 5 randomly generated $\mathcal{D}^{train}_{novel}$ are used in all the experiments performed in this work for a fairer comparison, avoiding the effect of selecting different fine-tuning sets for different models. The number of annotations per category K is set to 5 by default in all experiments to facilitate the comparison with previous methods. An analysis on the influence of K is also provided.

4.4.2 Implementation details

The feature extractor backbone for all experiments is a ResNet-101 pretrained on ImageNet [74]. First, we train the spatial part of FSVDet on $\mathcal{D}_{base}$. Then, we fine-tune both the spatial and

*FSVOD500 also contains a validation set with 770 annotated videos with 80 object categories. We do not use this set in the experimentation.

Extra data	Synthetic Data	AP_{50}
		$38.5_{\pm 0.7}$
✓		$38.9_{\pm 0.5}$
	✓	$38.8_{\pm 0.7}$

Table 4.1: Different support frames generation strategies. Results for FSVDet without CSO.

spatio-temporal parts of FSVDet on $\mathcal{D}_{novel}^{train}$ —for the spatial part, we start with the pretrained weights on $\mathcal{D}_{base}$, while the spatio-temporal weights are randomly initialized.

For the baseline training on $\mathcal{D}_{base}$, we set the batch size to 16 and the base learning rate to 2×10^{-2} for the first 20K iterations, reducing it to 2×10^{-3} for the next 5K iterations, and to 2×10^{-4} for the last 5K iterations. The baseline fine-tuning on $\mathcal{D}_{novel}^{train}$ is also performed with 16 images per batch with a learning rate set to 1×10^{-2} for the first 9K iterations, reducing it to 1×10^{-3} for the last 1K iterations. The spatio-temporal training is performed for 40K iterations with 1 image per batch and an initial learning rate of 2.5×10^{-4}, reducing it to 2.5×10^{-5} after the first 30K iterations. The classification loss function is implemented as cross entropy with label smoothing regularization [80].

Input images are resized with the shortest dimension randomly set to (640, 672, 704, 736, 768, 800) pixels for training and 800 pixels for test. For each video frame, 15 support frames are also considered to mine object relations at test time. For training, 2 synthetic support frames are generated for each input image with a maximum number of $\gamma = 5$ new objects inserted into each synthetic support frame. The hyperparameter β that modulates the influence of the Prototypical Calibration Block on the classification score is set to 0.5. For the relation module, the ratio of proposals selected for the advanced stage ϵ is 20%.

4.4.3 Ablation Studies

We have conducted a series of ablation studies to assess the performance of each component of the detection framework. In addition to the final AP, we also provide an error type analysis to evaluate how the proposed model improves the baseline results.

Tab. 4.1 shows the result of applying different strategies regarding the support frames selection for training. The first row contains the results for the simplest strategy in which the reference frame is replicated, generating a static video. In the second row, we also report the AP achieved training the proposed model with real support frames extracted from the original video to which the images in $\mathcal{D}_{novel}^{test}$ belong. This approach violates the data availability limitations, as extra data is used in the training process, and can only be applied in scenarios in which the whole training videos are available. Nevertheless, we have included it in the table for comparison purposes. The proposed synthetic support frame generation method —Tab. 4.1, third row— achieves results comparable to those obtained from selecting real frames from the



Figure 4.2: Examples of real reference frames and their corresponding synthetic support frames generated following Alg. 4. The synthetic support frames might appear flipped as the network data augmentation includes horizontal flipping.

IoU loss	CSO	AP_{50}
		$38.6_{\pm 0.4}$
✓		$38.8_{\pm 0.7}$
✓	✓	$41.9_{\pm 2.0}$

Table 4.2: Ablation Studies for FSVDet.

original videos. These results prove that the variability generated by placing objects on different backgrounds helps the training process without requiring any extra data. Fig. 4.2 shows some examples of the generated synthetic support frames.

Tab. 4.2 provides an analysis of the influence of the IoU loss. These results prove that sharing the spatio-temporal features for the spatio-temporal classification and IoU prediction does not harm classification. In fact, including the multi-objective loss function defined in Eq. 4.8 improves the baseline results reported in the first row of the table. This baseline is trained only with the classification loss as the learning optimization objective for the spatio-temporal branch. We argue that the more complex loss function reduces the overfitting caused by the data scarcity. By using the predicted IoU to generate object tubes and performing the final classification score optimization in the CSO module, the AP_{50} further improves 3.1 points.

We have performed an error type analysis based on the TIDE toolbox [2] to identify the source of errors. This framework defines six different error types: classification (Cls.), localization (Loc.), localization and classification (Both), duplicated detections (Dup.), background detected

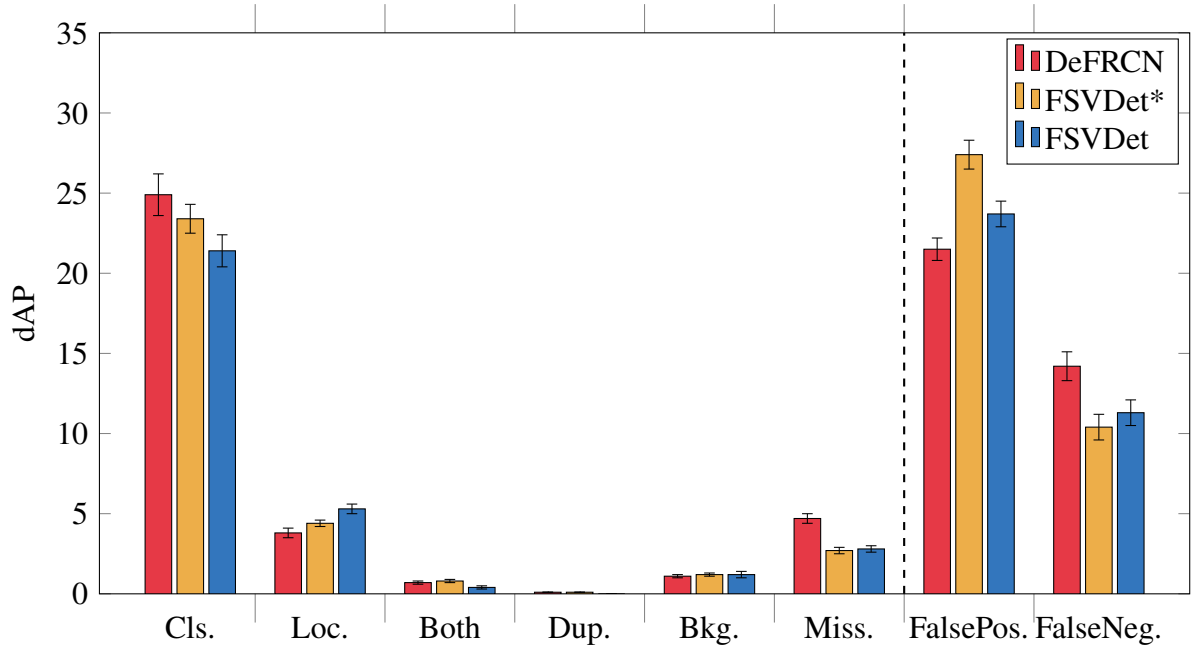


Figure 4.3: Error type analysis for the single image baseline (DeFRCN), FSVDet and FSVDet without CSO (FSVDet*). For five random training sets we show the mean and standard deviation of the dAP [2] —difference in AP, the lower the better.

as object of interest (Bkg.) and missing objects (Miss.). Moreover, it also analyzes false positive and false negative errors. The proposed methodology consists of defining oracles that solve each error type independently and report the difference in AP (dAP). Fig. 4.3 shows the comparison between the single image baseline (DeFRCN), FSVDet and FSVDet without CSO (FSVDet* in the figure).

As expected in a few-shot scenario, the most prominent error type is the classification error. That motivates the need to extract information from multiple frames in video sequences. We argue that combining spatial information from multiple input frames might not contribute to localize the object in the current frame, as the most valuable spatial information to localize each object proposal must come from the current frame. However, this spatio-temporal information might be crucial for object classification, in order to overcome issues such as motion blur or occlusions that have a greater impact in the classification task. Our spatio-temporal method (21.4 dAP) improves the single image baseline (24.9 dAP) by 3.5 dAP in classification error.

Fig. 4.3 also shows that the proposed method without CSO (FSVDet*) has a higher false positive error than the baseline but, on the other hand, it decreases the false negative error of the baseline. Including CSO, the whole system achieves a better balance between false positives and false negatives, significantly reducing the number of false positives.

Fig. 4.4 analyzes the influence of the shot size (K) for both DeFRCN and FSVDet. Our approach outperforms the spatial baseline in all the cases, from 1 annotated example per category up to 30 annotated examples per category. As expected, the results also show an improvement as the number of annotations in $\mathcal{D}^{train}_{novel}$ increases. DeFRCN improves 18.6 AP_{50} by increasing K

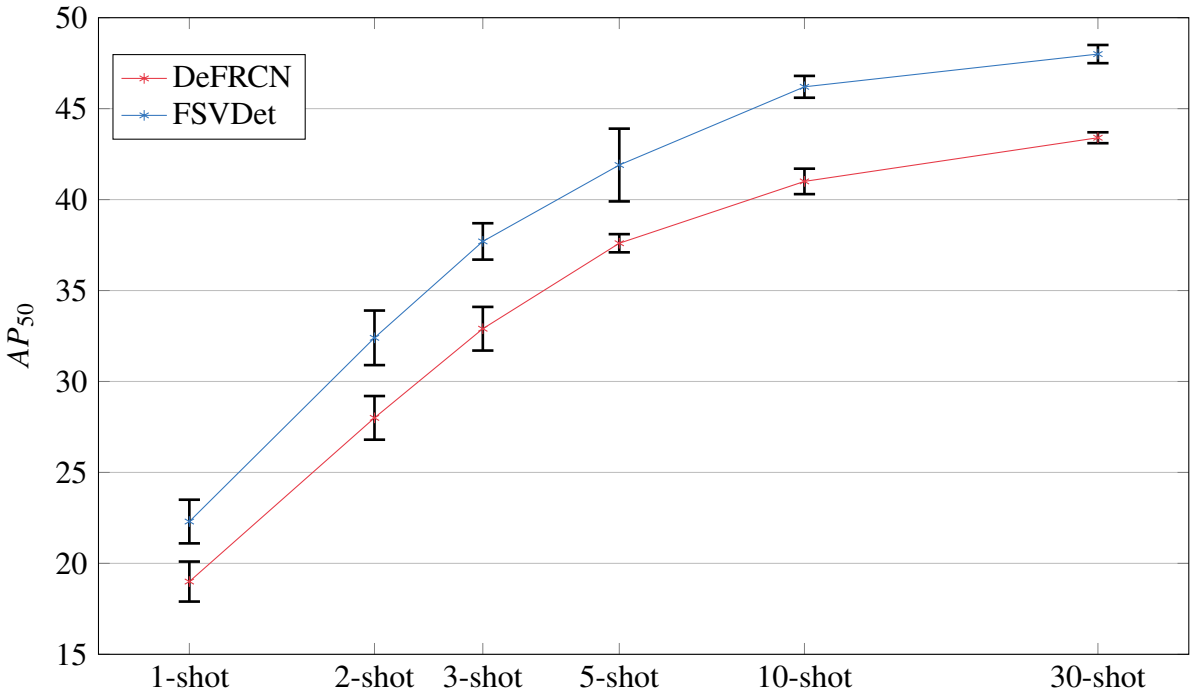


Figure 4.4: Influence of K. Mean AP and standard deviation for each value of K.

from 1 to 5 while FSVDet improves 19.6 AP_{50}. The maximum AP_{50} for FSVDet is 48.0, outperforming DeFRCN by 4.6 points (43.4 AP_{50}) and achieving a total improvement —from 1-shot to 30-shot— of 25.7 points, which is 1.3 points higher than DeFRCN. Thus, our method performs better in very limited data availability scenarios, and also exploits better new annotations with a higher increment in the final AP_{50}.

4.4.4 Experimental Results

Tab. 4.3 shows the comparison with the state-of-the-art for the FSVOD-500 dataset. As the few-shot video object detection problem remains almost unexplored with only one previous work, for comparison purposes we also include single image few-shot object detectors, traditional video object detectors and methods based on multiple object tracking (MOT). The results for traditional video object detectors and MOT-based methods were originally reported in [25]. In those experiments, authors extracted class agnostic object tubes and applied a meta-learner classifier to calculate the final detection set.

FSVDet outperforms previous approaches, improving our single image baseline (DeFRCN) by 4.3 points and previous few-shot video object detectors by 3.2 points. Traditional video object detectors that include attention mechanisms for mining proposal relationships [12, 21] fail to perform few-shot object detection falling behind single image few-shot detectors. This proves the need for algorithms specifically designed for few-shot video object detection. We also evaluate whether the differences are statistically significant through the Wilcoxon Signed Rank

Table 4.3: Results on the few-shot video object detection FSVOD-500 dataset. The comparision includes single image object detectors, traditional video object detectors, multiple object trackers + meta learning classifier, and few-shot video object detectors.

Type	Method	AP_{50}
Obj. Det.	Faster R-CNN [72]	$26.4_{\pm 0.4}$
Few-shot Obj. Det.	TFA [90]	$31.0_{\pm 0.8}$
	FSOD [26]	$31.3_{\pm 0.5}$
	DeFRCN [69]	$37.6_{\pm 0.5}$
Vid. Obj. Det.	MEGA [12]	$26.4_{\pm 0.5}$
	RDN [21]	$27.9_{\pm 0.4}$
Mult. Obj. Track.	CTracker [67]	$30.6_{\pm 0.7}$
	FairMOT [100]	$31.0_{\pm 1.0}$
	CenterTrack [101]	$30.5_{\pm 0.9}$
Few-shot Vid. Obj. Det.	FSVOD [25]	$38.7_{\pm 0.7}$
	FSVDet	$41.9_{\pm 2.0}$

Test on paired samples over multiple runs of DeFRCN and FSVDet. We have used the STAC platform [73] to perform this analysis. We obtain a probability of having statistically significant differences (1 – p-value) of 95.7%. We cannot replicate the same analysis for comparing our method with previous state-of-the-art approaches, as the authors only report the aggregated information. In summary, we have set a strong baseline that achieves competitive results compared to previous methods, and we can conclude that FSVDet outperforms this baseline and the differences are statistically significant.

4.5 CONCLUSIONS

We have proposed FSVDet, a new few-shot video object detection framework that, first, applies attention mechanisms to mine proposals relationships between different frames. Then, a spatio-temporal double head classifies object proposals leveraging spatio-temporal information, and it also predicts the overlap of each proposal with the ground truth. Finally, overlapped predictions are used in an object linking method to create long tubes and optimize classification scores. We have also defined a new training strategy to learn from single images while considering a group of input frames at inference. FSVDet outperforms previous solutions by a large margin, and establishes a new state-of-the-art result for the FSVOD-500 dataset for few-shot video object detection.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5 Conclusions

The main challenges of video object detection have been addressed in this PhD Thesis with the introduction of three different video object detectors. Chapters 2 and 3 propose new detection frameworks that leverage spatio-temporal information available in videos to boost the detection precision. Chapter 4 reimplements the main modules of these methods to solve the few-shot video object detection problem.

In Chapter 2 we have described a video object detection framework called FANet that extends a two-stage object detector to exploit spatio-temporal information available in videos. It includes a feature aggregation module that summarizes information from a set of input frames by aggregating information throughout short object tubelets. This module produces enhanced proposal features alleviating issues such as motion blur, out of focus or, in general, any image degradation in a particular frame. These spatio-temporal features are processed by a spatio-temporal double head that also considers spatial information from the current frame to boost the detection precision. A similar implementation of the network head was followed in Chapters 3 and 4 showing promising results in a wide range of scenarios. The confidence score of the final detection set is optimized, ensuring spatio-temporal consistency throughout long tubes. This method also considers short object tubelets to overcome missing detections, generating longer tubes. Results show that this approach effectively increases the score of low-confidence true positive detections, assigning them a higher score than false-positive detections. The proposed video object detector outperforms the single image baseline by 5.0% AP_{50} in the ImageNet VID dataset and by 5.4% AP_{50} in the USC-GRAD-STDdb dataset.

Chapter 3 presented SLTnet, an improved video object detector that modifies the short-term aggregation proposed in Chapter 2 and incorporates a long-term aggregation module. The new short-term aggregation strategy incorporates a similarity metric among features in nearby frames to modulate the contribution of each proposal feature map. Therefore, a proposal linking error in the short-term has less impact than in the previous approach. Long-term information is included in the network, performing feature aggregation based on attention mechanisms to relate features in different frames. This method calculates object trajectories based on previous attention weights to update positions of proposals from previous key frames. The updated positions can be compared with positions of proposals in the current frame allowing the attention module to

USC
DE COMPOSTELA

fully exploit appearance and geometry features. SLTnet outperforms the single image baseline in five different datasets and achieves better results on average than previous state-of-the-art video object detectors, proving to be more robust.

We have also explored the benefits of exploiting spatio-temporal information in data scarce scenarios with the introduction of FSVDet in Chapter 4. This chapter describes a new training strategy for the attention-based aggregation module aiming to learn from single images to mine detection relationships between proposals in different frames. A new double head architecture was also proposed, extending the spatio-temporal branch to also predict the overlap of each detection with the ground truth. The introduction of the multi-task learning objective in the spatio-temporal branch reduces the overfitting caused by the data scarcity. This overlap prediction was used by a long-term object linking algorithm to optimize the final confidence score, ensuring spatio-temporal consistency. Experimental results show that the proposed solutions lead to a reduction of the classification error, the main issue in few-shot object detection. Also, our system outperforms the single image baseline by 4.3% AP_{50}, and the best few-shot video object detector by 3.2% AP_{50} in the FSVOD500 dataset, achieving a 41.9% AP_{50} which is the new state-of-the-art for this dataset.

Overall, the thesis results highlight the importance of spatio-temporal information to solve the video object detection problem. The results of the proposed solutions exceed those from the single-image counterpart on multiple datasets. This supports the initial hypothesis that applying single-image techniques yields suboptimal results for video object detection in both data-rich and data-scarce scenarios, as single-image techniques do not fully exploit the available information. Still, the advent of novel techniques to deal with spatio-temporal information as visual transformers, or the mining of pseudo-annotations from unlabeled images for the few-shot approach, are promising lines of research to further improve our results in video object detection in the near future.

FUTURE WORK

Few-shot object detection is an emerging field of research that might dramatically extend the potential applications of object detection to new domains. The results obtained in few-shot video object detection in Chapter 4 open a very interesting line of research for future work. Specifically, the recent trend [7, 52] of automatically mining new annotations extending the training set to boost the detection precision might be of great importance in the video domain. The pipeline defined in [52] includes steps for detections filtering, verification and correction to generate pseudo-annotations in a set of unlabeled images. Hence, it assumes that a set of unlabeled images is available, and the main limitation is the cost of manually annotating new object instances. This may be a reasonable assumption for most applications, and even more likely in the video domain, as full video annotation is highly time-consuming: each object

instance appears in several video frames, requiring to fully annotate each of them, dramatically increasing the number of images that need to be annotated. However, automatically generating pseudo-annotations having one video frame annotated is more straightforward than mining new annotations in completely different images. This strategy can be repeated iteratively to extend the initial annotations set. Iteratively discovering new pseudo-annotations, increasing the variability of the training set, should also improve the object detector precision in each iteration.

The long-term object linking and score optimization, which increase the confidence score of detections that can be associated with a long tubes, have proved that leveraging spatio-temporal consistency significantly improves the detection precision. Applying a more robust tracker than an overlap based method should create more robust tubes. Object tracking techniques can be combined with few-shot detection methods to discover new pseudo-annotations throughout the training videos. Single Object Trackers (SOT) are class agnostic and do not require annotated instances in the target domain. Therefore, the combination of these techniques might be useful for label verification by removing detections that do not fit into a tracked trajectory. These filtering techniques have not been applied in the context of video object detection, as removing annotations that do not belong to any trajectory would damage the system recall. However, false positive pseudo-annotations have a higher negative impact than false negatives in the training process, making pseudo-annotations precision far more important than recall. Thus, opposite to a conventional object detector, the pseudo-annotations mining pipeline should focus on removing false negatives rather than on achieving a good precision-recall trade-off.

This new few-shot setting based on pseudo-annotations requires the design of new datasets to evaluate the models. The dynamic training set for novel categories proposed by the authors of FSVOD500 [25] is not feasible in this context. Rotating the training set to perform object detection inference in the entire dataset would require to run multiple iterations of the pseudo-annotation mining process for each new training set. Therefore, a new data partition with few annotations and a set of unlabeled video frames is required in addition to the test set. Moreover, recent work on few-shot detection in single images also focused on the precision of the model in base categories. This analysis shows the performance drop in base categories due to the fine-tuning in novel categories, which might be of great importance in applications that require to detect objects from both novel and base categories. Although FSVOD500 is a first attempt of defining a specific dataset for few-shot video object detector, a more complete dataset is needed to fully evaluate and compare the performance of future few-shot video object detectors.

The use of CNNs as feature extractor has been the standard approach in recent years for many computer vision tasks. However, architectures based on Visual Transformers are becoming more popular, achieving impressive results in different research areas. The Transformer architecture was originally designed in the Natural Language Processing (NLP) field to model relationships in sequences of words. Also inspired by the success achieved in NLP, masked autoencoders have

USC DE COMPOSTELA

been proposed to train Visual Transformers [41] following a self-supervised learning strategy. This approach perfectly fits in the context of mining pseudo-annotations from unlabeled images. It is expected that including this information in the training process will help to improve the pseudo-annotations extracted from these images.

The standard approach in video object detection is to apply a single-image backbone in each video frame independently, and then perform feature aggregation throughout a set of input frames. However, it might be worth exploring new feature extraction techniques that consider spatio-temporal information in earlier network stages. Visual transformer architectures have also proven to be highly effective for video classification [59], considering multiple frames. Visual Transformers generally tokenize the input image by dividing it into patches. Then, these patches are processed, applying self-attention blocks that consider patches' positions and appearance. The extension of this idea to the spatio-temporal domain requires a new position embedding calculation that also takes into account the time dimension.

Improving detection performance in terms of frames per second was out of the scope of this Thesis. Previous work [37] has explored the possibility of combining a strong feature extractor in some frames with a lighter model in other frames to improve the execution time with CNN-based backbones. The key idea is to exploit spatio-temporal redundancy present in videos to avoid unnecessary calculations, enhancing the low-quality features by aggregating them with more robust features extracted from previous frames. Backbones such as Visual Transformers (ViT) [22] define multiple models with different numbers of parameters ranging from 86M of ViT-Base to 632M of ViT-Huge. As a reference, the recent ViTDet object detector [58] requires 77ms per image using ViT-Base while the implementation based on ViT-Huge needs 189ms per image on inference, with a precession gain of 5.1% AP in the COCO dataset. Combining features extracted by ViT-Base in most frames with features extracted by ViT-Huge in specific frames might achieve a better trade-off between execution time and precision.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Bibliography

[1] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2018.

[2] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. TIDE: A general toolbox for identifying object detection errors. In *European Conference on Computer Vision (ECCV)*, pages 558–573. Springer, 2020.

[3] Brais Bosquet, Daniel Cores, Lorenzo Seidenari, Víctor M. Brea, Manuel Mucientes, and Alberto Del Bimbo. A full data augmentation pipeline for small object detection based on generative adversarial networks. *Pattern Recognition*, page 108998, 2022.

[4] Brais Bosquet, Manuel Mucientes, and Víctor M Brea. STDnet: A convnet for small target detection. In *British Machine Vision Conference (BMVC)*, 2018.

[5] Brais Bosquet, Manuel Mucientes, and Víctor M Brea. STDnet: Exploiting high resolution feature maps for small object detection. *Engineering Applications of Artificial Intelligence*, 91:103615, 2020.

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Yuhang Cao, Jiaqi Wang, Yiqi Lin, and Dahua Lin. Mini: Mining implicit novel instances for few-shot object detection. *arXiv preprint arXiv:2205.03381*, 2022.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229, 2020.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the Kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

USC DE COMPOSTELA

[10] Changrui Chen, Yu Zhang, Qingxuan Lv, Shuo Wei, Xiaorui Wang, Xin Sun, and Junyu Dong. Rrnet: A hybrid detector for object detection in drone-captured images. In *IEEE International Conference on Computer Vision Workshops (ICCV)*, 2019.

[11] Tung-I Chen, Yueh-Cheng Liu, Hung-Ting Su, Yu-Cheng Chang, Yu-Hsiang Lin, Jia-Fong Yeh, Wen-Chin Chen, and Winston Hsu. Dual-awareness attention for few-shot object detection. *IEEE Transactions on Multimedia*, 2021.

[12] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10337–10346, 2020.

[13] Daniel Cores, Victor M Brea, and Manuel Mucientes. Short-term anchor linking and long-term self-guided attention for video object detection. *Image and Vision Computing*, 110:104179, 2021.

[14] Daniel Cores, Víctor M Brea, and Manuel Mucientes. Spatio-temporal object detection from UAV on-board cameras. In *International Conference on Computer Analysis of Images and Patterns (CAIP)*, pages 143–152, 2021.

[15] Daniel Cores, Víctor M Brea, and Manuel Mucientes. Spatiotemporal tubelet feature aggregation and object linking for small object detection in videos. *Applied Intelligence*, pages 1–13, 2022.

[16] Daniel Cores, Manuel Mucientes, and Víctor M Brea. RoI feature propagation for video object detection. In *European Conference on Artificial Intelligence (ECAI)*, 2020.

[17] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[18] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, 2005.

[19] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

[20] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Object guided external memory network for video object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6678–6687, 2019.

USC
DE COMPOSTELA

[21] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 7023–7032, 2019.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020.

[23] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *European Conference on Computer Vision (ECCV)*, pages 370–386, 2018.

[24] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[25] Qi Fan, Chi-Keung Tang, and Yu-Wing Tai. Few-shot video object detection. In *European Conference on Computer Vision (ECCV)*, 2022.

[26] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-RPN and multi-relation detector. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4013–4022, 2020.

[27] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. Spatiotemporal residual networks for video action recognition. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[28] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[30] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[31] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

[32] Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region and semantic segmentation-aware CNN model. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[33] Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[34] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[35] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[36] Tao Gong, Kai Chen, Xinjiang Wang, Qi Chu, Feng Zhu, Dahua Lin, Nenghai Yu, and Huamin Feng. Temporal ROI align for video object recognition. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 1442–1450, 2021.

[37] Chaoxu Guo, Bin Fan, Jie Gu, Qian Zhang, Shiming Xiang, Véronique Prinet, and Chunhong Pan. Progressive sparse local attention for video object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3909–3918, 2019.

[38] Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. Mining inter-video proposal relations for video object detection. In *European Conference on Computer Vision (ECCV)*, pages 431–446. Springer, 2020.

[39] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. Seq-NMS for video object detection. *arXiv preprint arXiv:1602.08465*, 2016.

[40] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3018–3027, 2017.

[41] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.

[42] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.

USC
DE COMPOSTELA

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[45] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3588–3597, 2018.

[46] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[47] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Joint learning of object and action detectors. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[48] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8420–8429, 2019.

[49] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. Object detection in videos with tubelet proposal networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[50] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-CNN: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2896–2907, 2017.

[51] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[52] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Label, verify, correct: A simple few shot object detection method. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14237–14247, 2022.

[53] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 25, 2012.

[54] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

[55] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[56] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[57] Byungjae Lee, Enkhbayar Erdenee, Songguo Jin, Mi Young Nam, Young Giu Jung, and Phill Kyu Rhee. Multi-class multi-object tracking using changing point detection. In *European Conference on Computer Vision (ECCV)*, 2016.

[58] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022.

[59] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved multiscale vision transformers for classification and detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[60] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[61] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[62] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[63] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8759–8768, 2018.

[64] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 2016.

[65] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[66] Ala Mhalla, Thierry Chateau, and Najoua Essoukri Ben Amara. Spatio-temporal object detection by deep learning: Video-interlacing to improve multi-object tracking. *Image and Vision Computing*, 88:120–131, 2019.

[67] Jinlong Peng, Changan Wang, Fangbin Wan, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *European Conference on Computer Vision (ECCV)*, pages 145–161, 2020.

[68] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003.

[69] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. DeFRCN: Decoupled Faster R-CNN for few-shot object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8681–8690, 2021.

[70] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[71] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7263–7271, 2017.

[72] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[73] Ismael Rodríguez-Fdez, Adrián Canosa, Manuel Mucientes, and Alberto Bugarín. STAC: a web platform for the comparison of algorithms using statistical tests. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015.

[74] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[75] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. In *British Machine Vision Conference (BMVC)*, 2016.

[76] Mykhailo Shvets, Wei Liu, and Alexander C Berg. Leveraging long-range temporal relationships between proposals for video object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9756–9764, 2019.

USC DE COMPOSTELA

[77] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[78] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

[79] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017.

[80] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[81] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.

[82] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10781–10790, 2020.

[83] Peng Tang, Chunyu Wang, Xinggang Wang, Wenyu Liu, Wenjun Zeng, and Jingdong Wang. Object detection in videos by high quality object linking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[84] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9627–9636, 2019.

[85] Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017.

[86] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[87] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, 2017.

USC
DE COMPOSTELA

[88] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2001.

[89] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2018.

[90] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In *International Conference on Machine Learning (ICML)*, pages 9919–9928, 2020.

[91] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9217–9225, 2019.

[92] Shengkai Wu, Xiaoping Li, and Xinggang Wang. IoU-aware single-stage object detector for accurate localization. *Image and Vision Computing*, 97:103911, 2020.

[93] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[94] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *European Conference on Computer Vision (ECCV)*, 2018.

[95] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3415–3424, 2017.

[96] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[97] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta R-CNN: Towards general solver for instance-level low-shot learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9577–9586, 2019.

[98] Jing Yang, Hui Shuai, Zhengbo Yu, Rongrong Fan, Qiang Ma, Qingshan Liu, and Jiankang Deng. ILSVRC2016 object detection from video: Team NUIST, 2016.

[99] Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Asghar, and Brian Lee. A survey of modern deep learning based object detection models. *Digital Signal Processing*, page 103514, 2022.

[100] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. A simple baseline for multi-object tracking. *arXiv preprint arXiv:2004.01888*, page 6, 2020.

[101] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision (ECCV)*, pages 474–490. Springer, 2020.

[102] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018.

[103] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable ConvNets v2: More deformable, better results. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9308–9316, 2019.

[104] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[105] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

List of Figures

Fig. 1.1 Object detection output in different datasets. 2
Fig. 1.2 Different two-stage object detectors. 4

Fig. 2.1 FANet architecture with a single level backbone. Dotted backbone and RPN boxes represent components that are reused without new computations. Blocks labeled as 2MLP in the network head implement a multilayer perceptron with two fully connected layers. 22
Fig. 2.2 Temporal pooling example with a number of input frames $N = 4$. From left to right, there are the four input frames with an object proposal (the blue bounding box) in each one, belonging all of them to the same tubelet. Under each input frame, a subset of channels from the RoI Align output are represented. On the right, we show the tubelet proposal linking all box proposals. At the bottom right, we also show the aggregated feature map calculated by our Temporal Pooling method. We highlight one channel (surrounded in red) as an example of how the highest activations (lighter colors) in each frame contribute to the aggregated feature map. 24
Fig. 2.3 Long-term object linking. Green boxes are actual network detections. The network does not detect the two objects in f_4 breaking $\tilde{v}_j$ in two fragments. The last detection of the first fragment and the first detection of the second fragment belong to the same RPN tubelet (τ_k). 26
Fig. 2.4 Detection mAP when training and testing with the same tubelet lengths (N) on the ImageNet VID validation set without long-term information. 31

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Fig. 2.5 Detection mAP when training with $N = 2$ and testing with different tubelet lengths (N) on the ImageNet VID validation set without long-term information. 31

Fig. 3.1 Our approach has three main components: (a) short-term object linking based on anchor boxes, and box aggregation features throughout the nearby frames; (b) long-term self-guided attention module that enhances short-term aggregated features with key frame information ($k f_{t'}$); (c) spatio-temporal double head. . 43
Fig. 3.2 Images from ImageNet VID validation set. Red boxes represent the bounding box in the supporting frame. The green box represents the object location in the reference frame. Location information from the supporting frame (red box) is not accurate in the reference frame (green box). 45
Fig. 3.3 Average number of objects per frame for each test set. 50
Fig. 3.4 Box plot for the object size of each test set. 51
Fig. 3.5 Mean IoU for the same object bounding boxes separated by different number of frames, and for all the training sets. 51
Fig. 3.6 Examples from the different video datasets evaluated with state-of-the-art solutions and our approach. 52
Fig. 3.7 mAP varing the number of short-term input frames without considering long-term information. 55

Fig. 4.1 FSVDet architecture overview, without the CSO module. 65
Fig. 4.2 Examples of real reference frames and their corresponding synthetic support frames generated following Alg. 4. The synthetic support frames might appear flipped as the network data augmentation includes horizontal flipping. 73
Fig. 4.3 Error type analysis for the single image baseline (DeFRCN), FSVDet and FSVDet without CSO (FSVDet*). For five random training sets we show the mean and standard deviation of the dAP [2] —difference in AP, the lower the better. 74
Fig. 4.4 Influence of K. Mean AP and standard deviation for each value of K. 75

Fig. A.1 License information of publication [15]. 95
Fig. A.2 License information of publication [13]. 96
Fig. A.3 License information of publication [16]. 97
Fig. A.4 License information of publication [14]. 98

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

List of Tables

Tab. 2.1 Influence of each component on the framework precision on ImageNet VID dataset. . . . 32
Tab. 2.2 VID validation set results. We use a number of input frames $N = 6$ to test our framework. . . . 33
Tab. 2.2 Continued from previous page. . . . 34
Tab. 2.3 USC-GRAD-STDdb results 36

Tab. 3.1 Long-term strategy, without short-term information. . . . 54
Tab. 3.2 Influence of each component on the framework precision on ImageNet VID dataset. . . . 56
Tab. 3.3 Results on ImageNet VID dataset. . . . 57
Tab. 3.4 Results on UAVDT dataset. . . . 57
Tab. 3.5 Results on Visdrone dataset. . . . 58
Tab. 3.6 Results on STDdb dataset. . . . 58
Tab. 3.7 Results on MOT20 dataset. . . . 58
Tab. 3.8 Binomial Sign test. p is the calculated p-value comparing each method with our proposal. . . . 58

Tab. 4.1 Different support frames generation strategies. Results for FSVDet without CSO. . . . 72
Tab. 4.2 Ablation Studies for FSVDet. . . . 73

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Tab. 4.3 Results on the few-shot video object detection FSVOD-500 dataset. The comparision includes single image object detectors, traditional video object detectors, multiple object trackers + meta learning classifier, and few-shot video object detectors. 76

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

A Reproduction Rights

This appendix describes licensing information, including reproduction rights, of the publications from which the content of this PhD thesis has been partially extracted.

A.1 Journals

> Daniel Cores, Víctor M. Brea, Manuel Mucientes. Spatiotemporal tubelet feature aggregation and object linking for small object detection in videos. Applied Intelligence, 2022. DOI: 10.1007/s10489-022-03529-w



Figure A.1: License information of publication [15].

- Quality indicators:
  Impact factor: 5.019 (JCR 2021).
  Journal ranked Q2 in JCR 2021.
  Category: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (46/144)
- PhD candidate contribution (CRediT): Writing - original draft, Conceptualization, Data curation, Formal analysis, Investigation, Validation.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Reproduction rights : the contents of Chapter 2 are extracted from this publication. It is licensed under the Creative Commons CC BY (Figure A.1).

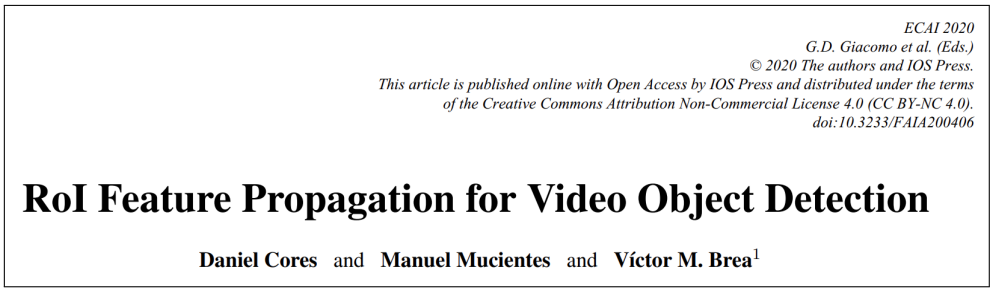
> Daniel Cores, Víctor M. Brea, Manuel Mucientes. Short-term anchor linking and long-term self-guided attention for video object detection. Image and Vision Computing, Vol. 110, 2021. DOI: 10.1016/j.imavis.2021.104179

- Quality indicators:
  Impact factor: 3.860 (JCR 2021).
  Journal ranked Q1 in JCR 2021.
  Category: COMPUTER SCIENCE, SOFTWARE ENGINEERING (22/110)
- PhD candidate contribution (CRediT): Writing - original draft, Conceptualization, Data curation, Formal analysis, Investigation, Validation.
- Reproduction rights: the contents of Chapter 2 are extracted from this publication. It is licensed under the Creative Commons CC BY (Figure A.2).



Figure A.2: License information of publication [13].

A.2 CONFERENCES



ECAI 2020
G.D. Giacomo et al. (Eds.)
© 2020 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms
of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA200406

RoI Feature Propagation for Video Object Detection

Daniel Cores and **Manuel Mucientes** and **Víctor M. Brea**[1]

Figure A.3: License information of publication [16].

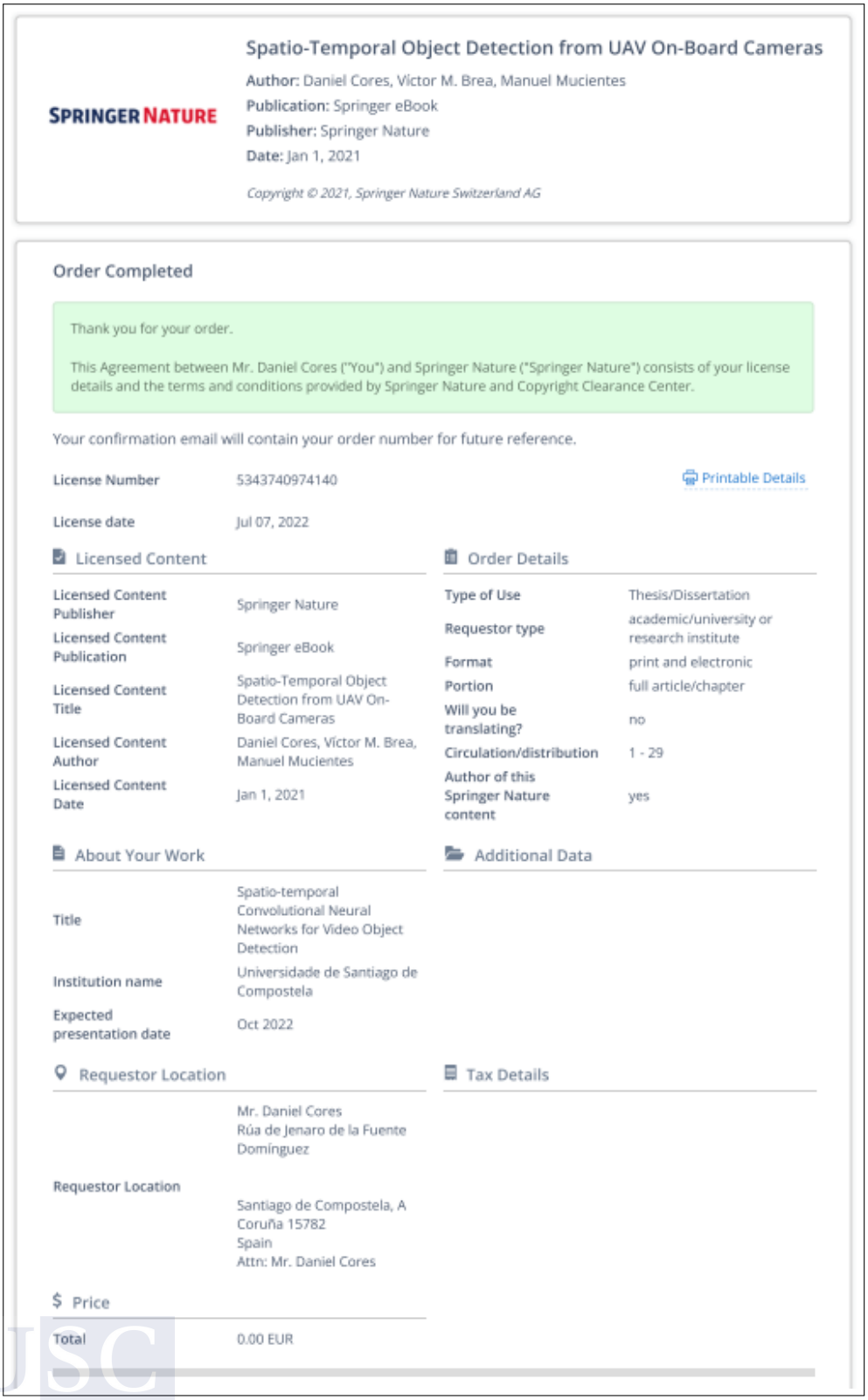
> Daniel Cores, Manuel Mucientes, Víctor M. Brea. RoI feature propagation for video object detection. *In Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, pp. 2680-2687, Santiago de Compostela (Spain), 2020. DOI: 10.3233/FAIA200406

- Quality indicators:
  Conference Ranking (GGS Rating): A-, Class 2.
- PhD candidate contribution (CRediT): Writing - original draft, Conceptualization, Data curation, Formal analysis, Investigation, Validation.
- Reproduction rights: the contents of Chapter 2 are extracted from this publication. It is licensed under the Creative Commons CC BY (Figure A.2).

> Daniel Cores, Víctor M. Brea, Manuel Mucientes. Spatio-Temporal Object Detection from UAV On-Board Cameras. *In Proceedings of the 19th International Conference on Computer Analysis of Images and Patterns (CAIP)*, pp. 143-152, Nicosia (Cyprus), 2021. DOI: 10.1007/978-3-030-89131-2_13

- Quality indicators:
  Conference Ranking (GGS Rating): B-, Class 3.
- PhD candidate contribution (CRediT): Writing - original draft, Conceptualization, Data curation, Formal analysis, Investigation, Validation.
- Reproduction rights: the contents of Chapter 3 are extracted from this publication. The permission granted by the publisher (Springer Nature) for the reproduction of this publication is shown in Figure A.4.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA



SPRINGER NATURE

Spatio-Temporal Object Detection from UAV On-Board Cameras

Author: Daniel Cores, Víctor M. Brea, Manuel Mucientes
Publication: Springer eBook
Publisher: Springer Nature
Date: Jan 1, 2021

Copyright © 2021, Springer Nature Switzerland AG

Order Completed

Thank you for your order.

This Agreement between Mr. Daniel Cores ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

License Number	5343740974140
License date	Jul 07, 2022

Printable Details

Licensed Content

Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Spatio-Temporal Object Detection from UAV On-Board Cameras
Licensed Content Author	Daniel Cores, Víctor M. Brea, Manuel Mucientes
Licensed Content Date	Jan 1, 2021

Order Details

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

About Your Work

Title	Spatio-temporal Convolutional Neural Networks for Video Object Detection
Institution name	Universidade de Santiago de Compostela
Expected presentation date	Oct 2022

Additional Data

Requestor Location

Requestor Location	Mr. Daniel Cores Rúa de Jenaro de la Fuente Domínguez Santiago de Compostela, A Coruña 15782 Spain Attn: Mr. Daniel Cores

Tax Details

Price

Total	0.00 EUR

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Figure A.4: License information of publication [14].

The detection precision in images has greatly improved with the use of Deep Learning techniques, especially with the adoption of Convolutional Neural Networks. However, object detection in videos presents new challenges such as motion blur, out-of-focus or object occlusions that make object detection more difficult. This thesis proposes new methods to exploit spatio-temporal information, establishing relations among detections from different frames and performing feature aggregation throughout time. This improves the detection precision in frames in which a single image object detector would not be able to provide the correct object category. This thesis also explores the utility of spatio-temporal information to reduce the number of training examples, keeping a competitive detection precision.