Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Aprendizaje basado en contexto para el control automático de calidad en segmentación de imágenes médicas

Tesis de Licenciatura en Ciencias de Datos

Matias Cosarinsky

Director: Enzo Ferrante

Buenos Aires, 2024

RESUMEN

El control de calidad en la segmentación automática de imágenes médicas es una tarea crucial que enfrenta grandes desafíos, principalmente debido a la frecuente ausencia de etiquetas de referencia (*ground truth*), lo cual dificulta una evaluación precisa de las predicciones.

En *Reverse Classification Accuracy* (RCA) [1], se propone una metodología que permite evaluar de forma automática la calidad de segmentaciones sin la necesidad de disponer de etiquetas de referencia. El enfoque que plantean consiste en utilizar la segmentación predicha de una nueva imagen para entrenar un clasificador inverso, el cual es evaluado en un conjunto de imágenes de referencia con *ground truth* disponible. La hipótesis es que, si la segmentación predicha es de buena calidad, el clasificador inverso debería tener un buen rendimiento en al menos algunas de las imágenes de referencia.

En este trabajo proponemos un enfoque que denominamos *In-Context RCA*, el cual expande el marco tradicional de RCA mediante el uso de modelos de segmentación basados en *in-context learning*, tales como UniverSeg [2] y SAM 2 [3], como clasificadores inversos. Este tipo de modelos permiten adaptar sus predicciones a partir de una cantidad limitada de ejemplos etiquetados, sin la necesidad de entrenamiento adicional, lo que los hace ideales en este contexto. Además, incorporamos técnicas de *retrieval augmentation* para seleccionar conjuntos de referencia más relevantes, lo que permite una ulterior mejora en las predicciones.

La metodología propuesta es evaluada sobre múltiples modalidades de imágenes médicas (Rayos-X, ultrasonido, tomografía computarizada, resonancia magnética, entre otras), ampliando el análisis de RCA realizado hasta el momento y demostrando a su vez la capacidad del método para adaptarse y ofrecer resultados robustos en distintos escenarios clínicos. Los resultados muestran que *In-Context RCA* es capaz de igualar e incluso superar el enfoque tradicional de RCA, siendo más eficiente computacionalmente. Esto lo convierte en una excelente opción para su integración en procesos de segmentación automática en la práctica clínica.

Palabras claves: Reverse classification accuracy (RCA), In-context learning, Segmentación automática de imágenes, Retrieval augmentation, Imágenes médicas, Control automático de calidad en segmentaciones, Evaluación en ausencia de GT.

ABSTRACT

Quality control in automatic segmentation of medical images is a critical task that presents significant challenges. This is mainly due to the frequent absence of ground truth, which makes accurate evaluation of predictions a difficult task.

Reverse Classification Accuracy (RCA) [1], presents a methodology that allows to automatically evaluate segmentation quality in the absence of ground truth. Their presented framework, consists in using the predicted segmentation of a new image to train a reverse classifier, which is then evaluated on a set of reference images with available ground truth. The hypothesis is that if the predicted segmentation is of good quality, then the reverse classifier should perform well on at least some of the reference images.

In this work, we propose a novel approach that we call *In-Context RCA*, which expands the traditional RCA framework by utilizing segmentation models based on in-context learning, such as UniverSeg [2] and SAM 2 [3], as reverse classifiers. These models can adapt their predictions based on a limited number of labeled examples without the need for additional training, making them ideal for this setting. Additionally, we incorporate retrieval augmentation techniques that allows to select more relevant reference sets, which further improves the predictions as a result.

The proposed methodology is evaluated across multiple different modalities in the field of medical imaging (X-rays, US, CT scan, MRI, among others), broadening the RCA analysis conducted thus far and demonstrating the method's ability to adapt and provide robust results in a wide variety of clinical scenarios. The results show that In-Context RCA matches or even surpasses the traditional RCA approach while being more computationally efficient. This makes it an excellent option for integration into automatic segmentation processes in clinical practice.

Keywords: Reverse classification accuracy (RCA), In-context learning, Automatic image segmentation, Retrieval augmentation, Medical imaging, Automatic quality control in segmentations, Evaluation in the absence of GT.

AGRADECIMIENTOS

Quiero expresar mi más profundo agradecimiento a mi familia, en especial a mis padres y a mi hermana, por su apoyo incondicional y la confianza que me brindaron a lo largo de estos años. Juntos, lograron crear un ambiente de motivación, apoyo y gratificación que me permitió llegar hasta este punto.

A mi mamá, que siempre mostró un interés genuino en mis estudios, preguntando constantemente por mis avances y esforzándose por entender mi trabajo con curiosidad, entusiasmo e ilusión.

A mi papá, le agradezco por impulsarme a seguir esta carrera y por estar presente en cada paso del camino.

A mi director, Enzo, le agradezco sinceramente por introducirme en el fascinante mundo de la investigación y por confiar en mí al incluirme en este proyecto. Su tiempo y dedicación fueron fundamentales para el desarrollo de este trabajo. A lo largo de este intenso período, aprendí enormemente bajo su orientación, lo que me ha inspirado a seguir explorando este campo.

No puedo dejar de mencionar a todos los profesores de la Facultad de Ciencias Exactas y Naturales, así como a la institución en sí misma, que proporcionan un ambiente lleno de potencial y oportunidades donde diariamente ocurren grandes cosas. Su dedicación y esfuerzo son una inspiración constante en mi camino académico.

Índice general

1. INTRODUCCIÓN

1.1. Motivación

La segmentación de imágenes médicas es una tarea fundamental en el análisis de imágenes. Esta consiste en identificar y delinear regiones de interés (ROI) tales como órganos, lesiones y tejidos en diferentes tipos de imágenes médicas. La generación de segmentaciones precisas resulta esencial para una amplia gama de aplicaciones clínicas, entre ellas el diagnóstico de enfermedades, la planificación de tratamientos y el monitoreo de la progresión de pacientes, etc.
Tradicionalmente, profesionales médicos solían realizar este tipo de tareas de forma manual. Sin embargo, estos procedimientos frecuentemente demandan mucho tiempo y requieren un gran grado de conocimiento especializado. Gracias a los avances tecnológicos de los últimos años y la rápida adopción de redes neuronales en el campo, métodos de segmentación automática comenzaron a ser ampliamente utilizados en entornos clínicos, facilitando y agilizando ampliamente la labor de los profesionales de salud.

De todas formas, es importante contar con métodos capaces de evaluar este tipo de modelos. Enfoques tradicionales, suelen emplear evaluación supervisada [4, 5], lo cual requiere el uso de bases de datos anotadas para medir el rendimiento, y se acostumbra estudiar diversas métricas en un contexto de validación cruzada. Estas métricas por lo general reflejan el nivel de coincidencia entre la segmentación generada por un modelo y una referencia conocida, denominada *ground truth*(GT), que fue anotada por un experto. Comúnmente estas medidas o bien evalúan el solapamiento entre la predicción y la referencia o buscan cuantificar la disimilaridad entre ambas.

Sin embargo, una vez que el modelo es desplegado en producción, resulta muy difícil prever con exactitud cuál será su desempeño real ante nuevos datos, principalmente porque en el ámbito clínico es complicado obtener etiquetas de *ground truth* contra las cuales poder realizar comparaciones. En la práctica, el rendimiento puede diferir significativamente de lo esperado, siendo a menudo inferior a la estimación obtenida mediante validación cruzada. Esto frecuentemente se debe a problemas de *overfitting*, donde el modelo ha sido ajustado demasiado a los datos del conjunto de entrenamiento, pero también a variaciones en los protocolos de adquisición de imágenes entre el entrenamiento y el despliegue del modelo, así como a diferencias inherentes entre la demografía de los pacientes, hospitales y equipos de imágenes utilizados. Por ende, resulta crucial contar con métodos capaces de evaluar con precisión el rendimiento de los modelos en producción. En particular, es esencial poder detectar fallos en la segmentación automática, ya que estos errores pueden comprometer la calidad de la atención médica y tener implicaciones significativas en las decisiones clínicas, afectando la salud de los pacientes.

El enfoque propuesto en este trabajo se basa en una técnica denominada RCA [1], que ofrece una solución a este problema al permitir determinar de forma automática la calidad de las segmentaciones de imágenes médicas a nivel individual, prescindiendo de datos etiquetados para realizar evaluaciones.

1.2. Trabajos previos

Diversos trabajos previos han propuesto métodos para evaluar la calidad de segmentaciones médicas en ausencia de etiquetas. Sin embargo, la mayoría de ellos requieren analizar y extraer características geométricas [6], lo cual no siempre garantiza buenos resultados en tareas más complejas, como la segmentación de imágenes médicas. Nuestro trabajo se basa fuertemente en el framework propuesto en RCA [1]. Una idea inspirada en *reverse validation* [7] y *reverse testing* [8], los cuales proponen entrenar un nuevo modelo denominado clasificador inverso sobre los datos de test mediante las predicciones realizadas por el modelo, para luego evaluarlo en el conjunto de entrenamiento. La diferencia clave es que RCA entrena un clasificador inverso para cada instancia individual, mientras que [7] y [8] entrenan clasificadores únicos sobre el conjunto de test y sus predicciones de manera conjunta para identificar cuál es el mejor predictor original. Gracias a este cambio RCA tiene la ventaja de permitir predecir la calidad de segmentaciones en instancias individuales.

Por otra parte, recientemente modelos como SAM 2 [3] y UniverSeg [2] han demostrado gran capacidad para segmentar imágenes médicas utilizando un enfoque denominado *in-context learning* [9]. Este concepto evita la necesidad de tener que entrenar un modelo desde cero, ahorrando grandes costos, y permite realizar segmentaciones a partir de unos pocos ejemplos etiquetados en lo que se conoce como *few-shot segmentation* (FSS). Motivados por el surgimiento y éxito de estos modelos, en este trabajo proponemos expandir el framework de RCA en lo que llamamos *In-Context RCA*. Incorporamos el uso de modelos capaces de realizar FSS como clasificadores inversos, junto con la implementación de técnicas de *retrieval-augmentation* (RAG) [10], que nos permiten mejorar la calidad de las segmentaciones aprovechando bases de datos de referencia. Además, ampliamos el análisis realizado en el trabajo original, probando el funcionamiento del método sobre una mayor variedad de conjuntos de datos que comprenden distintas modalidades y tareas de segmentación de imágenes médicas. Esto nos permite evaluar el desempeño del método en un contexto más amplio y realista, demostrando su robustez y capacidad de adaptación a nuevos dominios.

1.3. Estructura de la tesis

Las secciones restantes de este trabajo se organizan de la siguiente manera:

En el Capítulo 2 se introduce el marco teórico del trabajo, presentando los conceptos fundamentales del área de imágenes médicas. Se describen las distintas modalidades de imágenes, así como los diferentes métodos de segmentación, incluyendo métodos clásicos, basados en deep learning y por último basados en in-context learning con especial foco en UniverSeg [2] y SAM 2 [3] que son los que utilizaremos más adelante.

En el Capitulo 3 se detalla el método propuesto para evaluar la calidad de la segmentación en ausencia de etiquetas de GT. Se explica en detalle el funcionamiento clásico de RCA, así como los cambios que proponemos en *In-Context RCA*. También se describen los conjuntos de datos y métricas que utilizamos para la evaluación.

En el Capitulo 4 se presentan los resultados obtenidos mediante el método propuesto sobre los distintos conjuntos de datos. Se muestra el resultado del proceso de generación

de segmentaciones a evaluar y luego se realiza un análisis comparativo entre las distintas arquitecturas utilizadas para su evaluación, incluyendo UniverSeg, SAM 2 y Single-Atlas, analizando sus ventajas y limitaciones. Además, en el Apéndice se incluyen figuras adicionales que complementan los resultados presentados.

Por último en el Capítulo 5 se resumen las principales conclusiones del trabajo, se discuten ciertas limitaciones del método y se proponen futuras líneas de investigación.

2. MARCO TEÓRICO

2.1. Imágenes médicas

Las imágenes médicas son una herramienta esencial en la medicina moderna, ya que proporcionan una visualización detallada de las estructuras internas del cuerpo humano, asistiendo a los profesionales de la salud en tareas tales como el diagnóstico de enfermedades, la planificación tratamientos y el monitoreo de la evolución de sus pacientes. Con el avance de la tecnología, los algoritmos de procesamiento de imágenes y los métodos computacionales han mejorado significativamente el análisis y la interpretación de este tipo de datos, aumentando su impacto en la práctica e investigación médica.

En este trabajo, utilizamos distintos tipos de imágenes médicas, cada una con características particulares y aplicaciones clínicas específicas que describimos brevemente a continuación:

- **Rayos X (X-Ray)**: Las radiografías son una de las técnicas de imagen médica más antiguas y ampliamente utilizadas. Producen imágenes bidimensionales mediante la absorción diferencial de rayos X por los tejidos del cuerpo, permitiendo visualizar huesos, pulmones y otras estructuras. En su formato digital, las imágenes de rayos X suelen ser almacenadas en escala de grises, donde la intensidad del píxel corresponde al nivel de absorción de los rayos X.

- **Ultrasonido (US)**: El ultrasonido utiliza ondas sonoras de alta frecuencia para crear imágenes del interior del cuerpo [11]. Es particularmente útil para visualizar órganos blandos y durante el monitoreo fetal [12]. Las imágenes de ultrasonido también se presentan en formato bidimensional y, al igual que las radiografías, se representan en escala de grises, con variaciones de intensidad que reflejan las propiedades acústicas de los tejidos.

- **Tomografía Computarizada (CT)**: La tomografía computarizada utiliza rayos X para generar imágenes tridimensionales de las estructuras internas del cuerpo [13]. A través de múltiples cortes axiales, el escáner CT proporciona una representación detallada en forma de volumen, donde cada voxel (volumen-píxel) tiene una intensidad, normalmente medida en unidades Hounsfield (HU), que corresponde a la densidad del tejido en esa ubicación [14]. Esto permite una visualización más precisa de estructuras óseas y algunos órganos internos.

- **Resonancia Magnética (MRI)**: La resonancia magnética es una técnica no invasiva que utiliza campos magnéticos y ondas de radio para generar imágenes detalladas de los tejidos blandos del cuerpo [15]. A diferencia de la tomografía computada, la MRI no utiliza radiación ionizante, lo que la convierte en una opción segura para estudios repetidos. Las imágenes de MRI pueden ser bidimensionales o tridimensionales, y son especialmente útiles para el diagnóstico de afecciones neurológicas y musculoesqueléticas.

- **Histopatología**: La histopatología es una técnica que permite estudiar los tejidos a nivel microscópico para detectar enfermedades, como el cáncer [16]. Las imágenes

de histopatología se obtienen mediante microscopios ópticos o electrónicos, y suelen estar teñidas con colorantes especiales para resaltar diferentes estructuras celulares. Estas imágenes juegan un papel clave en el diagnóstico de tumores y otras enfermedades, permitiendo a los patólogos examinar la morfología celular y el entorno tisular. Las imágenes digitales de histopatología suelen ser de muy alta resolución, lo que permite el análisis detallado de características como el tamaño, forma y organización de las células y los tejidos.

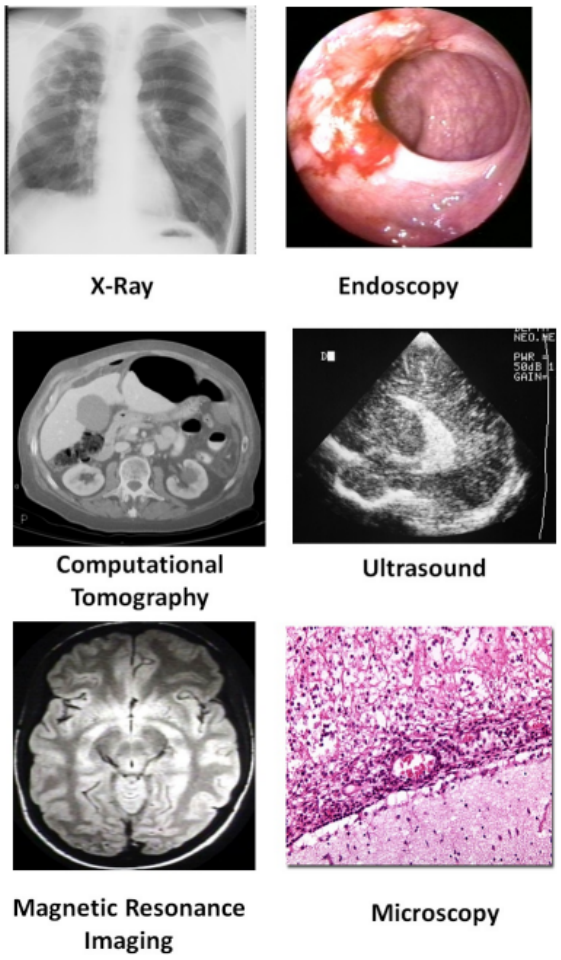

Fig. 2.1: Ejemplo de distintas modalidades de imágenes médicas. Imagen tomada de [17].

Cada uno de estos tipos de imágenes tiene un formato específico: las imágenes de Rayos X, Ultrasonido y de histopatología se almacenan comúnmente en 2D, mientras que las imágenes de CT y MRI suelen estar en formato 3D. Estas imágenes pueden presentarse en distintos formatos, como RGB, escala de grises o, en el caso de las imágenes de CT, en unidades Hounsfield (HU). Sin embargo, para asegurar la simplicidad y compatibilidad entre los diferentes modelos de segmentación a utilizar, en nuestro caso trabajaremos exclusivamente con imágenes en escala de grises, lo que facilita la integración de datos provenientes de distintas modalidades. En particular, en este trabajo se utilizarán: imágenes de Rayos X para segmentar estructuras como el corazón y los pulmones; imágenes de CT para segmentar tejido del hígado; imágenes de histopatología para detectar núcleos

de células cancerígenas en casos de cáncer de mama; imágenes de ultrasonido para la segmentación de cabezas fetales y estructuras cardíacas en ecocardiografías e imágenes de MRI para el análisis del ventrículo izquierdo. Los datasets utilizados, serán presentados y detallados en la sección 3.3.

2.2. Métodos de segmentación de imágenes

La segmentación de imágenes es una tarea fundamental en el área del procesamiento de imágenes y visión por computadora, que busca identificar y analizar áreas específicas dentro de una imagen, conocidas como regiones de interés (ROI, por sus siglas en inglés). Esta tarea se puede formular como un problema de clasificación de píxeles con etiquetas semánticas (segmentación semántica) o como una partición de objetos individuales (segmentación de instancia) [18]. Las etiquetas permiten distinguir entre las regiones de interés y el fondo o *background*, que corresponde a las áreas de la imagen a ser ignoradas durante el análisis.

La segmentación semántica realiza un etiquetado a nivel de píxel, asignando una categoría de objeto (por ejemplo, persona, coche, árbol, cielo) a cada píxel de la imagen. Este enfoque es generalmente más complejo que la clasificación de imágenes, que predice una sola etiqueta para toda la imagen.

Por otro lado, la segmentación de instancia amplía el alcance de la segmentación semántica al detectar y delinear cada objeto de interés en la imagen, permitiendo la partición de objetos individuales, como la separación de diferentes personas en una escena. En este contexto, también existe la segmentación panóptica, muy utilizada en tareas de reconocimiento de escenas, que combina la segmentación semántica y de instancia para proporcionar una comprensión más completa de la imagen, asignando etiquetas a todos los píxeles y delimitando las instancias de objetos.

La segmentación puede enfocarse tanto en distinguir una sola clase, donde solo se identifica una categoría específica de interés, como también múltiples clases, donde se abordan diversas categorías simultáneamente. Esta flexibilidad permite adaptar las técnicas de segmentación a una amplia variedad de aplicaciones y contextos, tales como la comprensión de escenas, la percepción robótica, la vigilancia por video, la realidad aumentada, la compresión de imágenes, sistemas automáticos de control de tráfico y el análisis de imágenes médicas, entre otros. En particular, dentro de la medicina puede ser utilizada en diferentes áreas como el diagnóstico, la localización de patologías, el estudio de estructuras anatómicas, la planificación de tratamientos, en cirugías asistidas por computadoras, etc. [19].

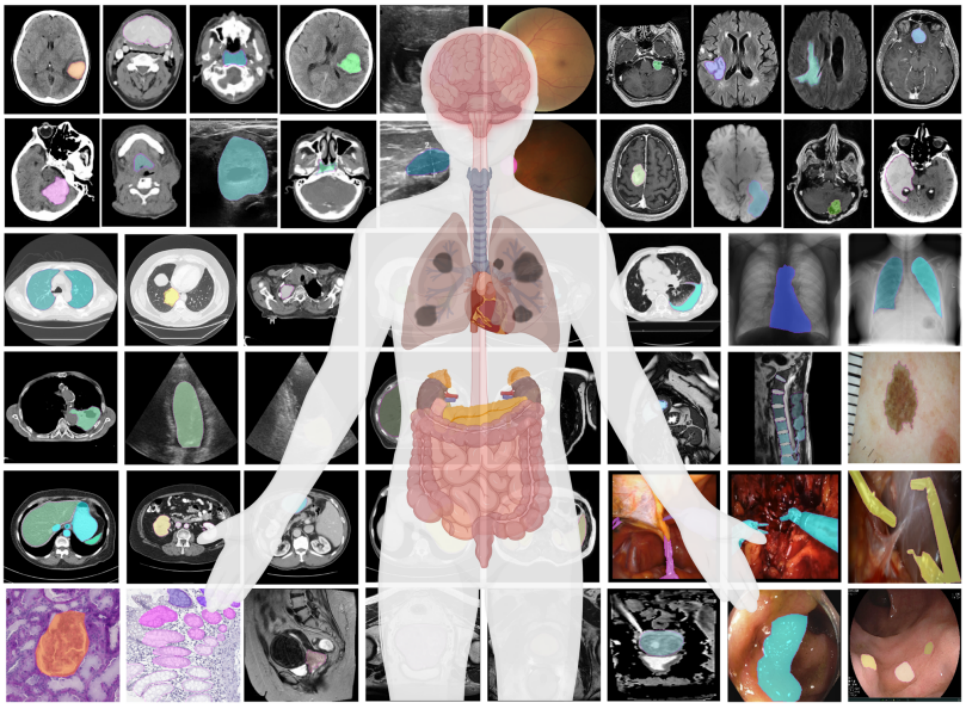

Fig. 2.2: Ejemplo de segmentaciones de distintas estructuras anatómicas. Imagen tomada de [20].

A continuación, veremos los métodos más utilizados en la segmentación de imágenes, abarcando desde enfoques clásicos hasta métodos basados en aprendizaje supervisado y técnicas de in-context learning, haciendo un especial énfasis en aquellos que son mayormente aplicados en el ámbito de imágenes médicas.

2.2.1. Métodos clásicos

Entre los métodos empleados clásicamente en la segmentación de imágenes [19, 21], distinguimos principalmente los siguientes

- **Umbralamiento**: los métodos de umbralamiento (*thresholding*) se encuentran entre los más simples y rápidos. Se basan en la suposición de que las imágenes están formadas por regiones con diferentes niveles de gris que se dividen a partir de un umbral. Este, subdivide las intensidades en dos partes: el "primer plano", con píxeles que tienen intensidades mayores o iguales al umbral, y el "fondo", con píxeles que tienen intensidades menores. Sin embargo, este tipo de métodos pueden ser muy sensibles al ruido y a la inhomogeneidad de intensidades, especialmente en imágenes clínicas.

- **Crecimiento de regiones**: estos métodos requieren la inicialización de uno o varios puntos semilla y se basan en la homogeneidad de intensidad de los píxeles vecinos para segmentar una región de la imagen. Utilizan un criterio de uniformidad para expandir la región a partir de los puntos iniciales, incorporando píxeles adyacentes que cumplen con este criterio [22]. Una gran desventaja que suelen presentar, es que el resultado depende significativamente de la elección de los puntos semilla.

- **Segmentación por bordes**: estos métodos buscan identificar transiciones bruscas en la intensidad de las imágenes. Para ello, analizan las variaciones en el gradiente de la intensidad combinado con la selección de umbrales.
- **Clustering**: los métodos de clustering dividen el espacio de características de los píxeles en grupos, donde los píxeles dentro de un grupo son más similares entre sí que a los de otros grupos. Si bien suelen ser eficientes, no tienen en cuenta las características espaciales, lo que los hace sensibles al ruido y a la variabilidad en las intensidades [23].
- **Clasificación**: este enfoque utiliza algoritmos de clasificación para asignar etiquetas a los píxeles de una imagen. Se entrenan modelos supervisados basados en características extraídas de los píxeles [24]. Al igual que los métodos de clustering, tienen la desventaja de que no consideran información espacial.

Otro enfoque relevante son los métodos de segmentación mediante registración, comúnmente conocidos como segmentación multi-atlas (MAS por sus siglas en inglés) [25, 26]. Este tipo de métodos son especialmente populares en el campo de las imágenes médicas [27]. Lo que propone la segmentación multi-atlas, es tratar la segmentación como un problema de registro de imágenes, buscando establecer una correspondencia espacial entre un atlas (una imagen de referencia con etiquetas ya segmentadas) y una nueva imagen. El proceso implica deformar una de las imágenes hasta que sea similar a la otra. A través de este mapeo entre los dos sistemas de coordenadas, se pueden transferir o "propagar"las etiquetas de segmentación del atlas a la nueva imagen. Este tipo de tarea, es computacionalmente costosa debido a los complejos cálculos de deformación necesarios. Aunque el uso de un solo atlas puede ser eficaz, generalmente no es suficiente para capturar toda la variabilidad anatómica presente en las imágenes. Por ello, se suelen usar múltiples atlas y luego fusionar las etiquetas resultantes, comúnmente mediante voto mayoritario. En este trabajo utilizaremos el enfoque de single-atlas, en el cual se emplea una única imagen de referencia. Este método ha demostrado grandes resultados en RCA [1], aunque presenta un alto costo computacional que puede resultar una gran limitación en la práctica.

2.2.2. Métodos basados en aprendizaje supervisado

El aprendizaje supervisado, un enfoque en el que un modelo aprende a partir de un conjunto de datos etiquetados, ha revolucionado el campo de la visión por computadora, proporcionando un marco poderoso para abordar tareas complejas como la clasificación de imágenes, la detección de objetos y la segmentación semántica. Con el creciente avance del *deep learning* [28], que permite el uso de redes neuronales profundas capaces de aprender representaciones jerárquicas, estos enfoques han demostrado ser particularmente efectivos en aplicaciones complejas como la segmentación de imágenes médicas [29]. A continuación, describiremos algunas de las arquitecturas más relevantes basadas en el uso de redes neuronales, con un enfoque particular en aquellas utilizadas para la segmentación de imágenes médicas.

- **CNNs**:
Las redes neuronales convolucionales fueron propuestas inicialmente por Fukushima [30] y más tarde popularizadas en el área de visión por computadora por Yann LeCun con su

arquitectura LeNet [31], diseñada para el reconocimiento de dígitos manuscritos. Aunque su adopción inicial fue limitada debido a las restricciones computacionales de la época, la acelerada evolución del hardware, especialmente de las GPUs, permitió su uso más extendido. Este proceso se vio acelerado particularmente a partir del 2012 cuando la arquitectura AlexNet [32], desarrollada por Alex Krizhevsky, gana el concurso ImageNet. Las CNNs están compuestas principalmente por tres tipos de capas:

I. **capas convolucionales**: en estas capas se aplica un filtro (o kernel) de pesos que se desplaza a través de la imagen para extraer características locales. Este proceso de convolución permite detectar patrones y características relevantes, como bordes y texturas, en las imágenes;
II. **capas no lineales**: aplican una función de activación a los mapas de características (o *feature-maps*) resultantes de las capas convolucionales. La activación se realiza de manera elemento a elemento, lo que permite que la red modele funciones no lineales, esencial para el aprendizaje de representaciones complejas
III. **capas de pooling**: en estas capas un vecindario pequeño de un mapa de características se reemplaza por información estadística, como el valor máximo o el promedio del vecindario. Esto reduce la resolución espacial de la representación y ayuda a mantener la invarianza a pequeñas transformaciones en las imágenes.

En las CNNs, las neuronas están conectadas localmente, recibiendo entradas de un pequeño campo receptivo en la capa anterior. A medida que se apilan más capas, los campos receptivos se amplían, permitiendo que las capas superiores capten características más complejas. El uso de *weight sharing* (pesos compartidos) en las capas convolucionales, junto con las capas de *pooling* que reducen la resolución espacial, hace que las CNNs sean altamente escalables y más eficientes en términos de parámetros y entrenamiento en comparación con redes *fully-connected*.

- **Modelos encoder-decoder**:
Los modelos encoder-decoder son una familia de redes neuronales que aprenden a mapear datos de un dominio de entrada a un dominio de salida mediante una arquitectura compuesta por dos etapas: el encoder, que se encarga de comprimir la entrada en una representación latente $z = f(x)$ y el decoder $y = g(z)$, que busca predecir la salida a partir de la representación latente que busca capturar la información semántica necesaria para realizar la predicción.
Estos modelos, suelen ser entrenados minimizando la pérdida en la reconstrucción $L(y, \hat{y})$, que mide la diferencia entre la reconstrucción $\hat{y} = g(z)$ y la salida verdadera y.

La mayor parte de los modelos utilizados para la segmentación de imágenes médicas basados en deep leraning, utilizan esta arquitectura.

- **U-Net**:
Es una de las arquitecturas más destacadas en el ámbito de la segmentación de imágenes médicas, fue inicialmente propuesta por Ronneberger et al. [33] para segmentar imágenes de microscopía biológica.
La arquitectura sigue un enfoque encoder-decoder y combina capas de convolución. Se compone de dos partes principales: una de contracción y una de expansión. Ambas partes

se combinan conformando una estructura simétrica con un cuello de botella en el medio, lo que caracteriza su forma de U. En la Figura 2.3, podemos visualizar la arquitectura como fue originalmente propuesta en [33].

En el camino de contracción, se aplican sucesivas capas de convolución, seguidas por una unidad lineal rectificada (ReLU) y una operación de *max-pooling* para reducir la dimensionalidad (*downsampling*). Además, en cada paso de downsampling se duplican la cantidad de capas convolucionales.

El camino de expansión consiste en una sucesión de capas de *upsampling* seguidas de una operación de deconvolución (o *up-convolution*) que reduce a la mitad la cantidad de *feature maps* mientras aumenta sus dimensiones. La última capa utiliza una convolución de 1x1 para procesar los *feature maps* finales, generando un mapa de segmentación que categoriza cada uno de los píxeles de la imagen de entrada según la cantidad de clases deseadas.

Una característica distintiva de la U-Net, es el uso de conexiones residuales o *skip-connections*, donde los mapas de características de la parte de contracción se copian a su par correspondiente en la parte de expansión. Esto evita la pérdida de información y permite combinar características de bajo y alto nivel, lo cual es crucial en el contexto de imágenes médicas.

Desde su introducción, la arquitectura U-Net se ha transformado en un modelo estándar para la segmentación de imágenes médicas y también ha ganado popularidad en otras áreas. Diversas mejoras han sido propuestas, entre ellas se destacan UNet++ [34] y 3D U-Net junto a V-Net [35, 36] que permiten trabajar con imágenes 3D, muy frecuentes en medicina.

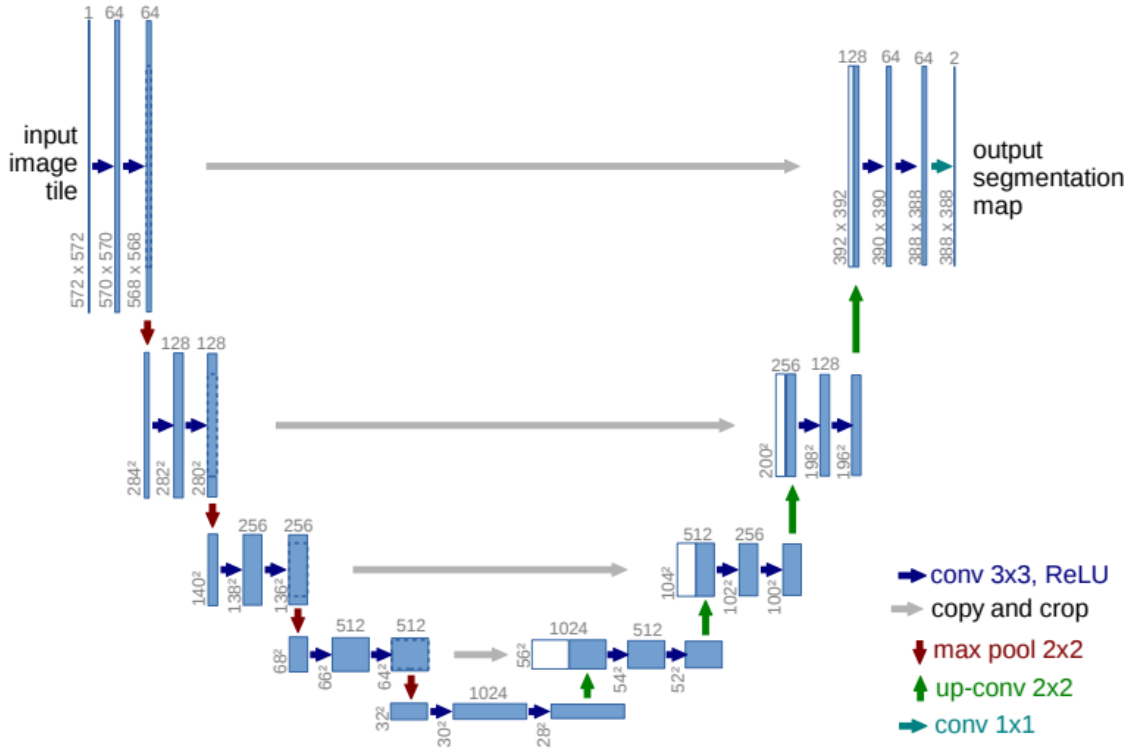

Fig. 2.3: Arquitectura de U-Net (ejemplo para 32x32 píxeles en la resolución más baja). Cada recuadro azul corresponde a un mapa de características multi-canal. El número de canales se indica en la parte superior de cada recuadro, mientras que el tamaño en el la esquina inferior izquierda. Los recuadros blancos representan mapas de características copiados. Las flechas indican las diferentes operaciones realizadas. Fuente: Ronneberger et al. [33].

2.2.3. Métodos basados en in-context learning

El aprendizaje basado en contexto o *in-context learning*, introducido inicialmente en el ámbito de modelos de lenguaje por Brown et al. [9], es un paradigma que denota la capacidad de un modelo de adaptarse nuevas tareas simplemente a partir de una cantidad limitada de ejemplos de entrada, sin la necesidad de modificar los pesos del modelo. En lugar de ajustar sus parámetros, la información proporcionada en los ejemplos (el contexto) condiciona al modelo para que sea capaz de inferir sobre nuevas entradas. Este enfoque, en donde se adapta un modelo en base a una cantidad limitada de ejemplos etiquetados, es también muy comúnmente conocido como *few-shot learning*. Sin embargo, a diferencia del aprendizaje basado en contexto, este último admite la modificación de parámetros del modelo para adaptarlo a la nueva tarea [37].

En el ámbito de visión por computadora, las redes prototípicas, introducidas por Snell et al. [38], representan un enfoque tradicional de *few-shot learning* (FSL). Este método se basa en aprender un prototipo (o representación promedio) para cada clase en un espacio latente conocido como *embedding*. Este espacio, consiste en una representación vectorial que captura las características más relevantes de los datos. Dada una nueva instancia, podemos calcular la distancia entre ella y los prototipos de cada clase para luego asignarle la clase cuyo prototipo sea más cercano, facilitando así la clasificación con un número

limitado de muestras.
Originalmente propuestas para tareas de clasificación, las redes prototípicas fueron rápidamente extendidas al campo de la segmentación de imágenes [39, 40], en lo que se conoce como *few-shot segmentation* (FSS). Modelos como ALPNet (Adaptive Local Prototypes) [41, 42] permiten una ulterior adaptación al ámbito de imágenes médicas.
Más recientemente surgen otros modelos como UniverSeg [2] y SAM 2 [3] que permiten la segmentación de imágenes médicas en el contexto de FSS. Estos dos modelos resultan de particular interés, ya que fueron utilizados en este trabajo en el contexto de RCA. A continuación, pasamos a explicar su arquitectura.

UniverSeg:
Este modelo, introducido en Butoi et al. [2], propone un método que permite la segmentación de imágenes médicas en tareas no vistas durante el entrenamiento sin ajustes adicionales en el modelo ni la necesidad de entrenamiento complementario específico en la nueva tarea. Dada una imagen como query, su propuesta permite segmentarla en base a un conjunto soporte compuesto de pares imágen-etiqueta.
En la Figura 2.4 podemos ver como a diferencia de métodos tradicionales que requieren ajustes o entrenamientos adicionales para cada nueva tarea, UniverSeg es capaz de generalizar de manera efectiva a tareas no vistas previamente, optimizando así el proceso de segmentación y facilitando su aplicación en entornos clínicos y de investigación.

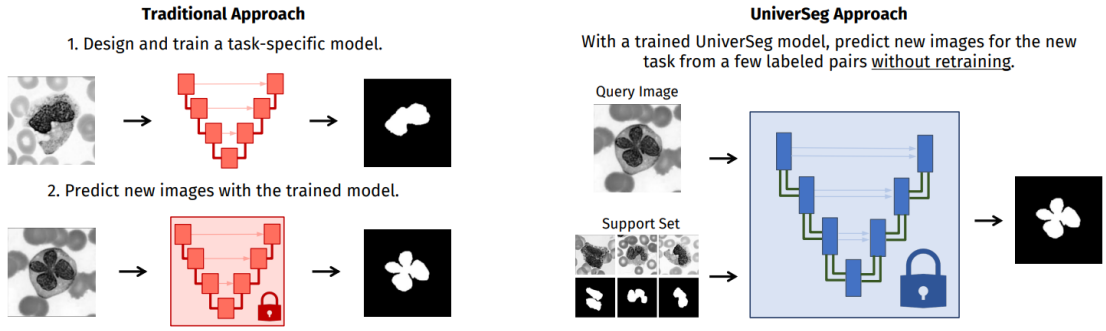

Fig. 2.4: **Flujo de trabajo para la inferencia en una nueva tarea, a partir de un conjunto de datos no visto**. Dada una nueva tarea, los modelos tradicionales (**izquierda**) son entrenados antes de realizar predicciones. UniverSeg (**derecha**) emplea un *único* modelo entrenado capaz de realizar predicciones para imágenes (queries) de la nueva tarea con unos pocos ejemplos etiquetados como entrada (conjunto soporte), sin necesidad de ajuste adicional, siguiendo el enfoque de *in-context learning*. Fuente: Butoi et al. [2].

Formalmente, el método se puede definir de la siguiente manera: data una tarea de segmentación t compuesta de pares imágen-etiqueta $\{(x_i^t, y_i^t)\}_{i=1}^N$, el objetivo es aprender una función universal $\hat{y} = f_\theta(x^t, S^t)$ capaz de predecir una segmentación para la imagen de entrada x^t (query) en la tarea t en base a un conjunto soporte $S^t = \{x_j^t, y_j^t)\}_{j=1}^n$, el cual comprende una cantidad limitada de datos etiquetados sobre la tarea t. Para ello, se sigue el enfoque de aprendizaje basado en contexto, donde el modelo es capaz de adaptarse a la tarea no vista t en base a un conjunto de ejemplos S^t que le proporcionan el contexto necesario. La calidad del resultado, va a ser dependiente del conjunto soporte utilizado.

Para la implementación de f_θ, la arquitectura que introducen sigue el enfoque encoder-decoder y utiliza conexiones residuales, imitando el enfoque de una U-Net [33]. La principal innovación se encuentra en la incorporación de *Cross-Blocks*. Estos son módulos que permiten la interacción entre la imagen de consulta (*query*) y el conjunto soporte mediante una capa de convolución cruzada (*cross-convolution*). Esta capa combina el mapa de características de la imagen de consulta con los mapas de características del conjunto soporte, generando así nuevas representaciones que se propagan a través de las sucesivas capas del modelo, facilitando de esta forma la transferencia de información entre ambas fuentes.
Cada nivel en el camino del codificador consiste en un *CrossBlock* seguido de una operación de reducción espacial (*down-sampling*) tanto de las representaciones de la imagen de consulta como del conjunto soporte.
Por su parte, cada nivel en el camino expansivo consiste en aumentar las representaciones (*up-sampling*) duplicando sus resoluciones espaciales y concatenándolas con la representación de igual tamaño en el camino de codificación, seguida de un *CrossBlock*.
En la Figura 2.5 podemos ver una ilustración detallada de la arquitectura.

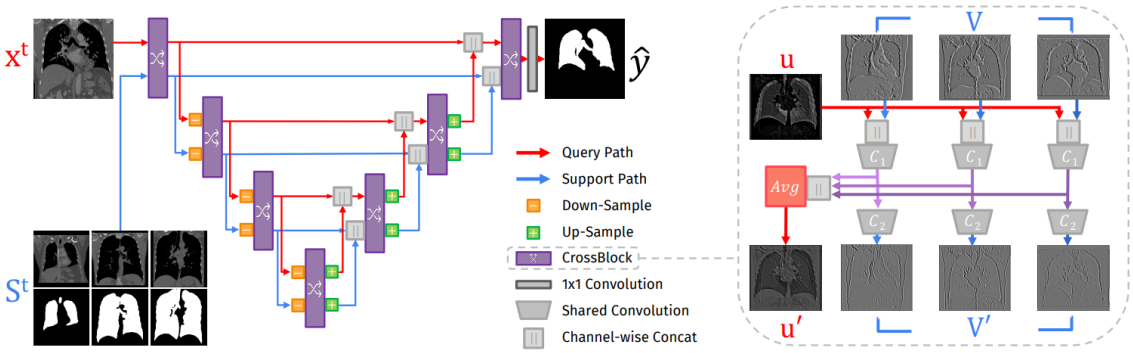

Fig. 2.5: Arquitectura del modelo. La red de UniverSeg (**izquierda**) toma como entrada una imagen a segmentar (query) X^t y un conjunto soporte S^t y los combina en un *CrossBlock* multi-escala para generar la segmentación. Un *CrossBlock* (**derecha**) toma como entrada representaciones de la query u y un conjunto soporte $V = \{v_i\}$. Cada entrada v_i interactúa con u produciendo u' y V'. Fuente: Butoi et al. [2].

Para el entrenamiento del modelo, se compiló un gran conjunto de datos, denominado *MegaMedical*, que incluye una amplia variedad de datos de segmentación médica, abarcando diferentes anatomías y modalidades de imagen. Además, se aplicaron diversas técnicas de aumentación de datos (*data-augmentation*) para incrementar tanto la cantidad de tareas como de datos de entrenamiento. Esto, en combinación con la diversidad de datos utilizados, contribuye a evitar el sobreajuste y mejora la robustez del modelo, aumentando en consecuencia su capacidad de generalización hacia nuevas tareas.

SAM 2:
Este modelo es una extensión de SAM [43], un modelo fundacional reconocido por su gran capacidad de segmentar imágenes a partir de un *prompt* inicial proporcionado por el usuario. Ambos modelos fueron entrenados con millones de imágenes de dominio general, lo que les provee un gran grado de generalización, permitiéndoles realizar segmentaciones precisas en una amplia variedad de contextos.
SAM 2 [3] introduce mejoras significativas respecto a su predecesor, como la capacidad de segmentar videos y aceptar máscaras de segmentación como *prompts*, funcionalidades que

no estaban presentes en su anterior versión. El cambio principal, se encuentra en la introducción de un banco de memoria. En este, se almacenan las interacciones con el objeto o región de interés a lo largo de los diversos fotogramas del video. Además, cuenta con un módulo de atención [44] que condiciona el resultado de la segmentación del fotograma actual en base a las memorias guardadas en el banco. SAM 2 también puede ser aplicado para segmentar imágenes, ya que estas son videos de un único fotograma. En estos casos la memoria se encuentra vacía y el framework es idéntico al de SAM.

A continuación, pasamos a describir las principales componentes del modelo:

- **Image encoder**: utilizado para obtener *feature embeddings* representativos de las imágenes, consiste en un *Vision Transformer* (ViT) [45] pre-entrenado que se basa en un *masked autoencoder*(MAE) [46] para generar representaciones robustas.
- **Memory attention**: este módulo es utilizado para condicionar los features del fotograma actual (y su posible prompt) en base a los de fotogramas anteriores y sus respectivas predicciones. Consiste en bloques transformers apilados con *self-attention* (para el fotograma actual) seguidos de *cross-attention*, lo cual permite combinarlo con la memoria de los fotogramas precedentes.
- **Prompt encoder y mask decoder**: el *prompt encoder* permite utilizar diferentes tipos de *prompts* como clicks (positivos o negativos), cajas delimitadoras o máscaras para definir el objeto de interés en el fotograma actual. Los *prompts* dispersos se representan mediante *positional encodings* que se suman a embeddings aprendidos, mientras que las máscaras mediante convoluciones.
  El *mask decoder* emplea bloques *transformer* que actualizan los embeddings tanto del *prompt* dado por el usuario como del fotograma, y es capaz de predecir múltiples máscaras en casos de ambigüedad. También incluye una cabeza de atención que se encarga de determinar si el objeto de interés se encuentra presente en el fotograma actual.
- **Memory encoder**: se encarga de generar una memoria, reduciendo la resolución de la máscara de salida mediante convoluciones y combinándola con el embedding del fotograma (sin condicionar) del *image encoder*. Posteriormente, capas convolucionales fusionan esta información para generar la memoria final del fotograma.
- **Memory bank**: se encarga de retener información sobre las predicciones anteriores del objeto en el video, manteniendo una cola FIFO con los fotogramas más recientes y sus características espaciales.

Si bien estos modelos no fueron diseñados específicamente para la segmentación de imágenes médicas, sus grandes capacidades han impulsado el desarrollo de propuestas para adaptarlos a este dominio. Entre ellas, se destacan MedSAM [20] y Medical SAM 2 [47], que se basan en SAM y SAM 2 respectivamente. Estos modelos ajustan los pesos de la versión original mediante *fine-tuning* para adaptarlo al entorno médico, un proceso que puede resultar en una limitación en muchos casos. No obstante, SAM 2 ha demostrado una mayor capacidad de generalización respecto a su predecesor. Esto ha dado lugar a enfoques como FS-MedSAM2 [48], que propone su uso en la segmentación de imágenes médicas siguiendo el paradigma *few-shot*, sin necesidad de realizar ajustes sobre los pesos originales del modelo. Para ello, las imágenes médicas se tratan como una secuencia de fotogramas que conforman un video, ignorando por completo la ausencia de relación

temporal entre ellas. El flujo de trabajo que proponen (ver Figura 2.6), consiste en tratar al conjunto soporte como los primeros fotogramas de un video, mientras que la imagen que queremos segmentar (query) viene a ser el fotograma siguiente en esta secuencia. El banco de memoria se encarga de almacenar la información del conjunto soporte que luego se utiliza para condicionar, mediante el módulo de atención, la imagen query, permitiendo así su segmentación. Este proceso incluso permite segmentar varias imágenes query a la vez, ya que los resultados de cada paso se almacenan en el banco de memoria y son propagados sucesivamente.
A pesar de que el enfoque anterior presenta limitaciones, particularmente ante imágenes de CT y MRIs que son muy distintas a las del conjunto original de entrenamiento, los resultados son muy prometedores, especialmente si utilizamos imágenes lo más parecidas posibles a la query como conjunto soporte. Esto es posible siguiendo enfoques como el de [49], que replicaremos en este trabajo.

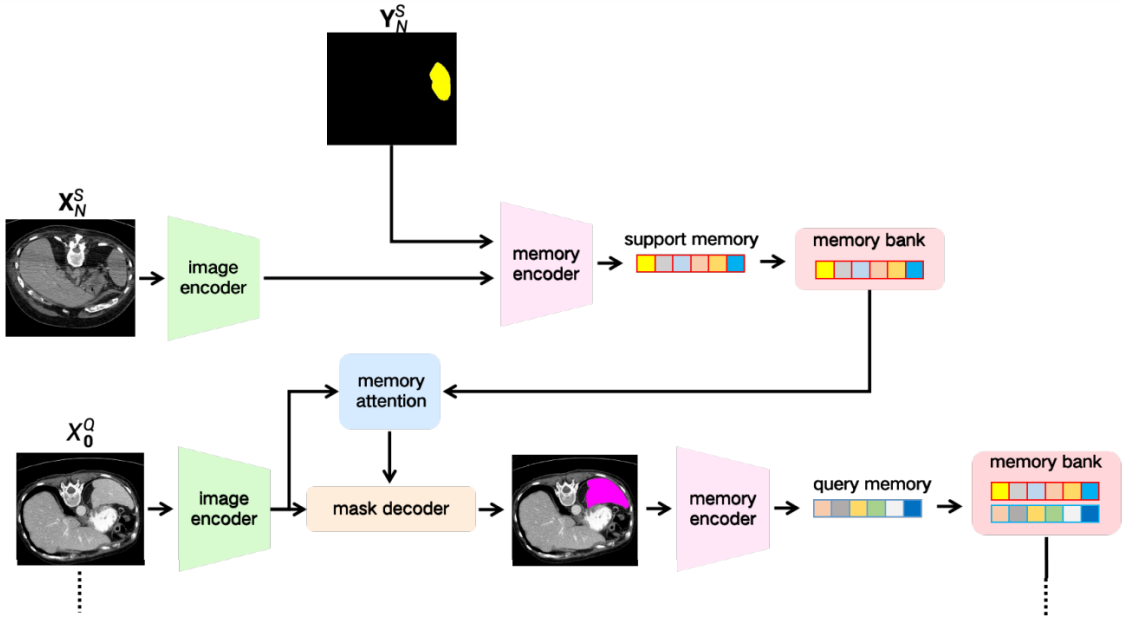

Fig. 2.6: Flujo de trabajo para segmentación FSS mediante SAM 2. La figura ilustra un caso de segmentación *one-shot*, donde (X_N^S, Y_N^S) son la imagen y máscara soporte y X_0^Q la imagen query. La imagen query es condicionada por el conjunto soporte mediante el módulo de memory attention antes de entrar al mask decoder, el cual producirá la segmentación de salida. Una vez que se realiza la segmentación esta se utiliza para actualizar el banco de memoria en un proceso iterativo. Fuente: Bai et al. [48].

3. METODOLOGÍA

3.1. Framework clásico de RCA

El concepto de *reverse classification accuracy* (RCA) introducido por Valindria et al.[1], permite evaluar la calidad de la segmentación de una imagen en ausencia de GT. Es un método aplicable para evaluar el rendimiento de cualquier método de segmentación, particularmente para imágenes médicas.
Dada una imagen y su respectiva máscara de segmentación, la cual buscamos evaluar, el enfoque consiste en entrenar un clasificador utilizando la imagen junto a la segmentación, que actúa como pseudo GT. El clasificador resultante, que llamamos *reverse classifier* o clasificador RCA, es luego evaluado en un conjunto de datos de referencia con etiquetas de GT disponible. Este conjunto puede ser el conjunto de datos utilizados para entrenar el modelo original (cuyas segmentaciones queremos evaluar) o alguna otra base de datos etiquetados disponible.
La hipotésis del método es que, si la calidad de la segmentación utilizada como pseudo GT es alta, entonces el clasificador RCA dará buenos resultados para al menos algunas de las imágenes del dataset de referencia. Análogamente, si la segmentación es de baja calidad, las predicciones del clasificador sobre los datos de referencia también lo serán. Se espera que el mejor puntaje (evaluado mediante alguna métrica como DSC) del clasificador RCA sobre el conjunto de referencia correlacione bien con el valor real que obtendríamos si dispusiéramos de GT para la imagen evaluada.
Este enfoque está inspirado en *reverse validation* [7] y *reverse testing* [8], con la diferencia de que RCA entrena un clasificador específico para cada imagen, lo que permite predecir la precisión de forma individual.

Formalmente, definimos el proceso de la siguiente manera: dada una imagen I que fue segmentada por algún método de segmentación produciendo S_I, aspiramos aprender una función $f_{I,S_I}(x) : \mathbb{R}^n \mapsto C$ que actúe de clasificador, mapeando el vector de features $x \in \mathbb{R}^n$ extraído de una imagen a las etiquetas de cada clase $c \in C$ que queremos segmentar. Este clasificador, entrenado únicamente mediante (I, S_I), nos servirá para nuestro objetivo final que es estimar la calidad de S_I ante la ausencia de GT. Para ello, definimos una función de segmentación $F_{I,S_I}(J) = S_J$ que simplemente aplica el clasificador RCA f_{I,S_I} sobre una imagen J produciendo S_J como resultado. Para la imagen de referencia J contamos con GT de referencia S_J^{GT}. Esto nos permite computar una métrica de evaluación ρ sobre el par (S_J, S_J^{GT}). La hipótesis subyacente en el proceso es que hay una correlación entre $\rho(S_J, S_J^{GT})$ y $\rho(S_I, S_I^{GT})$, siendo este último el valor que realmente nos interesa pero que en la práctica no podemos computar porque no contamos con S_I^{GT}. Sin embargo, es poco probable que esta hipótesis se cumpla para una imagen arbitraria de referencia J, es por ello que utilizamos un conjunto de datos de referencia $T = \{(J_k, S_{J_k}^{GT})\}_{k=1}^m$ que en nuestro caso serán los datos utilizados para el entrenamiento del modelo de segmentación.
A partir del clasificador RCA y los datos de referencia, definimos la siguiente medida para estimar la calidad de S_I:

$$\hat{\rho}(S_I) = \max_{1 \leq k \leq m} \rho(F_{I,S_I}(J_k), S_{J_k}^{GT}) \tag{1}$$

donde asumimos que valores más altos de ρ corresponden a segmentaciones de mejor calidad como es en el caso del Dice-score (DSC), la métrica de evaluación que utilizaremos por defecto. Otras medidas estádisticas podrían utilizarse como estimación, no obstante se vio en [1] que el máximo es el que suele funcionar mejor. Medidas tales como el promedio o la mediana, por lo general no son muy útiles en este escenario, ya que es esperable que el clasificador RCA no funcione bien en la mayoría de las imágenes, pues después de todo este fue entrenado mediante un único ejemplo.

3.2. In-context RCA

Nuestro método, que denominamos *In-Context RCA*, propone como principal cambio el uso de modelos de segmentación basados en *in-context learning* como clasificadores inversos f_{I,S_I} dentro del marco de RCA [1]. En particular emplearemos los modelos SAM 2 y UniverSeg presentados anteriormente para ello.

Originalmente, en Valindria et al. [1] se experimenta utilizando varios modelos como clasificadores inversos: Atlas Forests (AFs) [50], un método de segmentación basado en Random Forests (RFs) [51]; redes neuronales convolucionales (CNNs), y segmentación mediante registración con un atlas único [27]. Sin embargo, los resultados que obtienen con AFs y CNNs no son muy buenos, principalmente porque estos modelos requieren entrenamiento y una única imagen con su segmentación como pseudo-GT no es suficiente para generalizar bien sobre el dataset de referencia. Por otro lado, el enfoque basado en single-atlas, que no requiere entrenamiento sino que realiza la segmentación mediante registración, resulta ser muy efectivo como clasificador RCA, aunque a un costo computacional considerablemente alto.

La introducción de modelos basados en *in-context learning* como clasificadores inversos ofrece una alternativa mucho más eficiente. A diferencia de enfoques que requieren entrenamiento previo, como es el caso de AFs o CNNs, estos modelos no necesitan entrenar el clasificador inverso, ya que utilizan modelos preentrenados como UniverSeg y SAM 2 capaces de adaptar su comportamiento al conjunto de referencia utilizando la imagen I y su segmentación S_I (pseudo GT) como conjunto soporte. Esto permite obtener resultados comparables, e incluso en algunos casos superiores, a los del método de single atlas, como se verá en el Capítulo 4, pero con un costo computacional significativamente menor. Esta característica convierte a *In-Context RCA* en una opción ideal para integrarse en pipelines automáticos de procesamiento de imágenes en rutinas clínicas y en estudios de análisis de imágenes a gran escala.

La otra novedad de nuestro método es la selección del conjunto de referencia mediante técnicas de *retrieval augmentation* [10]. Este enfoque, que explicaremos con mayor detalle en 3.2.1, nos permite elegir imágenes de referencia lo más similares posibles a la imagen I que utilizamos como conjunto soporte. De esta forma, las segmentaciones producidas por el clasificador RCA serán más parecidas a S_I, lo que nos permitirá mejorar la calidad de las predicciones utilizando un subconjunto más pequeño de datos de referencia, compuesto únicamente por las imágenes más semejantes a I. En consecuencia, la variabilidad entre I y las imágenes de referencia J_k se verá reducida en este subconjunto, obteniendo como consecuencia segmentaciones más relevantes.

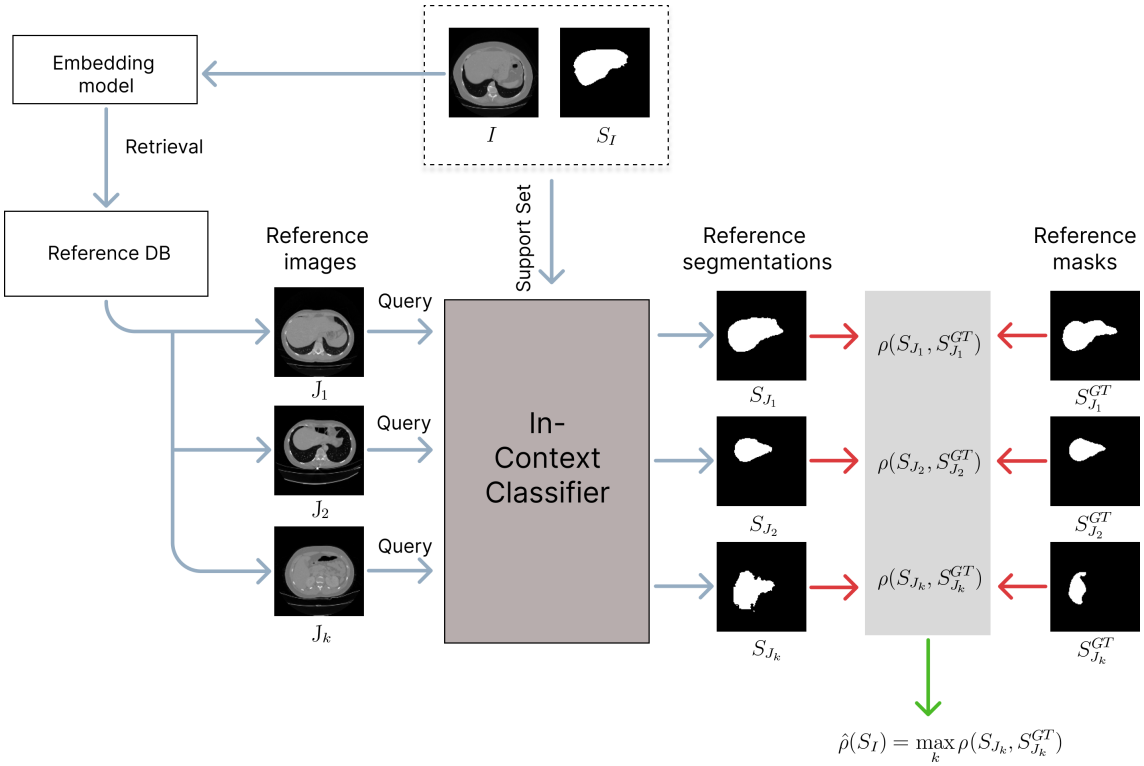

Fig. 3.1: Ilustración general del funcionamiento de In-Context RCA. La calidad de la segmentación S_I generada sobre una imagen I se estima mediante un clasificador RCA basado en In-Context Learning. Este clasificador utiliza S_I como pseudo *ground-truth* y se aplica a imágenes J_k, seleccionadas mediante *retrieval augmentation*, para las cuales se dispone de segmentaciones de referencia S^{GT}_{Jk}. El mejor puntaje de segmentación ρ, calculado sobre las imágenes de referencia, se utiliza como predicción del valor real desconocido $\rho(S_I, S^{GT}_I)$.

El flujo de trabajo que proponemos para incorporar *In-Context RCA* en la práctica se puede resumir de la siguiente manera. Dada una imagen I a segmentar:

1. Corremos un método de segmentación automática obteniendo la segmentación predicha S_I. Esta será la segmentación cuya calidad buscamos evaluar.

2. Utilizamos algún método basado en *in-context learning* como clasificador inverso f_{I,S_I}, donde $\{(I, S_I)\}$ constituye el conjunto soporte que guiará a este clasificador para realizar segmentaciones sobre nuevas imágenes J_k. Esto nos permite generar un segmentador de imágenes F_{I,S_I}.

3. Seleccionamos las n imágenes más similares a I en nuestro conjunto de referencia con imágenes cuyo GT se encuentra disponible $T = \{(J_k, S^{GT}_{J_k})\}_{k=1}^{m}$ mediante retrieval augmentation obteniendo como resultado un nuevo subconjunto de referencia $T' = \{(J_{k_l}, S^{GT}_{J_{k_l}})\}_{l=1}^{n}$ que será con el que trabajaremos de ahora en adelante.

4. Evaluamos el clasificador RCA en la base de datos de referencia de T', obteniendo como resultado segmentaciones $F_{I,S_I}(J_{k_l})\ \forall l$.

5. Estimamos la calidad de S_I mediante $\hat{\rho}(S_I)$ siguiendo la Ecuación 1, lo que nos proporciona una medida de cuán precisa fue la segmentación predicha por el modelo de segmentación automática.

En la Figura 3.1 se puede ver ilustrado el framework que sigue nuestro enfoque.

3.2.1. Implementación de Retrieval Augmentation mediante DINOv2

El concepto de *Retrieval-Augmented Generation* (RAG) se propuso originalmente para mejorar el desempeño de grandes modelos de lenguaje (LLMs) en diversas tareas del área de procesamiento del lenguaje natural (NLP) [10]. La idea principal detrás de RAG consiste en recuperar información pertinente para la tarea objetivo a partir de una base de datos de referencia. Esta información externa le proporciona contexto relevante al modelo, incorporando nuevos datos con el objetivo final de enriquecer sus respuestas y aumentar su precisión.

En este trabajo, nos basamos en el concepto de *retrieval augmentation* para mejorar la calidad de las segmentaciones de nuestro clasificador inverso. En lugar de utilizar un conjunto de referencia sin criterios específicos como se hacía tradicionalmente, este enfoque nos permite seleccionarlo de forma dinámica en base a las imágenes más relevantes para la tarea de segmentación. Estas, serán aquellas que son más similares a la imagen utilizada en el conjunto soporte, ya que se espera obtener mejores segmentaciones en este caso.

La forma más simple para calcular la similaridad entre dos imágenes es hacerlo píxel a píxel, utilizando alguna métrica de comparación como la norma L^2. Sin embargo, este enfoque suele dar resultados insatisfactorios, ya que no tiene en cuenta la estructura espacial y el contexto de las imágenes. Una alternativa más efectiva y comúnmente utilizada es calcular la similaridad en un espacio de embedding, donde las características relevantes de las imágenes se representan de manera más compacta y significativa. Esto permite capturar relaciones más complejas a nivel semántico entre las imágenes, mejorando la calidad de las predicciones en nuestro clasificador inverso.

Como espacio de embedding para calcular la similaridad entre imágenes, proponemos utilizar el generado por el modelo DINOv2 [52]. Este modelo es un *vision transformers* (ViT) [45] ampliamente utilizado en el área de visión por computadora. DINOv2 fue entrenado sobre millones de imágenes con el objetivo de que sea un robusto extractor de características para ser utilizado en tareas de diversa índole sin la necesidad de aplicar *fine-tuning*. Esto lo convierte en una opción ideal para computar la similaridad entre imágenes a la hora de realizar el retrieval. Si bien DINOv2 no fue diseñado específicamente para trabajar con imágenes médicas, su gran capacidad de abstracción y extracción de características hace que aún así resulte eficaz. No obstante, para el caso de imágenes de rayos X, optamos por utilizar RAD-DINO [53], una adaptación de DINOv2 mediante *fine-tuning* sobre imágenes específicamente de esta modalidad.

Cabe destacar que este mismo enfoque ya ha sido utilizado previamente en trabajos como [49], donde también se emplearon embeddings generados por DINOv2 para mejorar el proceso de selección de imágenes relevantes.
Para la implementación del módulo de *retrieval*, utilizamos FAISS (*Facebook AI Similarity*

Search) [54], una herramienta comúnmente utilizada para realizar búsquedas eficientes de similitud entre vectores. FAISS nos permite construir una base de datos vectorial indexada a partir de los *embeddings* de las imágenes del conjunto de referencia, generados mediante DINOv2. Luego, utilizamos el *embedding* de nuestra imagen soporte como consulta en el índice de FAISS, lo que nos permite identificar y recuperar las imágenes más similares. En nuestro caso probamos tres distintas métricas de comparación: la norma L^2, el producto interno y la similitud coseno, siendo esta última la que produce los mejores resultados, como veremos más adelante.

3.3. Descripción de los datos

En Valindria et al. [1], RCA fue evaluado exclusivamente en el conjunto de datos MALIBO [55], el cual aborda la tarea de segmentación de múltiples órganos en imágenes de resonancia magnética (MRI) del cuerpo completo. En nuestro trabajo, proponemos ampliar este análisis a una variedad mucho más amplia de conjuntos de datos que abarcan diversas modalidades y tareas de segmentación de imágenes médicas. De esta manera, buscamos demostrar la robustez y capacidad de *In-Context RCA* ante diferentes escenarios. Las modalidades que testeamos incluyen Rayos X, ultrasonido, MRI, tomografía computarizada (CT scan) e imágenes histopatológicas. A continuación, presentamos los conjuntos de datos utilizados en el proceso:

- **SCD**: este dataset fue propuesto por Radau et al. [56] como parte de la competición de segmentación automática del ventrículo izquierdo en imágenes de resonancia magnética cardíaca, realizada en el contexto de MICCAI 2009. El conjunto de datos está compuesto por 45 imágenes 2D construidas a partir de cortes de eje corto de MRI. Las imágenes corresponden a pacientes con diversas condiciones patológicas, incluyendo individuos sanos, pacientes con hipertrofia, insuficiencia cardíaca con infarto e insuficiencia cardíaca sin infarto.

- **HC18**: este dataset fue propuesto como parte de un desafío centrado en la medición automática de la circunferencia de la cabeza fetal (HC por sus siglas en inglés) a partir de imágenes de ultrasonido [57]. Los datos comprenden un total de 1334 imágenes bidimensionales 2D de ultrasonido, tomadas en un plano estándar del cráneo fetal. Esta tarea resulta relevante, dado que la circunferencia de la cabeza fetal es un indicador clave para estimar la edad gestacional y monitorear el desarrollo adecuado del feto a lo largo del embarazo.

- **PSFHS**: este dataset también consiste en imágenes de ultrasonido fetales y fue introducido como parte de un desafío realizado en MICCAI 2023 [58]. Los datos incluyen 5,101 imágenes de ultrasonido intraparto, recolectadas en dos máquinas de ultrasonido en tres hospitales. La tarea en cuestión es un caso de segmentación multiclase, donde buscamos segmentar dos estructuras distintas: la sínfisis púbica y la cabeza fetal (PS y FH respectivamente por su nombre en inglés). Este tipo de segmentación es crucial para monitorear el progreso del trabajo de parto y predecir el modo de parto más adecuado, una tarea que actualmente es compleja y costosa en tiempo.

- **PH^2**: fue desarrollado para la investigación y la evaluación comparativa de algoritmos de segmentación y clasificación de imágenes dermatoscópicas en respuesta al

creciente aumento de la incidencia de melanomas. Este conjunto de datos comprende un total de 200 imágenes dermatoscópicas de lesiones melanocíticas [59]. Incluye 80 lesiones cancerosas, 80 atípicas y 40 melanomas, con anotaciones médicas detalladas realizadas por dermatólogos expertos. Estas anotaciones abarcan la segmentación manual de las lesiones, diagnósticos clínicos e histológicos, así como la evaluación de varios criterios dermatoscópicos.

- **3D-IRCADb**: está compuesto por tomografías computarizadas (CT) en 3D de 20 pacientes, 10 hombres y 10 mujeres, de los cuales el 75 % presenta tumores hepáticos [60]. Sin embargo, en este trabajo, nos centramos únicamente en la segmentación del tejido del hígado en lugar de los tumores.
- **JSRT**: desarrollado por el Japanese Society of Radiological Technology (JSRT) [61] incluye 247 radiografías de tórax convencionales, de las cuales 154 presentan un nódulo pulmonar (100 malignos y 54 benignos), mientras que 93 no tienen nódulos. En nuestro trabajo, utilizamos este dataset para abordar la segmentación multiclase de dos estructuras anatómicas: el corazón y los pulmones.
- **NuCLS**: contiene más de 220,000 anotaciones generadas a través de la colaboración entre patólogos, residentes de patología y estudiantes de medicina. Se centra en la clasificación, localización y segmentación de núcleos celulares (NuCLS por sus siglas en inglés) en imágenes histopatológicas de cáncer de mama [62]. En nuestro trabajo, nos concentramos en la segmentación de los núcleos de las células cancerígenas, lo cual es una tarea clave para el análisis de la progresión del cáncer.

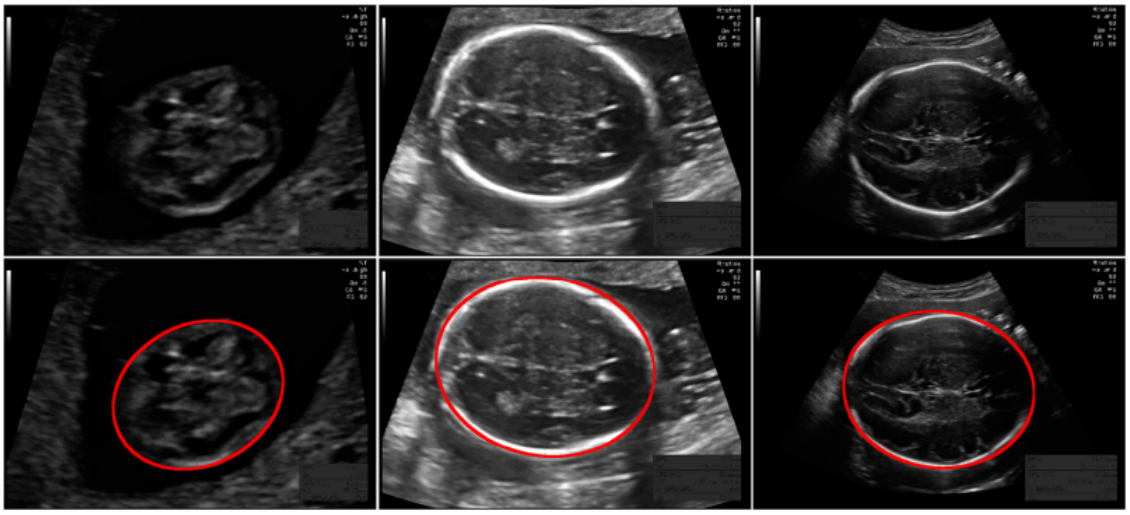

Fig. 3.2: Ejemplo de imágenes de ultrasonido del dataset HC18, centrado en la segmentación de cabezas fetales. De arriba hacia abajo: sin anotación y con la segmentación delineada en rojo. De izquierda a derecha: primer trimestre con una circunferencia de la cabeza (HC) de 65.1 mm, segundo trimestre con HC de 167.9 mm y tercer trimestre con un HC de 278.4 mm. Notar que en el primer trimestre, el cráneo aún no es visible como una estructura brillante. Fuente: [57].

En todos los casos, las imágenes fueron convertidas a escala de grises y redimensionadas a un formato de 256x256 píxeles. En casos de tener que reescalar la imagen original,

utilizamos interpolación bilineal para las imágenes e interpolación por vecino más cercano para las máscaras de segmentación. Para asegurar que las proporciones originales se mantuvieran, aplicamos padding simétrico en los dos bordes de menor tamaño como paso previo a la interpolación. En todas, el procesamiento se realizó con la ayuda de la librería Python Imaging Library (PIL) [63], que facilita enormemente este tipo de tareas.

Trabajamos siempre con imágenes 2D y en escala de grises, ya que es el formato común que soportan todos los modelos utilizados como clasificadores RCA. En los casos en que la imagen original se encontraba en formato 3D, como es el caso del dataset 3D-IRCADb, donde las imágenes corresponden a CT scans, realizamos cortes axiales y descartamos aquellos en los que no hubiera nada relevante para segmentar (determinamos esto asegurándonos de que una cierta proporción de los píxeles de la máscara no fueran 0).

Todos los conjuntos de datos fueron divididos en particiones de Train y Test, manteniendo un balance de aproximadamente 80 % para el entrenamiento y 20 % para el testeo. Se tomó especial cuidado para asegurarse de que no hubiera mezcla de datos de un mismo paciente entre ambas particiones. En el Capítulo 4, veremos que el conjunto de datos de entrenamiento se utilizó para entrenar una U-Net y luego sirvió como base de datos de referencia en el método de *In-Context RCA*. Por otro lado, los datos de prueba se utilizaron para generar segmentaciones que posteriormente fueron evaluadas siguiendo nuestro framework.

3.4. Métricas de evaluación

La evaluación de segmentaciones semánticas puede resultar bastante compleja, ya que se requiere medir tanto la precisión en la clasificación como la corrección de su localización. El objetivo consiste siempre en puntuar la similitud entre la segmentación predicha y la segmentación anotada (GT).

Las medidas más comúnmente utilizadas se basan en cuantificar el solapamiento entre la predicción y la segmentación de referencia o GT. Estas métricas se fundamentan en el cálculo de una matriz de confusión para una tarea de segmentación binaria. Esta contiene el número de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN). Entre estas métricas, la más frecuente es el índice Sørensen–Dice, que lleva este nombre en honor a Thorvald Sørensen y Lee Raymond Dice, quienes la introdujeron por primera vez [64, 65]. Esta métrica, más comúnmente conocida como coeficiente Dice o Dice Similarity Coefficient (DSC), se calcula de la siguiente manera:

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} \tag{2}$$

donde los conjuntos A y B representan la predicción y la anotación de referencia. El resultado es un valor entre 0 y 1, donde 1 indica total coincidencia entre ambos conjuntos. Cabe destacar que el coeficiente DSC puede ser reescrito en términos de la matriz de confusión como

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} \tag{3}$$

Esta fórmula, en el contexto de la segmentación binaria, coincide justamente con la media armónica entre *Precision* y *Recall* ampliamente conocida bajo el nombre de *F1-score*.

El coeficiente Dice es la métrica más utilizada en la gran mayoría de las publicaciones científicas dedicadas a la evaluación de la segmentación de imágenes médicas [66, 67, 68], por ello será la principal métrica de evaluación que consideraremos durante nuestro trabajo. Sin embargo, limitarse a una única métrica puede introducir sesgos inherentes y ofrecer una visión incompleta del rendimiento del modelo. Por esta razón, también exploramos métricas basadas en distancia espacial, ampliamente utilizadas como medidas de similaridad para evaluar aspectos más específicos, como la precisión en la delineación de bordes o contornos de las segmentaciones [67, 69]. Estas métricas complementan las medidas basadas en solapamiento y ayudan a proporcionar una evaluación más exhaustiva [70].
En particular, consideraremos dos métricas de este tipo: la distancia de Hausdorff (HD) [71] y la distancia promedio a superficie (ASSD por sus siglas en inglés). Estas medidas valen 0 en caso de coincidencia absoluta, aunque pueden tomar valores muy grandes dada su sensibilidad ante outliers [67, 70]. A continuación, presentamos las fórmulas para calcular HD y ASSD:

$$HD(A,B) = \max\left\{\max_{a\in A} d(a,B), \max_{b\in B} d(A,b)\right\} \tag{6}$$

$$ASSD(A,B) = \frac{\sum\limits_{a\in A} d(a,B) + \sum\limits_{b\in B} d(A,b)}{|A| + |B|} \tag{7}$$

$$d(x,A) = \min_{a\in A} d(x,A) \tag{8}$$

donde (8) denota la distancia de un punto a un conjunto. En todos los casos para el cálculo de las distancias se suele utilizar la norma euclídea.

Si bien métricas como DSC pueden ser extendidas fácilmente al caso de segmentación multiclase, los resultados en estos casos pueden resultar sesgados, sobre todo si hay un gran desbalance entre las clases. Por ende, incluso para los datasets de segmentación multiclase, computamos las métricas para cada clase por separado (considerando siempre el problema como de segmentación binaria) y reportamos los puntajes por clase.

En el contexto de RCA, siempre tomamos el mejor puntaje sobre los datos de referencia como predicción final; es decir que utilizamos el máximo en el caso de las medidas de solapamiento (en nuestro caso, el DSC) y el mínimo para las medidas de distancia (HD y ASSD en este caso).

4. RESULTADOS

4.1. Generación de datos mediante una U-Net

Para evaluar *In-Context RCA*, decidimos generar segmentaciones de diferentes calidades entrenando una U-Net [33] durante 10 iteraciones o *epochs*. Este enfoque nos permite probar el rendimiento de nuestro método ante una amplia variedad de escenarios, asegurando su efectividad independientemente de la calidad de las segmentaciones.
Para este proceso en todos los casos utilizamos el split de Train para entrenar la red y guardamos las segmentaciones generadas sobre las imágenes de la partición de Test al final de cada iteración. De esta manera, buscamos obtener segmentaciones de menor calidad en las primeras *epochs*, que mejoren progresivamente hasta alcanzar una mayor calidad en las últimas iteraciones. Nuestro objetivo en mente es que la distribución del DSC de las segmentaciones generadas sea lo más uniforme posible. La evolución de los scores es fácil de monitorear evaluando los resultados en cada iteración (o *epoch*) del proceso de entrenamiento de la red.

La U-Net utilizada tiene una estructura estándar, con tres bloques de *downsampling* y tres de *upsampling*, con un cuello de botella en el medio. Comienza con 32 canales en la primera capa, que se duplican progresivamente hasta llegar a 256 canales en el cuello de botella. Luego en la fase de *upsampling*, se reduce la cantidad de canales de manera simétrica hasta regresar a los 32 canales iniciales antes de la capa de salida. La cantidad de canales de salida estará determinada por el número de clases que buscamos segmentar. Dentro de cada bloque, utilizamos lo que se conoce como *batch normalization* después de cada operación de convolución. Esta incorporación ayuda a mejorar la convergencia y a reducir el overfitting [72]. Como funciones de activación, empleamos siempre ReLU y, en la salida, utilizamos una sigmoidea en el caso de segmentación binaria y softmax para segmentación multiclase.

La red es entrenada mediante *Mini-batch gradient descent* [73] utilizando *backpropagation* [74] con un *batch-size* de entre 8 y 24. Como optimizador empleamos Adam [75] con *weight-decay* de 10^{-5} y un *learning-rate* inicial oscilando entre 10^{-5} y 10^{-3}. La elección del *batch size* y el *learning rate* dependerá en cada caso del dataset utilizado y se realiza con el objetivo de que el modelo no aprenda ni demasiado rápido ni demasiado lento, buscando así obtener una distribución de scores lo más uniforme posible a lo largo de las 10 iteraciones.

Por último como función de pérdida utilizamos una Dice loss [36], comúnmente utilizada en problemas de segmentación [76, 29, 77]. Esta se puede expresar de la siguiente manera:

$$DL(y, \hat{y}) = 1 - 2\frac{y\hat{y} + \epsilon}{y + \hat{y} + \epsilon} \tag{9}$$

donde ϵ es un factor de smoothing que establecemos en 1. La misma función de pérdida puede ser utilizada para el caso de segmentación multi-clase; en este caso, se calcula el Dice de cada clase por separado partiendo siempre de la Ecuación 9.

En la Figura 4.1 vemos un ejemplo que ilustra como la calidad de las segmentaciones mejora a medida que avanza el entrenamiento de la U-Net para el dataset PSFHS.

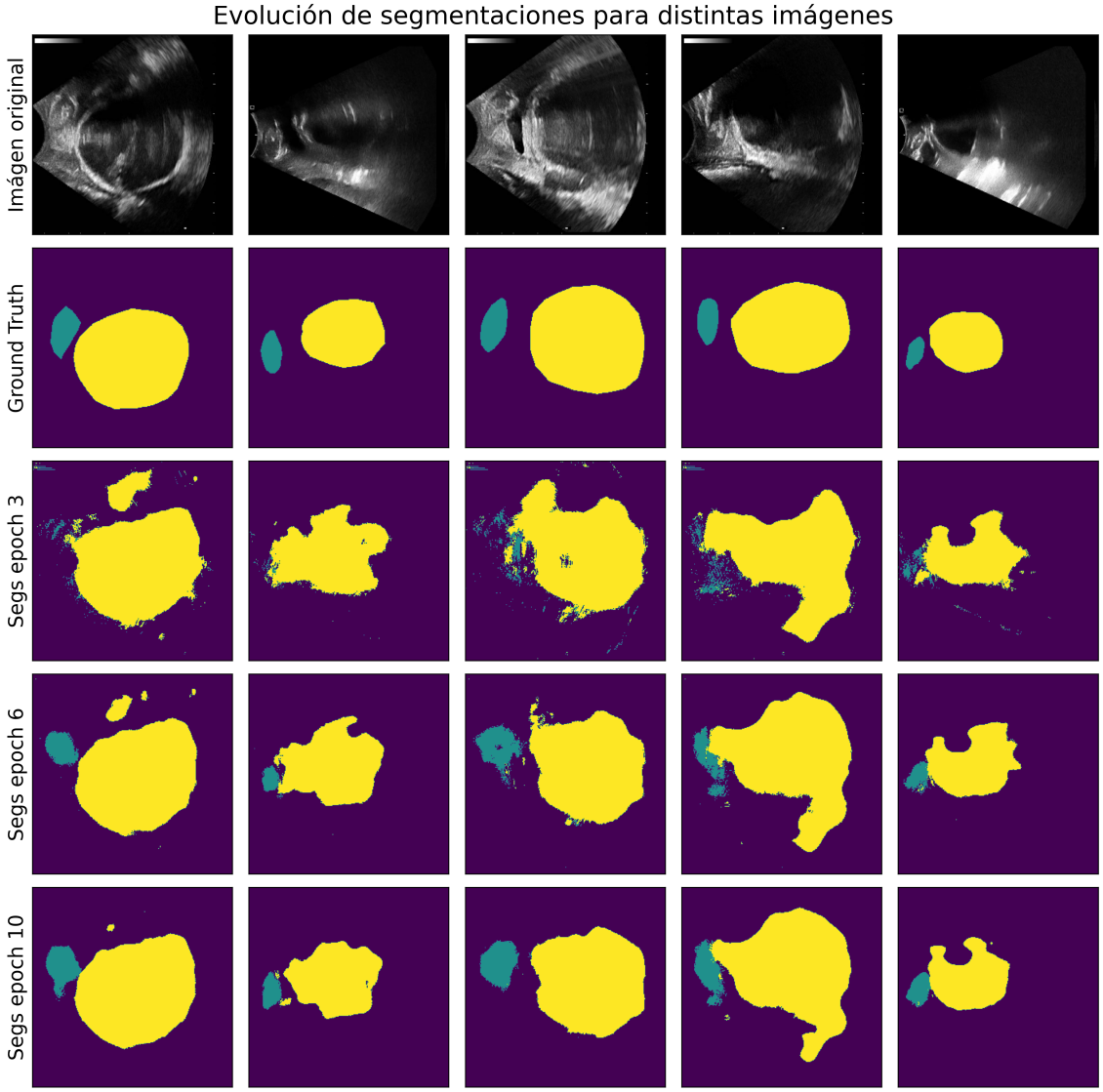

Fig. 4.1: Segmentaciones generadas por la U-Net en distintas epochs para 5 imágenes tomadas del dataset PSFHS.

En la Figura 4.2 vemos la distribución de los Dice scores finales para este mismo dataset (PSFHS). En casos de segmentación multiclase, es más difícil obtener un buen balance entre las segmentaciones de las distintas clases. Como podemos ver, en este caso la clase PS es más difícil de segmentar que la clase FH, además de presentar una mayor variabilidad en sus resultados. Es por esto que la distribución resultante no es del todo uniforme; cuando comenzamos a obtener buenas segmentaciones para la clase PS, las segmentaciones de FH ya presentan resultados bastante satisfactorios. Este efecto se aprecia mejor en la Figura 4.3, donde se puede observar la evolución de los Dice scores a lo largo de las distintas epochs de entrenamiento.

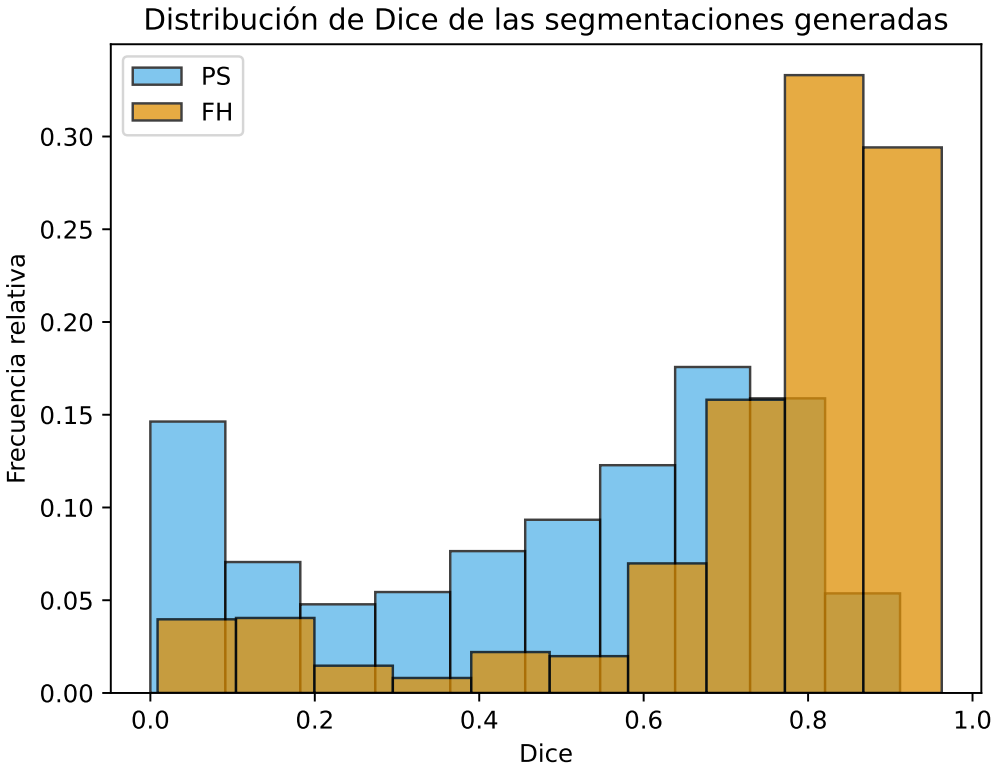

Fig. 4.2: Distribución del DSC de ambas clases para las segmentaciones generadas sobre el dataset PSFHS.

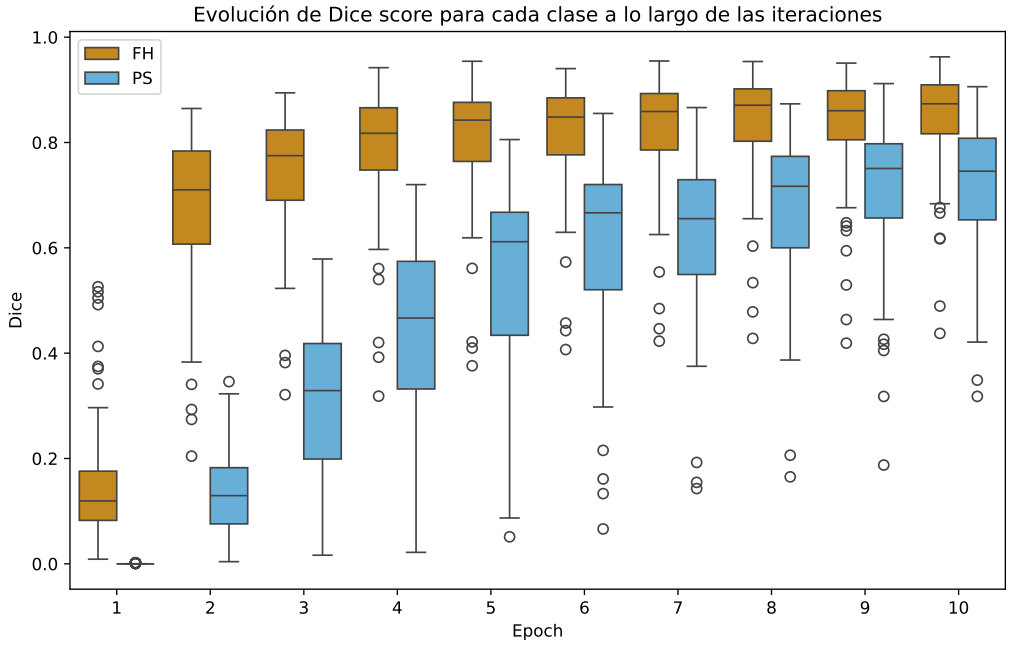

Fig. 4.3: Evolución de la distribución de los DSC a lo largo de las epochs para ambas clases del dataset PSFHS.

4.2. Experimentos utilizando UniverSeg y retrieval augmentation

En esta serie de experimentos, aplicamos *In-Context RCA* utilizando UniverSeg como clasificador RCA para evaluar las segmentaciones generadas. El objetivo es comparar tanto el impacto de la cantidad de datos de referencia utilizados, así como el uso de *retrieval augmentation* para seleccionarlos. En cuanto al número de datos de referencia empleados, se probó con 2, 5, 10, 25 y 50 ejemplos. Utilizar conjuntos de cardinalidad más allá de estos valores no resulta beneficioso, ya que el tiempo de inferencia por imagen aumenta considerablemente y puede haber cierto riesgo de sobreestimación del puntaje real. Esto se debe a que, al utilizar siempre el mejor puntaje como predictor, las predicciones son monótonas crecientes (o monótonas decrecientes en el caso de medidas de distancia).
Dado que UniverSeg solo soporta segmentación binaria, para los casos de segmentación multiclase segmentamos cada clase por separado, incluyendo la clase *background*, y luego elegimos la clase más probable para cada píxel utilizando *softmax*, replicando de esta forma el proceso descripto por Butoi et al. [2].

En la Figura 4.4 se muestran los resultados obtenidos sobre el dataset HC18 al seleccionar k datos de referencia al azar para distintos valores de k. Para la evaluación de las segmentaciones en este caso mantenemos siempre fijo el conjunto elegido, siguiendo el proceso tradicional de RCA con la salvedad de que UniverSeg es nuestro clasificador inverso. Para cada una de las segmentaciones vemos graficado en el eje horizontal el DSC predicho, mientras que en el vertical se ve el valor real. Vemos que a medida que incrementamos el tamaño de la base de datos de referencia mejoran los resultados, aumentando la correlación y disminuyendo el error absoluto medio (MAE).
Por otro lado, en la figura 4.5 vemos que los resultados mejoran considerablemente al seleccionar de forma dinámica el conjunto de referencia. Esta selección se basa a partir de los k elementos más similares según la similitud coseno en el espacio de embedding generado por DINOv2. El enfoque de retrieval augmentation nos permite obtener buenos resultados con una menor cantidad de datos de referencia lo que es muy útil en la práctica, ya que permite reducir costos de inferencia agilizando el proceso. Un efecto similar se ve también en casos de segmentación multiclase como puede observarse en las Figuras 4.6 y 4.7.

La estrategia de *retrieval augmentation*, no obstante, presenta limitaciones. Esto se puede notar en las Figuras 4.8 y 4.9, que corresponden al dataset SCD. En este caso, seleccionar las imágenes de referencia mediante *retrieval augmentation* no presenta mejoras significativas en comparación a realizarlo de forma aleatoria. Esto podría deberse a una limitación de DINOv2 que, si bien es muy eficaz para extraer características robustas, no fue entrenado imágenes médicas, lo que podría representar un inconveniente frente a ciertos tipos de modalidades o imágenes de este tipo. Como podemos ver, en este escenario es necesario aumentar el tamaño de los datos de referencia para mejorar los resultados.

Cabe destacar que para computar la similaridad entre imágenes, otras medidas como el producto interno o la distancia euclídea también fueron consideradas. Sin embargo, en todos los casos, la similitud coseno produjo los mejores resultados.

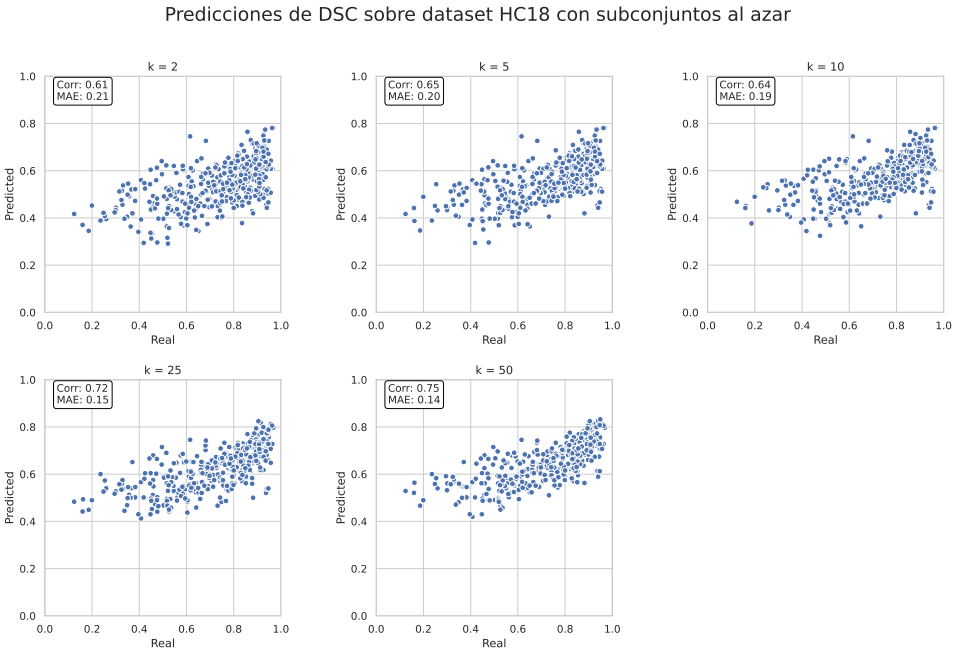

Fig. 4.4: Predicciones de DSC sobre el dataset HC18 seleccionando al azar conjuntos de referencia de tamaño k. En todos los casos se indica tanto la correlación (Corr) como el error absoluto medio (MAE) entre las predicciones y los valores reales.

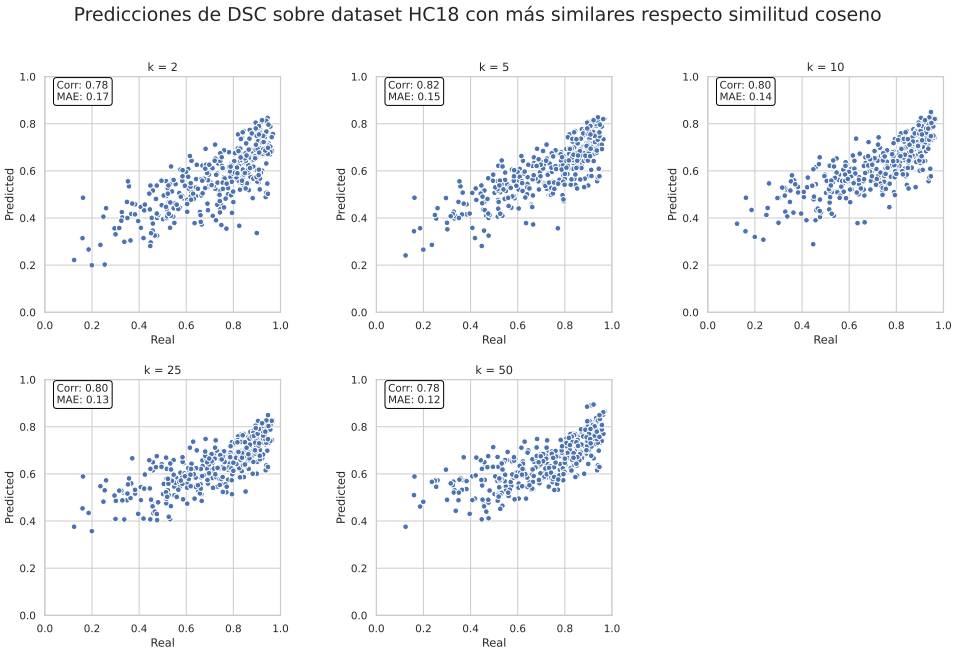

Fig. 4.5: Predicciones de DSC sobre el dataset HC18 seleccionando conjuntos de referencia en base a las k imágenes más similares, siguiendo un enfoque de retrieval augmentation. La similitud entre imágenes es determinada en base a la similitud coseno en el espacio de embedding dado por DINOv2. En todos los casos se indica tanto la correlación (Corr) como el error absoluto medio (MAE) entre las predicciones y los valores reales.

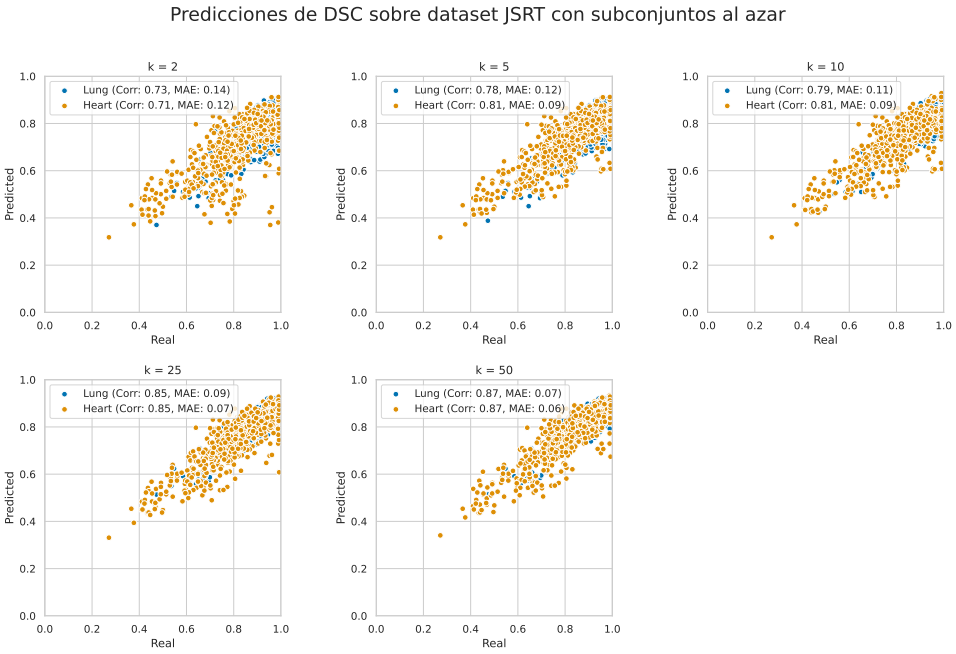

Fig. 4.6: Predicciones de DSC sobre el dataset JSRT seleccionando al azar conjuntos de referencia de tamaño k. Vemos distinguidas por color las predicciones para las clases pulmón y corazón. En todos los casos se indica tanto la correlación (Corr) como el error absoluto medio (MAE) entre las predicciones y los valores reales para cada clase.

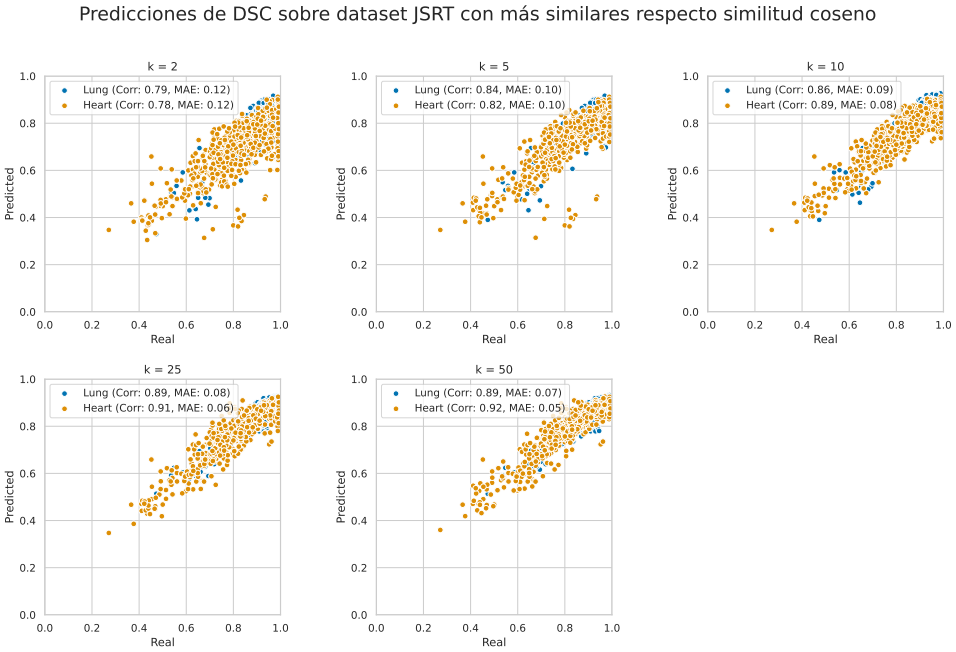

Fig. 4.7: Predicciones de DSC sobre el dataset JSRT seleccionando conjuntos de referencia en base a las k imágenes más similares, siguiendo un enfoque de retrieval augmentation. El parecido entre imágenes se determina a partir de la similitud coseno en el espacio de embedding dado por RAD-DINO. En todos los casos se indica tanto la correlación (Corr) como el error absoluto medio (MAE) entre las predicciones y los valores reales para cada clase.

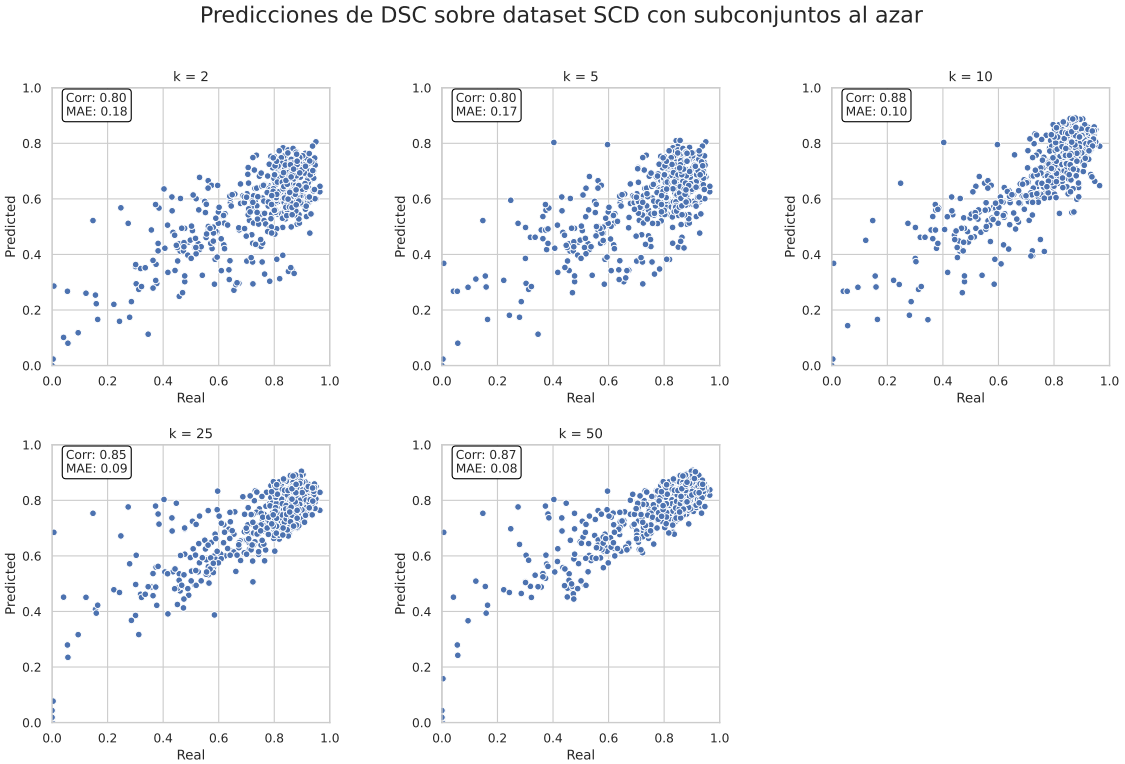

Fig. 4.8: Predicciones de DSC sobre el dataset SCD seleccionando al azar conjuntos de referencia de tamaño k. En todos los casos se indica tanto la correlación (Corr) como el error absoluto medio (MAE) entre las predicciones y los valores reales.

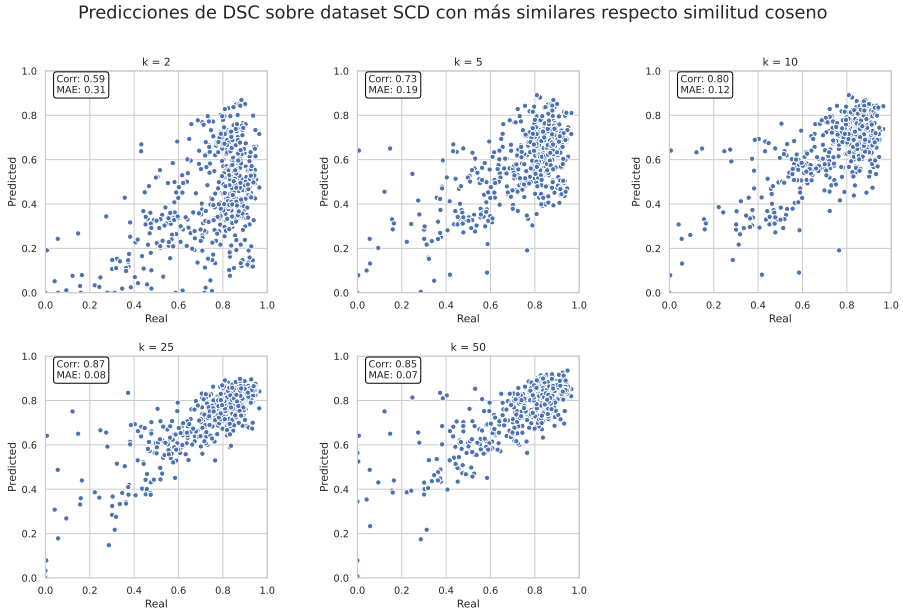

Fig. 4.9: Predicciones de DSC sobre el dataset SCD seleccionando conjuntos de referencia en base a las k imágenes más similares. En todos los casos se indica tanto la correlación (Corr) como el error absoluto medio (MAE) entre las predicciones y los valores reales.

4.3. Análisis comparativo

Para explorar la efectividad de *In-Context RCA* y sus posibles limitaciones, realizamos un análisis exhaustivo evaluando el método con UniverSeg y SAM 2 como clasificadores inversos en todos los conjuntos de datos propuestos. Para ello utilizamos principalmente el DSC pero también incluimos la distancia Hausdorff y ASSD como métricas de evaluación. Además, elaboramos una comparación con el enfoque single-atlas, que ofrecía los mejores resultados en el framework clásico de RCA, replicando la propuesta de Valindria et al. [1], salvo por la selección del conjunto de referencia mediante *retrieval augmentation*.
Dado que el método Atlas es muy lento de correr, decidimos limitar la cantidad de datos a evaluar a 75 en todos los casos y, a su vez, elegimos siempre las 16 imágenes más similares para conformar el conjunto anotado de referencia. Este número fue propuesto en base a las observaciones previas, donde notamos que aumentar la cantidad de datos de referencia es beneficioso, aunque al seguir el enfoque de *retrieval augmentation* no es necesario que el número sea muy grande, lo que nos permite reducir este valor y consecuentemente el tiempo requerido para la inferencia.

En las Figuras 4.10, 4.11 y 4.12 vemos respectivamente las predicciones de los tres modelos para el DSC, la distancia Hausdorff y la distancia ASSD para el caso del dataset JSRT. Las predicciones tanto para el DSC como para la distancia ASSD son buenas en todos los casos, con los mejores resultados obtenidos mediante el método single-atlas. Sin embargo, vemos que los tres modelos presentan grandes dificultades al estimar la distancia Hausdorff. Esta limitación en la estimación de medidas basadas en distancias es una problemática que también se menciona en Valindria et al. [1] y que nosotros observamos en muchos de los datasets, particularmente para la distancia Hausdorff aunque en algunos casos también para la distancia ASSD. Esta problemática podría deberse a que las medidas basadas en distancias suelen ser sensibles a outliers y dan mayor importancia a la exactitud en la delineación de los bordes, los cuales a menudo varían considerablemente entre imágenes médicas.

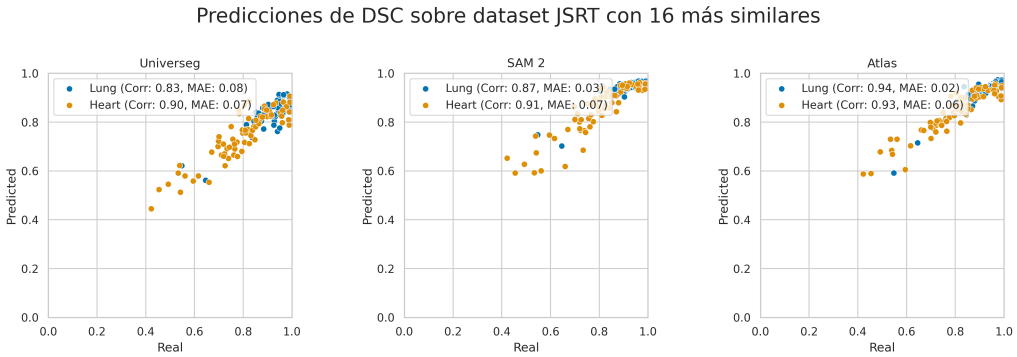

Fig. 4.10: Predicciones de DSC sobre el dataset JSRT para distintos modelos. Se utilizaron 75 segmentaciones y las respectivas 16 imágenes más similares como referencia en cada caso, tomando el puntaje máximo sobre estas como predicción. Para cada modelo se indican la correlación (Corr) y error absoluto medio (MAE) entre las predicciones y valores reales de cada clase.

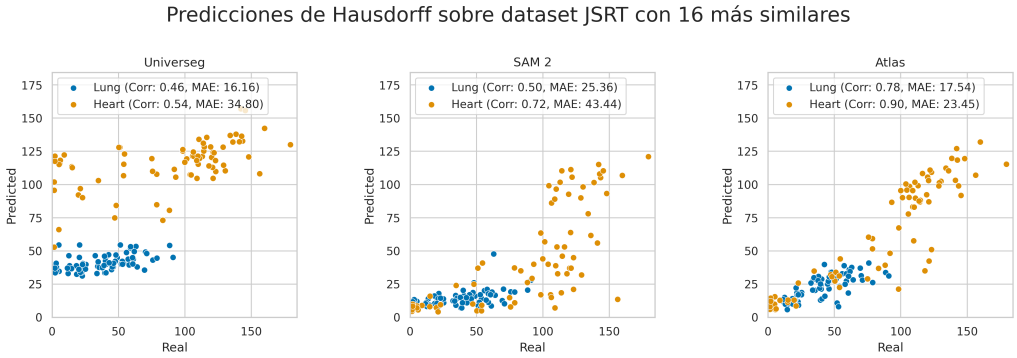

Fig. 4.11: Predicciones de distancia Hausdorff sobre el dataset JSRT para distintos modelos. Se utilizaron 75 segmentaciones y 16 imágenes de referencia en cada caso, tomando el puntaje mínimo como predicción. Para cada modelo se indican la correlación (Corr) y error absoluto medio (MAE) entre las predicciones y valores reales de cada clase.

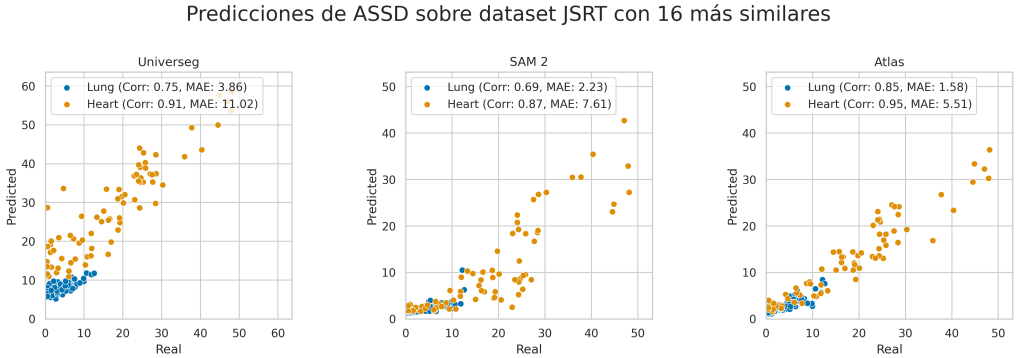

Fig. 4.12: Predicciones de distancia ASSD sobre el dataset JSRT para distintos modelos. Se utilizaron 75 segmentaciones y 16 imágenes de referencia por caso, tomando el puntaje mínimo como predicción. Para cada modelo se indican la correlación (Corr) y error absoluto medio (MAE) entre las predicciones y valores reales de cada clase.

Por último, en la Figura 4.13 podemos ver resumidos los resultados obtenidos por cada uno de los modelos sobre los distintos datasets al estimar el DSC. Una tabla resumiendo estos mismos valores junto a resultados análogos para el resto de métricas pueden encontrarse en Apéndice.

Un caso interesante es el de las predicciones sobre el dataset NuCLS (ubicado en la columna 5), que corresponde a la segmentación de núcleos de células de cáncer de mama, caracterizados por una gran variabilidad entre sí. En este escenario, se hace más evidente la limitación del método Atlas, que intenta establecer una correspondencia directa entre la imagen de soporte y la imagen a segmentar, lo cual resulta especialmente complicado debido a la gran heterogeneidad de las estructuras celulares. Incluso UniverSeg muestra dificultades en este contexto, lo que sugiere que en estos casos, una única imagen como contexto no proporciona suficiente información para capturar de manera efectiva la variabilidad de la tarea. Sin embargo, SAM 2 muestra resultados significativamente mejores, demostrando su gran capacidad de adaptación y generalización, a pesar de haber sido entrenado con imágenes de dominios muy diferentes al de la tarea objetivo.

En general, los resultados de SAM 2 y UniverSeg son comparables e incluso en muchos casos superiores a los obtenidos con el enfoque tradicional de single-atlas, pero con un costo computacional mucho menor. Esto se debe al hecho de que no solo SAM 2 y UniverSeg son inherentemente más rápidos, sino que también son fácilmente paralelizables en GPUs, mientras que el método de Atlas es ejecutado mediante CPUs.
Además, es muy posible que Atlas sea el principal beneficiario de la selección dinámica del conjunto de referencia mediante *retrieval augmentation*, dado que su enfoque de segmentación basado en registro, intenta replicar la estructura de las imágenes de referencia. Por lo tanto, si siguiéramos estrictamente el enfoque clásico descripto en la Figura 3.1, es muy probable que los resultados de Atlas resulten inferiores.

En muchos casos, se observa una correlación muy alta entre el puntaje predicho y el real, aunque con un MAE relativamente elevado. Sin embargo, esto también suele revelar un sesgo sistemático, donde los modelos muestran un desvío constante respecto a la diagonal. Una posible solución para este tipo de problemas podría ser seleccionar una muestra de datos de evaluación para realizar un ajuste lineal y, a partir de la función aprendida, corregir las evaluaciones subsiguientes.

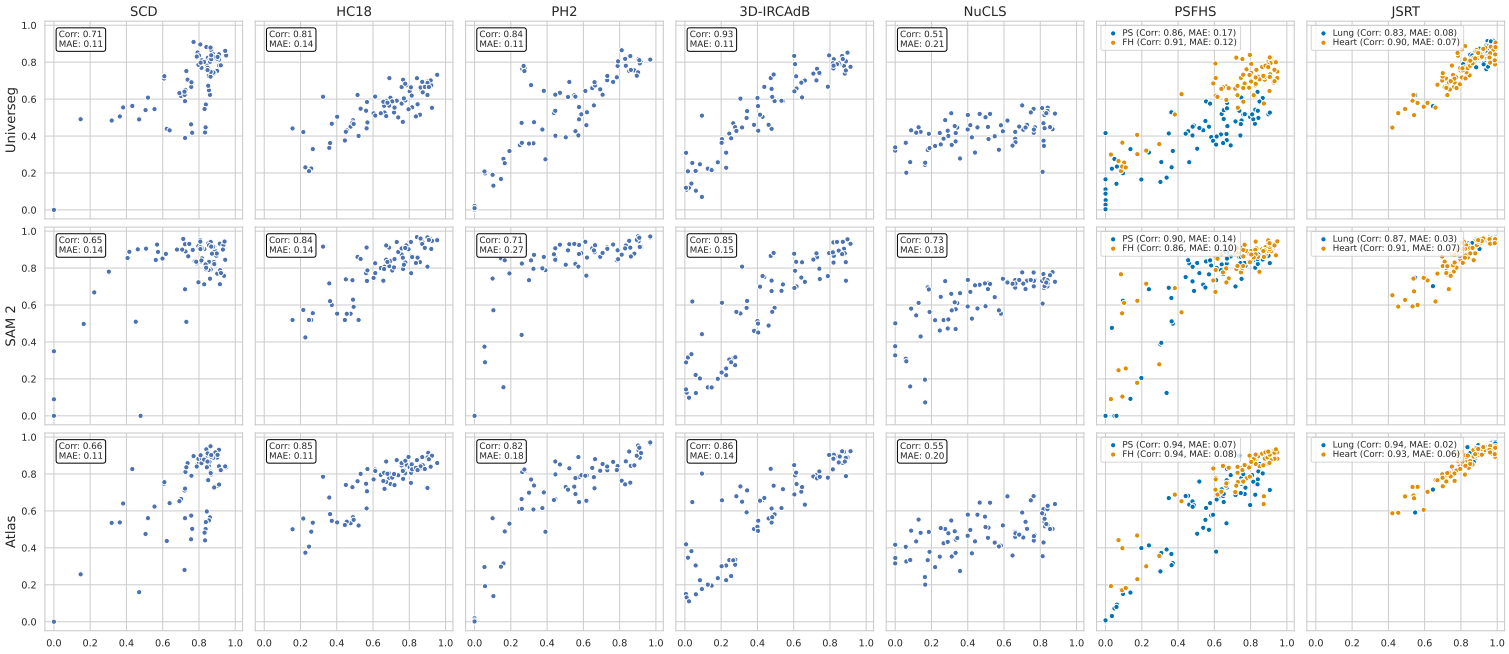

Fig. 4.13: Comparación de los resultados obtenidos al estimar el Dice score en todos los datasets para cada modelo. En el eje horizontal se encuentra el valor real mientras que en el vertical el estimado. En todo caso se utilizaron las 16 imágenes más similares como conjunto de referencia y se evaluaron un total de 75 imágenes. Además, se incluye la correlación (Corr) y el error absoluto medio (MAE) entre los valores de Dice reales y estimados para cada una de las clases a segmentar.

5. CONCLUSIONES

En este trabajo, nos propusimos abordar el control automático de la segmentación de imágenes médicas en ausencia de etiquetas de GT, partiendo del enfoque de RCA [1] para hacer esto posible. Nuestra investigación propuso varias mejoras significativas respecto al trabajo original. En primer lugar, introdujimos la posibilidad de utilizar modelos de segmentación automática basados en in-context learning como clasificadores inversos, con UniverSeg [2] y SAM 2 [3] como modelos de ejemplo. En segundo lugar, integramos al framework clásico de RCA técnicas de retrieval augmentation para seleccionar de forma dinámica mejores conjuntos de referencia en cada caso. Como se vio, estas mejoras permitieron mejorar la eficiencia en el proceso de evaluación y alcanzar resultados similares o incluso superiores respecto al enfoque tradicional.

Sin embargo, encontramos también que nuestra propuesta puede presentar también ciertas limitaciones: el enfoque de retrieval augmentation puede no ser efectivo en todos los casos, especialmente cuando se aplica a cierto tipo de imágenes o a estructuras muy diversas. En ciertas situaciones, modelos tales como DINOv2 [52] pueden no ser capaces de capturar la complejidad de las imágenes, lo que puede afectar la selección de los conjuntos de referencia y consecuentemente la calidad de las predicciones. La estimación de medidas basadas en distancia, principalmente la distancia Hausdorff, también sigue siendo un desafío para ambos métodos en varios de los escenarios evaluados.

En cuanto a la segmentación de estructuras muy diversas, como en el caso del dataset NuCLS [62], el enfoque de RCA tradicional con métodos como Atlas revelan grandes dificultades. Al utilizar modelos basados en *in-context learning*, por otro lado, las predicciones pueden mejorar ante este tipo de circunstancias, como se observó con SAM 2 que demuestra una gran capacidad de generalización. No obstante, todavía existen desafíos para capturar la complejidad de estas estructuras. Posiblemente emplear una única imágen como conjunto soporte resulte en un contexto muy limitado para los modelo en estos casos. Para este tipo de problemas, donde la variabilidad de la estructura es muy alta, se podrían explorar alternativas; una de ellas es utilizar un conjunto de soporte más amplio, lo que permitiría a los modelos basados en in-context learning tener una visión más completa de la tarea objetivo. Sin embargo, esta alternativa implicaría perder la capacidad de estimar la calidad de las segmentaciones a nivel individual, obteniendo una estimación a nivel global en su lugar.

Para analizar la eficacia del método se amplió el análisis original (realizado únicamente sobre el dataset MALIBO), evaluando tanto UniverSeg y SAM 2 como Single-Atlas sobre una amplia variedad de conjuntos de datos médicos. Para ello, se utilizaron datasets que abarcan diversas modalidades de imágenes médicas como Rayos X, Ultrasonido, MRI, CT e imágenes histopatológicas. En casi todos los casos, se observó un gran desempeño del método para evaluar la calidad de las segmentaciones, especialmente al usar retrieval augmentation.

En línea general, este trabajo aporta una importante contribución al avance del control

automático de la calidad de la segmentación de imágenes médicas en ausencia de GT, demostrando la efectividad de utilizar modelos de in-context learning y técnicas de retrieval augmentation en el proceso, lo que lo hace ideal en la práctica para incorporar a grandes rutinas de segmentación de imágenes médicas.

Finalmente, este trabajo ha servido como punto de partida para futuras investigaciones que busquen superar las limitaciones del método propuesto y explorar nuevas posibilidades de aplicación. Es necesario analizar con mayor profundidad los modelos de *in-context learning* para entender mejor las ventajas y limitaciones de su uso como clasificadores inversos. Asimismo, se requiere profundizar en las técnicas de retrieval augmentation para mejorar ulteriormente la selección de los conjuntos de referencia y así obtener mejores predicciones. Sobre todo, ayudaría contar con grandes modelos especializados en imágenes médicas y capaces de extraer características robustas, como es el caso de RAD-DINO [53], que adapta DINOv2 [52] para el caso de imágenes radiológicas. Por otro lado, se debe buscar mejorar las predicciones para medidas basadas en distancia, ya que esta es una limitación que poseen tanto el enfoque tradicional de RCA como *In-Context RCA*.

Bibliografía

[1] Vanya V. Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, and Ben Glocker. Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth, 2017.

[2] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation, 2023.

[3] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.

[4] Y.J. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.

[5] Luren Yang, Fritz Albregtsen, Tor Lønnestad, and Per Grøttum. A supervised approach to the evaluation of image segmentation methods. In Václav Hlaváč and Radim Šára, editors, *Computer Analysis of Images and Patterns*, pages 759–765, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg.

[6] Hui Zhang, Jason E. Fritts, and Sally A. Goldman. Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2):260–280, 2008.

[7] Erheng Zhong, Wei Fan, Qiang Yang, Olivier Verscheure, and Jiangtao Ren. Cross validation framework to choose amongst models and datasets for transfer learning. In José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 547–562, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[8] Wei Fan and Ian Davidson. Reverse testing: an efficient framework to select amongst classifiers under sample selection bias. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 147–156, New York, NY, USA, 2006. Association for Computing Machinery.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[11] Vincent Chan and Anahi Perlas. Basics of ultrasound imaging. *Atlas of ultrasound-guided procedures in interventional pain management*, pages 13–19, 2011.

[12] Jim Neilson. Ultrasound for fetal assessment in early pregnancy (cochrane review). *Cochrane database of systematic reviews (Online)*, 2:CD000182, 02 2000.

[13] Zainab T. Al-Sharify, Talib A. Al-Sharify, Noor T. Al-Sharify, and Husam Yahya naser. A critical review on medical imaging techniques (ct and pet scans) in the medical field. *IOP Conference Series: Materials Science and Engineering*, 870(1):012043, jun 2020.

[14] Thomas D DenOtter and Jordan Schubert. *Hounsfield Unit*. StatPearls Publishing, Treasure Island (FL), 2023. March 6, 2023.

[15] Girish Katti, Syeda Arshiya Ara, and Ayesha Shireen. Magnetic resonance imaging (mri)–a review. *International journal of dental clinics*, 3(1):65–70, 2011.

[16] Mitko Veta, Josien P. W. Pluim, Paul J. van Diest, and Max A. Viergever. Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411, 2014.

[17] Glauco Vitor Pedrosa, Agma J.M. Traina, and Caetano Traina. Using sub-dictionaries for image representation based on the bag-of-visual-words approach. In *2014 IEEE 27th International Symposium on Computer-Based Medical Systems*, pages 165–168, 2014.

[18] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

[19] Alireza Norouzi, Mohd Shafry Mohd Rahim, Ayman Altameem, Tanzila Saba, Abdolvahab Ehsani Rad, Amjad Rehman, and Mueen Uddin. Medical image segmentation methods, algorithms, and applications. *IETE Technical Review*, 31(3):199–213, 2014.

[20] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), January 2024.

[21] Robert M. Haralick and Linda G. Shapiro. Image segmentation techniques. In *Other Conferences*, 1984.

[22] Maria Kallergi, Kevin Woods, Laurence P Clarke, Wei Qian, and Robert A Clark. Image segmentation in digital mammography: comparison of local thresholding and region growing algorithms. *Computerized medical imaging and graphics*, 16(5):323–331, 1992.

[23] Bharath Sathya and R Manavalan. Image segmentation by clustering methods: performance analysis. *International Journal of Computer Applications*, 29(11):27–32, 2011.

[24] W. M. Wells, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz. Adaptive segmentation of mri data. In Nicholas Ayache, editor, *Computer Vision, Virtual Reality and Robotics in Medicine*, pages 59–69, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg.

[25] Juan Eugenio Iglesias and Mert R. Sabuncu. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1):205–219, 2015.

[26] Lucas Mansilla and Enzo Ferrante. Segmentación multi-atlas de imágenes médicas con selección de atlas inteligente y control de calidad automático. In *XXIV Congreso Argentino de Ciencias de la Computación (Tandil, 2018).*, 2018.

[27] Wenjia Bai, Wenzhe Shi, Declan P O'regan, Tong Tong, Haiyan Wang, Shahnaz Jamil-Copley, Nicholas S Peters, and Daniel Rueckert. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac mr images. *IEEE transactions on medical imaging*, 32(7):1302–1315, 2013.

[28] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[29] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: A survey. *IET image processing*, 16(5):1243–1267, 2022.

[30] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[31] 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.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[34] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation, 2018.

[35] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016.

[36] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *CoRR*, abs/1606.04797, 2016.

[37] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024.

[38] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning, 2017.

[39] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, page 4, 2018.

[40] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019.

[41] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 762–780. Springer, 2020.

[42] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(7):1837–1848, 2022.

[43] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[46] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[47] Jiayuan Zhu, Yunli Qi, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2, 2024.

[48] Yunhao Bai, Qinji Yu, Boxiang Yun, Dakai Jin, Yingda Xia, and Yan Wang. Fs-medsam2: Exploring the potential of sam2 for few-shot medical image segmentation without fine-tuning, 2024.

[49] Lin Zhao, Xiao Chen, Eric Z. Chen, Yikang Liu, Terrence Chen, and Shanhui Sun. Retrieval-augmented few-shot medical image segmentation with foundation models, 2024.

[50] Darko Zikic, Ben Glocker, and Antonio Criminisi. Encoding atlases by randomized classification forests for efficient multi-atlas label propagation. *Medical image analysis*, 18(8):1262–1273, 2014.

[51] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[52] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[53] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Rad-dino: Exploring scalable medical image encoders beyond text supervision, 2024.

[54] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2024.

[55] I Lavdas, B Glocker, D Rueckert, SA Taylor, EO Aboagye, and AG Rockall. Machine learning in whole-body mri: experiences and challenges from an applied study using multicentre data. *Clinical radiology*, 74(5):346–356, 2019.

[56] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright. Evaluation framework for algorithms segmenting short axis cardiac mri. *The MIDAS Journal*, July 2009.

[57] Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. *PloS one*, 13(8):e0200412, 2018.

[58] Jieyun Bai, Zihao Zhou, Zhanhong Ou, Gregor Koehler, Raphael Stock, Klaus Maier-Hein, Marawan Elbatel, Robert Martí, Xiaomeng Li, Yaoyang Qiu, Panjie Gou, Gongping Chen, Lei Zhao, Jianxun Zhang, Yu Dai, Fangyijie Wang, Guénolé Silvestre, Kathleen Curran, Hongkun Sun, Jing Xu, Pengzhou Cai, Lu Jiang, Libin Lan, Dong Ni, Mei Zhong, Gaowen Chen, Víctor M. Campello, Yaosheng Lu, and Karim Lekadir. Psfhs challenge report: Pubic symphysis and fetal head segmentation from intrapartum ultrasound images, 2024.

[59] Teresa Mendonca, Pedro M. Ferreira, Jorge S. Marques, Andre R. S. Marcal, and Jorge Rozeira. PH^2 - a dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5437–5440. IEEE, 2013.

[60] Luc Soler, Alexandre Hostettler, Vincent Agnus, Arnaud Charnoz, J Fasquel, Johan Moreau, A Osswald, Mourad Bouhadjar, and Jacques Marescaux. 3d image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. *IRCAD, Strasbourg, France, Tech. Rep*, 1(1), 2010.

[61] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists'

detection of pulmonary nodules. *American journal of roentgenology*, 174(1):71–74, 2000.

[62] Mohamed Amgad, Lamees A Atteya, Hagar Hussein, Kareem Hosny Mohammed, Ehab Hafiz, Maha A T Elsebaie, Ahmed M Alhusseiny, Mohamed Atef AlMoslemany, Abdelmagid M Elmatboly, Philip A Pappalardo, Rokia Adel Sakr, Pooya Mobadersany, Ahmad Rachid, Anas M Saad, Ahmad M Alkashash, Inas A Ruhban, Anas Alrefai, Nada M Elgazar, Ali Abdulkarim, Abo-Alela Farag, Amira Etman, Ahmed G Elsaeed, Yahya Alagha, Yomna A Amer, Ahmed M Raslan, Menatalla K Nadim, Mai A T Elsebaie, Ahmed Ayad, Liza E Hanna, Ahmed Gadallah, Mohamed Elkady, Bradley Drumheller, David Jaye, David Manthey, David A Gutman, Habiba Elfandy, and Lee A D Cooper. NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. *GigaScience*, 11:giac037, 05 2022.

[63] Alex Clark. Pillow (pil fork) documentation, 2015.

[64] Lee Raymond Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, July 1945.

[65] T. Sørensen. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. Biologiske skrifter. Munksgaard in Komm., 1948.

[66] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation, 2022.

[67] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, 2015.

[68] Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021.

[69] Araon Fenster and Bernard Chiu. Evaluation of segmentation algorithms for medical imaging. In *2005 IEEE engineering in medicine and biology 27th annual conference*, pages 7186–7189. IEEE, 2006.

[70] Annika Reinke, Minu D. Tizabi, Carole H. Sudre, Matthias Eisenmann, Tim Rädsch, Michael Baumgartner, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Peter Bankhead, Arriel Benis, Matthew Blaschko, Florian Buettner, M. Jorge Cardoso, Jianxu Chen, Veronika Cheplygina, Evangelia Christodoulou, Beth Cimini, Gary S. Collins, Sandy Engelhardt, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Ben Glocker, Patrick Godau, Robert Haase, Fred Hamprecht, Daniel A. Hashimoto, Doreen Heckmann-Nötzel, Peter Hirsch, Michael M. Hoffman, Merel Huisman, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, A. Emre Kavur, Hannes Kenngott, Jens Kleesiek, Andreas Kleppe, Sven Kohler, Florian Kofler, Annette Kopp-Schneider, Thijs Kooi, Michal Kozubek, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, David Moher, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, M. Alican Noyan,

Jens Petersen, Gorkem Polat, Susanne M. Rafelski, Nasir Rajpoot, Mauricio Reyes, Nicola Rieke, Michael Riegler, Hassan Rivaz, Julio Saez-Rodriguez, Clara I. Sánchez, Julien Schroeter, Anindo Saha, M. Alper Selver, Lalith Sharan, Shravya Shetty, Maarten van Smeden, Bram Stieltjes, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, Manuel Wiesenfarth, Ziv R. Yaniv, Paul Jäger, and Lena Maier-Hein. Common limitations of image processing metrics: A picture story, 2023.

[71] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

[72] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[73] Sebastian Ruder. An overview of gradient descent optimization algorithms, 2017.

[74] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[75] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[76] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, pages 1–7. IEEE, 2020.

[77] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*, page 240–248. Springer International Publishing, 2017.

5. APÉNDICE

Dataset	Correlation			MAE		
	Universeg	SAM2	Atlas	Universeg	SAM2	Atlas
SCD	**0.71**	0.65	0.66	**0.11**	0.14	0.11
HC18	0.81	0.84	**0.85**	0.14	0.14	**0.11**
PH2	**0.84**	0.71	0.82	0.14	0.14	**0.11**
3D-IRCAdB	**0.93**	0.85	0.86	**0.11**	0.15	0.14
NuCLS	0.51	**0.73**	0.55	0.21	**0.18**	0.20
PSFHS (PS)	0.86	0.90	**0.94**	0.17	0.14	**0.07**
PSFHS (FH)	0.91	0.86	**0.94**	0.12	0.10	**0.08**
JSRT (Lung)	0.83	0.87	**0.94**	0.08	0.03	**0.02**
JSRT (Heart)	0.90	0.91	**0.93**	0.07	0.07	**0.06**

Fig. 5.1: Comparación de la correlación y el MAE obtenidos al estimar el Dice score en todos los datasets para cada modelo.

Dataset	Correlation			MAE		
	Universeg	SAM2	Atlas	Universeg	SAM2	Atlas
SCD	0.75	0.36	**0.89**	**14.80**	38.08	16.66
HC18	0.01	**0.14**	0.07	70.12	80.65	**60.57**
PH2	0.31	**0.42**	0.19	59.57	67.96	**44.29**
3D-IRCAdB	0.79	0.74	**0.84**	24.95	30.59	**22.72**
NuCLS	0.20	**0.27**	0.14	59.36	55.96	**55.06**
PSFHS (PS)	0.16	0.28	**0.86**	40.62	108.50	**26.01**
PSFHS (FH)	0.77	0.52	**0.82**	**18.61**	51.77	22.66
JSRT (Lung)	0.46	0.50	**0.79**	**16.16**	25.36	17.17
JSRT (Heart)	0.54	0.72	**0.92**	34.80	43.44	**22.42**

Fig. 5.2: Comparación de la correlación y el MAE obtenidos al estimar la distancia Hausdorff en todos los datasets para cada modelo.

Dataset	Correlation			MAE		
	Universeg	SAM2	Atlas	Universeg	SAM2	Atlas
SCD	0.62	0.63	**0.89**	**4.77**	9.64	4.87
HC18	0.85	**0.91**	0.89	**6.40**	18.52	15.76
PH2	**0.81**	0.48	0.61	**8.22**	30.89	19.35
3D-IRCAdB	**0.91**	0.84	0.88	**7.05**	10.31	10.29
NuCLS	0.22	0.21	**0.26**	18.48	18.80	**18.09**
PSFHS (PS)	0.68	0.52	**0.97**	18.78	13.39	**4.54**
PSFHS (FH)	0.81	0.83	**0.89**	5.37	7.21	**3.87**
JSRT (Lung)	0.75	0.69	**0.87**	3.86	2.23	**1.46**
JSRT (Heart)	0.91	0.87	**0.95**	11.02	7.61	**4.77**

Fig. 5.3: Comparación de la correlación y el MAE obtenidos al estimar la distancia ASSD en todos los datasets para cada modelo.

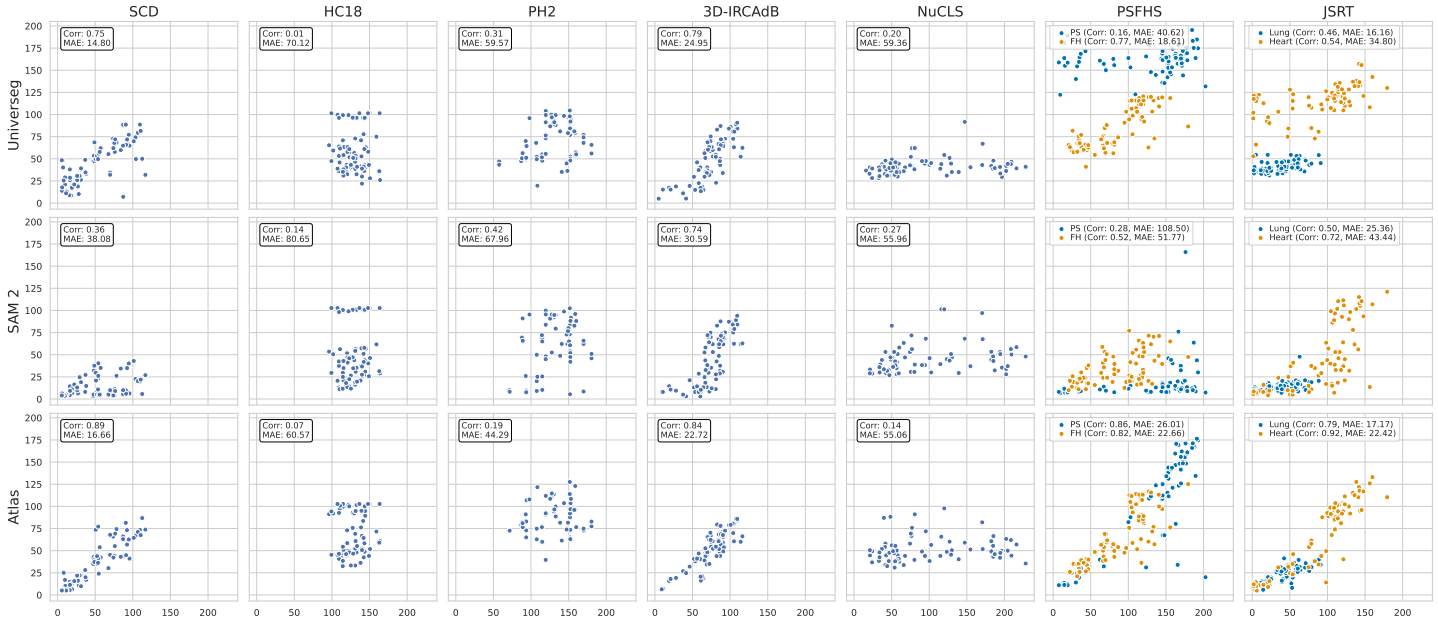

Fig. 5.4: Comparación de los resultados obtenidos al estimar la distancia Hausdorff en todos los datasets para cada modelo. En el eje horizontal se encuentra el valor real mientras que en el vertical el estimado. En todo caso se utilizaron las 16 imágenes más similares como conjunto de referencia y se evaluaron un total de 75 imágenes. Además, se incluye la correlación (Corr) y el error absoluto medio (MAE) entre los valores reales y estimados para cada una de las clases a segmentar.

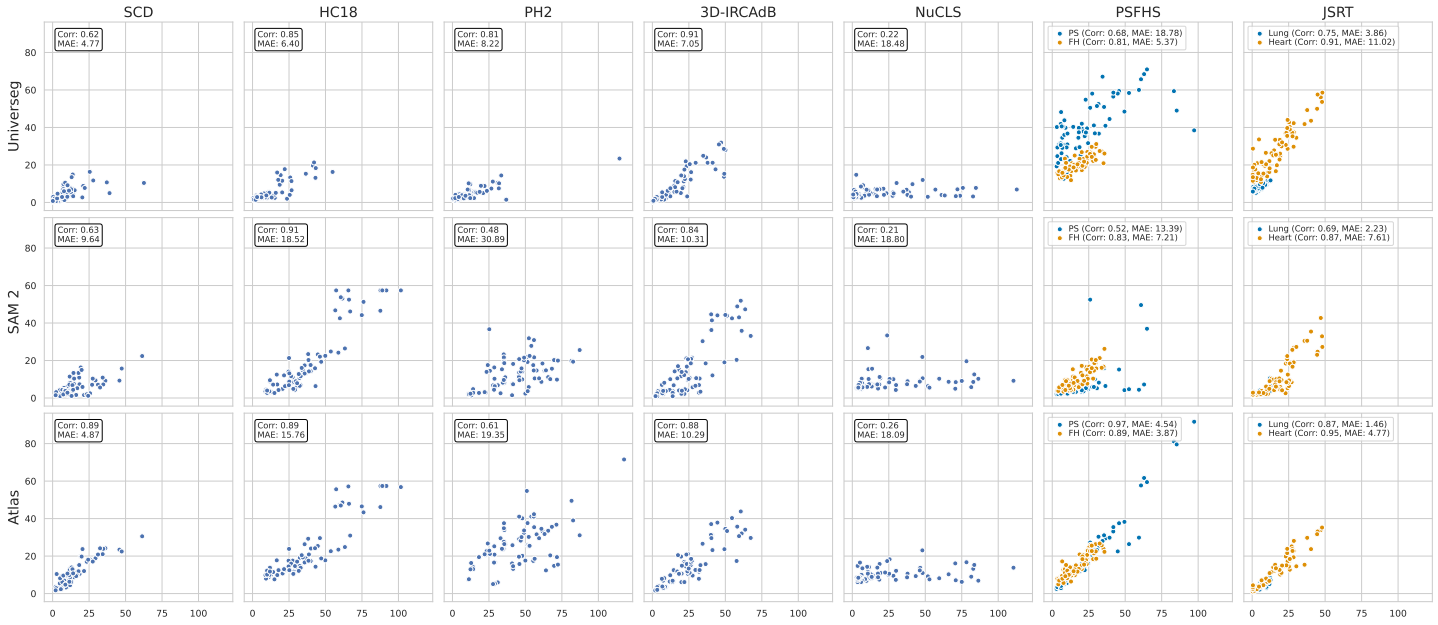

Fig. 5.5: Comparación de los resultados obtenidos al estimar la distancia ASSD en todos los datasets para cada modelo. En el eje horizontal se encuentra el valor real mientras que en el vertical el estimado. En todo caso se utilizaron las 16 imágenes más similares como conjunto de referencia y se evaluaron un total de 75 imágenes. Además, se incluye la correlación (Corr) y el error absoluto medio (MAE) entre los valores reales y estimados para cada una de las clases a segmentar.