UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Sistemas Informáticos y Computación

Segmentación automática de estructuras cardíacas en imágenes médicas usando aprendizaje profundo

Trabajo Fin de Máster

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

AUTOR/A: Marhuenda Tendero, Luis Jesús

Tutor/a: Monserrat Aranda, Carlos

CURSO ACADÉMICO: 2023/2024

Universidad Politécnica de Valencia

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Segmentación automática de estructuras cardíacas en imágenes médicas usando aprendizaje profundo

Luis Jesús Marhuenda Tendero

2024

Segmentación automática de estructuras cardíacas en imágenes médicas usando aprendizaje profundo

Autor
Luis Jesús Marhuenda Tendero

Tutor
Carlos Monserrat Aranda
Sistemas Informáticos y Computación

VALENCIA, Septiembre 2024

Resumen

La detección precisa de estructuras cardíacas a partir de imágenes médicas es esencial para el diagnóstico preciso y oportuno de enfermedades coronarias, lo que puede mejorar significativamente la atención médica y la calidad de vida de los pacientes. Este Trabajo de Fin de Máster se centra en el desarrollo de un modelo de aprendizaje profundo que pueda llevar a cabo la identificación semántica de estas estructuras con precisión y eficiencia. Para lograr este objetivo, se realizará un exhaustivo entrenamiento de diversos modelos, optimizando sus configuraciones según las recomendaciones y hallazgos de la literatura especializada en este ámbito. La comparación entre la propuesta desarrollada y los modelos preexistentes permitirá destacar las fortalezas y debilidades de la solución propuesta, proporcionando así una evaluación integral de su rendimiento y utilidad clínica potencial.

Dada la escasez de muestras etiquetadas disponibles para el entrenamiento, se explorará la viabilidad de emplear técnicas de aumento de datos y aprendizaje por transferencia para mejorar la capacidad predictiva del modelo y su generalización a diferentes conjuntos de datos. Este análisis exhaustivo de enfoques de mejora de datos y transferencia de conocimientos servirá para enriquecer la comprensión sobre la idoneidad y eficacia de estas estrategias en el contexto específico de la detección de estructuras cardíacas en imágenes médicas. En conjunto, este trabajo busca contribuir al avance de la investigación en el campo de la detección y diagnóstico de enfermedades cardíacas mediante el uso de técnicas de aprendizaje automático y análisis de imágenes médicas.

Palabras clave: Aprendizaje profundo · Segmentación de imágenes · Imagenología médica

Abstract

Accurate detection of cardiac structures from medical images is essential for accurate and timely diagnosis of coronary heart disease, which can significantly improve medical care and patients' quality of life. This Master Thesis focuses on the development of a deep learning model that can perform semantic identification of these structures accurately and efficiently. To achieve this goal, an exhaustive training of different models will be carried out, optimising their configurations according to the recommendations and findings of the specialised literature in this field. The comparison between the developed proposal and pre-existing models will highlight the strengths and weaknesses of the proposed solution, thus providing a comprehensive assessment of its performance and potential clinical utility.

Given the paucity of labelled samples available for training, the feasibility of employing data augmentation and transfer learning techniques to improve the predictive capability of the model and its generalisability to different datasets will be explored. This comprehensive analysis of data enhancement and knowledge transfer approaches will serve to enrich the understanding of the suitability and effectiveness of these strategies in the specific context of cardiac structure detection in medical imaging. Overall, this work aims to contribute to the advancement of research in the field of cardiac imaging.

Keywords: Deep learning · Image segmentation · Medical imaging

Resum

La detecció precisa d'estructures cardíaques a partir d'imatges mèdiques és essencial per a un diagnòstic precís i oportú de la cardiopatia coronària, que pot millorar significativament l'atenció mèdica i la qualitat de vida dels pacients. Aquest Treball de Fi de Màster se centra en el desenvolupament d'un model d'aprenentatge profund que pugui realitzar la identificació semàntica d'aquestes estructures de manera precisa i eficient. Per assolir aquest objectiu, es durà a terme una formació exhaustiva de diferents models, optimitzant les seves configuracions segons les recomanacions i troballes de la literatura especialitzada en aquest àmbit. La comparació entre la proposta desenvolupada i els models preexistents posarà de manifest els punts forts i febles de la solució proposada, proporcionant així una avaluació integral del seu rendiment i potencial utilitat clínica.

Donada l'escassetat de mostres etiquetades disponibles per a l'entrenament, s'explorarà la viabilitat d'utilitzar tècniques d'augmentació i transferència de dades per millorar la capacitat predictiva del model i la seva generalització a diferents conjunts de dades. Aquesta anàlisi exhaustiva de les aproximacions de millora de dades i transferència de coneixement servirà per enriquir la comprensió de la idoneïtat i l'eficàcia d'aquestes estratègies en el context específic de la detecció d'estructura cardíaca en imatge mèdica. En general, aquest treball pretén contribuir a l'avanç de la recerca en el camp de la imatge cardíaca.

Paraules clau: Aprenentatge profund · Segmentació d'imatges · Imagenologia mèdica

Agradecimientos

Quiero expresar mi profundo agradecimiento a todas las personas que han contribuido de manera significativa a la realización de este trabajo.

En primer lugar, agradezco enormemente el apoyo incondicional de mi familia. Especialmente a mis padres, quienes han depositado su confianza en mí durante estos últimos años, permitiéndome seguir mi camino según mis deseos. También quiero enviar mis mejores deseos a mi hermano, quien continúa su trayectoria estudiantil; su apoyo ha sido fundamental.

Agradezco sinceramente a mis amigos por brindarme su compañía y apoyo, aliviando la carga mental que conlleva el día a día. Para mí, son parte de mi familia y su respaldo es invaluable.

Asimismo, quiero reconocer y agradecer el compromiso de todos los profesores que he tenido a lo largo de mi formación. Su dedicación a la enseñanza es inspiradora, y espero que continúen compartiendo su sabiduría durante muchos años más. Especialmente, agradezco a los profesores de este máster por su excelencia y aliento. Tampoco puedo dejar de mencionar a mi tutor, Carlos, quien confió en mí para llevar a cabo este proyecto. A pesar de los desafíos, hemos logrado sacarlo adelante, y estoy profundamente agradecido por su guía y confianza en mí.

A todas estas personas, mi más sincero agradecimiento. Sin su apoyo y aliento, este trabajo no habría sido posible.

En este curso se me ha otorgado una beca para estudios de máster por parte de *ValgrAI - Valencian Graduate Shcool and Research Network of Artificial Intelligence*. Es por ello que quiero expresar mi profundo agradecimiento tanto a esta institución como a la propia Generalitat Valenciana por la concesión de la misma.

valgrAI Valencian Graduate School and Research Network of Artificial Intelligence

GENERALITAT VALENCIANA

El genio se hace con un 1% de talento
y con un 99% de trabajo

Albert Einstein.

Preámbulo

"Hace exactamente un año, me embarqué en la redacción de un texto similar para mi Trabajo de Fin de Grado. En aquella ocasión, mi enfoque se centró en la clasificación de acciones domésticas a partir de vídeos, con el objetivo de extraer métricas que pudieran indicar signos de deterioro cognitivo. A pesar de los desafíos que enfrenté, abordé el proyecto con determinación y sin temor. Hoy, un año después, me encuentro nuevamente frente a un nuevo desafío académico con la misma confianza y compromiso.

La elección de este Trabajo de Fin de Máster sigue la misma línea de mi trabajo anterior: contribuir al campo de la medicina para mejorar la labor de los profesionales de la salud. En un mundo donde la tecnología y la ciencia convergen para crear soluciones innovadoras, me motiva profundamente utilizar mis habilidades y conocimientos para apoyar el avance de la medicina. Este proyecto no solo representa una oportunidad para ampliar mi experiencia y conocimientos en el ámbito de la salud, sino también un compromiso personal con la mejora continua y el servicio a la sociedad.

Al igual que en mi trabajo anterior, enfrento este desafío con determinación y entusiasmo, consciente del impacto positivo que puede tener en la comunidad médica y, en última instancia, en la calidad de vida de las personas."

Índice general

Índice de figuras

Índice de cuadros

1 Introducción

La medicina ha experimentado una evolución sin precedentes a lo largo de la historia, impulsada por descubrimientos científicos, avances tecnológicos y cambios en las prácticas médicas. Desde los primeros métodos rudimentarios de diagnóstico y tratamiento hasta las complejas técnicas y terapias actuales, el campo de la medicina ha sido testigo de una transformación continua.

En los últimos años, el papel de la tecnología, especialmente la Inteligencia Artificial (IA), ha revolucionado aún más la práctica médica. La IA ha demostrado ser una herramienta invaluable en el diagnóstico, tratamiento y gestión de enfermedades, permitiendo análisis más precisos, pronósticos más certeros y una atención personalizada a los pacientes.

Una de las áreas donde la IA ha dejado una marca significativa es en la segmentación de imágenes médicas mediante técnicas de Deep Learning (DL). Esta modalidad se ha convertido en una nueva moda, impulsada por la capacidad de los algoritmos de DL para analizar grandes conjuntos de datos de imágenes médicas y extraer características relevantes con una precisión sin precedentes. La segmentación de imágenes con DL ha demostrado ser especialmente útil en la identificación y delimitación de regiones de interés en imágenes médicas, como tomografías computarizadas, resonancias magnéticas y ecografías, facilitando así el diagnóstico y el tratamiento de diversas enfermedades.

Este avance en la segmentación de imágenes con DL está transformando la forma en que los médicos interpretan y utilizan la información visual en el ámbito clínico. Desde la detección temprana de anomalías hasta la planificación de intervenciones quirúrgicas precisas, la segmentación de imágenes con DL promete revolucionar aún más la práctica médica, mejorando la precisión diagnóstica, la eficiencia del tratamiento y, en última instancia, la atención al paciente. En este sentido, explorar y comprender el potencial de esta tecnología emergente es fundamental para aprovechar al máximo sus beneficios en la mejora de la salud y el bienestar de las personas.

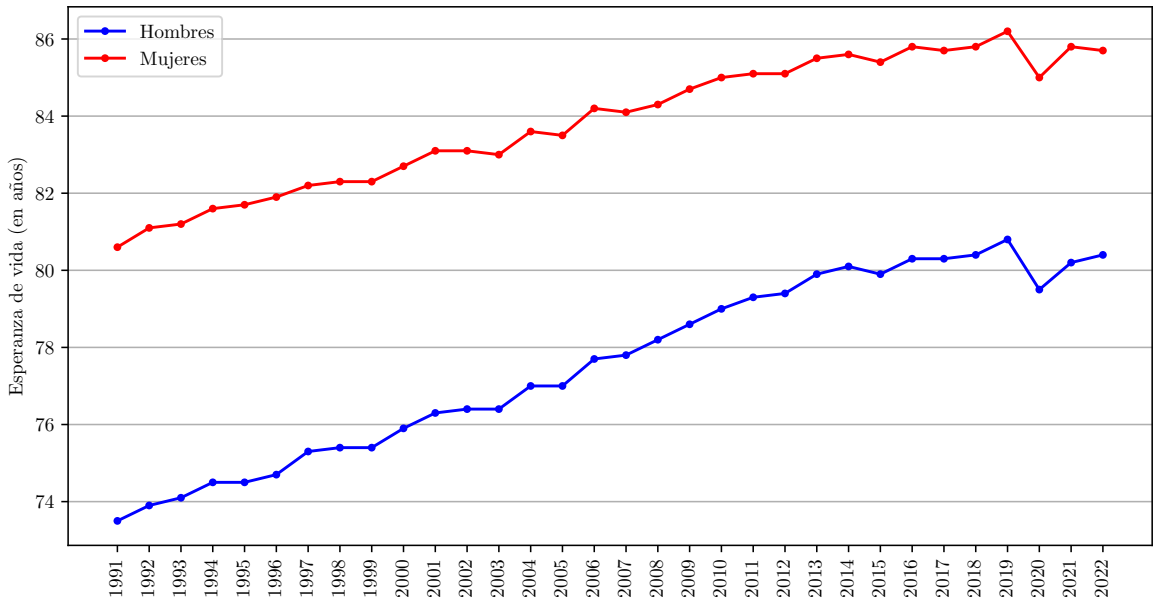
En este contexto de rápida evolución tecnológica y transformación en la práctica médica, es crucial examinar la intersección entre la historia de la medicina y su impacto en la sociedad actual. En los próximos apartados, se explorará cómo los avances históricos han sentado las bases para los desarrollos contemporáneos en el campo de la salud, destacando cómo los conocimientos y técnicas heredadas continúan influyendo en la medicina moderna y se analizará más a fondo cómo las innovaciones tecnológicas, en particular la IA y el DL, están transformando radicalmente la práctica médica. Se examinarán casos de estudio y ejemplos concretos que ilustran cómo estas tecnologías están revolucionando el diagnóstico, tratamiento y gestión de enfermedades, así como su impacto en la relación médico-paciente y en la atención

sanitaria en general.

1.1 Evolución de la esperanza de vida

De la mano del Instituto Nacional de Estadística (INE) se obtienen los datos contenidos en la siguiente figura y cuadro, que recogen la evolución de la esperanza de vida que ha sufrido España en las últimas tres décadas. El gráfico presenta una comparativa entre géneros donde se ve la proyección que se va desarrollando año tras año.

Figura 1.1: Esperanza de vida al nacer según el sexo en España



La figura anterior muestra un notable aumento en la esperanza de vida al nacer en los últimos años, tanto para hombres como para mujeres. Para los hombres, este incremento representa un aumento de más de 3 años, mientras que para las mujeres, el aumento es ligeramente superior a dos años. Es importante tener en cuenta que, inicialmente, las mujeres ya contaban con una esperanza de vida más alta en comparación con los hombres.

Del mismo modo que se presenta la esperanza de vida desde el nacimiento, se puede hacer dependiendo de la edad. También de la mano del INE, se puede observar este aumento en la siguiente tabla:

Cuadro 1.1: Esperanza de vida a diferentes edades desde 2004 a 2021 en España

Edad	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Al nacer	80.3	80.3	80.9	81.0	81.3	81.7	82.1	82.3	82.3	82.8	82.9	82.7	83.1	83.1	83.2	83.6	82.3	83.1
10 años	70.7	70.7	71.3	71.3	71.7	72.0	72.4	72.6	72.6	73.1	73.2	73.0	73.4	73.4	73.5	73.9	72.6	73.3
20 años	60.9	60.8	61.5	61.5	61.8	62.1	62.5	62.7	62.7	63.2	63.3	63.1	63.5	63.5	63.6	63.9	62.7	63.4

Continúa en la siguiente página

Cuadro 1.1: Esperanza de vida a diferentes edades desde 2004 a 2021 en España (*continuación*)

Edad	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
30 años	51.2	51.1	51.7	51.7	52.0	52.3	52.7	52.9	52.9	53.3	53.5	53.2	53.6	53.6	53.7	54.1	52.8	53.6
40 años	41.6	41.5	42.1	42.1	42.3	42.6	43.0	43.1	43.1	43.5	43.7	43.4	43.9	43.8	43.9	44.3	43.1	43.8
50 años	32.3	32.2	32.8	32.7	33.0	33.3	33.6	33.7	33.7	34.1	34.2	33.9	34.3	34.3	34.4	34.8	33.5	34.3
60 años	23.5	23.4	24.0	23.9	24.2	24.4	24.7	24.9	24.8	25.2	25.3	25.0	25.4	25.4	25.4	25.8	24.6	25.3
70 años	15.4	15.3	15.9	15.8	16.0	16.2	16.5	16.6	16.6	17.0	17.1	16.8	17.2	17.1	17.2	17.5	16.4	17.2
80 años	8.7	8.5	9.0	8.9	9.0	9.2	9.4	9.5	9.4	9.8	9.8	9.5	9.9	9.8	9.9	10.2	9.3	10.0
90 años	4.2	4.1	4.3	4.3	4.3	4.5	4.6	4.6	4.5	4.8	4.8	4.5	4.8	4.7	4.7	4.9	4.4	4.8

En la tabla anterior se presenta la esperanza de vida según la edad actual. Al comparar los valores, se observa que en 2004, a los 10 años, la esperanza de vida era de 70.7 años, mientras que en 2021, para la misma edad, se registra un incremento en la esperanza de vida de 2.6 años, alcanzando así un valor mayor. Y es que en las últimas dos décadas, la medicina ha experimentado un vertiginoso avance que ha tenido un impacto significativo en la esperanza de vida a nivel mundial. Gracias a una combinación de descubrimientos científicos, innovaciones tecnológicas y mejoras en la atención médica, se ha presenciado un aumento notable en la longevidad y calidad de vida de la población.

La implementación de nuevas terapias farmacológicas, como los medicamentos biológicos y las terapias génicas, ha revolucionado el tratamiento de enfermedades crónicas como el cáncer, las enfermedades cardiovasculares y las enfermedades autoinmunes, permitiendo a los pacientes gestionar mejor sus condiciones y prolongar su supervivencia. Además, los avances en técnicas quirúrgicas mínimamente invasivas han reducido los riesgos asociados con procedimientos médicos importantes, mientras que la mejora en los métodos de diagnóstico, como la imagenología[1] de alta resolución y los biomarcadores[2], ha permitido una detección temprana y un tratamiento más efectivo de enfermedades graves. En conjunto, estos avances han contribuido a un aumento sustancial en la esperanza de vida en las últimas dos décadas, mejorando la salud y el bienestar de millones de personas en todo el mundo.

Sin embargo, a pesar de estos logros, persisten desafíos importantes, como las desigualdades en el acceso a la atención médica y el creciente impacto de enfermedades no transmisibles, lo que destaca la necesidad continua de innovación y colaboración en el campo de la medicina para abordar los desafíos de salud actuales y futuros.

1.2 Inteligencia Artificial en la medicina

Con la llegada de la IA, se ha producido una revolución sin precedentes en la forma en que la medicina se diagnostica, trata y gestiona. La IA ha demostrado ser una herramienta invaluable en el campo de la salud, ofreciendo capacidades de análisis de datos y patrones que sobrepasan las habilidades humanas en muchos aspectos. En la medicina, la IA ha encontrado numerosas aplicaciones, desde la interpretación de imágenes médicas hasta el desarrollo de

[1] En el ámbito de la medicina, estudio y utilización clínica de las imágenes producidas por los rayos X, el ultrasonido, la resonancia magnética, etc.

[2] Sustancia que indica la presencia de material biológico o de un proceso fisiológico, y que se emplea para diagnosticar una enfermedad

terapias personalizadas.

Uno de los avances más significativos ha sido en el ámbito del diagnóstico médico. Los algoritmos de IA pueden analizar grandes conjuntos de datos de imágenes médicas, como Tomografía Computarizada (TC), Resonancia Magnética (RM) y radiografías, para detectar patrones y anomalías que podrían pasar desapercibidas para el ojo humano. Esta capacidad ha llevado a una mejora sustancial en la precisión y rapidez del diagnóstico, permitiendo a los médicos identificar enfermedades en etapas tempranas cuando son más tratables.

Además del diagnóstico, la IA también está transformando el tratamiento médico. Los sistemas de IA pueden analizar datos genómicos y clínicos de pacientes para identificar terapias personalizadas y predecir la eficacia de diferentes tratamientos. Esto permite una atención médica más precisa y efectiva, reduciendo los riesgos y efectos secundarios asociados con tratamientos estándar.

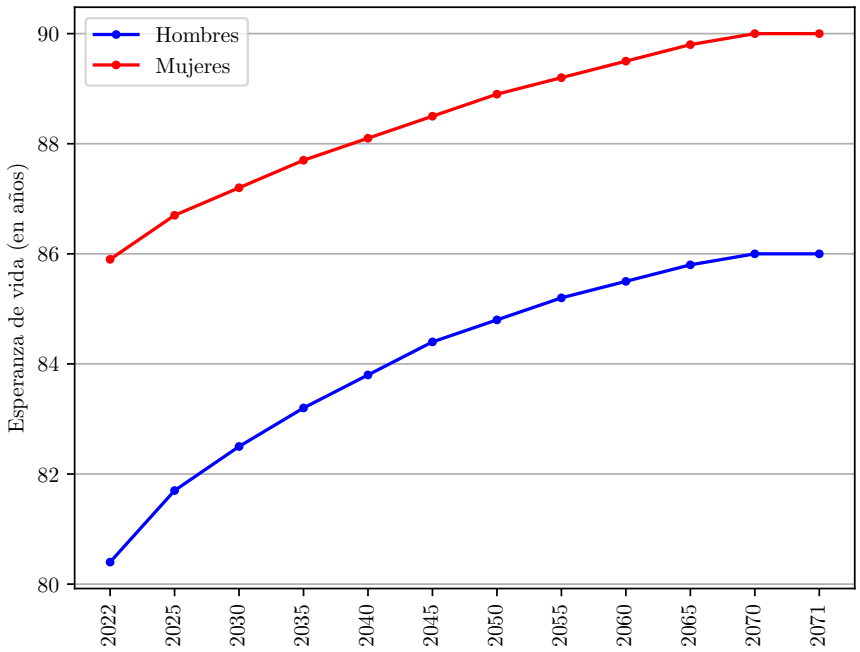
Otra área en la que la IA está teniendo un impacto significativo es en la gestión de la salud. Los sistemas de IA pueden analizar grandes cantidades de datos de pacientes para identificar tendencias y patrones que podrían indicar un mayor riesgo de enfermedad o complicaciones. Esto permite a los proveedores de atención médica intervenir de manera proactiva, ofreciendo intervenciones preventivas y gestionando mejor los recursos sanitarios.

En resumen, la conexión entre las nuevas tecnologías, especialmente la IA, y la medicina ha abierto un nuevo horizonte de posibilidades en el campo de la salud. Desde el diagnóstico hasta el tratamiento y la gestión de la salud, la IA está transformando la forma en que se practica la medicina, mejorando la precisión, eficiencia y accesibilidad de la atención médica para pacientes en todo el mundo.

Proyección de la esperanza de vida en los próximos 50 años

Como ya se ha comentado anteriormente, la esperanza de vida es un indicador clave que refleja la calidad de vida y el estado de salud de una población. En el contexto de la evolución de la medicina y el impacto de las nuevas tecnologías la figura proporcionada por el INE ofrece una visión prospectiva de la esperanza de vida en España para los próximos 50 años, brindando información sobre las tendencias demográficas y el impacto potencial de los avances médicos y tecnológicos en la longevidad de la población.

Figura 1.2: Proyección de la esperanza de vida al nacer según el sexo en España para los próximos 50 años



La figura destaca un aumento significativo para los hombres de casi 6 años y una proyección de que las mujeres puedan llegar a alcanzar una esperanza de vida al nacer de 90 años. Como se ha ido diciendo, estas cifras reflejan no solo reflejan los cambios en el estilo de vida y las condiciones socioeconómicas, si no también los avances en la medicina, principalmente gracias a la IA.

1.3 Motivación

El presente Trabajo de Fin de Máster (TFM) nace de mutuo acuerdo con el tutor Carlos Monserrat y el motivo principal por el que se ha optado por realizar este proyecto es el gran interés y deseo de adentrarse en el mundo de la medicina desde un punto de vista informático. El marco de este trabajo parte de un proyecto Fondos de Investigaciones Sanitarias (FIS) del Hospital La Fe (Valencia, España) en el ámbito de la cirugía hepática. Desde dicho ámbito, el hospital se encuentra en la situación de analizar biomarcadores del corazón, debido a que muchos trasplantes de hígado acaban con fallos cardíacos y se quiere analizar el porqué con el fin de analizar si merecerá la pena dicho trasplante o por el contrario, no.

1.4 Estructuración

La estructura del presente TFM ha sido cuidadosamente diseñada para abordar de manera exhaustiva el estudio de la segmentación automática de estructuras cardíacas en imágenes médicas mediante el uso de técnicas de DL. A continuación, se presenta un resumen de

cada capítulo, destacando su enfoque y contribuciones específicas. Esta organización no solo facilita la comprensión del flujo y la cohesión del trabajo realizado sino que también subraya la rigurosidad y la profundidad del análisis llevado a cabo.

1. **Introducción**: Este capítulo establece el contexto del estudio, destacando la importancia de las técnicas de DL en la detección precisa de estructuras cardíacas para diagnósticos médicos eficientes.

2. **Marco Teórico**: Se detallan los fundamentos teóricos de la Visión Por Computador (VPC) y técnicas de segmentación de imágenes, proporcionando una base sólida para el entendimiento de las tecnologías empleadas.

3. **Estado del Arte**: Revisa los desarrollos recientes en segmentación de imágenes médicas, comparando el trabajo actual con otros esfuerzos en el campo y resaltando la contribución única del TFM.

4. **Metodología**: Describe las herramientas y técnicas utilizadas, incluyendo software y métodos de evaluación de modelos de aprendizaje profundo, esencial para entender el desarrollo práctico del estudio.

5. **Conjuntos de Datos**: Explica las características de los datos utilizados, incluyendo su procedencia y estructura, subrayando su importancia para los resultados del estudio.

6. **Experimentación con Modelos de Segmentación**: Presenta los resultados experimentales, analizando el rendimiento de los modelos desarrollados en aplicaciones médicas.

7. **Extracción de Biomarcadores**: Discute la utilización de los modelos para extraer biomarcadores de las imágenes, con implicaciones directas en diagnóstico y tratamiento de enfermedades.

8. **Conclusiones y Trabajo Futuro**: Evalúa el trabajo realizado, sus logros y limitaciones, y sugiere áreas para futuras investigaciones.

1.5 Objetivos

Los objetivos de un TFM son fundamentales para establecer el alcance y las metas específicas que guiarán la investigación. Los objetivos delinean claramente las intenciones del estudio y proporcionan un marco para evaluar los resultados y el impacto del trabajo realizado. En este apartado, se describen las metas concretas que se pretenden alcanzar mediante la aplicación de técnicas de DL para la segmentación de estructuras cardíacas en imágenes médicas. Estos objetivos no solo reflejan las aspiraciones académicas y técnicas del proyecto, sino que también destacan su relevancia práctica y potencial contribución al avance del diagnóstico y tratamiento en el ámbito de la medicina cardiovascular. Cada objetivo ha sido cuidadosamente diseñado para asegurar que el proyecto sea exhaustivo, innovador y de gran impacto en el campo de la imagenología médica y la IA.

1. **Realizar una revisión exhaustiva de la literatura actual sobre técnicas de segmentación cardíaca**. El primer objetivo consiste en realizar un estudio detallado de la literatura existente para comprender las metodologías actuales, los avances recientes y las limitaciones en el campo de la segmentación cardíaca utilizando DL. Este análisis crítico ayudará a identificar las brechas de conocimiento y las oportunidades de innovación, proporcionando una base sólida para el desarrollo del proyecto.

2. **Optimizar modelos de DL para la segmentación de estructuras cardíacas**. Centrarse en la optimización de modelos avanzados de aprendizaje profundo que sean capaces de segmentar de manera eficiente las estructuras cardíacas en diversas modalidades de imágenes médicas. Esto incluirá la experimentación con diferentes hiperparámetros, técnicas de regularización y ajuste fino de modelos pre-entrenados para mejorar su precisión y capacidad de generalización en conjuntos de datos clínicos.

3. **Validar y Comparar la Eficacia de los Modelos Optimizados con los Estándares de Oro y Métodos Tradicionales**. Evaluar exhaustivamente los modelos optimizados comparándolos entre sí y el estado del arte.

4. **Extracción de Biomarcadores a Partir de las Imágenes Segmentadas**. El cuarto objetivo es utilizar los modelos de segmentación para extraer biomarcadores relevantes de las imágenes cardíacas. Esta extracción busca proporcionar datos cuantitativos que puedan ser utilizados para un diagnóstico más preciso y personalizado, así como para la evaluación del riesgo de enfermedades cardíacas. Esto implica no solo identificar características morfológicas de las estructuras cardíacas sino también analizar cambios dinámicos y funcionales a lo largo del tiempo.

2 Marco teórico

El marco teórico desempeña un papel fundamental en cualquier investigación o proyecto, ya que proporciona el contexto necesario para comprender el estado actual del conocimiento en un área específica. En este caso particular, comprender la VPC y la segmentación de imágenes será esencial para establecer los fundamentos teóricos y conceptuales que sustentan estas disciplinas. En los siguientes apartados, se explorará en detalle el marco teórico relacionado con la VPC y la segmentación de imágenes, proporcionando una base sólida para comprender los principios, métodos y aplicaciones en estas áreas de estudio.

La VPC se refiere al campo de la IA que se centra en el desarrollo de algoritmos y técnicas para permitir a los ordenadores interpretar y comprender el contenido visual de las imágenes o vídeos, de manera similar a cómo lo hacen los seres humanos. Por otro lado, el DL es una subdisciplina del Machine Learning (ML) que se basa en redes neuronales artificiales con múltiples capas interconectadas para aprender representaciones jerárquicas de datos. A modo resumen, si la IA permite a los ordenadores pensar, la VPC permite ver.

La VPC opera de manera muy similar a la visión humana, aunque los humanos tienen una ventaja inicial. La vista humana se beneficia de toda una vida de experiencias y contexto para discernir objetos, estimar distancias, detectar movimientos y reconocer anomalías en una imagen. La VPC capacita a las máquinas para ejecutar estas tareas, pero lo hace en un tiempo considerablemente menor utilizando cámaras, datos y algoritmos en lugar de retinas, nervios ópticos y una corteza visual. Por ejemplo, un sistema entrenado para inspeccionar productos o supervisar activos de producción puede analizar miles de productos o procesos por minuto, identificando defectos o anomalías que podrían pasar desapercibidos para los seres humanos, lo que permite superar rápidamente las capacidades humanas en términos de velocidad y precisión.

La VPC requiere una gran cantidad considerable de datos y realiza análisis repetidos hasta detectar diferencias y, en última instancia, reconocer imágenes. Por ejemplo, para entrenar a un ordenador en el reconocimiento de animales, se necesita proporcionar una gran cantidad de imágenes de animales y objetos relacionados para que pueda aprender a distinguir y reconocerlos.

Principalmente, para lograr esta tarea se emplean dos tecnologías fundamentales: el ML, específicamente el DL y las Convolutional Neural Network (CNN). El ML utiliza modelos algorítmicos que permiten a un ordenador por sí mismo aprender el contexto de los datos visuales. Si se le suministran suficientes datos a través del modelo, el ordenador observará los datos y aprenderá a distinguir una imagen de otra. Estos algoritmos permiten que el ordenador aprenda de forma autónoma, en lugar de ser programada para reconocer una imagen.

Por otro lado, una CNN ayuda a un modelo de ML o DL a observar mediante la descomposición de las imágenes en píxeles a los que se les asignan etiquetas. Utilizando estas etiquetas, realiza convoluciones[1] y efectúa predicciones sobre lo que ve. La red neuronal lleva a cabo convoluciones y verifica la precisión de sus predicciones en iteraciones repetidas hasta que sus predicciones coinciden con la realidad, lo que permite reconocer o ver imágenes de manera similar a los seres humanos.

Al igual que un humano que observa una imagen a distancia, una CNN primero distingue los bordes más definidos y las formas simples, y luego rellena los detalles a medida que avanza en sus predicciones. Por otro lado, mientras que una CNN se utiliza para comprender imágenes individuales, una Recurrent Neural Network (RNN) se emplea de manera similar en aplicaciones de vídeo para ayudar a las computadoras a entender cómo se relacionan entre sí las imágenes en una secuencia de fotogramas.

La VPC abarca una amplia gama de tareas y aplicaciones, todas ellas centradas en permitir a las máquinas interpretar y comprender datos visuales. Algunas de las tareas más comunes en el campo de la VPC son: clasificación de imágenes, detección de objetos y segmentación. En este TFM, se enfocará específicamente en la segmentación de imágenes como una técnica fundamental de VPC y se analizará su aplicación en el ámbito médico. La segmentación de imágenes desempeña un papel crucial en la extracción de características y en la comprensión de la estructura visual de los objetos, lo que la convierte en un área de investigación relevante y de gran interés en este proyecto.

2.1 Segmentación de imágenes

La segmentación de imágenes es una técnica de VPC que fragmenta una imagen digital en grupos discretos de píxeles, conocidos como segmentos de imagen, con el objetivo de facilitar la detección de objetos y otras tareas afines. Esta división de los datos visuales complejos de una imagen en segmentos con formas específicas posibilita un procesamiento más rápido y avanzado.

Las técnicas de segmentación de imágenes varían desde análisis heurísticos simples e intuitivos hasta avanzadas implementaciones de DL. Los algoritmos convencionales de segmentación procesan características visuales como el color o el brillo de cada píxel para determinar los límites de los objetos y las áreas de fondo. Por otro lado, el ML, mediante conjuntos de datos anotados, se emplea para entrenar modelos que clasifiquen con precisión los objetos y regiones específicas en una imagen.

La segmentación de imágenes es un método altamente versátil y práctico en VPC, utilizado en diversos campos de la IA. Desde el diagnóstico médico con imágenes hasta la automatización de la conducción en vehículos autónomos, e incluso en la identificación de objetos en imágenes satelitales, la segmentación de imágenes desempeña un papel esencial en una amplia

[1] Operaciones matemáticas entre dos funciones para generar una tercera función

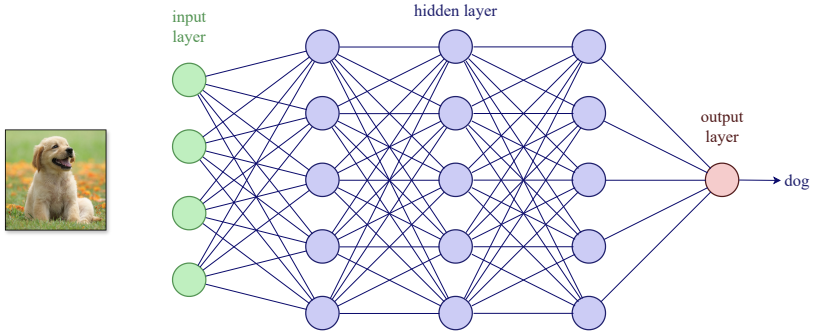
gama de aplicaciones.

2.1.1 Evolución de la detección de objetos y clasificación de imágenes

La segmentación de imágenes marca una evolución avanzada tanto en la clasificación de imágenes como en la detección de objetos, ofreciendo un conjunto de capacidades únicas en el ámbito de la VPC.

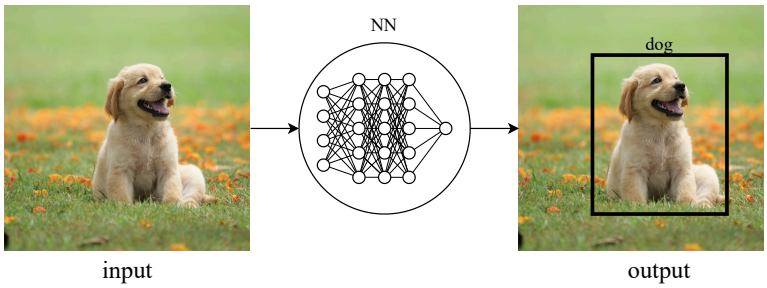
La clasificación de imágenes asigna una etiqueta de clase a toda la imagen. Por ejemplo, un modelo básico de clasificación de imágenes puede ser entrenado para etiquetar imágenes de animales como `perro` o `gato`. Los sistemas tradicionales de clasificación de imágenes son menos sofisticados, ya que no analizan las características de cada imagen de manera individualizada.

Figura 2.1: Red neuronal para clasificar entre perros y gatos



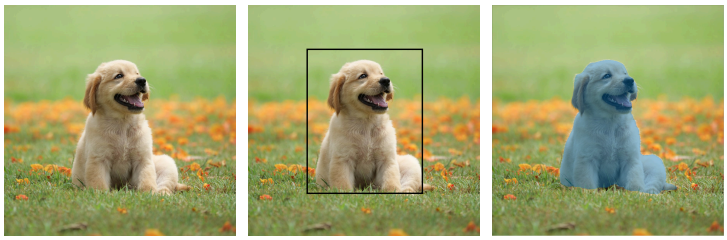
La detección de objetos fusiona la clasificación de imágenes con la localización de objetos al generar regiones rectangulares, conocidas como *bounding boxes*, que señalan la ubicación de los objetos. En lugar de simplemente etiquetar una imagen de un animal como `perro` o `gato`, un modelo de detección de objetos puede indicar dónde se encuentran específicamente los perros o gatos en la imagen. Sin embargo, aunque la detección de objetos puede clasificar múltiples elementos en una imagen y proporcionar aproximaciones de su anchura y altura, no puede precisar límites o formas con exactitud. Esta limitación afecta la capacidad de los modelos tradicionales de detección de objetos para delinear objetos estrechamente agrupados con cajas delimitadoras superpuestas.

Figura 2.2: Ejemplo de salida general tras la detección de objetos



La segmentación de imágenes implica el procesamiento de datos visuales a nivel de píxel, empleando diversas técnicas para etiquetar cada píxel como parte de una clase o instancia específica. Las técnicas clásicas de segmentación se basan en la análisis de propiedades intrínsecas de los píxeles, como el color y la intensidad (llamadas "heurísticas"), mientras que los modelos de DL utilizan redes neuronales complejas para un reconocimiento avanzado de patrones. El resultado de este proceso son las máscaras de segmentación, que representan de manera precisa, píxel a píxel, los límites y la forma de cada clase en la imagen, correspondiente a diferentes objetos, características o regiones.

Figura 2.3: Clasificación vs. Detección vs. Segmentación



(a) Clasificación **(b)** Detección **(c)** Segmentación

2.1.2 Tareas de la segmentación de imágenes

La diferencia principal entre cada tipo de tarea de segmentación de imágenes reside en como se trata a las clases semánticas: las categorías específicas a las que puede determinarse que pertenece un píxel dado. En el ámbito de la VPC, se distinguen dos tipos de clases semánticas, cada una de las cuales requiere enfoques específicos para lograr una segmentación precisa y eficiente.

Los elementos contables (*things*) son categorías con formas características, como `coche`, `árbol` o `persona`. Por lo general, estos objetos tienen instancias claramente definidas que se pueden contar. Su tamaño suele variar poco de una instancia a otra, y cada objeto tiene partes constituyentes distintas; por ejemplo: todos los coches tienen ruedas, pero una rueda por sí sola no constituye un coche.

Los elementos no contables (*stuff*) son categorías semánticas con formas amorfas y tamaños muy variables, como `cielo`, `agua` o `hierba`. Generalmente, estos elementos no tienen instancias individuales claramente definidas y contables. A diferencia de los objetos, los elementos no tienen partes diferenciadas: tanto una brizna de hierba como un campo de hierba se consideran `hierba`.

En ciertas imágenes, algunas categorías pueden ser interpretadas tanto elementos contables como no contables. Por ejemplo, un grupo numeroso de personas puede ser percibido como varias `personas`, cada una de las cuales es un elemento contable con una forma distinta, o como una `multitud` singular y sin forma aparente.

Aunque la mayoría de los esfuerzos de detección de objetos se centran principalmente en las

categorías de objetos, es crucial tener en cuenta que los elementos como el cielo, las paredes, el suelo o la tierra, constituyen la mayor parte del contexto visual. Los elementos son datos esenciales para identificar objetos, y viceversa: una superficie metálica en una carretera suele representar un coche; el fondo azul detrás de un barco probablemente sea agua, mientras que el fondo azul detrás de un avión seguramente sea cielo. Esto cobra especial relevancia para los modelos de DL.

Segmentación semántica

La segmentación semántica es el tipo más básico de segmentación de imágenes, donde un modelo asigna una clase semántica a cada píxel sin proporcionar otro contexto o información, como objetos individuales. En la segmentación semántica, todos los píxeles se consideran objetos; no se hace distinción entre objetos individuales y elementos.

Figura 2.4: Ejemplo de segmentación semántica

(a) Imagen original **(b)** Seg. semántica

Segmentación de instancias

La segmentación de instancias cambia el enfoque de la segmentación semántica: mientras que los algoritmos de segmentación semántica solo predicen la clasificación semántica de cada píxel (sin considerar las instancias individuales), la segmentación por instancias delinea la forma precisa de cada instancia de objeto de manera separada. La segmentación por instancias distingue los objetos unos de otros, a diferencia de la segmentación semántica, que los agrupa, y por lo tanto, puede considerarse como una evolución de la detección de objetos que produce máscaras de segmentación precisas en lugar de cajas delimitadoras aproximadas.

Es una tarea más desafiante que la segmentación semántica: incluso cuando los objetos de la misma clase se tocan o se superponen, los modelos de segmentación por instancias deben ser capaces de separar y determinar la forma de cada uno de ellos, mientras que los modelos de segmentación semántica pueden simplemente agruparlos.

Figura 2.5: Ejemplo de segmentación de instancias

(a) Imagen original **(b)** Seg. semántica **(c)** Seg. de instancias

Los algoritmos de segmentación de instancias generalmente abordan el problema en dos etapas o en una sola. Los modelos de dos etapas, como las Region-based Convolutional Neural Network (R-CNN) [Girshick y cols. (2014)], primero realizan una detección de objetos convencional para generar *bounding boxes* para cada instancia propuesta. Luego, realizan una segmentación y clasificación más refinadas dentro de cada *bounding box*. Por otro lado, los modelos *one-shot*, como You Only Look Once (YOLO) [Redmon y cols. (2016)], logran la segmentación de instancias en tiempo real al llevar a cabo la detección de objetos, la clasificación y la segmentación simultáneamente.

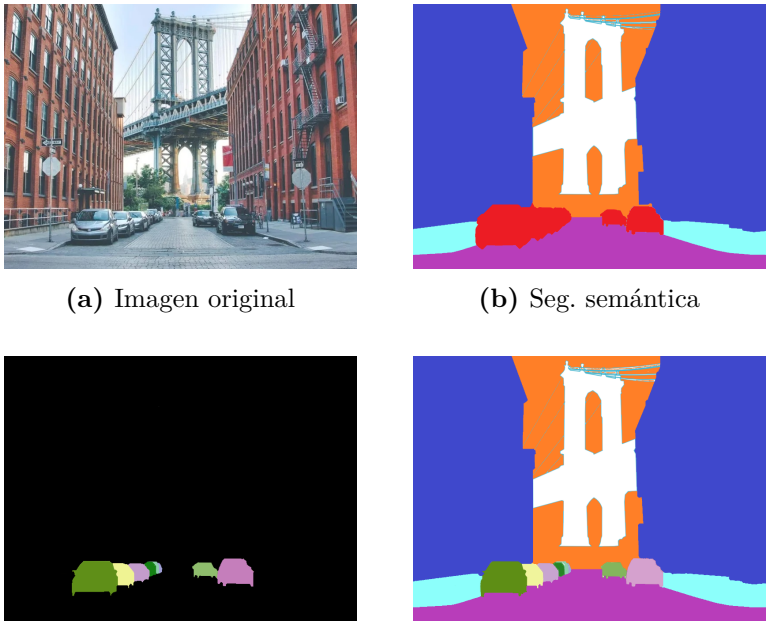
Los enfoques de un solo paso ofrecen una mayor velocidad a costa de una reducción en la precisión, mientras que los enfoques de dos pasos proporcionan una mayor precisión, aunque a expensas de una menor velocidad.

Segmentación panóptica

Los modelos de segmentación panóptica combinan tanto la clasificación semántica de todos los píxeles como la identificación de cada instancia de objeto en una imagen, aprovechando las fortalezas de la segmentación semántica y de la segmentación de instancias. En la tarea de segmentación panóptica, cada píxel debe etiquetarse con una clase semántica y un ID de instancia. Los píxeles con la misma etiqueta y el mismo ID pertenecen al mismo objeto, mientras que los píxeles clasificados como elementos no contables tienen un ID de instancia ignorado.

Así, la segmentación panóptica proporciona una comprensión integral y global de una imagen dada para los sistemas de visión por computadora. Aunque su potencial es evidente, lograr una segmentación panóptica de manera coherente y eficiente desde el punto de vista computacional representa un desafío considerable.

Figura 2.6: Ejemplo de segmentación panóptica



(a) Imagen original **(b)** Seg. semántica

(c) Seg. de instancias **(d)** Seg. panóptica

El desafío radica en reconciliar dos enfoques contradictorios: los modelos de segmentación semántica tratan todos los píxeles como elementos no contables, ignorando las instancias individuales de las cosas, mientras que los modelos de segmentación por instancias aíslan los elementos contables, dejando de lado los que no lo son. Ninguno de estos modelos puede cumplir adecuadamente las funciones del otro.

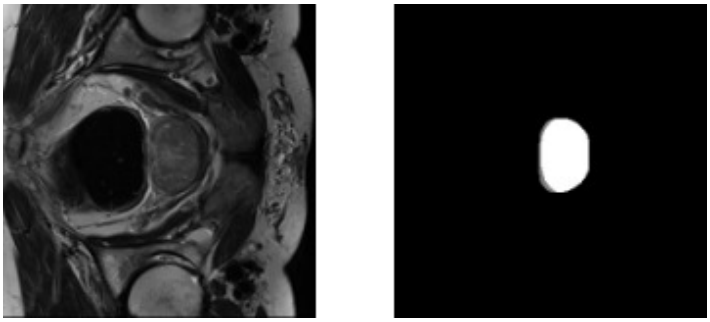
2.1.3 Casos de uso de la segmentación de imágenes

La segmentación de imágenes ha pasado a ser una herramienta fundamental en diversas áreas:

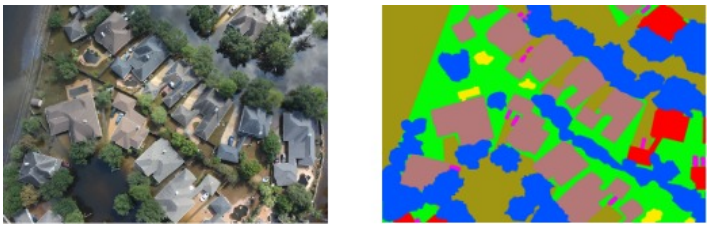
- **Imágenes médicas**.En radiografías, RM y TC, la segmentación de imágenes juega un papel crucial en la detección de tumores, la segmentación cerebral, el diagnóstico de enfermedades y la planificación quirúrgica.
- **Vehículos autónomos**. La segmentación de imágenes permite a los vehículos autónomos sortear obstáculos como peatones y otros vehículos, así como identificar carriles y señales de tráfico. Este proceso también se emplea en la navegación en robótica.
- **Imágenes por satélite**. Tanto la segmentación semántica como la de instancias automatizan la identificación de diferentes características del terreno y topográficas.
- **Ciudades inteligentes**. La segmentación de imágenes potencia tareas como el control y la vigilancia del tráfico en tiempo real.
- **Fabricación**. Además de las aplicaciones en robótica, la segmentación de imágenes facilita la clasificación de productos y la detección de defectos.

Figura 2.7: Ejemplos de casos de uso de la segmentación de imágenes

(a) Vehículos autónomos



(b) Imágenes médicas



(c) Imágenes por satélite

2.2 Técnicas basadas en aprendizaje profundo en segmentación de imágenes

Los modelos de segmentación basados en DL son entrenados con conjuntos de datos de imágenes etiquetadas, lo que les permite descubrir patrones subyacentes en los datos visuales y discernir características relevantes para la clasificación, detección y segmentación. Aunque requieran más recursos computacionales y tiempo de entrenamiento en comparación con los modelos tradicionales, los modelos de DL consistentemente superan a estos últimos y son fundamentales para los avances actuales en VPC. Algunas de las técnicas basadas en DL se explican a continuación.

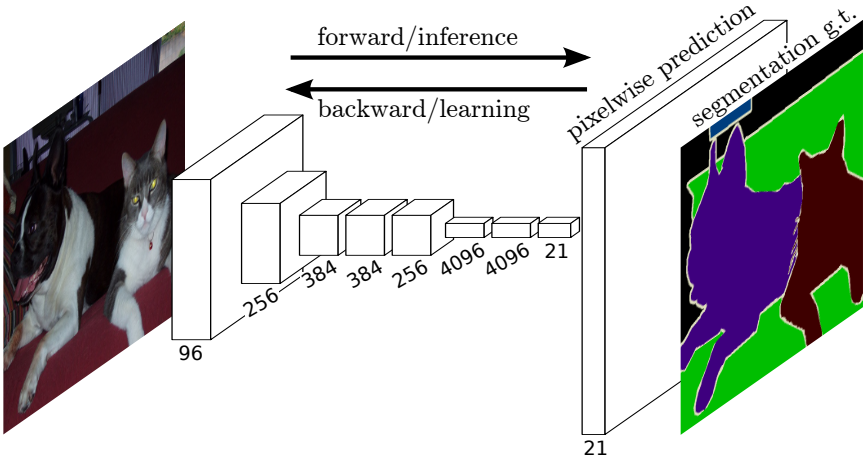
2.2.1 Redes totalmente convolucionales

Una Fully Convolutional Network (FCN) [Long y cols. (2015)] es una clase de red neuronal profunda diseñada para abordar tareas de VPC como la segmentación de imágenes, en las que la entrada y salida son de tamaño variable. En el contexto en que se encuentra

este TFM, este aspecto es especialmente relevante en la segmentación de imágenes debido a que habilitas a la red a que funcione de manera efectiva en imágenes de cualquier tamaño, lo que la hace altamente adaptable a una variedad de escenarios y aplicaciones. Las FCNs eliminan las capas totalmente conectadas de una CNN tradicional y las reemplazan por capas convolucionales, lo que permite que la red acepte imágenes de cualquier tamaño y produzca mapas de características con la misma resolución que la entrada.

La principal diferencia entre una FCN y una CNN radica en su estructura y su capacidad para manejar entradas de tamaño variable. Mientras que una CNN clásica está compuesta por capas convolucionales seguidas de capas completamente conectadas, las FCNs eliminan estas últimas capas y las reemplazan por convoluciones con *kernel* de tamaño 1×1, lo que permite producir mapas de características en la misma resolución que la entrada.

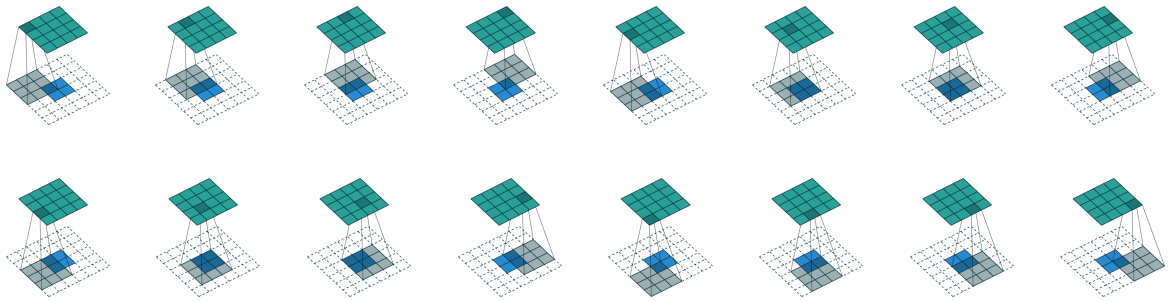
Figura 2.8: Arquitectura original de una Red Completamente Conectada



La convolución de bloques 1x1 en una FCN es crucial para mantener la información espacial durante el proceso de segmentación. Estas convoluciones actúan como un mecanismo de reducción de dimensionalidad, permitiendo que la red reduzca la cantidad de canales de características sin perder información espacial. Esto ayuda a preservar la resolución de los mapas de características durante la propagación hacia atrás y permite que la red capture características a múltiples escalas, lo que mejora la precisión de la segmentación.

Luego de aplicar dicha convolución de bloques 1×1, la red utiliza capas de *upsampling* o convoluciones transpuestas para aumentar el tamaño de los mapas de características a la misma resolución que la imagen original. El *upsampling* es una técnica utilizada en las FCN para aumentar la resolución espacial de las características de la salida. Esto es fundamental, sobre todo, en aplicaciones donde se necesita una salida detallada que coincida con la resolución inicial. Estas convoluciones transpuestas, a diferencia de las convoluciones convencionales que reducen la resolución, funcionan tomando un pequeño parche de datos y aplicando una transformación lineal para asignar los datos a una región más grande en la salida.

Figura 2.9: Secuencia de operaciones de una convolución transpuesta

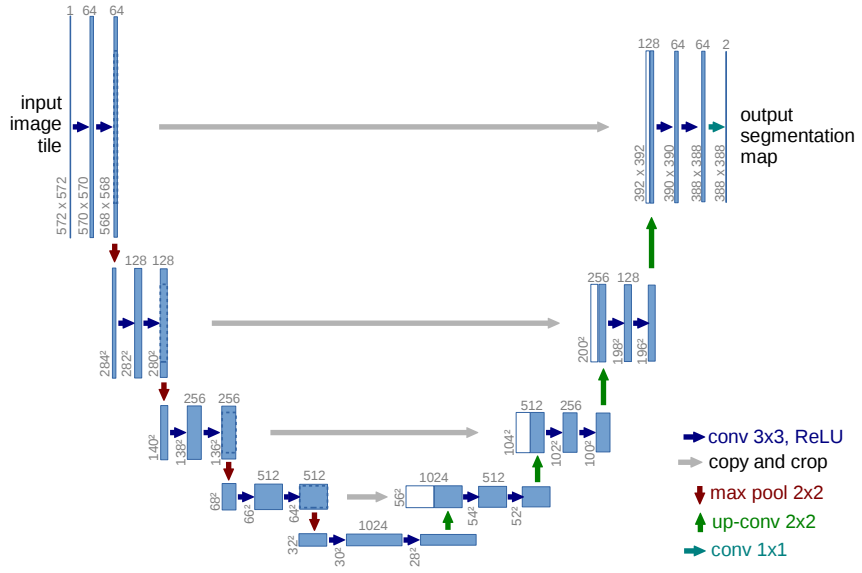


La figura 2.9 muestra la secuencia de operaciones de una convolución transpuesta donde los mapas azules son la entrada y los mapas cian son la salida. La animación visualmente demuestra que, a diferencia de la convolución estándar que reduce el tamaño espacial de la entrada, la convolución transpuesta expande el tamaño espacial de la salida. Cada posición del filtro en la entrada se multiplica y se suma a las ubicaciones correspondientes en la salida, ilustrando cómo se reconstruye una imagen más grande a partir de una más pequeña.

2.2.2 U-Net

U-Net [Ronneberger y cols. (2015)] es una arquitectura CNN ampliamente utilizada en este tipo de tareas, especialmente en el campo de la medicina para aplicaciones como la segmentación de órganos en imágenes médicas. Esta arquitectura fue presentada por primera vez en el paper *U-Net: Convolutional Networks for Biomedical Image Segmentation*.

Figura 2.10: Arquitectura original de una U-Net



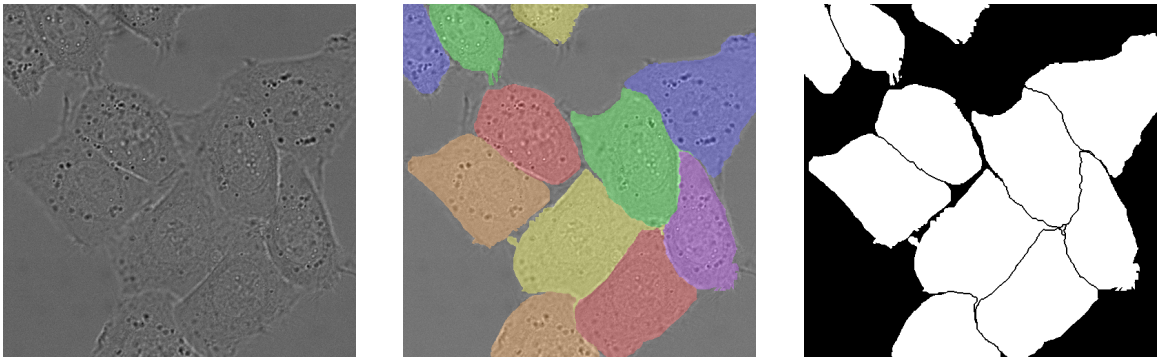
La característica distintiva de la U-Net es su estructura en forma de U, de la cual deriva su nombre. Esta estructura incluye un camino de contracción (*encoder*) seguido de un camino

de expansión (*decoder*). El *encoder* está compuesto por capas de convolución y *pooling*, que disminuyen progresivamente la resolución espacial de la entrada mientras aumentan la profundidad de las características. Por otro lado, el *decoder* utiliza convoluciones transpuestas (o *upsampling*) para aumentar la resolución espacial de las características de salida y generar una máscara de segmentación que tenga la misma resolución que la imagen de entrada.

El *encoder* de U-Net captura características de diferentes niveles de abstracción a medida que se profundiza en la red, lo que permite que la red aprenda características contextuales de la imagen a diferentes escalas. Esta información contextual es crucaial para la segmentación precisa de objetos de la imagen. Además, U-Net utiliza conexiones residuales entre capas en el camino de contracción y expansión, lo que facilita el flujo de información detallada y preserva la información de alta resolución.

En la arquitectura U-Net, las características extraídas en el camino de contracción se fusionan con las características de la capa correspondiente en el camino de expansión mediante conexiones concatenadas. Esta fusión de características permite que la red utilice tanto la información contextual de alto nivel como los detalles de baja resolución para producir una segmentación precisa.

Figura 2.11: Aplicación de una U-Net en segmentación de imágenes



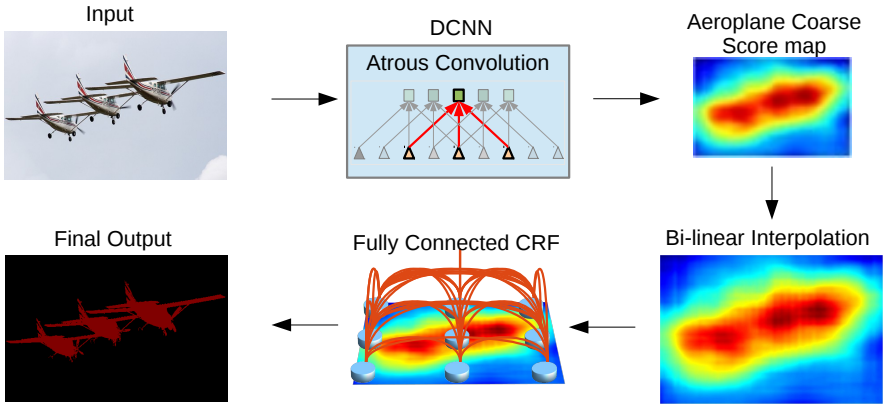
(a) Imagen original **(b)** Ground truth **(c)** Segmentación

La figura 2.11 ilustra la aplicación de una red U-Net en la segmentación de imágenes. En la imagen **(a)**, se presenta la imagen original que muestra un conjunto de células. La imagen **(b)** muestra el *ground truth*, donde cada célula está coloreada de manera diferente para indicar las regiones que deben ser reconocidas y segmentadas por el algoritmo. La imagen **(c)** muestra el resultado de la segmentación automática realizada por la U-Net, donde se observa que las células han sido correctamente identificadas y segmentadas, separadas por líneas blancas. Esta figura destaca la eficacia de la U-Net en tareas de segmentación de imágenes biomédicas, proporcionando una segmentación precisa y detallada de las estructuras celulares.

2.2.3 DeepLab

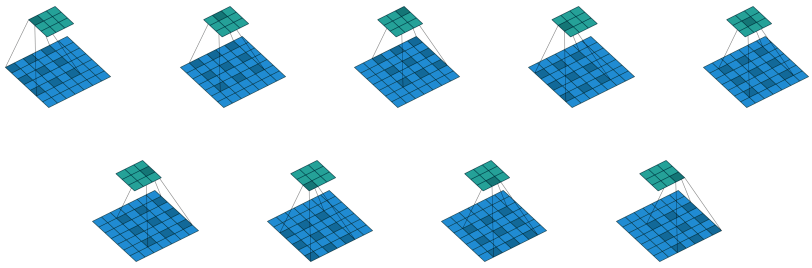
DeepLab [Chen y cols. (2017)] es una arquitectura CNN desarrollada principalmente para la segmentación semántica de imágenes. Es conocida por su capacidad para capturar detalles finos y producir segmentaciones precisas incluso en imágenes de alta resolución. La arquitectura DeepLab ha sido desarrollada y refinada a lo largo de los años por Google Research.

Figura 2.12: Ilustración original del modelo DeepLab



DeepLab se beneficia del uso de las Deep Convolutional Neural Networks (DCNN), que son esenciales para el aprendizaje de representaciones jerárquicas y abstractas de las imágenes. Las DCNN están compuestas por múltiples capas de convolución que aprenden características de la imagen en diferentes niveles de abstracción, desde bordes y texturas en las primeras capas hasta objetos completos y escenas en las capas más profundas. Este aprendizaje permite a DeepLab capturar patrones complejos y contextuales en las imágenes, mejorando la precisión de la segmentación. Una característica distintiva de DeepLab es su uso de *atrous convolutions*, también conocidas como convoluciones dilatadas. Estas convoluciones permiten aumentar el tamaño del campo receptivo de una neurona convolucional sin aumentar el número de parámetros entrenables, lo que es crucial para capturar contextos más amplios y detalles finos en imágenes de alta resolución. Esta técnica es especialmente útil para la segmentación semántica, donde se requiere tanto la contextualización global como la precisión local.

Figura 2.13: Secuencia de operaciones de una convolución dilatada



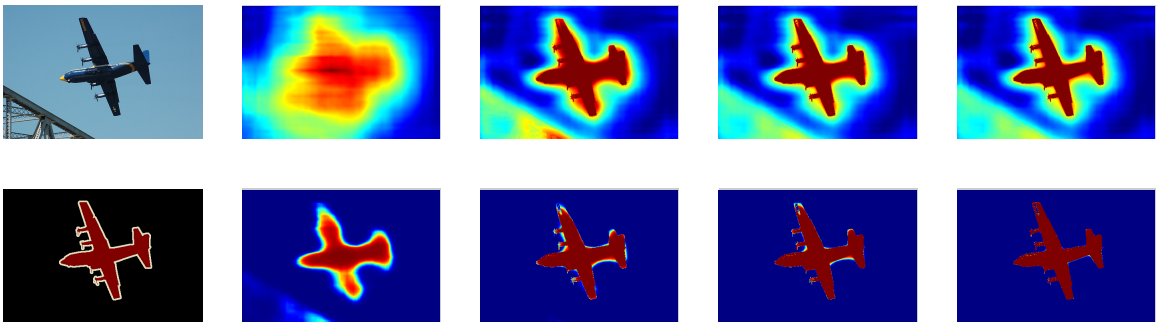
La arquitectura DeepLab también incorpora técnicas de *spacial attention* para resaltar re-

giones importantes de la imagen durante el proceso de segmentación. Estas técnicas ayudan a la red a enfocarse en áreas relevantes de la imagen y a suprimir el ruido o las regiones menos informativas. Esto mejora la capacidad de la red para capturar características relevantes y producir segmentaciones más precisas.

Otra característica clave de DeepLab es su capacidad para manejar imágenes de diferentes tamaños mediante el uso de *atrous convolutions* a diferentes tasas de dilatación. Esto permite que la red segmente imágenes de cualquier tamaño sin necesidad de reescalado previo, lo que la hace más flexible y adaptable a diferentes aplicaciones y escenarios.

Finalmente, para refinar aún más los resultados de segmentación, DeepLab integra Fully Connected Conditional Random Fields (FCCRF). Los FCCRF se utilizan para mejorar la coherencia espacial de las segmentaciones, corrigiendo errores y refinando los bordes de los objetos segmentados. Esta técnica postprocesa las predicciones de la red neuronal considerando tanto las relaciones de cercanía como las características de apariencia entre los píxeles de la imagen, lo que resulta en segmentaciones más precisas y detalladas.

Figura 2.14: Aplicación de DeepLab en segmentación de imágenes



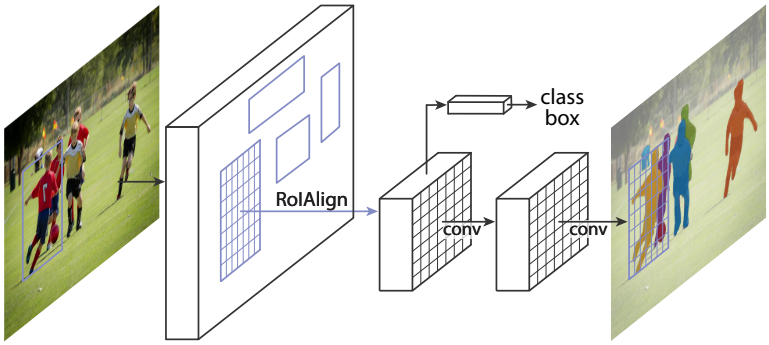
La figura 2.14 ilustra el proceso de segmentación de una imagen de un avión. Se presentan los mapas de puntuación (antes de la *softmax*) y de creencia (después de la *softmax*) en varias etapas. La primera columna muestra la imagen original y su *ground truth*. La segunda columna muestra la salida inicial de la DCNN. Las siguientes columnas muestran los resultados después de 1, 2 y 10 iteraciones del FCCRF, respectivamente. Con cada iteración del FCCRF, los mapas se refinan, mejorando la precisión y delimitación de la segmentación del avión. El texto destaca que la salida de la última capa de la DCNN se utiliza como entrada para la inferencia del campo medio en el FCCRF.

2.2.4 Mask R-CNNs

Mask R-CNN [He y cols. (2018)] es una arquitectura CNN desarrollada para la detección de objetos y la segmentación de imágenes. Se basa en la popular red neuronal Faster R-CNN, que se utiliza para la detección de objetos, y extiende su funcionalidad para incluir la generación de máscaras precisas alrededor de cada objeto detectado. Mask R-CNN ha sido

ampliamente utilizada en aplicaciones relacionadas con la segmentación precisa de objetos, como el reconocimiento facial, la robótica y la conducción autónoma.

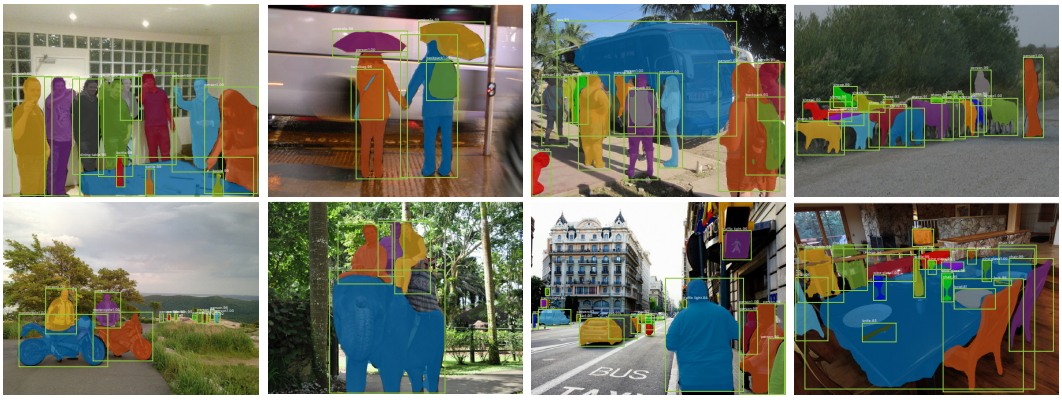
Figura 2.15: Ilustración original del modelo Mask R-CNN



La arquitectura Mask R-CNN consta de tres componentes principales: la red de detección de objetos, la red de segmentación de instancias y el generador de máscaras. La red de detección de objetos utiliza una serie de capas convolucionales para localizar y clasificar objetos en la imagen, produciendo *bounding boxes* que rodean cada objeto detectado. La red de segmentación de instancias, basada en la arquitectura Feature Pyramid Network (FPN) [Lin y cols. (2017)], genera características a nivel de píxel para cada objeto detectado, lo que permite una segmentación precisa de instancias incluso en objetos pequeños o superpuestos. Finalmente, el generador de máscaras utiliza una serie de capas convolucionales para producir máscaras binarias que delinean la forma exacta de cada objeto detectado.

Una de las principales fortalezas de Mask R-CNN es su capacidad para generar máscaras precisas alrededor de objetos en imágenes, lo que permite una segmentación detallada de instancias. Esto es especialmente útil en escenarios donde se requiere una precisión alta, como en aplicaciones médicas o de reconocimiento de objetos. Además, Mask R-CNN es capaz de detectar y segmentar múltiples objetos en una sola imagen, lo que lo hace adecuado para aplicaciones en las que se necesita analizar múltiples objetos simultáneamente.

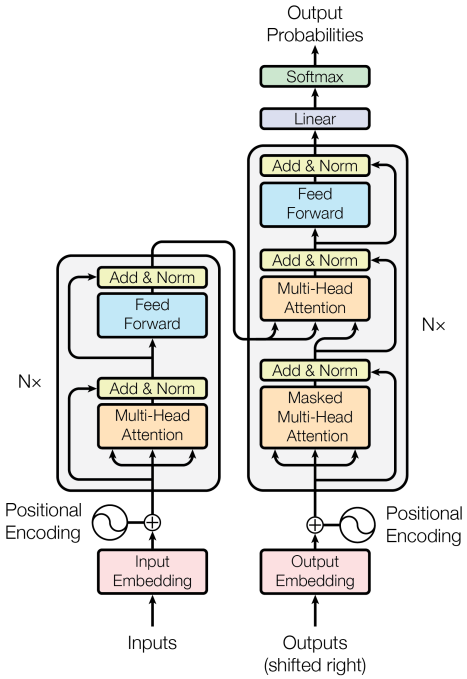
Figura 2.16: Aplicación de una Mask R-CNN en segmentación de imágenes



2.2.5 Transformers

La arquitectura Transformers [Vaswani y cols. (2017)] presentada en el famoso trabajo *Attention Is All You Need* es una innovadora estructura de redes neuronales que ha revolucionado el campo del Procesamiento del Lenguaje Natural (NLP) y ha sido adaptada con éxito a una variedad de tareas en el ámbito de la VPC. Originalmente introducida por Google Research en el contexto del NLP, los Transformers se han convertido en una opción popular debido a su capacidad para capturar relaciones de largo alcance en secuencias de datos, lo que les permite modelar eficazmente tanto la dependencia local como la global entre elementos.

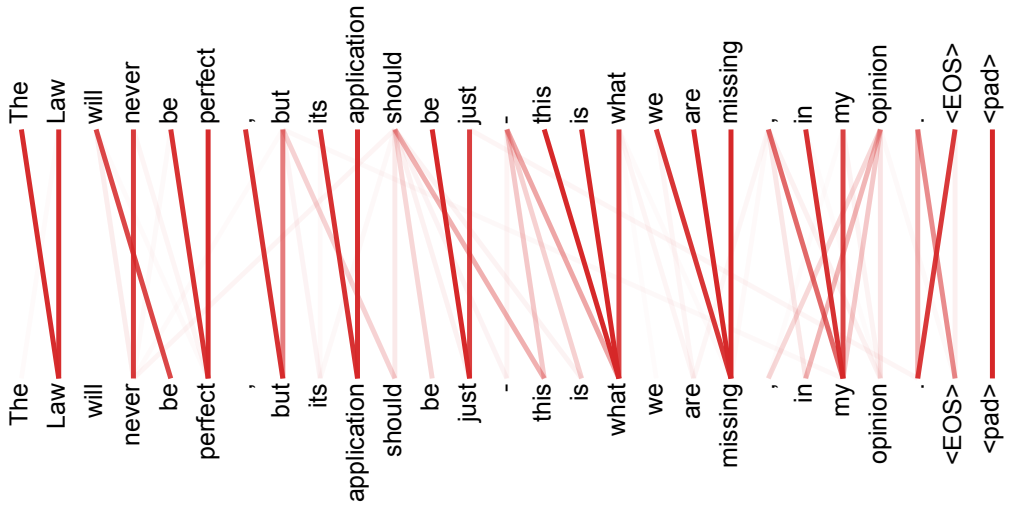
Figura 2.17: Arquitectura original de un Transformer



La arquitectura empieza con un *input embedding* y un *output embedding*, que convierten los *tokens* de entrada y salida en vectores de alta dimensión. Dado que los Transformers no procesan datos secuencialmente como las RNN, necesitan incorporar información sobre la posición de cada *token* en la secuencia. Esto se realiza mediante un *positional encoding*, que se suma a los *embeddings* de entrada y salida para proporcionar información posicional a los vectores.

El siguiente elemento de la arquitectura es el *encoder* y se compone de N capas idénticas, cada una con dos subcapas principales. La primera subcapa es una capa *multi-head self-attention*, que permite al modelo enfocarse en diferentes partes de la secuencia de entrada simultáneamente. Este mecanismo de atención múltiple mejora la capacidad del modelo para capturar dependencias a largo plazo de los datos. La segunda subcapa es una capa *feed forward* que aplica una red neuronal completamente conectada por separado a cada posición en la secuencia.

Figura 2.18: Aplicación de la atención en una secuencia de palabras



Cada subcapa en el *encoder* se complementa con una operación *add & norm*. La salida de cada subcapa se normaliza y se suma a la entrada de la subcapa mediante una conexión residual, lo que ayuda a estabilizar el entrenamiento y facilita el flujo de gradientes a través de la red.

Por otro lado, el *decoder* también se compone de N capas idénticas, pero tiene una estructura ligeramente más compleja con tres subcapas. La primera subcapa es una capa *masked multi-head self-attention*, que impide que el modelo vea las futuras palabras en la secuencia durante el entrenamiento, asegurando que la predicción de una palabra solo dependa de las palabras anteriores. La segunda subcapa es una capa *multi-head attention* que enfoca la salida del *encoder*, permitiendo que el *decoder* atienda a diferentes partes de la secuencia de entrada relevante. La tercera subcapa es una capa *feed forward* similar a la del *encoder*.

Al igual que en el *encoder*, cada subcapa se acompaña de una operación *add & norm*. La salida de cada subcapa se normaliza y se suma a la entrada de la subcapa, manteniendo la estabilidad del modelo durante el entrenamiento.

Finalmente, la salida del *decoder* pasa por una capa lineal seguida de una capa *softmax*, que genera las posibilidades de las próximas palabras en la secuencia de salida. Este mecanismo permite que el modelo prediga la siguiente palabra en función de las palabras anteriores y la información procesada por el *encoder*.

Transformers aplicados a imágenes

Los Transformers en su inicio se crearon para la tarea de traducción y es por ello que lo que se contempla como es entrada es una secuencia de palabras, pero en este TFM se trabaja sobre imágenes, por lo que se debe adaptar los datos. Aplicar la atención y la propia arquitectura a la segmentación de imágenes implica adaptar los principios utilizados en el NLP

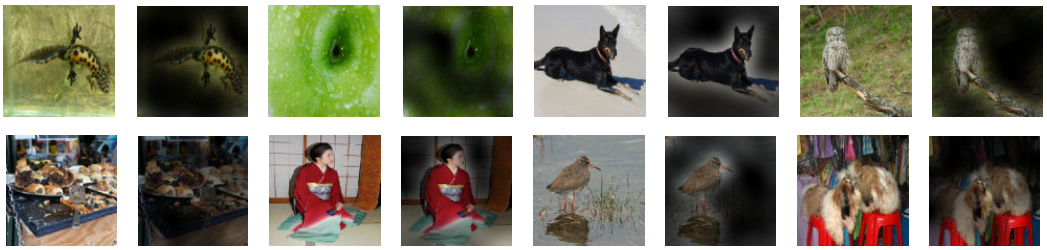
para manejar datos espaciales en imágenes.

En lugar de palabras, en la segmentación de imágenes, es necesario convertir la imagen en una secuencia de elementos comprensibles para el modelo Transformers. En lugar de palabras, se trabaja con parches de la imagen. Estos parches son pequeñas secciones de la imagen (por ejemplo, 16×16 píxeles) que se aplanan y se proyectan linealmente a una dimensión fija para crear una representación vectorial de cada parche.

Cada uno de estos parches no tiene un orden secuencial natural, por lo que se agrega una codificación posicional a cada *embedding* de parche, al igual que sucedía con las palabras, para retener la información sobre la ubicación espacial de cada parche en la imagen. Este paso es crucial para que el modelo entienda la disposición espacial de los parches y, por ende, de la imagen completa. La combinación de *embeddings* de parches y sus codificaciones posicionales se introduce en el *encoder* del Transformer.

El *encoder* del Transformer procesa los *embeddings* de los parches mediante el mecanismo de *multi-head attention*, que permite que el modelo preste atención a diferentes partes de la imagen simultáneamente, capturando relaciones espaciales complejas. Este mecanismo de atención, conocido como *self-attention*, permite que cada parche influya en otros parches, lo que es fundamental para que el modelo entienda el contexto completo de la imagen. La *multi-head attention* mejora este proceso al permitir que el modelo se enfoque en diferentes características espaciales de la imagen simultáneamente.

Figura 2.19: Aplicación de la atención en imágenes

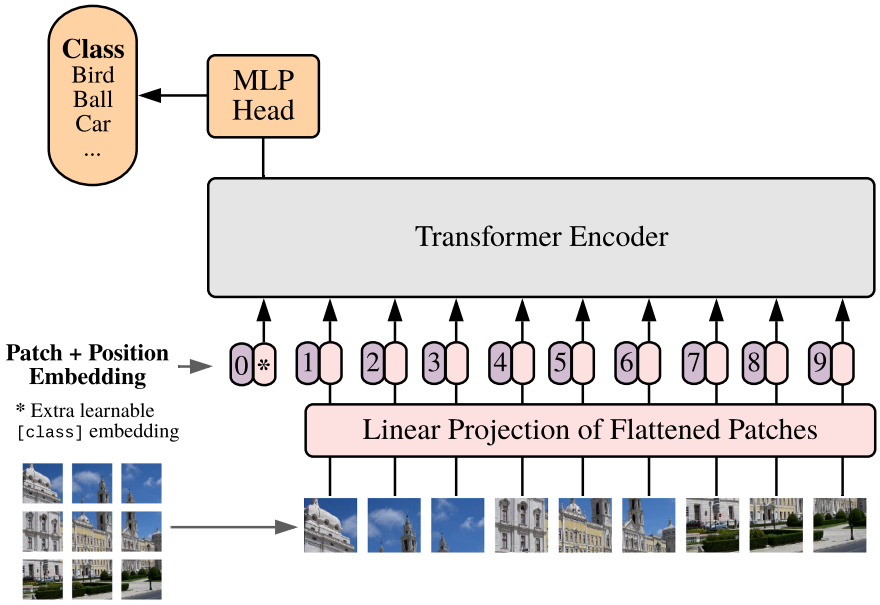


Para la segmentación de imágenes, se emplea un *decoder* que toma la salida del *encoder* y la refina para producir la segmentación. El *decoder* genera una máscara de segmentación que indica la clase de cada píxel. Este enfoque a menudo se combina con CNNs para aprovechar las capacidades de extracción de características de las convoluciones junto con la capacidad de modelado de relaciones globales de los Transformers.

Implementaciones populares como los Vision Transformer (ViT) [Dosovitskiy y cols. (2020)] y Segmentation Transformer (SETR) [Li y cols. (2023)] combinan Transformers con cabezas de segmentación para producir mapas de segmentación precisos. ViT, por ejemplo, divide la imagen en parches y aplica un Transformer directamente a la secuencia de parches, mientras que SETR combina ViT con una cabeza de segmentación. Estos enfoques híbridos han demostrado ser altamente efectivos en tareas de segmentación de imágenes, beneficiándose

tanto de las ventajas de las CNNs como de los Transformers.

Figura 2.20: Arquitectura original de un Vision Transformer



La figura 2.20 muestra la arquitectura original de un ViT. En esta arquitectura, una imagen de entrada se divide en varios parches más pequeños, los cuales son aplanados y luego proyectados linealmente. A cada parche se le añade una codificación posicional para mantener la información sobre la posición de cada parche en la imagen original. Estos parches transformados se alimentan al Transformer Encoder, que aplica múltiples capas de *self-attention* y *feed-forward* para procesar la información. Finalmente, la salida del *token* de clase se pasa a una cabeza de Multilayer Perceptron (MLP) para realizar la clasificación de la imagen en categorías como `Bird`, `Ball`, `Car`, entre otras.

3 Estado del arte

El estado del arte juega un papel crucial en cualquier investigación o proyecto, ya que proporciona un panorama completo y actualizado sobre el tema de estudio. Esta exhaustiva revisión de la literatura existente y los avances recientes en el campo permitirán identificar el contexto en el que se trabaja y las tendencias actuales. Además, el estado del arte no solo sirve como punto de partida para el investigador, sino también como una herramienta para contextualizar los resultados.

3.1 Segment Anything

El artículo *Segment Anything* [Kirillov y cols. (2023)] introduce un nuevo modelo y conjunto de datos para la segmentación de imágenes. La tarea principal es desarrollar un modelo de segmentación que sea capaz de generalizar a nuevas distribuciones de datos y tareas mediante el uso de técnicas de *prompting*. Para ello, los autores construyeron el mayor conjunto de datos de segmentación hasta la fecha. El modelo propuesto, Segment Anything (SAM), está diseñado para ser *promptable*, lo que le permite transferir su aprendizaje de manera directa a nuevas tareas de segmentación sin necesidad de reentrenamiento.

SA-1B

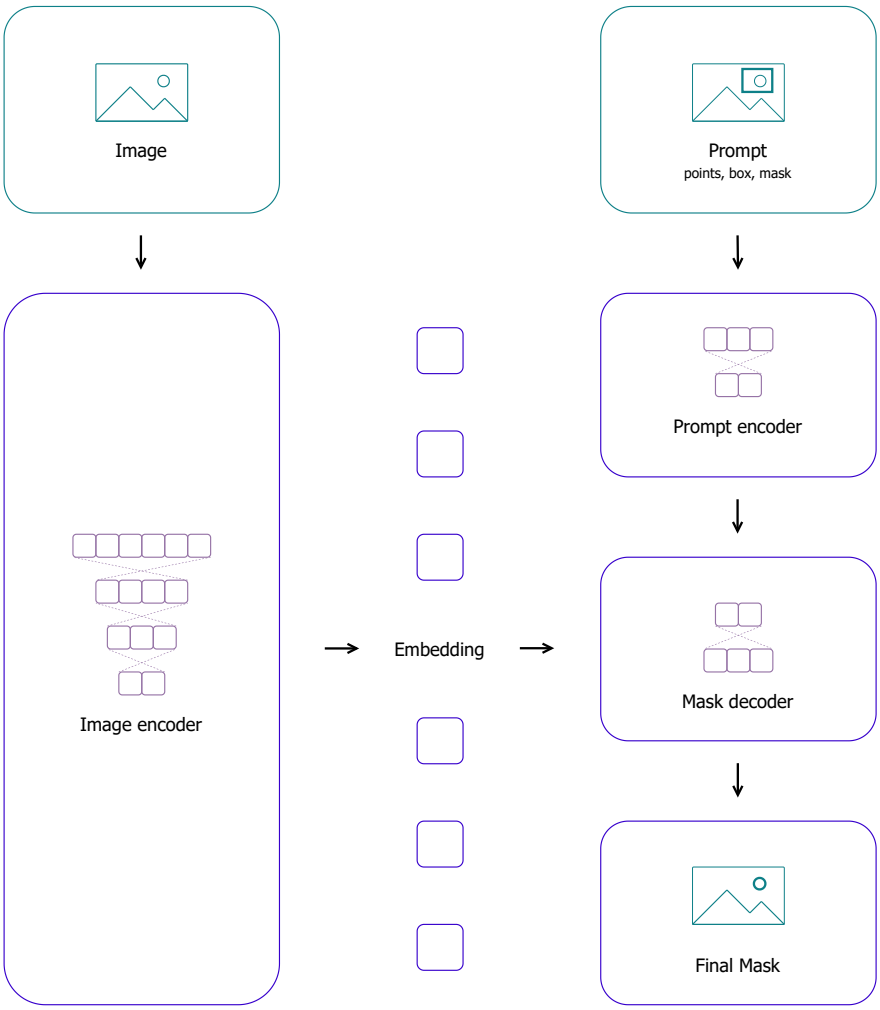
SAM ha sido entrenado con el conjunto de datos SA-1B, que incluye más de mil millones de máscaras provenientes de 11 millones de imágenes. Este conjunto de datos tiene 400 veces más máscaras que cualquier otro conjunto de datos de segmentación existente, y se ha verificado extensivamente que las máscaras son de alta calidad y diversidad.

Figura 3.1: Pares de muestras imagen-máscara del dataset SA-1B

Modelo

El modelo se compone de tres componentes principales: un *encoder* de imágenes, un *encoder* de *prompts* flexible y un *decoder* de máscaras rápido. La siguiente figura muestra un esquema simplista de la arquitectura del modelo.

Figura 3.2: Arquitectura original del modelo SAM



El *encoder* de imágenes puede ser cualquier red que produzca un *embedding* de la imagen con dimensiones $C \times H \times W$, donde C representa el número de canales, H la altura y W el ancho. En este caso, se opta por un ViT preentrenado con Masked Autoencoders (MAE) debido a su escalabilidad y robustez en el preentrenamiento, haciendo solo las adaptaciones mínimas necesarias para procesar entradas de alta resolución. En particular, se utiliza un ViT-H/16 que implementa atención en ventanas de 14×14 y cuenta con cuatro bloques de atención global distribuidos uniformemente.

La salida del *encoder* de imágenes es un *embedding* que ha sido reducido 16 veces respecto a la imagen de entrada. Por ejemplo, si la imagen de entrada tiene una resolución de 1024×1024,

el *embedding* resultante será de 64×64. Este proceso de reducción permite que el modelo sea más eficiente en términos de procesamiento.

Una vez obtenido el *embedding* de la imagen con dimensiones 64×64, se utiliza una convolución de 1×1 para reducir la dimensión del canal a 256. Posteriormente, se aplica una convolución de 3×3, también con 256 canales, para refinar aún más las características. Cada una de estas convoluciones va seguida de un normalización de capa (en inglés, *layer normalization*), lo cual ayuda a estabilizar y mejorar el entrenamiento del modelo.

Se consideran dos conjuntos de *prompts*: esparcidos (en inglés, *sparse*) (puntos, cuadros, texto) y densos (máscaras). Los puntos y cuadros se representan mediante codificaciones posicionales sumadas con *embeddings* aprendidos para cada tipo de *prompt*. El texto libre se codifica utilizando un *encoder* de texto del modelo Contrastive Language-Image Pre-Training (CLIP). Los *prompts* densos (es decir, las máscaras) son codificados utilizando convoluciones y se suman de manera *element-wise*[1] con el *embedding* de la imagen.

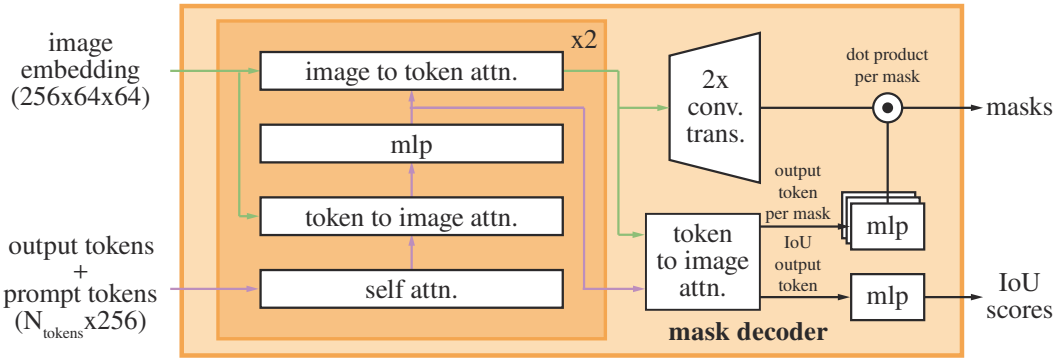
Los *sparse prompts* se mapean a *embeddings* vectoriales de 256 dimensiones de la siguiente manera: un punto se representa como la suma de una codificación posicional de la ubicación del punto y uno de dos *embeddings* aprendidos que indican si el punto está en primer plano o en el fondo. Un cuadro se representa mediante un par de *embeddings*: (1) la codificación posicional de su esquina superior izquierda sumada con un *embedding* aprendido que representa "esquina superior izquierda" y (2) la misma estructura pero utilizando un *embedding* aprendido que indica "esquina inferior derecha". Finalmente, para representar texto libre se utiliza el *encoder* de texto de CLIP (en general, cualquier *encoder* de texto es posible).

Los *prompts* densos (es decir, las máscaras), tiene una correspondencia espacial con la imagen. Se introducen las máscaras a una resolución $4x$ menor que la imagen de entrada, luego se reducen adicionalmente $4x$ utilizando dos convoluciones de 2×2 con *stride*, con canales de salida 4 y 16, respectivamente. Una última convolución de 1×1 mapeala dimensión del canal a 256. Cada capa está separada por activaciones Gaussian Error Linear Unit (GELU) y normalización de capa. El *embedding* de la máscara y el *embedding* de la imagen se suman de manera *element-wise*. Si no hay un *prompt* de máscara, se añade un *embedding* aprendido que representa "sin máscara" a cada ubicación del *embedding* de la imagen.

El *decoder* de máscaras mapea eficientemente los *embeddings* de la imagen, los *embeddings* de los *prompts* y un *token* de salida a una máscara. Este diseño se inspira en bloques de *decoders* de los Transformers, empleando *self-attention* y *cross-attention* en dos direcciones (de *prompt* al *embedding* de la imagen y viceversa) para actualizar todos los *embeddings*. Después de ejecutar dos bloques, se realiza un *upsampling* del *embedding* de la imagen y un MLP mapea el *token* de salida a un clasificador lineal dinámico que luego calcula la probabilidad de primer plano de la máscara en cada ubicación de la imagen.

[1] Término utilizado en matemáticas y en computación para describir una operación que se aplica de manera individual a cada elemento correspondiente de dos matrices o vectores.

Figura 3.3: Ilustración detallada sobre el decoder de máscaras de SAM



La figura 3.3 representa la arquitectura del *decoder* de máscaras. Comienza con el *embedding* de la imagen, que es el resultado del procesamiento de la imagen de entrada a través del *encoder* de imágenes. Este *embedding* tiene dimensiones $256 \times 64 \times 64$. Además, los *tokens* de salida y los *tokens* de *prompt* (esparcidos y densos) se combinan en una representación con dimensiones $N_{tokens} \times 256$, donde N_{tokens} es el número total de *tokens*.

El siguiente paso es la *attention* de imagen a token (*Image to Token Attention*). Este bloque aplica *cross-attention*, permitiendo que el *embedding* de la imagen influencie las representaciones de los *tokens*. Después de esta *cross-attention*, los resultados pasan por un MLP para transformaciones adicionales.

Luego, se aplica la *attention* de *token* a imagen (*Token to Image Attention*), que realiza *cross-attention* en la dirección opuesta, permitiendo que los *tokens* influencien el *embedding* de la imagen. Posteriormente, se aplica *self-attention* a los *tokens* para capturar relaciones internas entre ellos. Estos pasos de *attention* de imagen a *token*, MLP, *attention* de *token* a imagen y *self-attention* se repiten dos veces para mejorar las representaciones de proceder al siguiente paso.

El *embedding* de la imagen se somete a dos transformaciones por convoluciones para ajustar las dimensiones y mejorar la resolcuión de las características. Después de estas transformaciones, se realiza un *dot product* por cada máscara entre los *embeddings* resultantes y un *token* de salida específico para cada máscara. Esta operación genera las máscaras segmentadas finales, indicando las áreas segmentadas en la imagen. Finalmente, se utiliza un MLP adicional para calcular las puntuaciones Intersection over Union (IoU) de las máscaras, que miden la precisón de la segmentación. Los *tokens* de salida específicos por máscara se utilizan tanto para generar las máscaras como para calcular las puntuaciones IoU.

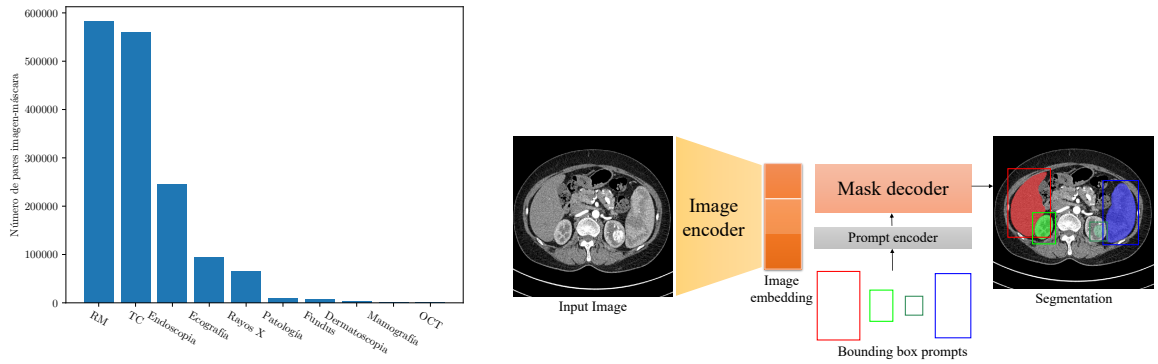
3.1.1 Segment Anything in Medical Images

Los enfoques actuales suelen estar diseñados para modalidades particulares o tipos de afecciones, lo que limita su capacidad para abordar la diversidad de desafíos que presenta la segmentación de imágenes médicas en su conjunto. Investigadores de la Universidad de Toronto presentaron Segment Anything in Medical Images (MedSAM) [Ma, He, y cols. (2024)],

un modelo de base diseñado para cerrar la brecha en la segmentación de imágenes médicas al permitir su aplicación de manera universal.

MedSAM es una versión refinada de SAM, diseñada específicamente para mejorar la segmentación de imágenes médicas. Se entrenó en un conjunto de datos sin precedentes que incluye más de un millón y medio de pares imágenes-máscaras médicas que abarcan 10 modalidades de imágenes, más de 30 tipos de cáncer y una variedad de protocolos de imagen. Esta gran cantidad de datos permite que MedSAM aprenda una representación rica de imágenes médicas, capturando un amplio espectro de anatomías y lesiones en diferentes modalidades. Sigue la arquitectura de red utilizada en SAM que incluye un *encoder* de imagen, un *encoder* de *prompts* y un *decoder* de máscaras. El *encoder* de imagen mapea la imagen de entrada a un espacio de *embeddings* de alta dimensión, mientras que el *encoder* de *prompts* transforma el *prompt* proporcionado por el usuario en representación de características. Finalmente, el *decoder* de máscaras fusiona el *embedding* de la imagen y las características de los *prompts* mediante un proceso de *cross-attention*.

Figura 3.4: Datos variados sobre MedSAM



(a) Número de pares imagen-máscara por modalidad

(b) Arquitectura original del modelo MedSAM

Para el desarrollo y validación del modelo, se recopilaron 1,570,263 pares de imágenes médicas y sus correspondientes máscaras. Durante la validación interna, los datos se dividieron aleatoriamente en 80% para entrenamiento, 10% para ajuste y 10% para validación. Para la validación externa, todos los conjuntos de datos se mantuvieron separados y no aparecieron durante el entrenamiento del modelo. Estos conjuntos de datos proporcionan una prueba rigurosa de la capacidad de generalización del modelo, ya que representan nuevos pacientes, condiciones de imagen y potencialmente nuevas tareas de segmentación que el modelo no había encontrado antes.

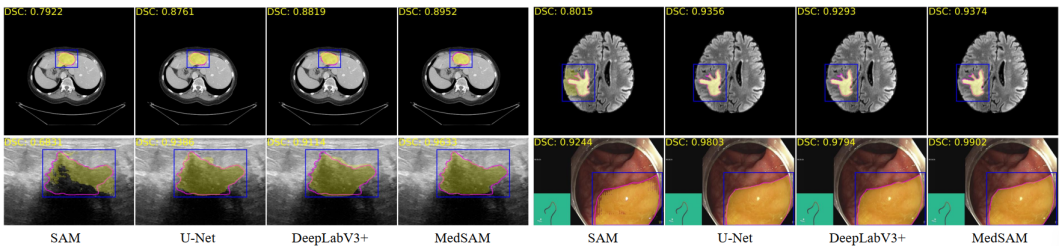
El modelo se inicializó con el modelo SAM preentrenado usando el modelo ViT-Base. El *prompt encoder* se mantuvo fijo, ya que ya puede codificar los *bounding boxes* (único *prompt* que se admitirá en MedSAM). Todos los parámetros entrenables en el *encoder* de imágenes y el *decoder* de máscaras se actualizaron durante el entrenamiento. Específicamente, el número de parámetros entrenables para el *encoder* de imágenes y el *decoder* de máscaras es de

89,670,912 y 4,058,340, respectivamente. El *bounding box* se simuló a partir de de las anotaciones de expertos con una perturbación aleatoria de 0-20 píxeles. La función de pérdida es la suma no ponderada entre la pérdida DICE y la entropía cruzada, que ha demostrado ser robusta en varias tareas de segmentación.

En el estudio se comparó el modelo MedSAM con otros modelos de segmentación líderes en el campo, incluyendo SAM, U-Net y DeepLabV3+ [Chen y cols. (2018)]. Cada uno de estos modelos fue evaluado en dos fases: validación interna y validación externa. La validación interna consistió en 86 tareas de segmentación, mientras que la validación externa abarcó 60 tareas adicionales, incluyendo nuevos conjuntos de datos y objetivos de segmentación no vistos anteriormente.

Durante la validación interna, los modelos especializados U-Net y DeepLabV3+ se entrenaron específicamente en imágenes de modalidades correspondientes, resultando en diez modelos dedicados para cada método. En contraste, SAM y MedSAM se utilizaron para segmentar imágenes en todas las modalidades. Los resultados mostraron que SAM tuvo un rendimiento inferior en la mayoría de las tareas de segmentación, aunque se destacó en algunas tareas específicas de segmentación de imágenes Red-Green-Blue (RGB), como la segmentación en imágenes de endoscopia. No obstante, los modelos U-Net, DeepLabV3+ y MedSAM, superaron ampliamente a SAM en la mayoría de las tareas. MedSAM mostró una distribución más estrecha de los puntajes en comparación con los modelos especializados, indicando su robustez en diversas tareas de segmentación. Además, MedSAM obtuvo el primer lugar en la mayoría de las tareas, demostrando un rendimiento superior en coparación con las los modelos especializados, que frecuentemente ocupaban el segundo y tercer lugar.

Figura 3.5: Resultados varios de la validación interna entre MedSAM y el estado del arte

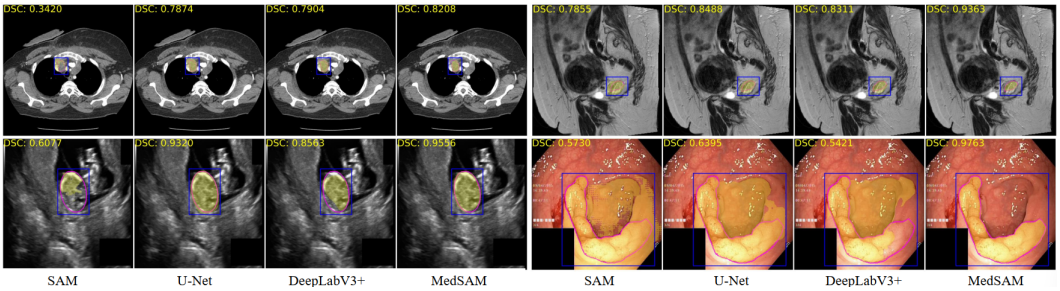


La figura 3.5 muestra una comparación visual de los resultados de segmentación de cada método mediante los valores de Dice Similarity Coefficient (DSC). En azul, se encuentra el *bounding box*, el rosa el *ground truth* y en amarillo la segmentación del modelo. Destaca que *medsam* y U-Net parecen ofrecer una precisión significativamente superior en la mayoría de los casos, especialmente notable en las imágenes de endoscopia.

En la validación externa, que incluyó tareas de segmentación de conjuntos de datos nuevos y objetivos no vistos, MedSAM continuó mostrando un rendimiento sobresaliente. A diferencia de los modelos especializados que no lograron un rendimiento superior en estas nuevas tareas, MedSAM mostró una capacidad de generalización notable. También demostró un rendimiento

superior en modalidades no vistas previamente.

Figura 3.6: Resultados varios de la validación externa entre MedSAM y el estado del arte



La figura 3.6 muestra también una comparativa visual de validación externa para las diferentes técnicas de segmentación. En general, MedSAM muestra un rendimiento superior en la mayoría de las imágenes, especialmente notable en las imágenes endoscópicas, destecándose significativamente sobre los otros métodos.

3.2 U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation

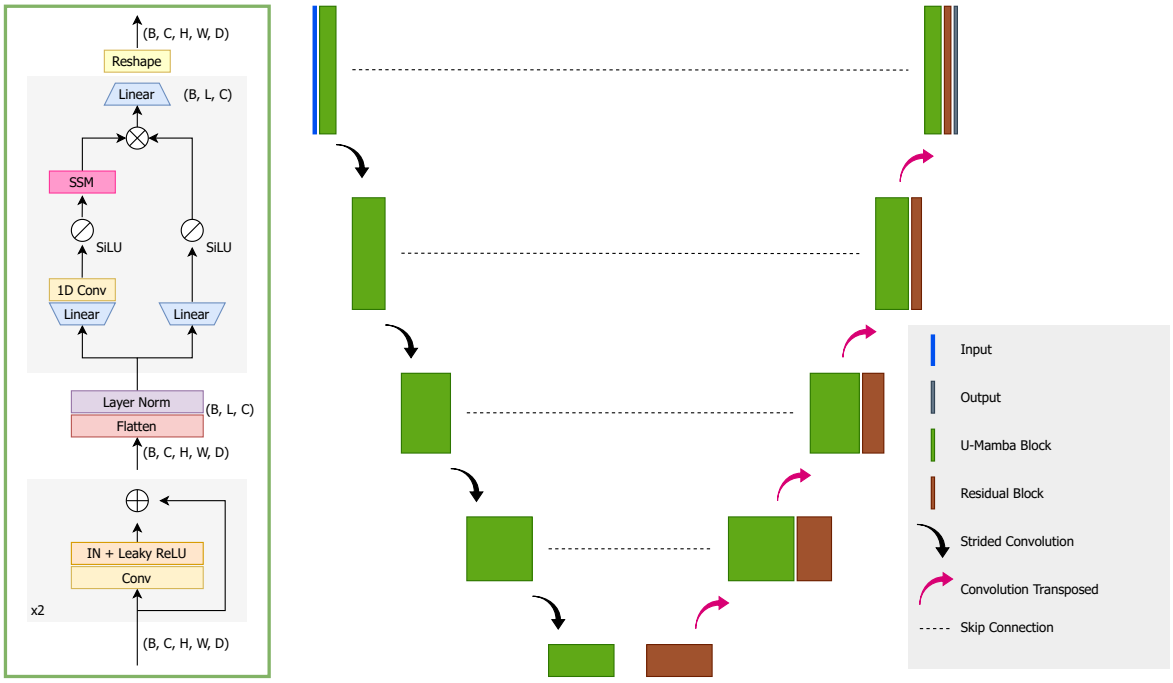
Los Transformers han mejorado la capacidad para manejar dependencias a largo plazo, pero suelen ser muy costosos en términos computacionales, especialmente para imágenes biomédicas de alta resolución. Por lo tanto, cómo mejorar de manera eficiente la capacidad de las CNNs para manejar dependencias a largo plazo sigue siendo un caso abierto. Recientemente, los modelos State Space Sequence Model (SSM), y en particular los modelos Structured State Space Sequence Model (S4), han surgido como una capa eficiente y efectiva para construir redes profundas, obteniendo un rendimiento de vanguardia en el análisis continuo de datos de secuencias largas.

Mamba [Gu y Dao (2024)] mejoró aún más S4 con un mecanismo selectivo, permitiendo que el modelo seleccione información relevante de manera dependiente de la entrada. En ciertas tareas, como en el lenguaje o la genómica, ha sido capaz de igualar a modelos libres de Transformers. Por otro lado, los SSM también han mostrado resultados prometedores en tareas de visión, como la clasificación de imágenes y vídeos. Dado que los parches de imágenes y las características de imágenes pueden tratarse como secuencias, estas características atractivas de los SSMs motivaron a explorar el potencial de usar este tipo de bloques Mamba para mejorar la capacidad de modelado de largo alcance en las CNNs. En este artículo, presentan U-Mamba [Ma, Li, y Wang (2024)], una red de propósito general, tanto para la segmentación de imágenes biomédicas 3D como 2D, que es capaz de capturar características locales detalladas y dependencias de largo alcance en las imágenes.

U-Mamba sigue la estructura de red *encoder-decoder* que captura tanto características locales como contextos de largo alcance de manera eficiente. La siguiente figura muestra un diagrama de cómo está compuesto el bloque Mamba y una estructura completa de la

arquitectura.

Figura 3.7: Arquitectura original de U-Mamba



La arquitectura híbrida de U-Mamba está diseñada para integrar bloques Mamba dentro del marco tradicional de U-Net. Los bloques Mamba son esencialmente bloques de SSM que se insertan después de dos bloques residuales sucesivos en cada unidad de la red. Esta configuración permite a U-Mamba procesar dependencias de largo alcance de manera más efectiva, superando una de las limitaciones típicas de las redes convencionales que suelen centrarse principalmente en relaciones espaciales locales.

En el diseño del bloque Mamba, las características de imagen se manipulan para adaptarse a un modelo de secuencia, lo que implica aplanar y transponer las características antes de procesarlas a través de capas lineales y una capa SSM. Este enfoque trata las relaciones entre características a lo largo de toda la imagen como secuencias, permitiendo al modelo reconocer y utilizar patrones complejos y dependencias a lo largo de grandes distancias espaciales.

U-Mamba también hereda la característica de autoconfiguración del modelo nnU-Net [Isensee y cols. (2018)], lo que permite que la red se ajuste dinámicamente a variaciones en los conjuntos de datos sin intervención manual. Esta adaptabilidad es fundamental para aplicar el modelo a diferentes tareas de segmentación biomédica, facilitando su uso en diversos entornos clínicos y de investigación sin la necesidad de ajustes detallados por parte del usuario.

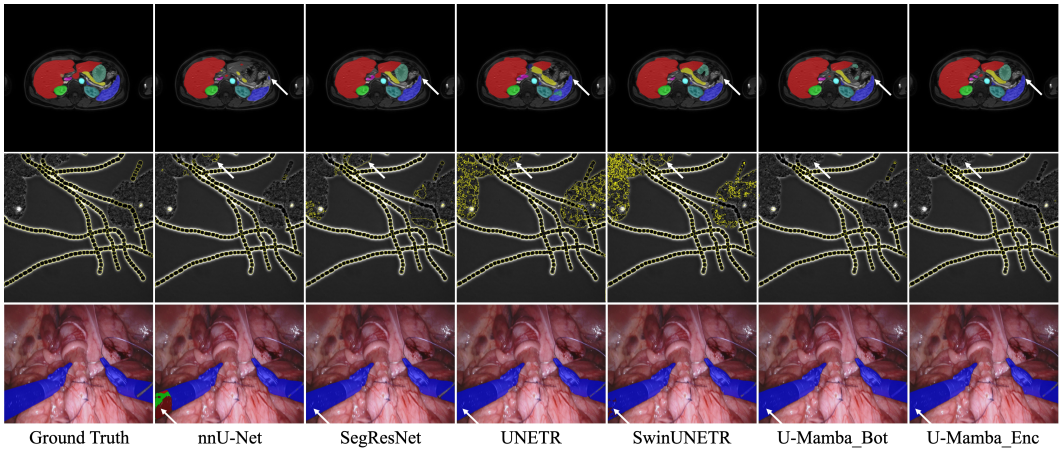
La estructura del *encoder-decoder* en U-Mamba está diseñada para capturar eficazmente características a múltiples escalas. Los bloques U-Mamba en el *encoder* extraen característi-

cas, mientras que los bloques residuales en el *decoder* se enfocan en la recuperación de detalles y la resolución de la imagen. Las conexiones de salto entre el *encoder* y el *decoder* ayudan a preservar información importante a través de capas, asegurando que la red pueda utilizar tanto el contexto global como los detalles locales para realizar la segmentación de manera efectiva.

U-Mamba se compara con dos redes de segmentación basadas en CNNs como son la nnU-Net y SegResNet [Myronenko (2018)] y dos redes basadas en Transformers como son UNETR [Hatamizadeh y cols. (2021)] y SwinUNETR [Hatamizadeh y cols. (2022)], que son ampliamente utilizadas en competiciones de segmentación de imágenes médicas. Para una evaluación exhaustiva, se realizaron pruebas no solo con imágenes médicas en formato 3D, sino también en imágenes 2D y conversiones de imágenes 3D a 2D.

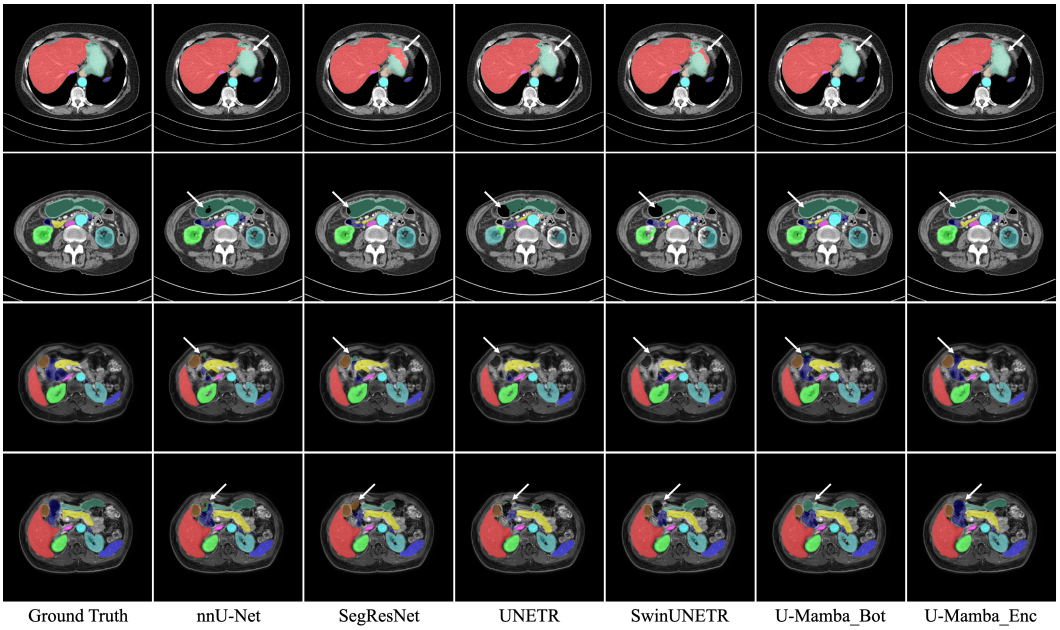
En el escenario de las imágenes 2D, U-Mamba demostró tener ventajas significativas sobre los métodos existentes, alcanzando el mejor promedio de DSC. U-Mamba también destacó en la segmentación celular dado que se requería realizar una segmentación de instancias, donde se debe asignar etiquetas únicas a cada instancia de célula.

Figura 3.8: Resultados varios de la segmentación 2D de U-Mamba



La figura 3.8 muestra ejemplos visualizados de segmentación en diferentes tipos de imágenes médicas. En la primera fila, se muestran ejemplos de segmentación de órganos abdominales. En la segunda fila, se visualizan ejemplos de segmentación celular. En la tercera fila, se presentan imágenes de segmentación de instrumentos en endoscopia. U-Mamba, con sus variantes *U-Mamba_Bot* y *U-Mamba_Enc*, muestra un rendimiento superior, evidenciado por su robustez frente a apariencias heterogéneas y la menor de atípicos en la segmentación.

Figura 3.9: Resultados varios de la segmentación 3D de U-Mamba



La figura 3.9 presenta ejemplos de segmentación de órganos abdominales. Las imágenes correspondientes al *ground truth* sirven como referencia estándar contra la cual se comparan los resultados de los modelos. De nuevo, se observa cómo U-Mamba, en sus dos variantes, demuestra manejar mejor los contornos de los órganos abdominales.

3.3 Customized Segment Anything Model for Medical Image Segmentation

El problema principal que motiva la creación de un nuevo modelo es la limitación de los modelos de VPC a gran escala, como SAM, en la segmentación de imágenes médicas. Uno de los obstáculos más significativos es la falta de datos médicos etiquetados. Los modelos a gran escala, como SAM, no pueden ser utilizados directamente para la segmentación de imágenes médicas debido a la ausencia de datos médicos y sus correspondientes etiquetas semánticas. Esto es crucial ya que la segmentación médica requiere identificar estructuras anatómicas o patológicas específicas que no están presentes en las imágenes naturales utilizadas para entrenar estos modelos.

Además, los modelos a gran escala deciden los límites entre diferentes regiones de segmentación basándose en la variación de la intensidad, lo cual es razonable en imágenes naturales pero no en imágenes médicas. En la segmentación médica, el análisis de estructuras anatómicas o patológicas es fundamental, y los modelos grandes carecen de esta capacidad. Por lo tanto, es esencial adaptar estos modelos para que puedan interpretar y segmentar correctamente las características específicas de las imágenes médicas.

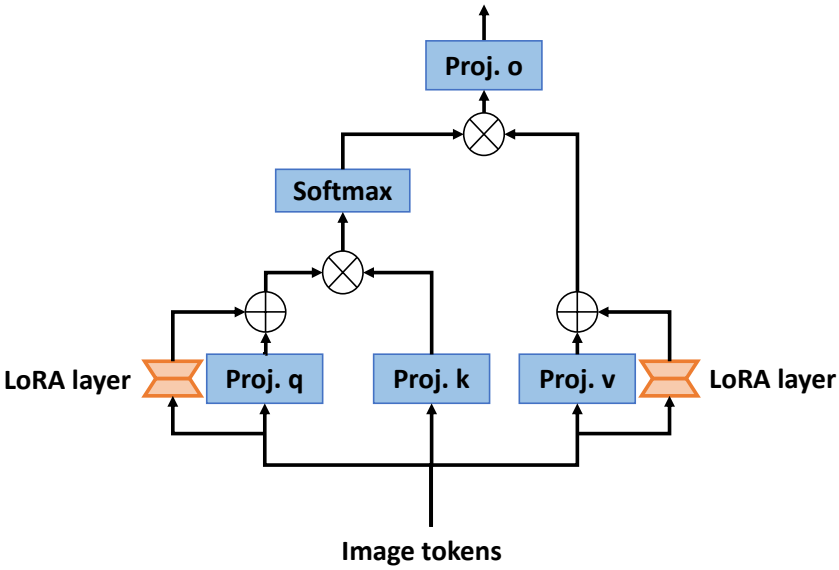
Otro desafío es la capacidad de los modelos a gran escala para realizar segmentación semántica en imágnees médicas. Estos modelos no pueden asociar las regiones de segmentación a clases semánticas significativas, lo que les impide realizar una segmentación semántica completa en imágenes médicas. Esta limitación es un obstáculo importante para su uso en aplicaciones de diagnóstico asistido por computadora, donde la precisión y la claridad de la segmentación son esenciales.

Finalmente, el costo de despliegue y almacenamiento de modelos grandes para usos específicos presenta un desafío considerable debido a su gran tamaño. Personalizar estos modelos para tareas específicas, como la segmentación de imágenes médicas, puede reducir significativamente los costos de despliegue y almacenamiento. Así pues, se propone SAMed [Zhang y Liu (2023)] como una solución que adapta SAM para la segmentación de imágenes médicas, abordando estos problemas y mejorando así su capacidad para analizar estructuras anatómicas y realizar segmentaciones semánticas precisas con un menor costo de despliegue y almacenamiento. *A prior*, parece una idea semejante a MedSAM, pero la implementación y arquitectura son distintas.

El objetivo de SAMed es predecir el mapa de segmentación correspondiente $\hat{S}$ de una imagen médica x con resolución espacial $H \times W$ y número de canales C. Cada píxel del mapa de segmentación debe pertenecer a un elemento de una lista de clases predefinidas $Y = y_0, y_1, ..., y_k$, donde y_0 es la clase de fondo y el resto son las clases de diferentes órganos. La arquitectura general de SAMed hereda de SAM. Se congela todos los parámetros en el *encoder* de imágenes y se diseña un *bypass* entrenable para cada bloque Transformer. Estos *bypasses* condensan las características del Transformer en el espacio de baja dimensión y reproyectan las características comprimidas para alinearlas con los canales de las características de salida en los bloques Transformers congelados. En cuando al codificador de *prompts*, SAMed no necesita ningún *prompt* durante la inferencia para realizar la segmentación automática, lo que beneficia enormemente al diagnóstico médico automático. Durante el entrenamiento, también se ajusta el *embedding* por defecto utilizado cuando no se proporcionan *prompts*.

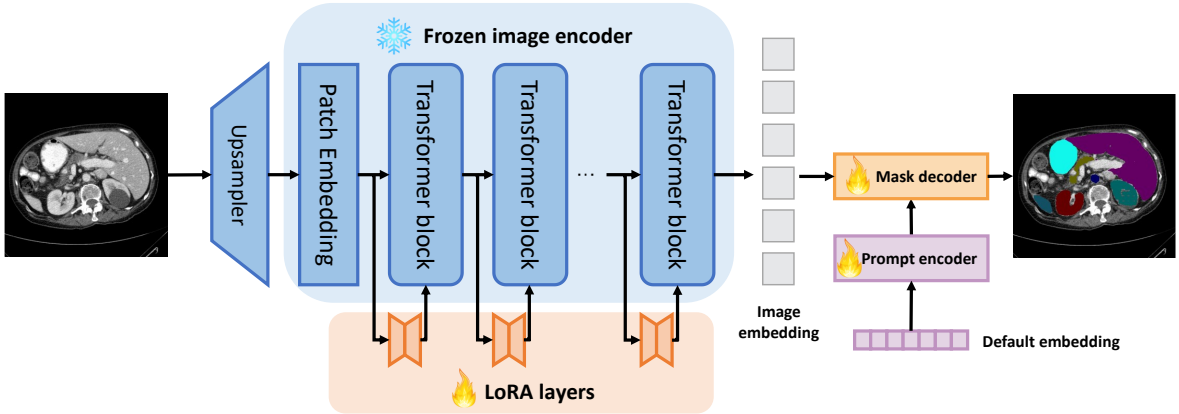
Comparado con ajustar todos los parámetros en SAM, Low-Rank Adaptation (LoRA) permite a SAM actualizar solo una pequeña fracción de los parámetros durante el entrenamiento en imágenes médicas, lo que no solo ahorra costos computacionales sino que también reduce la dificultad en el despliegue y almacenamiento de los modelos ajustados mientras se garantiza simultáneamente el rendimiento de segmentación. LoRA aplica una aproximación de rango bajo para delinear esta actualización gradual. Primero, SAMed congela las capas del Transformer para mantener los pesos W fijos, y luego agrega un *bypass* para lograr la aproximación de rango bajo. Este *bypass* contiene dos capas lineales $A \in \mathbb{R}^{r \times C_{in}}$ y $B \in \mathbb{R}^{C_{out} \times r}$, donde $r << min\{C_{in}, C_{out}\}$. La actualización de los pesos $\hat{W}$ se describe como $\hat{W} = W + BA$. Se observa que SAMed puede lograr un mejor rendimiento al aplicar LoRA a las capas de proyección *query* y *value*.

Figura 3.10: Aplicación de LoRA utilizada en SAMed



Para lograr un diagnóstico médico rápido y automático, SAMed no necesita ningún *prompt* durante la inferencia. El *encoder* de *prompts* en SAM utiliza un *embedding* por defecto cuando no se proporciona ningún *prompt*, por lo que SAMed mantiene este *embedding* por defecto y lo hace entrenable durante el proceso de ajuste fino. El *decoder* de máscaras en SAM consiste en una capa ligera de Transformer y una cabeza de segmentación o ajustar todos los parámetros en el *decoder* de máscaras directamente. SAMed modifica ligeramente la cabeza de segmentación de SAM para personalizar la salida para cada clase semántica en Y. A diferencia de la predicción ambigua de SAM, SAMed predice cada calse semántica de Y de manera determinista.

Figura 3.11: Arquitectura original de SAMed



4 Metodología

La metodología es crucial en cualquier TFM, ya que detalla los procedimientos y técnicas empleados para llevar a cabo la investigación. Los siguientes apartados describirán el *software* utilizado, proporcionando información sobre las herramientas y plataformas que facilitan el análisis de datos y el desarrollo de los modelos. Además, se explican los métodos de evaluación de los modelos de DL, incluyendo métricas específicas utilizadas para medir su eficacia y precisión. La claridad en la exposición de estos elementos no solo valida la rigurosidad del estudio, sino que también permite a otros investigadores reproducir o basarse en el trabajo realizado, asegurando así la transparencia y la contribución efectiva al campo de estudio.

4.1 Software utilizado

En los siguientes apartados, se describirá el uso de Python y sus librerías esenciales como PyTorch, Numpy y Matplotlib, entre otras, que han sido fundamentales para el desarrollo y evaluación de modelos de DL. Adicionalmente, se abordará el uso de Slurm para la asignación de recursos y 3D Slicer, una plataforma avanzada para la visualización y análisis de imágenes médicas.

4.1.1 Lenguaje de programación y librerías

Python es un lenguaje de programación de alto nivel, interpretado y de propósito general, que ha ganado popularidad en diversos campos, especialmente en la ciencia de datos y la IA. Se destaca por su sintaxis clara y legible que facilita el desarrollo rápido de aplicaciones complejas. Además, la extensa disponibilidad de bibliotecas y el soporte comunitario robusto lo convierten en una elección ideal para propósitos rápidos y desarrollo de proyectos científicos y analíticos.

PyTorch es un *framework*[1] de DL que proporciona flexibilidad y velocidad en la construcción de modelos complejos de ML. Es especialmente conocido por su facilidad de uso y eficiencia en la experimentación rápida. PyTorch ofrece un enfoque dinámico para la definición de grafos computacionales, lo que permite modificaciones en tiempo de ejecución y un depurado intuitivo, características especialmente útiles en la investigación académica y en el desarrollo de prototipos.

Nibabel es una librería para leer y escribir archivos de imágenes médicas en diversos formatos, proporcionando un acceso sencillo a datos complejos. Nilearn amplía las funcionalidades de Nibabel, ofreciendo herramientas avanzadas para el análisis y visualización de imágenes

[1] Conjunto de herramientas y librerías que proporciona una estructura básica para facilitar el desarrollo de software en áreas específicas

médicas. Juntas, estas librerías facilitan la manipulación y el análisis de imágenes médicas, apoyando extensamente los procesos de investigación.

Medical Open Network for Artificial Intelligence (MONAI) es un *framework* basado en PyTorch diseñado específicamente para el ámbito de la sanidad, que proporciona herramientas y preconfiguraciones para el desarrollo de aplicaciones de DL en imágenes médicas. Esta librería está optimizada para facilitar la reproducibilidad, la eficiencia y la precisión en aplicaciones médicas, apoyando desde la carga de datos hasta la implementación de modelos avanzados y su evaluación.

Torchvision es una librería que forma parte del ecosistema de PyTorch, diseñada para simplificar las tareas de procesamiento y manipulación de imágenes para aplicaciones de DL. Proporciona cargadores de datos predefinidos, modelos pre-entrenados y transformaciones comunes, lo que ayuda a acelerar considerablemente el desarrollo de aplicaciones de VPC.

Numpy es fundamental en el ecosistema de Python para la computación científica. Proporciona soporte para grandes y multidimensionales *arrays* y matrices, junto con una colección de funciones matemáticas de alto nivel para operar con éstos. Su eficiencia y versatilidad la hacen indispensable para el procesamiento de datos y operaciones matemáticas complejas en la ciencia y la ingeniería.

Scikit-image es una librería de procesamiento de imágenes que proporciona algoritmos y utilidades para análisis de imágenes en Python. Ofrece herramientas para la segmentación, transformaciones geométricas, análisis de color y espacio, y muchas otras operaciones, siendo una opción robusta para proyectos que requieren manipulaciones detalladas de imágenes.

Pandas es una librería de análisis de datos que ofrece estructuras de datos flexibles y herramientas de manipulación de datos diseñadas para facilitar la limpieza y el análisis rápido de datos de Python. Es ampliamente utilizado para la manipulación de datos tabulares y series temporales, siendo esencial en cualquier flujo de trabajo de ciencia de datos.

Matplotlib es una librería de trazado de gráficos que ofrece una amplia variedad de formatos y es capaz de producir gráficos de calidad de publicación. Es altamente personalizable y puede crear gráficos estáticos, animados e interactivos, lo que la hace muy valorada tanto en la academia como en la industria para la visualización de datos.

OpenCV-Python es un enlace de Python a la popular librería de VPC OpenCV. Proporciona herramientas para la captura de vídeo, manipulación de imágenes y VPC, incluyendo reconocimiento facial, detección de objetos y seguimiento de movimientos, siendo crucial para aplicaciones que requieren procesamiento avanzado de imágenes y vídeo en tiempo real.

4.1.2 Slurm

Simple Linux Utility for Resource Management (SLURM) es un sistema de gestión de colas de trabajos y administración de recursos ampliamente reconocido, especialmente diseñado para entornos de computación de alto rendimiento. Su diseño flexible y altamente configurable

lo hace una opción predilecta en muchos de los superordenadores del mundo, facilitando una gestión efectiva de la carga de trabajo y optimizando la utilización de recursos computacionales.

SLURM permite a los usuarios encolar trabajos especificando los recursos necesarios como Central Processing Unit (CPU), Graphics Processing Unit (GPU), memoria y tiempo de ejecución. Su motor de planificación asigna recursos a los trabajos de manera eficiente, maximizando la utilización de los recursos del sistema mientras se adhiere estrictamente a los requisitos especificados por los usuarios. Además, SLURM administra y monitorea activamente los recursos del sistema, permitiendo la implementación de políticas avanzadas de gestión que incluyen priorización de trabajos, *preemption*[2] y configuraciones de uso compartido o exclusivo de nodos.

4.1.3 3D Slicer

3D Slicer [Fedorov y cols. (2012)] es una plataforma avanzada y de código abierto diseñada para la visualización, análisis y procesamiento de imágenes médicas, Utilizando ampliamente en la investigación biomédica y la educación clínica, esta plataforma ofrece herramientas exhaustivas para la visualización de imágenes tridimensionales, análisis cuantitativo, simulación de procedimientos quirúrgicos y creación de modelos anatómicos. Su capacidad de extensión mediante *plugins* facilita la integración con tecnologías emergentes, como los modelos de DL para segmentación de imágenes.

Una de las características más destacadas de 3D Slicer es su capacidad para integrar modelos de DL a través de extensiones. Estas extensiones permite incorporar *frameworks* de ML como TensorFlow o PyTorch dentro del entorno de 3D Slicer. Esto permite a los usuarios emplear técnicas de segmentación automatizadas y altamente eficientes que son capaces de identificar patrones complejos, mejorando significativamente la precisión y reduciendo los tiempos de procesamiento.

Además, 3D Slicer proporciona funcionalidades para el ajuste, entrenamiento y evaluación de modelos de DL directamente desde su interfaz. Esto incluye herramientas para la anotación manual y la validación de los resultados de segmentación, lo que es esencial para el desarrollo y la mejora de los modelos. Tal integración hace que 3D Slicer no solo sea una herramienta para aplicar modelos de DL, sino también un entorno robusto para la investigación y desarrollo de nuevas técnicas de segmentación.

4.2 Medidas de evaluación

En la segmentación semántica, la clasificación se realiza a nivel de píxel. Sin embargo, no es recomendable aplicar simplemente las métricas de clasificación estándar a la colección completa de píxeles en un conjunto de datos por dos razones principales. En primer lugar, los píxeles de una misma imagen están altamente correlacionados. Por lo tanto, para respetar la estructura jerárquica de los datos, los valores de las métricas deben calcularse primero por

[2] Acto de interrumpir temporalmente una tarea en ejecución, con la intención de reanudarla en otro momento.

imagen y luego agregarse sobre el conjunto de imágenes. En segundo lugar, en problemas de segmentación, típicamente existe un interés inherente en los bordes de las estructuras, centros o volúmenes de las estructuras. Por ello, la familia de métricas basadas en bordes (subconjunto de métricas basadas en distancia) requiere la extracción de bordes de estructuras a partir de las máscaras de segmentación binaria como base para la evaluación de la segmentación. Basándose en estas consideraciones y dadas todas las fortalezas y debilidades complementarias de las métricas de segmentación comunes, Maier-Hein y cols. (2024) recomienda el uso de las métricas DSC y Normalized Surface Distance (NSD).

4.2.1 Dice-Sørensen Coefficient

La métrica DSC es ampliamente utilizada para evaluar la precisión de la segmentación en imágenes médicas y otros campos de VPC donde la segmentación de objetos es crítica. Este coeficiente es especialmente prevalente en tareas de segmentación semántica, donde se compara la similitud entre dos conjuntos de datos: las anotaciones de referencia y las predicciones de un modelo.

$$DSC = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{4.1}$$

La métrica DSC se define como el doble del área de la intersección entre la predicción y el *ground truth* dividida por la suma del número total de píxeles de ambos, donde X representa el conjunto de píxeles del *ground truth* y Y el conjunto de píxeles de la predicción. En relación a la interpretación, un valor de 1 indica una coincidencia perfecta entre la segmentación predicha y la anotación de referencia, del mismo modo que un valor 0 indica todo lo contrario. El valor DSC siempre debe estar entre 0 y 1.

A diferencia de la IoU, el DSC cuenta dos veces la intersección en el numerador, lo cual tiende a dar un valor ligeramente más alto en comparación para el mismo nivel de superposición. Esto se debe a que el DSC es esencialmente una medida de la media armónica entre la precisión y la sensibilidad, mientras que la IoU es una medida directa de la superposición relativa entre los dos conjuntos. En contextos donde es crucial penalizar tanto los falsos positivos como los falsos negativos, el DSC puede ser preferible debido a su simetría en tratar estos dos tipos de error.

4.2.2 Normalized Surface Distance

La métrica NSD es una herramienta de evaluación esencial en el campo de la segmentación de imágenes médicas, particularmente utilizada para medir la precisión en la delimitación de bordes de estructuras anatómicas tridimensionales. La NSD ofrece una evaluación detallada de las diferencias entre los contornos de segmentación generados por los modelos y los contornos de referencia anotados manualmente, proporcionando un indicativo claro de la precisión espacial de los contornos predichos. La fórmula se expresa como:

$$NSD_{b,c}\left(Y_{b,c}, \hat{Y}_{b,c}\right) = \frac{|\mathcal{D}'_{Y_{b,c}}| + |\mathcal{D}'_{\hat{Y}_{b,c}}|}{|\mathcal{D}_{Y_{b,c}}| + |\mathcal{D}_{\hat{Y}_{b,c}}|} \tag{4.2}$$

donde $Y_{b,c}$ y $\hat{Y}_{b,c}$ son los contornos de referencia y los contornos de la segmentación predicha, respectivamente. Los términos $\mathcal{D}_{Y_{b,c}}$ y $\mathcal{D}_{\hat{Y}_{b,c}}$ representan conjuntos de distancias de los vecinos más cercanos, calculadas desde la segmentación predicha hacia la referencia y viceversa. Los subconjuntos $\mathcal{D}'_{Y_{b,c}}$ y $\mathcal{D}'_{\hat{Y}_{b,c}}$ contienen las distancias que son menores o iguales a un umbral de distancia aceptable τ_c reflejando las distancias consideradas precisas y aceptables.

$$\mathcal{D}'_{Y_{b,c}} = \{d \in \mathcal{D}_{Y_{b,c}} | d \leq \tau_c\} \tag{4.3}$$

El uso de un umbral de distancia τ_c es crucial en esta métrica, ya que permite definir qué se considera una distancia "aceptable" entre los contornos predichos y los contornos de referencia. Esto es especialmente importante en aplicaciones clínicas donde pequeñas desviaciones pueden ser críticas. Un NSD cercano a 1 indica una alta precisión, donde la mayoría de las distancias están dentro del umbral aceptable, mientras que un NSD cercano a 0 indica que muchas de las distancias superan este umbral, reflejando una baja precisión en la segmentación. Esta implementación de NSD es compatible con tareas de segmentación multiclase y admite imágenes en 2D y 3D. Está basada en el artículo Seidlitz y cols. (2022) y el cálculo de los bordes sigue la implementación de Google DeepMind.

5 Conjuntos de datos

En este capítulo, se exploran diversos conjuntos de datos fundamentales que se usarán en los entrenamientos de los diversos modelos. La imagenología médica es una disciplina que depende en gran medida de la calidad y diversidad de los datos para desarrollar y validar algoritmos de segmentación avanzados. Los conjuntos de datos descritos a continuación han sido seleccionados por su relevancia y utilidad específica en diferentes áreas de la imagenología cardíaca y del corazón entero. Cada uno de estos conjuntos de datos proporciona oportunidades únicas para abordar desafíos específicos en la segmentación automática, contribuyendo a mejorar las técnicas diagnósticas y terapéuticas en el cuidado de la salud.

5.1 Imagenología médica

La imagenología médica es una rama de la medicina que utiliza diversas tecnologías de imágenes para obtener representaciones visuales del interior del cuerpo humano. Estas imágenes son esenciales para el diagnóstico, la monitorización y el tratamiento de enfermedades. Entre las tecnologías más comunes se encuentran la TC y la RM, ambas ampliamente utilizadas para obtener imágenes detalladas de estructuras internas.

La TC utiliza rayos X para crear imágenes detalladas de estructuras internas del cuerpo. Durante una exploración de TC, la máquina de rayos X gira alrededor del cuerpo y envía múltiples rayos X desde diferentes ángulos, creando secciones transversales del tejido corporal. Estas imágenes pueden ser reconstruidas en un modelo 3D del área escaneada. La TC es particularmente útil para visualizar huesos, órganos internos y otros tejidos densos.

La RM utiliza campos magnéticos y ondas de radio para generar imágenes de los órganos y tejidos del cuerpo. A diferencia de la TC, no utiliza radiación ionizante. En un examen de RM, el cuerpo del paciente se coloca dentro de un tubo grande que contiene imanes poderosos; estos campos magnéticos y las ondas de radio provocan que los protones en el cuerpo se alineen y luego emitan señales que son captadas para formar imágenes. La RM es altamente efectiva para visualizar tejidos blandos como el cerebro, músculos y cartílagos, así como estructuras internas complejas.

En el ámbito del DL aplicado a la segmentación de imágenes médicas, las diferencias entre la TC y la RM tienen implicaciones significativas en la eficacia de los modelos desarrollados. Estas diferencias se manifiestan principalmente en aspectos como el contraste de imagen, la resolución espacial y temporal, y la susceptibilidad a artefactos.

La TC se destaca por su capacidad para proporcionar un alto contraste entre tejidos de diferentes densidades, lo que la hace especialmente útil para visualizar estructuras óseas y

calcificaciones dentro del cuerpo. Esta característica la convierte en una herramienta valiosa para detectar y segmentar áreas donde los tejidos blandos y duros coexisten o están próximos, como es el caso de las áreas alrededor del corazón. Sin embargo, para el DL, esta alta capacidad de contraste puede resultar en desafíos cuando el objeto es diferenciar entre tipos similares de tejidos blancos, ya que las sutilezas entre ellos pueden no ser tan evidentes como en las imágenes de RM.

Por otro lado, la RM es preferida para estudios donde la visualización detallada de tejidos blandos es crucial, como en el análisis de la estructura cardíaca misma. La RM no utiliza radiación ionizante, lo que la hace más segura para los pacientes que requieren seguimientos frecuentes o que son particularmente susceptibles a los riesgos asociados con la radiación. Desde la perspectiva del DL, las imágenes de RM proporcionan una gran cantidad de detalles a nivel de textura de los tejidos, lo que permite a los algoritmos de segmentación trabajar con datos que destacan diferencias sutiles en la composición del tejido.

La calidad y los detalles de las imágenes también juegan un papel crucial en la eficacia de los algoritmos de DL para la segmentación. Mientras que las imágenes de TC son invaluables para identificar estructuras y patologías que involucran tejidos duros o calcificados, las RM ofrecen una resolución superior en la visualización de la composición y patología de los tejidos blandos. Esta capacidad es de especial interés en la segmentación del corazón, donde la definición precisa de las estructuras del tejido cardíaco es esencial para diagnósticos precisos y la planificación del tratamiento.

En conclusión, la elección entre TC y RM para la segmentación del corazón mediante técnicas de DL debe basarse en una evaluación cuidadosa de los objetivos del estudio, la necesidad de detalle en la imagen y las consideraciones clínicas del paciente. Cada tecnología ofrece ventajas únicas que pueden ser maximizadas según el contexto específico de su aplicación.

5.1.1 Formato de archivo NifTI

El formato Neuroimaging Informatics Technology Initiative (NifTI) es un estándar de archivo diseñado específicamente para el almacenamiento y el manejo de datos de imágenes médicas, especialmente en el campo de la neuroimagen. Este formato es esencialmente una evolución del formato ANALYZE desarrollado originalmente por la Clínica Mayo, que fue ampliado para mejorar la interoperabilidad y para incorporar una mayor cantidad de metadatos necesarios para el análisis avanzado de imágenes.

Una de las características mas destacadas del formato NifTI es su capacidad para almacenar imágenes tanto tridimensionales (3D) como cuatridimensionales (4D). Esto permite no solo visualizar estructuras espaciales sino también incorporar una dimensión temporal o de parámetro adicional, como puede ser el caso en estudios de resonancia magnética funcional donde el tiempo es un factor crítico.

En términos de estructura de archivos, NifTI ofrece flexibilidad al permitir que los datos se almacenen en un archivo único o en dos archivos separados. Un único archivo NifTI (`.nii`) combina la información del encabezado y los datos de imagen en un solo lugar, lo

que simplifica el manejo de los datos. Alternativamente, los datos pueden dividirse en un archivo de encabezado (`.hdr`) y un archivo de datos de imagen (`.img`), una característica heredada del formato ANALYZE, pero que no será el usado en los siguientes conjuntos de datos.

El encabezado de un archivo NifTI es rico en metadatos y contiene información crucial como las dimensiones de la imagen, el tipo de datos almacenado, la orientación de la imagen, escalas de intensidad, y unidades de medida tanto espaciales como temporales. Estos metadatos son fundamentales para el procesamiento y análisis correctos de las imágenes, ya que proporciona el contexto necesario para interpretar adecuadamente los datos visuales.

Además, el formato NifTI admite información detallada sobre la orientación espacial de las imágenes, lo que es crucial para el correcto alineamiento y comparación de imágenes en diferentes estudios o aplicaciones de *software*. Los códigos de orientación en el encabezado especifican cómo se deben interpretar los ejes de la imagen, asegurando así que las estructuras anatómicas se presenten de manera coherente y precisa en diferentes investigaciones.

La extensibilidad es otra ventaja importante del formato NifTI. Los archivos pueden incluir extensiones que permiten la adición de información adicional o datos relacionados con la imagen, lo que facilita la adaptación del formato a necesidades específicas de investigación o clínicas. Esta capacidad de adaptabilidad hace que NifTI sea un formato valioso y versátil en el ámbito de la imagenología médica, donde los requerimientos de datos pueden variar significativamente entre diferentes especialidades y aplicaciones.

5.1.2 Presentación de las muestras

Como se ha mencionado en la sección anterior, el encabezado de un archivo NifTI almacena metadatos cruciales que facilitan la interpretación y el procesamiento de las imágenes médicas. En particular, campos como `dim`, que indica las dimensiones de la imagen, y `pixdim`, que proporciona las dimensiones de los píxeles en cada dirección, son fundamentales para entender la escala y el tamaño del objeto escaneado. Además, `datatype` y `bitpix` informan sobre el tipo de datos y el número de bits por píxel respectivamente, lo cual es esencial para el manejo correcto de la intensidad de los píxeles durante el procesamiento de la imagen.

El `qform_code` y `sform_code` especifican las transformaciones geométricas que alinean la imagen con un espacio estandarizado, siendo crucial para comparaciones y análisis entre sujetos o a lo largo del tiempo. Los parámetros `quatern_b`, `quatern_c`, `quatern_d`, y los desplazamientos `qoffset_x`, `qoffset_y`, `qoffset_z` definen cómo la imagen debe ser rotada y trasladada para alinearla correctamente.

Cuadro 5.1: Campos destacados del *header* del archivo NifTI

Campo	Valor
`dim`	[3 512 512 363 1 1 1 1]
`pixdim`	[-1.0, 0.355, 0.355, 0.450, 0, 1, 1, 43155]

Continúa en la siguiente página

Cuadro 5.1: Campos destacados del *header* del archivo NifTI (*continuación*)

Campo	Valor
`datatype`	int16
`bitpix`	16
`qform_code`	scanner
`sform_code`	scanner
`qoffset_x`	45.5
`qoffset_y`	-228.58453
`qoffset_z`	-271.88

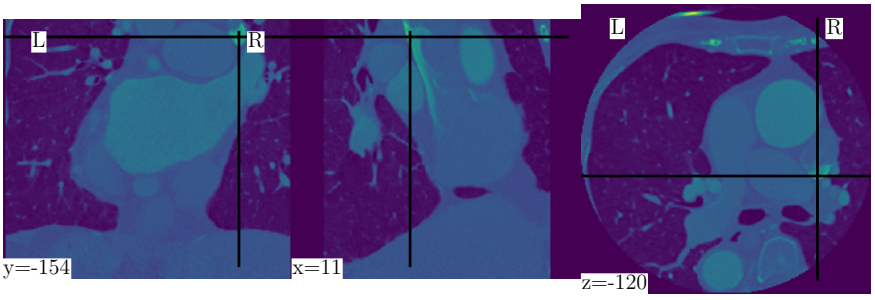
En el ámbito de la imagenología médica, el formato y el modo de visualización de las imágenes también son cruciales para una interpretación diagnóstica precisa. Librerías como Nilearn generan a partir de archivos NifTI imágenes anatómicas o imágenes Echo-Planar Imaging (EPI), además de mostrar la original, ejemplos destacados de cómo se presentan estas imágenes en contextos clínicos y de investigación.

El formato anatómico se centra en proporcionar una representación detallada de la anatomía del paciente. Este tipo de imagen es invaluable para observar con precisión las estructuras internas, permitiendo identificar anomalías estructurales o cambios morfológicos significativos. Las imágenes anatómicas son fundamentales en muchas áreas de diagnóstico médico, desde la evaluación de lesiones traumáticas hasta la detección de tumores o anomalías en el desarrollo.

Por otro lado, el formato EPI se refiere a imágenes que capturan episodios específicos o funcionalidades dentro del cuerpo, como estudios de perfusión que muestran cómo la sangre fluye a través de los tejidos. Estas imágenes son esenciales para diagnosticar que requieren una comprensión de la función de los órganos o sistemas, además de su estructura.

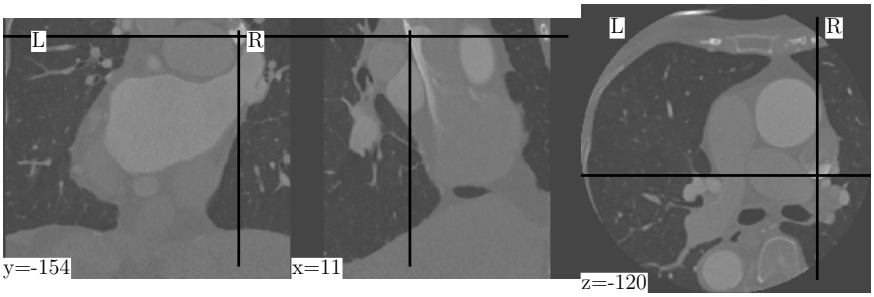
El formato básico de imagen representa una categoría más general que incluye cualquier imagen obtenida por TC o RM. Este formato abarca una amplia gama de aplicaciones diagnósticas, proporcionando vistas detalladas que son vitales para el diagnóstico médico general y la planificación del tratamiento.

Figura 5.1: Ejemplos del formato original, anatómico y EPI para una tomografía computarizada

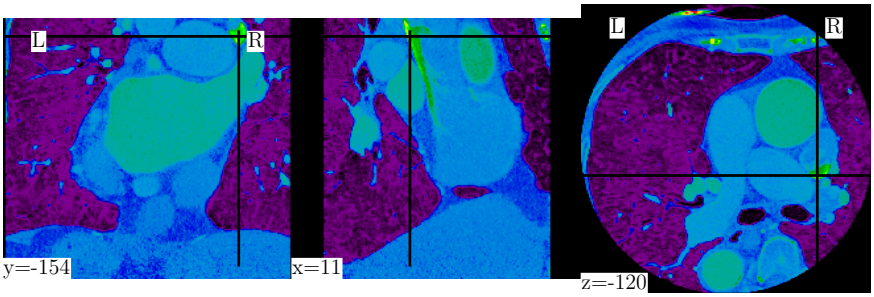


(a) Básico

Figura 5.1: Ejemplos del formato original, anatómico y EPI para una tomografía computarizada (*continuación*)



(b) Anatómico

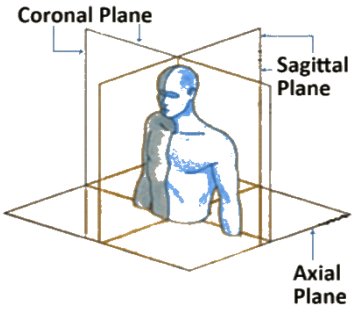


(c) EPI

La figura 5.1 muestra ejemplos del formato original, anatómico y EPI para una TC. En **(a)** se observan imágenes coloreadas artificialmente para destacar estructuras específicas, ideal para una identificación rápida en estudios iniciales. El panel **(b)** presenta imágenes en escala de grises que ofrecen detalles finos de la anatomía interna, típicamente utilizadas en el ámbito médico para diagnósticos precisos. Por último, el panel **(c)** utiliza colores intensos para realzar contrastes funcionales, útil en estudios de perfusión o funcionales. Aunque el formato básico pudiese ser el mejor debido a su simplicidad y adaptabilidad a la tarea de segmentación, lo cual facilitaría la visualización y análisis de resultados en este contexto tecnológico, de ahora en adelante se optará por el uso del formato anatómico, ya que es el formato que se emplea en los ámbitos médicos.

El modo de visualización ortogonal es particularmente útil en la visualización de imágenes médicas. Proporciona tres vistas distintas –coronal, sagital y axial– que son fundamentales para un examen completo. La vista axial corta el cuerpo en planos horizontales, proporcionando perspectivas desde arriba o desde abajo, ideal para evaluar la simetría y las estructuras horizontales. La vista coronal muestra el cuerpo en un plano vertical de lado a lado, como si se observara de frente, lo que es crucial para la visualización de la disposición frontal de los órganos. Finalmente, la vista sagital ofrece un plano vertical que divide el objeto en izquierda y derecha, mostrando vistas laterales que son importantes para entender las relaciones anteroposteriores dentro del cuerpo.

Figura 5.2: Ilustración de los planos axial, sagital y coronal sobre un cuerpo humano



La figura 5.2 ilustra los tres planos principales utilizados en la imagenología médica, como se ha mencionado antes. Aunque para tener una verdadera comprensión en un estudio son necesarias todas las vistas, para el entrenamiento de los modelos, únicamente se hará uso del plano axial, debido a su capacidad para proporcionar una vista integral y detallada de las estructuras cardíacas cruciales en una sola imagen. Por lo tanto, de ahora en adelante, únicamente se mostrará este plano en las visualizaciones y análisis.

5.2 MM-WHS: Multi-Modality Whole Heart Segmentation

Multi-Modality Whole Heart Segmentation (MM-WHS) [Zhuang y cols. (2019)] incluye un total de 120 imágenes de corazón entero de múltiples modalidades y procedentes de varios centros, dividido equitativamente en 60 TCs y 60 RMs en 3D. Estas imágenes cubren detalladamente las subestructuras del corazón completo y han sido aprobadas éticamente por las instituciones correspondientes, además de haber sido anonimizadas para asegurar la privacidad y la seguridad de los datos.

Los datos recopilados proceden de un entorno clínico *in vivo* y han sido utilizados en clínicas, lo que implica que presentan diversas calidades de imagen; algunas de estas imágenes son de calidad relativamente pobre. No obstante, la inclusión de estos conjuntos de datos es esencial para validad la robustez de los algoritmos desarrollados en condiciones reales de uso clínico.

En cuanto a la adquisición, las imágenes de TC cardíacas se obtuvieron mediante angiografía por TC cardíaca de rutina. Todas las imágenes abarcan el corazón completo desde la parte superior del abdomen hasta el arco aórtico, capturadas en vista axial con una resolución en el planto aproximadamente de 0.78×0.78 mm y un grosor promedio de corte de 1.60 mm. Por otro lado, los datos de RM se adquieren utilizando secuencias de precisión libre balanceada en estado estable 3D (b-SSFP), con una resolución de adquisición de aproximadamente 2 mm en cada dirección y reconstruidas (remuestreadas) a aproximadamente 1 mm.

Los conjuntos de datos se han dividido en datos de entrenamiento (20 TC y 20 RM representativos) y de test (40 TC y 40 RM). Para los datos de entrenamiento, se proporcionan

segmentaciones manuales de las siete subestructuras cardíacas enteras, que incluyen las cavidades sanguíneas de los ventrículos izquierdo y derecho, las cavidades de las aurículas izquierda y derecha, el miocardio del ventrículo izquierdo, la aorta ascendente y la arteria pulmonar. Las definiciones precisas de estas estructuras están codificadas con valores específicos de etiquetas.

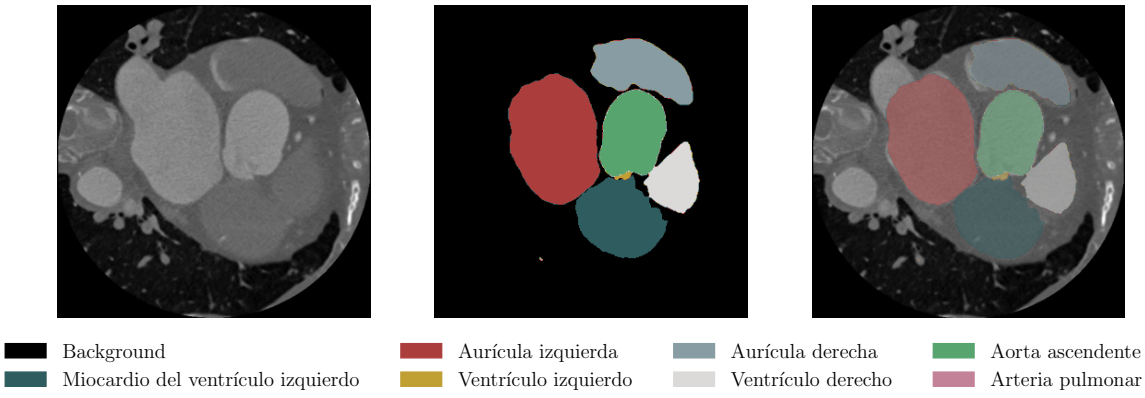
Los vasos grandes de interés, que incluyen la aorta ascendente y la arteria pulmonar, se definen de manera específica para garantizar una definición consistente a través de diferentes sujetos y análisis. La segmentación manual tanto para los datos de entrenamiento como para los de test generalmente abarca más allá de la longitud definida de estos vasos. En la evaluación, se utilizará un método para recortar los resultados de las segmentaciones de los vasos a un límite de longitud específico.

Especificaciones de las muestras

Este conjunto de datos emplea imágenes detalladas que cada una se acompaña de una correspondiente máscara de segmentación. Las imágenes y sus máscaras están estructuradas en un formato de 512×512 píxeles, agregando una dimensión adicional que corresponde al volumen de la muestra, proporcionando una perspectiva tridimensional enriqueciendo el contexto clínico de cada caso.

Como se ha mencionado previamente, todas las muestras se presentan desde una perspectiva axial. Esta orientación es esencial para mantener una consistencia que facilita la interpretación automática y manual, asegurando que los patrones identificados sean aplicables a través de todo el conjunto de datos de manera uniforme.

Figura 5.3: Corte, máscara de segmentación y superposición de ambas de una muestra del conjunto de datos MM-WHS



La figura 5.3 muestra un corte aleatorio de una muestra aleatoria junto su segmentación y la superposición de ambas para una mayor comprensión de las formas. Las áreas diferenciadas con un color específico para facilitar su identificación son el miocardio del ventrículo izquierdo

(MVI, 205)[1], las aurículas izquierda (AI, 420) y derecha (AD, 550), los ventrículos izquierdo (VI, 500) y derecho (VD, 600), la aorta ascendente (AA, 820) y la arteria pulmonar (AP, 850).

En conjuntos de datos médicos donde se contemplan varias etiquetas, es importante realizar una evaluación preliminar que revele variaciones significativas en la frecuencia de las etiquetas correspondientes a las diferentes estructuras anatómicas segmentadas. Cada etiqueta denota una región específica dentro de las imágenes, y una distribución desigual de estas etiquetas podría indicar un conjunto de datos desbalanceado. Por ejemplo, algunas estructuras como el miocardio del ventrículo izquierdo pueden estar presentes en una mayor proporción comparadas con regiones como la arteria pulmonar o la aorta ascendente, que podrían manifestarse con menos frecuencia.

Este desbalance en las etiquetas es una consideración crítica para el entrenamiento de modelos de DL, ya que podría influir en la capacidad del modelo para aprender representaciones equitativas de todas las categorías anatómicas. Por lo tanto, es crucial aplicar técnicas adecuadas de aumento de datos, para mitigar posibles sesgos y mejorar la generalización del modelo sobre datos nuevos y no vistos.

Cuadro 5.2: Distribución de etiquetas en el conjunto de datos MM-WHS

Modalidad	Etiqueta	MVI	AI	VI	AD	VD	AA	AP
TC	Presencia en cortes	2524	2047	2035	2685	3309	2643	2080
	% presencia en cortes	47.58	38.59	38.36	50.61	62.38	49.82	39.21
	Cortes por muestra	126.20	102.35	101.75	134.25	165.45	132.15	104.00

Modalidad	Etiqueta	MVI	AI	VI	AD	VD	AA	AP
RM	Presencia en cortes	384	255	296	245	319	194	348
	% presencia en cortes	55.01	36.53	42.41	35.10	45.70	27.79	49.86
	Cortes por muestra	76.80	51.00	59.20	49.00	63.80	38.80	69.60

El cuadro 5.2 ofrece un análisis detallado de la distribución de etiquetas en el conjunto de datos MM-WHS, comparando las modalidades de TC y RM. Detalla cada una de las estructuras cardíacas, mostrando la presencia en cortes, el porcentaje respecto al total de cortes y los cortes por muestra para cada estructura. En TC, el VD es notablemente prevalente con una presencia en 3309 cortes (62.38%), mientras que en RM, el MVI predomina, apareciendo en 384 cortes (55.01%). Esta variabilidad destacada entre las modalidades subraya cómo las técnicas de imagen específicas influyen en la visualización de estructuras cardíacas. La desigual distribución de muestras puede afectar significativamente al entrenamiento de modelos de DL, necesitando técnicas de aumento de datos para evitar sesgos y asegurar la precisión. Este análisis profundo no solo es crucial para entender el conjunto de datos, sino también para preparar los datos adecuadamente, garantizando que los modelos desarrollados sean equitativos y generalizables.

[1]El primer valor corresponde con una abreviatura y el segundo con el valor correspondiente a su etiqueta en la máscara.

5.3 ACDC: Automated Cardiac Diagnosis Challenge

El conjunto de datos Automatic Cardiac Diagnosis Challenge (ACDC) [Bernard y cols. (2018)] presentado en el Medical Image Computing and Computer Assisted Intervention Society (MICCAI) 2017 fue meticulosamente compilado a partir de exámenes clínicos reales adquiridos en el Hospital de Dijon (Francia). Incluye 150 pacientes divididos equitativamente en cinco grupos distintos, cada uno definido por características fisiológicas claras y basadas en parámetros como el volumen diastólico[2] y la fracción de eyección[3] obtenidos. La selección de pacientes fue cuidadosamente realizada, excluyendo aquellos con índices clínicos ambiguos y clasificando inicialmente según informes médicos.

Los cinco grupos dentro del ACDC están definidos de la siguiente manera:

- **NOR (Normal):** Pacientes con anatomía y función cardíaca normales, caracterizados por una fracción de eyección superior al 50%, grosor de pared diastólica menor a 12 mm y un volumen diastólico del ventrículo izquierdo inferior a lo sumbrales especificados según el género.
- **MINF (Infarto con insuficiencia sistólica):** Incluye pacientes con una fracción de eyección por debajo del 40% y contracciones miocárdicas anormales, típicamente como resultado de un infarto y el subsecuente remodelado del ventrículo izquierdo.
- **DCM (Cardiomiopatía dilatada):** Pacientes que muestran una fracción de eyección baja y un volumen aumentado del ventrículo izquierdo, con un engrosamiento de pared diastólica menor a 12 mm.
- **HCM (Cardiomiopatía hipertrófica):** Abarca a pacientes con función cardíaca normal pero con segmentos miocárdicos engrosados por encima de 15 mm en diástole.
- **ARV (Anomalías del ventrículo derecho):** Pacientes con volumen del ventrículo derecho y fracción de eyección anormalmente altos o bajos, respectivamente, aunque con un ventrículo izquierdo normal.

El conjunto de entrenamiento incluye 100 pacientes, con un 20 por grupo, y el de test 50 pacientes, con 10 por grupo. Cada conjunto de datos se convirtió a un formato de imagen 4D (NifTI) sin pérdida de resolución con un grosor de corte general de 5 mm.

Especificaciones de las muestras

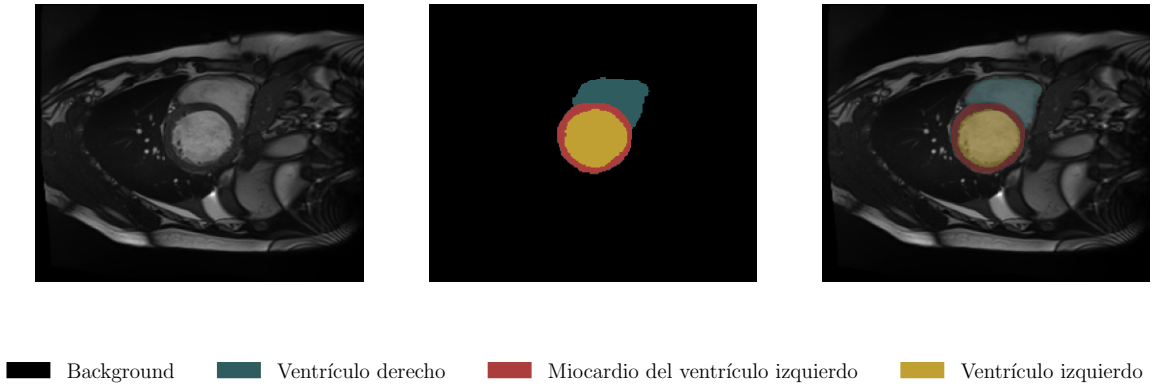
La figura 5.4 ilustra un corte de RM y su correspondiente máscara de segmentación, mostrando tres regiones de interés más el *background*. Las áreas delimitadas por colores son el ventrículo derecho (VD, 1), el miocardio del ventrículo izquierdo (MVI, 2) y el ventrículo izquierdo (VI, 3). Cada imagen tiene dimensiones de 216×256 píxeles, añadiendo una dimensión

[2] Cantidad total de sangre en el ventrículo al final de la fase de llenado, justo antes de que el corazón se contraiga.

[3] Medida clave en cardiología que indica el porcentaje de sangre que el ventrículo expulsa hacia la aorta en cada contracción del corazón respecto al volumen total de sangre presente en el ventrículo al final de la fase de llenado o diástole.

adicional que representa el volumen, permitiendo visualizar la precisión de la segmentación en relación con las estructuras anatómicas relevantes observadas en las imágenes de RM.

Figura 5.4: Corte, máscara de segmentación y superposición de ambas de una muestra del conjunto de datos ACDC



El conjunto de datos ACDC vuelve a tener varias regiones de interés en cada corte, por lo que es conveniente de nuevo realizar un estudio de la distribución de etiquetas. El cuadro 5.3 presenta una visión detallada de, al igual que en el cuadro 5.2, la distribución de etiquetas en el conjunto de datos ACDC, enfocándose en las estructuras cardíacas de interés.

Cuadro 5.3: Distribución de etiquetas en el conjunto de datos ACDC

Etiquetas	VD	MVI	VI
Presencia en cortes	2439	2817	2785
% presencia en cortes	81.90	94.59	93.52
Cortes por muestra	8.13	9.39	9.28

Los datos revelan que el MVI es la estructura más frecuentemente observada con 2817 apariciones, seguido muy de cerca por el VI con 2785, mientras que el VD se presenta con 2439 apariciones. En términos de presencia porcentual de los cortes, todas las regiones de interés aparecen en más del 80% de los cortes de imagen. Finalmente, la frecuencia promedio de todas las regiones es cercana a 10, dado que todas las imágenes tenían 10 cortes. Estas estadísticas vuelven a subrayar la prevalencia y visibilidad de estas estructuras cardíacas dentro del conjunto de datos, proporcionando información crucial para el desarrollo y ajuste de modelos.

5.4 M&Ms: Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge

El conjunto de datos M&Ms [Campello y cols. (2021)], denominado oficialmente como *Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge*, abarca un total de 375 estudios de RM. Estos estudios incluyen una variedad de patologías cardíacas, tanto en individuos saludables como en aquellos con cardiopatías, recolectadas en centros clínicos de España, Alemania y Canadá.

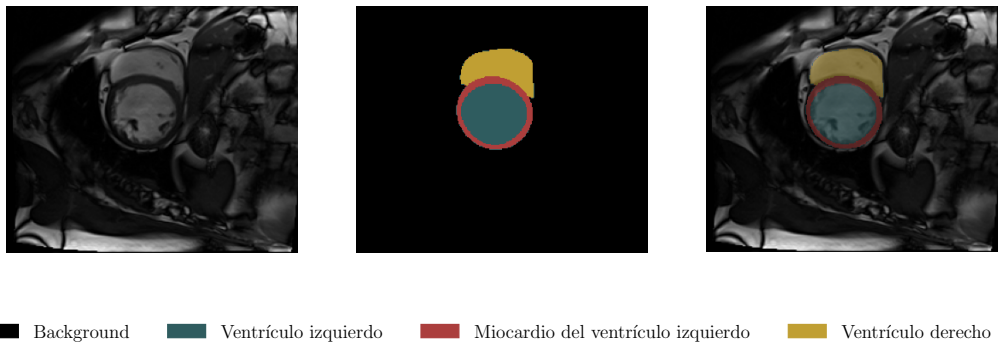
Para construir un conjunto de datos útil para la comunidad, se decidió basarse en el Procedimiento Operativo Estándar (SOP) del desafío ACDC MICCAI 2017 y corregir los contornos de acuerdo con este estándar. En particular, los contornos clínicos fueron corregidos por dos anotadores internos que debían llegar a un acuerdo sobre el resultado final.

El conjunto de entrenamiento incluye 150 imágenes anotadas de dos proveedores diferentes (75 cada uno) y 25 imágenes no anotadas de un tercer proveedor. El conjunto de test contiene 200 casos de prueba correspondiendo a 50 nuevos estudios de cada uno de los proveedores incluidos en el conjunto de entrenamiento y 50 estudios adicionales de un cuarto proveedor no visto anteriormente, que se utiliza para evaluar la generalización del modelo.

Especificaciones de las muestras

La figura 5.5 muestra un corte, una máscara de segmentación y la superposición de ambas de una muestra del conjunto de datos M&Ms, destacando diferentes estructuras del corazón mediante RM. En la máscara de segmentación resaltan el ventrículo izquierdo (VI, 1), el miocardio del ventrículo izquierdo (MVI, 2) y el ventrículo derecho (VD, 3). Este conjunto de datos incluye una amplia variedad de dimensiones de imagen, así como una cuarta dimensión que corresponde a diferentes instantes de tiempo para cada corte de cada muestra. Sin embargo, para evitar la saturación tanto espacial como temporal durante el entrenamiento, se ha optado por utilizar únicamente un instante temporal para cada corte, simplificando así el proceso y concentrando los esfuerzos de segmentación en una única fase del ciclo cardíaco por muestra.

Figura 5.5: Corte, máscara de segmentación y superposición de ambas de una muestra del conjunto de datos M&Ms



Al igual que en los dos anteriores conjuntos de datos, la máscara de segmentación contiene etiquetas distintas, por lo que se vuelve a requerir de nuevo un análisis de la distribución de etiquetas. El cuadro 5.5 ofrece una visión detallada de la distribución de etiquetas para el conjunto de datos M&Ms, enfocándose nuevamente en las estructuras cardíacas.

Cuadro 5.4: Distribución de etiquetas en el conjunto de datos M&Ms

Etiquetas	VI	MVI	VD
Presencia en cortes	3060	3072	2640
% presencia en cortes	77.39	77.69	66.77
Cortes por muestra	8.87	8.90	7.65

Los datos revelan que tanto el VI como el MVI tienen una presencia bastante equilibrada con más de 3000 apariciones en cortes y un porcentaje de presencia cercano al 77%, indicando una frecuencia de aparición similar entre estas dos estructuras. En contraste, el VD presenta una presencia menor con 2640 apariciones y un porcentaje del 66.77%, lo cual sugiere que el VD es menos frecuentemente capturado o identificado en los cortes analizados. Los cortes por muestra también reflejan esta tendencia, con el VI y el MVI mostrando valores más altos (8.87 y 8.90, respectivamente) en comparación con el VD (7.65), todo sobre una media de 11 cortes por muestra. De nuevo, este análisis subraya la variabilidad de la representación de las estructuras cardíacas dentro del conjunto de datos y, aunque no tan necesario como en anteriores casos, un aumento de datos para favorecer la generalización del modelo.

5.5 MSD: Medical Segmentation Decathlon

El conjunto de datos del Medical Segmentation Decathlon (MSD) [Simpson y cols. (2019a)] consta de 30 escaneos de RM monomodales del corazón completo, adquiridos durante una sola fase cardíaca, utilizando técnicas de respiración libre con puertas respiratorias y electrocardiogramas. Este conjunto de datos fue elegido específicamente para el 2013 Left Atrial Segmentation Challenge (LASC) debido a la combinación de un pequeño tamaño de conjunto de datos de entrenamiento y una gran variabilidad anatómica. Proporciona una variedad de niveles de calidad en las siguientes proporciones: 9 de alta calidad, 10 de calidad moderada, 6 con artefactos locales y 5 con alto nivel de ruido. La adquisición de las imágenes se realizó utilizando un escáner mediante una secuencia de adquisición de estado que abarcaba todo el corazón de una resolución de vóxel de $1.25 \times 1.25 \times 2.7$ mm^3.

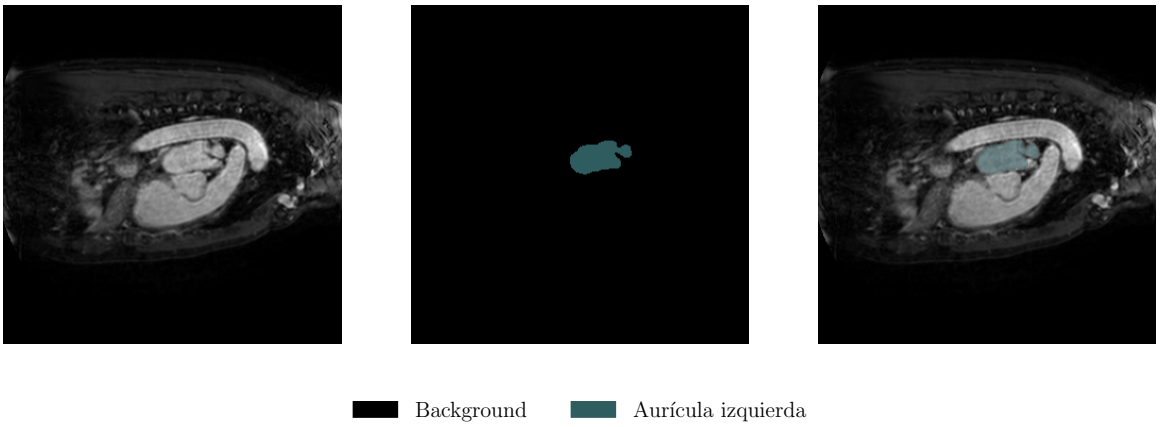
Especificaciones de las muestras

El conjunto de datos consta de muestras con dimensiones de 320×320, cada una complementada con un canal adicional que representa el volumen de la muestra. En este caso específico, las imágenes están segmentadas de manera que solo se diferencia una etiqueta con el valor 1 correspondiente a la AI, sobre un *background* con valor 0. Dado que el enfoque del estudio está centrado únicamente en la identificación de la aurícula izquierda en contraposición al fondo, no se requiere un análisis sobre el desbalanceo de datos, simplificando así el

proceso de evaluación y análisis de las imágenes segmentadas.

La siguiente figura 5.6 muestra un corte aleatorio de una muestra aleatoria junto su segmentación y la superposición de ambas para una mayor comprensión de las formas:

Figura 5.6: Corte, máscara de segmentación y superposición de ambas de una muestra del conjunto de datos MSD-Heart



6 Experimentación con modelos de segmentación

Este capítulo se enfoca en la evaluación de tres modelos avanzados de segmentación que han sido destacados en el estado del arte: MedSAM, U-Mamba y SAMed. Cada uno de estos modelos ha sido diseñado para abordar desafíos específicos en la segmentación de imágenes médicas, aprovechando técnicas de aprendizaje profundo para mejorar la precisión y la eficiencia.

El siguiente cuadro muestra cómo se han utilizado diferentes conjuntos de datos en los procesos de entrenamiento y test de los modelos mencionados.

Cuadro 6.1: Uso de los distintos conjuntos de datos en cada uno de los modelos

Conjunto de datos	MedSAM		U-Mamba		SAMed	
	Entrenamiento	Test	Entrenamiento	Test	Entrenamiento	Test
MM-WHS	✓	✓	✓	✓	✓	✓
ACDC	✓	✓	✓	✓	✓	✓
M&Ms	✗	✗	✓	✗*	✓	✗*
MSD	✗	✗	✓	✗	✓	✗

* Incluido como parte del conjunto de entrenamiento

Se ha decidido que MedSAM no empleará los conjuntos de datos M&Ms y MSD ni para el entrenamiento ni para los tests. La razón de esta decisión se fundamenta en que MedSAM ya ha sido expuesto a estos conjuntos de datos durante su fase de entrenamiento previa, lo cual hace que un reentrenamiento con los mismos datos no solo resulte en un uso ineficiente de recursos computacionales y tiempo, sino también en una evaluación poco fiable si estos datos se usaran para tests, dado que no medirían adecuadamente la capacidad del modelo para generalizar a nuevos datos. En contraste, los modelos U-Mamba y SAMed sí utilizarán estos conjuntos para entrenamiento con el fin de equiparar las condiciones de entrenamiento, pero, al igual que MedSAM, no los emplearán para tests para asegurar una evaluación justa y objetiva de su rendimiento.

Por otro lado, dadas las particularidades del conjunto MM-WHS, que carece de un conjunto de test público con máscaras disponibles, se ha optado por separar una parte del conjunto de entrenamiento original para utilizarla como conjunto de test, lo que ha impedido la creación de un conjunto de validación adicional sin comprometer significativamente el tamaño del conjunto de datos de entrenamiento. Considerando que MM-WHS es el único conjunto que incluye todas las estructuras cardíacas relevantes, reducir las muestras para formar un

conjunto de validación independiente limitaría excesivamente los recursos disponibles, restringiendo el entrenamiento del modelo con una cantidad adecuada de datos. Por ello, se ha decidido que el conjunto de test también funcione como conjunto de validación durante el entrenamiento, permitiendo ajustes y optimizaciones del modelo mientras se monitoriza su rendimiento en un conjunto que simula una situación de test real. Esta solución pragmática, aunque no ideal, es necesaria para maximizar el uso de todas las muestras disponibles y desarrollar un modelo robusto y efectivo, destacando la importancia de adaptar las estrategias de evaluación a las limitaciones específicas del conjunto de datos.

Finalmente, para los modelos MedSAM y SAMed, se aplicará Transfer Learning (TL), una técnica en la que un modelo desarrollado para una tarea específica se reutiliza como punto de partida para un modelo en una segunda tarea relacionada. Es especialmente útil en el ámbito de la IA, donde los modelos preentrenados pueden reducir significativamente el tiempo y los recursos computacionales necesarios para entrenar un nuevo modelo. MedSAM y SAMed aprovechan esta técnica, utilizando *checkpoints* (que también serán usados como puntos de partida de cada arquitectura) proporcionados por los desarrolladores que permiten a los usuarios iniciar sus modelos desde un estado avanzado, facilitando así una rápida convergencia y mejora en el rendimiento de nuevos conjuntos de datos o tareas.

6.1 MedSAM

Los creadores de MedSAM han desarrollado también una versión más ligera denominada LiteMedSAM, diseñada como modelo base para el desafío *CVPR 2024 MedSAM on Laptop Challenge*. Este modelo *lite* está optimizado para un entrenamiento y una inferencia rápidos, convirtiéndolo en una solución ideal para aplicaciones que requieren un menor consumo de recursos computacionales. La versión ligera fue desarrollada en dos etapas principales: en la primera, se destiló un *encoder* de imágenes más ligero, TinyViT [Wu y cols. (2022)], a partir del *encoder* de imágenes ViT utilizado en MedSAM, asegurándose de que los resultados de los *embeddings* de imágenes fueran consistentes. En la segunda etapa, se reemplazó el *encoder* de imágenes ViT de MedSAM por el TinyViT, y se afinó toda la cadena de procesamiento para optimizar el rendimiento.

Cuadro 6.2: Número de parámetros de MedSAM y LiteMedSAM

Modelo	Image encoder	Prompt encoder	Mask decoder	# Parámetros
MedSAM	89,670,912	6,220	4,058,340	93,735,472
LiteMedSAM	5,726,740			9,791,300

Resulta interesante comparar los modelos no solo en términos de su rendimiento sino también en relación con el número de parámetros que cada uno posee. Esta comparación puede revelar si los incrementos en la cantidad de parámetros se traducen en mejoras significativas en el rendimiento, o si modelos más ligeros como LiteMedSAM pueden ofrecer resultados comparables con una eficiencia computacional considerablemente mayor. Estos análisis son cruciales para determinar la viabilidad de implementar modelos más avanzados en platafor-

mas con recursos limitados, en el contexto de aplicaciones en tiempo real.

Protocolo de entrenamiento y entorno experimental

El preprocesamiento de imágenes y máscaras médicas es un paso esencial en la cadena de procesamiento para aplicaciones de VPC, especialmente en la segmentación médica. Todo este procedimiento se realiza mediante un primer *script* capaz de manejar imágenes de TC y RM, realizando una serie de transformaciones detalladas para optimizar las imágenes para el entrenamiento de modelos de DL.

Antes de proceder, verifica la consistencia de los datos asegurando que cada imagen tenga su correspondiente máscara, evitando discrepancias que podrían llevar a errores durante el entrenamiento. El procesamiento de imágenes y máscaras sigue los siguientes pasos:

1. **Carga y manipulación de máscaras:** Las máscaras se cargan y se convierten a *arrays* de Numpy. En primer lugar, incluye la eliminación de objetos pequeños mediante conectividad. Luego, se identifican los cortes con etiquetas con valores únicos distintos del 0, para más tarde recorta la máscara incluyendo únicamente estas secciones relevantes.

2. **Carga y ajuste de imágenes:**
   - **Para imágenes TC:** Se aplica un ajuste de ventana basado en los niveles y anchuras de ventana especificados, escalando las intensidades para mejorar la visibilidad de las estructuras relevantes. Las intensidades se normalizan posteriormente a la escala de 0-255.
   - **Para imágenes RM:** Se calculan los percentiles para robustecer el ajuste de intensidad contra valores atípicos. Después de recortar las intensidades según estos límites, se realiza una normalización similar a la aplicada en las imágenes TC.

3. **Recorte según zonas con etiquetas:** Tanto las imágenes como las máscaras se recortan para incluir solo los cortes que contienen etiquetas, eliminando así partes irrelevantes que podrían afectar la eficiencia del entrenamiento o sesgar el modelo.

4. **Guardado de datos preprocesados:** Los datos preprocesados se guardan en un formato comprimido para optimizar el uso del espacio de almacenamiento. Opcionalmente, para verificación, se pueden guardar también en formato NifTI.

Finalmente, se proporciona la opción de guardar las imágenes y máscaras procesadas en formato NifTI para comprobaciones de consistencia, permitiendo a los usuarios verificar visualmente el resultado del preprocesamiento antes de proceder con el entrenamiento del modelo. Una vez guardadas las imágenes en un formato comprimido, se hace uso de un segundo *script*.

El segundo *script* es una herramienta crucial para la conversión de archivos de imágenes y máscaras médicas desde el formato comprimido `.npz` hasta archivos individuales `.npy`. Este proceso es esencial para facilitar el manejo y procesamiento de datos en etapas posteriores de la *pipeline* de DL, como es el entrenamiento del modelo de segmentación. Los archivos `.npy` contendrán los cortes de las imágenes y máscaras, así cumpliendo con la necesidad de obtener

una entrada 2D para el modelo. En este segundo *script* se encuentra la primera diferencia entre ambos modelos. El tamaño de entrada de los modelos será distinta, en concreto, para el modelo MedSAM con mayor número de parámetros los pares de imágenes-máscaras serán de un tamaño de 1024×1024, mientras que para el modelo pequeño serán de un tamaño de 256×256. Tras esto, las imágenes se normalizarán entre 0 y 1 y se guardarán en memoria. Tras el proceso de preprocesamiento de los datos, finalmente se obtienen 230 pares de imágenes-máscaras que hacen un total de 6755 muestras 2D para el conjunto de entrenamiento y 110 pares para el conjunto de test.

El proceso de entrenamiento para los modelos de segmentación MedSAM y LiteMedSAM está meticulosamente diseñado para maximizar la efectividad utilizando diversas configuraciones técnicas. Este enfoque detallado permite adaptar cada aspecto del entrenamiento a las necesidades específicas de los datos y los objetivos del modelo.

En primer lugar, se configura el modelo para que tanto el *encoder* de imágenes como el *decoder* de máscaras estén activos durante el entrenamiento. Esta configuración completa permite que todo el sistema aprenda de manera integrada y se ajuste a las características complejas presentes en los datos médicos. Contrariamente, el *prompt encoder* tiene sus pesos congelados; este componente ya está optimizado para capturar y codificar información relevante del dominio, permitiendo así que el foco del entrenamiento se centre en las peculiaridades específicas de los nuevos conjuntos de datos.

El entrenamiento también incluye un paso de simulación de *bounding boxes* de expertos. Utilizando las máscaras de segmentación como base, se generan *bounding boxes* a los cuales se les añade un margen de error aleatorio entre 0 y 20 píxeles. Esta variabilidad imita las condiciones reales bajo las cuales los expertos determinan los límites de las estructuras de interés. Este paso es crucial para mejorar la robustez del modelo frente a las incertidumbres inherentes en la delimitación manual de las regiones de interés.

El entrenamiento se diversifica a través del uso de múltiples optimizadores: Adam [Kingma y Ba (2017)], AdamW [Loshchilov y Hutter (2019)], SGD [Kiefer y Wolfowitz (1952)], RMSprop [Ruder (2017)] y CLMR [Mortazi y cols. (2023)], con tasas de aprendizaje inicial variando entre 1e-05, 5e-05 y 1e-06 para LiteMedSAM. En el caso de MedSAM, únicamente se ha usado AdamW con tasas de aprendizaje inicial variando entre 1e-04, 1e-05 y 1e-06. Cada combinación de optimizador y tasa de aprendizaje se prueba en un modelo distinto para identificar cuál proporciona la mejor convergencia y eficacia. Los parámetros como β_1 y β_2, *momentum* y *alpha* se dejan en valores predeterminados para asegurar consistencia en los experimentos, mientras que todos los modelos aplican un *weight decay* de 1e-02 para prevenir el sobreajuste.

En cuanto a las funciones de pérdida, se otorga igual importancia a cada una, aunque se podría ponderar diferentemente. En MedSAM, se utilizan la DiceLoss y la CELoss, mientras que en LiteMedSAM se añade una tercera función, la MSELoss. Esta diversidad en las funciones de pérdida refleja la segunda diferencia clave entre los modelos. Cada modelo se entrena durante 15 épocas, y se selecciona para el conjunto de test el modelo que alcance la mejor pérdida en el conjunto de validación (test). Este enfoque garantiza que solo los modelos más

prometedores sean evaluados en la fase de test.

Resultados

A continuación se presentan los resultados obtenidos de los modelos de segmentación LiteMedSAM y MedSAM, los cuales han sido diseñados para evaluar la eficacia de diferentes configuraciones en la tarea de la segmentación médica. Los nombres de los modelos derivan de la combinación del nombre del modelo base, el tipo de optimizador y la tasa de aprendizaje inicial. Esta nomenclatura permite identificar de manera rápida y clara las especificaciones de cada configuración probada, por ejemplo:

`LMS_{Adam,AdamW,SGD,RMSprop,CLMR}_{1e-05,5e-05,1e-06}` o
`MS_AdamW_{1e-04,1e-05,1e-06}`.

Para cada arquitectura, se han elaborado dos cuadros de resultados: una para el DSC y otra para la NSD. Estas métricas son fundamentales para evaluar la precisión y eficacia de los modelos en la segmentación de estructuras anatómicas. Inicialmente, se presentan los resultados de los modelos basados en LiteMedSAM. Dado que LiteMedSAM es un modelo más ligero y compacto, facilita la experimentación con una amplia gama de configuraciones de manera más eficiente. Esta estrategia permite identificar las configuraciones más efectivas que posteriormente pueden aplicarse al modelo más grande y complejo, MedSAM.

Los cuadros están estructuradas de manera que las filas representan cada modelo entrenado y las columnas muestran los valores obtenidos de DSC y NSD para cada una de las etiquetas en ambos conjuntos de datos. Al final de cada fila, se calcula el promedio de estos valores para proporcionar una medida general de rendimiento del modelo. Además, para facilitar la visualización y comparación de resultados, los valores de cada columna están codificados por colores: el mejor valor se destaca en azul, el segundo mejor en rojo y el tercer mejor en morado. Esta codificación visual ayuda a identificar rápidamente qué configuraciones y modelos superan a otros en términos de rendimiento en las métricas seleccionadas.

Cuadro 6.3: Resultados con el modelo preentrenado de LiteMedSAM (DSC ↑)

Modelo	MM-WHS							ACDC			DSC
	MVI	AI	VI	AD	VD	AA	AP	VD	MVI	VI	
(L) Checkpoint base	54.963	92.259	95.837	93.327	95.091	84.153	71.352	72.119	12.013	70.918	74.203
LMS_Adam_1e-05	83.404	88.168	92.102	90.052	92.459	91.347	82.384	71.925	64.823	69.855	82.652
LMS_Adam_5e-05	67.760	83.022	84.686	86.776	84.619	86.009	69.331	63.401	58.643	66.339	75.059
LMS_Adam_1e-06	86.050	89.634	94.098	90.558	93.606	90.907	83.555	71.249	66.060	70.002	83.572
LMS_AdamW_1e-05	88.908	91.705	94.964	92.565	93.809	91.573	86.364	71.258	66.705	69.801	84.765
LMS_AdamW_5e-05	88.066	89.046	93.896	91.278	93.266	90.171	84.592	72.986	67.135	70.638	84.107
LMS_AdamW_1e-06	87.526	91.862	94.803	92.493	94.354	90.488	84.490	71.435	66.573	70.625	84.465
LMS_RMSprop_1e-05	83.147	87.766	92.157	89.610	90.748	89.490	80.871	70.688	65.521	69.282	81.928
LMS_RMSprop_5e-05	74.390	80.187	86.683	85.450	88.110	87.129	65.339	67.401	59.526	66.970	76.118
LMS_RMSprop_1e-06	85.643	89.508	93.569	90.419	93.520	89.912	82.488	70.568	65.716	69.355	83.070
LMS_SGD_1e-05	85.613	89.990	94.511	91.272	94.164	90.261	81.549	70.957	65.965	69.979	83.426
LMS_SGD_5e-05	86.319	89.597	93.878	90.768	93.180	89.821	82.737	70.859	66.172	69.763	83.309
LMS_SGD_1e-06	76.284	88.653	93.884	89.953	93.034	87.083	71.847	69.785	61.473	68.823	80.082
LMS_CLMR_1e-05	87.511	91.982	94.890	92.157	93.786	91.112	85.543	71.726	66.691	70.755	84.615
LMS_CLMR_5e-05	88.728	91.331	94.903	92.467	93.884	90.686	86.675	71.864	67.023	70.376	84.794
LMS_CLMR_1e-06	84.405	88.718	94.078	90.278	93.775	87.914	77.400	69.296	64.190	69.211	81.927

Cuadro 6.4: Resultados con con el modelo preentrenado de LiteMedSAM (NSD ↑)

Modelo	MM-WHS							ACDC			NSD
	MVI	AI	VI	AD	VD	AA	AP	VD	MVI	VI	
(L) Checkpoint base	69.211	83.340	90.313	82.250	88.933	69.813	56.490	78.549	35.058	74.146	72.810
LMS_Adam_1e-05	75.179	74.302	73.691	70.042	76.010	85.039	70.733	78.989	75.892	74.478	75.435
LMS_Adam_5e-05	61.286	60.922	55.341	57.966	51.542	72.222	49.390	75.516	75.520	73.398	63.310
LMS_Adam_1e-06	80.305	76.878	81.111	72.234	79.552	85.144	69.680	78.441	76.215	74.432	77.399
LMS_AdamW_1e-05	86.690	81.540	85.930	77.894	81.312	87.712	76.894	77.931	74.587	73.888	80.438
LMS_AdamW_5e-05	84.980	75.633	80.717	73.871	78.759	86.503	76.076	79.320	75.849	74.162	78.587
LMS_AdamW_1e-06	84.597	81.224	84.637	78.232	84.406	83.708	71.574	78.169	75.548	74.433	79.653
LMS_RMSprop_1e-05	76.724	71.959	74.655	68.250	70.599	83.289	67.551	78.225	76.358	74.140	74.175
LMS_RMSprop_5e-05	66.726	59.904	57.900	56.434	59.518	76.434	49.531	77.140	75.832	73.859	65.328
LMS_RMSprop_1e-06	79.686	76.271	77.709	71.891	79.391	83.989	69.241	77.947	76.218	74.081	76.642
LMS_SGD_1e-05	82.576	77.912	81.891	75.527	82.546	83.876	67.118	77.972	76.282	74.386	78.009
LMS_SGD_5e-05	80.626	77.118	80.112	71.966	77.867	84.988	69.785	78.243	75.991	74.157	77.085
LMS_SGD_1e-06	73.362	72.811	79.400	71.031	76.481	75.768	54.492	77.821	75.667	74.132	73.096
LMS_CLMR_1e-05	85.065	81.750	85.793	77.219	81.013	86.169	72.592	78.386	75.289	74.416	79.769
LMS_CLMR_5e-05	87.146	80.437	84.604	78.496	82.309	84.798	76.455	78.516	75.714	73.908	80.238
LMS_CLMR_1e-06	80.215	75.263	81.476	73.178	80.819	78.199	60.392	77.470	75.768	74.165	75.695

En el cuadro 6.3 se observa que los modelos `AdamW_1e-05`, `CLMR_1e-05` y `CLMR_5e-05` son los más destacados tomando en cuenta todas las etiquetas de ambos conjuntos de datos. Sin embargo, se observa que el modelo con mayor cantidad de mejores valores para las etiquetas es `Checkpoint base`. En particular, el modelo destaca en las etiquetas AI (92.259), VI (95.837), AD (93.327) y VD (95.091) del conjunto MM-WHS y en la etiqueta VI (70.918) del conjunto ACDC. Pese a esto, su DSC (74.203) viene a ser uno de los peores de la tabla. Esto es debido a que ciertas estructuras cardíacas como son el MVI (54.963) y la AP (71.352) no son tan fáciles de segmentar para el modelo y baja la media general.

Los resultados reflejados en el cuadro 6.4 para el modelo LiteMedSAM utilizando la métrica NSD evidencian la variabilidad y eficacia de las diferentes configuraciones. Respecto a los mejores modelos, sobresalen los mismos tres modelos que para la métrica DSC, aunque con un orden distinto. Para la métrica DSC los mejores modelos ordenados de forma descendente eran `CLMR_5e-05`, `AdamW_1e-05` y `CLMR_1e-05`, mientras que para la métrica NSD, la posición del primer y segundo modelo se intercambia. También sucede exactamente lo mismo respecto al modelo `Checkpoint base`, donde destaca en las mismas etiquetas en el conjunto MM-WHS con AI (83.340), VI (90.313), AD (82.250) y VD (88.933).

En general, las etiquetas VI y AD del conjunto MM-WHS parecen ser las mejor segmentadas, con muchos modelos alcanzando altos puntajes, posiblemente debido a características morfológicas más distintas que facilitan su segmentación. En contraste, las etiquetas como AP y las del conjunto ACDC tienden a ser las peor segmentadas, lo cual podría reflejar diferencias en la calidad de los datos, la complejidad de las estructuras o la menor cantidad de datos de entrenamiento para estas regiones específicas.

La variabilidad en el rendimiento entre diferentes configuraciones de optimizadores y tasas de aprendizaje muestra una alta sensibilidad del modelo a estos parámetros, lo que puede indicar la necesidad de una calibración más precisa. Esta sensibilidad sugiere posibles inesta-

bilidades en la arquitectura del modelo ante variaciones en los datos de entrada, resaltando la importancia de una cuidadosa selección y ajuste de los parámetros de entrenamiento para mejorar la precisión en la segmentación médica. La eficacia del optimizador CLMR, espacialmente a tasas de aprendizaje medias, es una fortaleza notable, ya que equilibra adecuadamente la exploración del espacio de parámetros sin caer en mínimos locales prematuramente.

Al analizar los resultados globales considerando ambas métricas, DSC y NSD, los tres modelos que destacan por su rendimiento son `CLMR_1e-05` (82.192), `CLMR_5e-05` (82.516) y `AdamW_1e-05` (82.602). Estos modelos no solo han mostrado ser superiores en términos de DSC, demostrando una excelente capacidad para capturar la extensión general de las estructuras cardíacas, sino que también exhiben altos valores de NSD, asegurando que los contornos de las estructuras segmentadas se ajustan de cerca a los contornos reales observados por los expertos. Estas combinaciones abre la base de las configuraciones para los modelos que parten de SAM y MedSAM, que serán presentados en los siguientes cuadros.

Cuadro 6.5: Resultados con los modelos preentrenados de SAM y MedSAM (DSC ↑)

Modelo	MM-WHS							ACDC			DSC
	MVI	AI	VI	AD	VD	AA	AP	VD	MVI	VI	
(S) Checkpoint base	53.044	85.456	89.054	84.856	86.037	86.400	66.989	90.048	74.549	90.454	80.689
S_AdamW_1e-04	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
S_AdamW_1e-05	56.718	75.216	80.145	81.105	76.972	72.030	58.143	80.597	64.777	83.123	72.883
S_AdamW_1e-06	76.437	86.715	91.121	88.959	89.561	89.965	76.811	91.514	81.106	92.825	86.501
(MS) Checkpoint base	72.141	89.127	95.138	88.610	93.328	84.920	69.426	96.199	33.462	97.184	81.953
MS_AdamW_1e-04	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MS_AdamW_1e-05	47.547	79.165	83.708	77.099	80.973	74.752	63.622	78.134	40.605	82.323	70.793
MS_AdamW_1e-06	89.015	90.798	94.837	91.784	93.616	90.734	87.524	94.530	89.049	95.898	91.778

Cuadro 6.6: Resultados con los modelos preentrenados de SAM y MedSAM (NSD ↑)

Modelo	MM-WHS							ACDC			NSD
	MVI	AI	VI	AD	VD	AA	AP	VD	MVI	VI	
(S) Checkpoint base	59.957	65.807	66.842	58.566	59.270	73.999	50.998	98.375	95.852	97.540	72.720
S_AdamW_1e-04	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
S_AdamW_1e-05	47.785	41.590	39.827	42.275	37.450	42.367	39.777	92.954	93.392	93.934	57.135
S_AdamW_1e-06	67.558	68.292	69.836	63.533	63.546	78.263	59.208	98.905	98.931	99.168	76.724
(MS) Checkpoint base	80.628	77.502	89.186	72.459	84.044	69.423	51.355	99.874	64.968	99.979	78.942
MS_AdamW_1e-04	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MS_AdamW_1e-05	50.403	48.613	50.204	40.045	42.933	48.898	47.763	91.920	82.715	93.642	59.714
MS_AdamW_1e-06	89.280	80.369	86.083	75.706	80.942	83.923	79.274	99.674	99.810	99.926	87.499

El cuadro 6.5 ilustra los modelos destacados tanto para SAM como para MedSAM. Estos modelos son `S_AdamW_1e-06` y `MS_AdamW_1e-06`, que exhiben desempeños superiores en varias etiquetas clave frente a los otros modelos entrenados dentro de la misma arquitectura. Se destaca especialmente en la etiqueta VI del conjunto MM-WHS con valores de 91.121 y 94.837, respectivamente. Otro modelo notable es el `Checkpoint base` de MedSAM, que logra la puntuación (97.184) más alta para la etiqueta VI del conjunto ACDC.

El cuadro 6.6 muestra que, al igual sucedía en el anterior cuadro, los modelos que incluyen la tasa de aprendizaje 1e-06 demuestran resultados sobresalientes respecto al resto, evidenciados por los altos valores de NSD en casi todas las etiquetas. En concreto, destaca el modelo `AdamW_1e-06` de MedSAM ya que para el conjunto de datos MM-WHS obtiene la mejor puntuación en todas las estructuras excepto en VI y VD, donde obtiene la segunda mejor puntuación. Además, para el conjunto ACDC, también obtiene grandes puntuaciones. Es por ello, que existe tanta diferencia frente al segundo mejor modelo.

En términos de las etiquetas mejor segmentadas según la métrica DSC, el VI generalmente muestra las puntuaciones más altas para ambos conjuntos de datos. Sucede lo contrario con la etiqueta MVI en el conjunto ACDC, aunque sigue siendo un valor alto. Para NSD, el VD y el MVI en el conjunto ACDC, y el VI en el conjunto MM-WHS han mostrado consistentemente altos valores, lo que indica una gran precisión en estas áreas. Por otro lado, algunas etiquetas como la AI y la AP en el conjunto MM-WHS tienen resultados más bajos, aunque siguen siendo valores altos.

Entre los modelos evaluados, los que utilizan AdamW han demostrado ser los más efectivos, destacándose especialmente los configurados con LiteMedSAM y MedSAM debido a su alta precisión en las métricas DSC y NSD. De nuevo, teniendo en cuenta que ambas métricas aportan lo mismo a la decisión final, los mejores modelos obtenidos han sido `CLMR_5e-05` (82.516) y `AdamW_1e-05` (82.602) de LiteMedSAM y `AdamW_1e-06` (89.639) de MedSAM. Para estos valores, también se debe tener en cuenta el número de parámetros que, LiteMedSAM con significativamente menos parámetros en comparación con MedSAM, muestra una eficacia notable, sugiriendo que la reducción de complejidad y el menor número de parámetros no comprometen considerablemente el rendimiento del modelo (siempre y cuando el experto lo considere así). Esto plantea una consideración valiosa sobre la eficiencia computacional contra la precisión en segmentación, indicando que modelos más ligeros como LiteMedSAM pueden ofrecer un equilibrio óptimo para aplicaciones que requieran menor carga computacional sin sacrificar demasiado en términos de precisión.

6.2 U-Mamba

A continuación, se presenta un análisis exhaustivo de los entrenamientos realizados utilizando la arquitectura U-Mamba. Dentro de esta arquitectura, los modelos se clasifican según la dimensión de entrada, ya sea en formatos 2D o 3D, y según la variante de la arquitectura específica empleada.

Existen dos configuraciones de la arquitectura principales en U-Mamba, discutidas anteriormente en la sección 3.2: `Bot` y `Enc`. La configuración `Bot` integra el bloque Mamba únicamente en el cuello de botella de la red, optimizando el procedimiento en la parte más crítica de la arquitectura, mientras que la configuración `Enc` distribuye bloques Mamba a lo largo de todo el *encoder*, permitiendo una integración más profunda de esta tecnología en el procesamiento de las imágenes. Esta diferenciación permite adaptar la arquitectura a necesidades específicas de segmentación, donde `UM_Bot` y `UM_Enc` se refieren a estas configuraciones únicas, y los sufijos `_2D` y `_3D`.

El siguiente cuadro, al igual que para MedSAM y su versión *lite*, muestra el número de parámetros de cada una de las variantes según el conjunto de datos empleado para su entrenamiento:

Cuadro 6.7: Número de parámetros de las variantes de la arquitectura U-Mamba

Modelo	# Parámetros	
	ACDC	MMWHS
Bot_2D	28,760,276	47,523,952
Enc_2D	27,561,556	47,666,464
Bot_3D	41,937,108	42,125,224
Enc_3D	43,056,788	42,880,784

Protocolo de entrenamiento y entorno experimental

El preprocesamiento de datos es una fase crucial en la preparación para el entrenamiento. Los conjuntos de datos para U-Mamba deben estar organizados meticulosamente para asegurar una integración sin problemas. Cada caso de entrenamiento dentro del conjunto de datos está asociado con un identificador único, que sirve para enlazar las imágenes con sus correspondientes mapas de segmentación. Estos datos incluyen múltiples canales de entrada, cada uno almacenado en un archivo de imagen por separado, lo cual es necesario para mantener la flexibilidad en el procesamiento de diferentes tipos de imágenes médicas. En este caso concreto, ningún conjunto de datos de los que se vaya a usar tiene varios tipos de imágenes, por lo que no es necesario tenerlo en cuenta.

Respecto a los conjuntos de datos, tienen que tener 3 componentes indispensables:

- **Imágenes en bruto:** Corresponden a las modalidades de entrada diversas y deben tener la misma geometríca para garantizar la coherencia y la exactitud en el entrenamiento.
- **Máscaras de segmentación:** Estos mapas son representaciones enteras donde cada valor corresponde a una clase semántica, siendo crucial que compartan la misma geometría que las imágenes a las que corresponden.
- **Archivo JSON del conjunto de datos:** Este archivo contiene metadatos esenciales como los nombres de los canales y las etiquetas de segmentación, proporcionando un contexto necesario para el procesamiento y entrenamiento. Esta configuración asegura que cada entrada y su correspondiente segmentación sean procesadas correctamente, aplicando las normalizaciones y ajustes adecuados según el tipo de imagen y las necesidades específicas del modelo.

El preprocesamiento a nivel de muestras se basa en el preprocesamiento de imagen realizado en nn-UNet e incluye varios pasos esenciales:

1. **Recorte:** El primer paso es recortar los datos para centrarse solo en regiones de interés, eliminando las áreas de valores no nulos que no contribuyen a la información relevante

para el análisis. Este enfoque reduce la carga computacional y mejora la eficiencia del procesamiento.

2. **Resampling:** Dado que las CNN no comprenden intrínsecamente los espaciamientos de los vóxeles, es esencial remuestrear todas las imágenes al espaciamiento de vóxeles mediano del conjunto de datos respectivo. Para los datos de imagen se utiliza la interpolación *spline* de tercer orden, mientras que para las máscaras de segmentación se emplea la interpolación del vercino más cercano. Este paso asegura que la red pueda aprender correctamente la semántica espacial sin ser confundida por diferencias en las escalas de resolución.

3. **Normalización:** Se normalizan las intensidades de las imágenes para que la red pueda procesar efectivamente las características a través de diversas escalas de intensidad. En el caso de las imágenes TC, la normalización se basa en estadísticas de todo el conjunto de datos correspondiente, ajustando los valores de intensidad a percentiles específicos seguidos de una normalización *z-score*. Para las imágenes de RM y otras modalidades, se aplica una normalización *z-score* individual.

Estos pasos de preprocesamiento son fundamentales para preparar los datos para el entrenamiento y la inferencia, asegurando que las redes puedan funcionar de forma óptima. En relación al entrenamiento, U-Mamba sigue la capacidad de nn-UNet para configurar los hiperparámetros automáticamente para diversos conjuntos de datos de segmentación. Como hiperparámetros básicos se destaca como optimizador SGD y tasa de aprendizaje inicial 1e-02. Cada modelo está entrenado durante 50 épocas y se utiliza para el conjunto de test el modelo de la época que menor pérdida haya obtenido para el conjunto de validación (test).

El ajuste de la tasa de aprendizaje se realiza mediante una clase propia como es `PolyLR` que ajusta esta tasa a lo largo del entrenamiento siguiendo una función de decaimiento polinomial. Esta función disminuye la tasa de aprendizaje desde un valor inicial hasta aproximadamente cero, basándose en el número total de épocas previstos para el entrenamiento y definiendo la agresividad de dicho decaimiento con un parámetro.

$$lr' = lr \times \left(1 - \frac{epoca\_actual}{total\_epocas}\right)^{exponente} \tag{6.1}$$

Resultados

Como se menciona anteriormente, existen cuatro variantes distintas de la arquitectura U-Mamba. Cada una de estas variantes será entrenada de manera independiente para los conjuntos de datos MM-WHS y ACDC, sin mezclar muestras entre ellos. Como resultado, aunque hay cuatro variantes de U-Mamba, se crean ocho modelos en total, dado que cada variante se entrena con dos conjuntos de datos distintos.

Además, para aprovechar al máximo los recursos y los datos disponibles, se plantea entrenar la variante que muestre el mejor desempeño en un "super" conjunto de datos que agrupe todas las muestras de todos los conjuntos disponibles. Esta estrategia, *a priori*, permite maximizar el aprendizaje y la generalización del modelo seleccionado, siguiendo las directrices

y configuraciones detalladas en el cuadro 6.1. Los siguientes cuadros muestran los resultados para el conjunto de test de ambos conjuntos de datos para cada variante de la arquitectura.

Cuadro 6.8: Resultados de las arquitecturas U-Mamba 'Bot' y 'Enc' entrenadas desde cero (DSC ↑)

Modelo	MM-WHS							ACDC			DSC
	MVI	AI	VI	AD	VD	AA	AP	VD	MVI	VI	
UM_Bot_2D	84.518	81.014	90.764	81.497	88.512	72.558	64.747	90.574	89.948	94.558	83.869
UM_Enc_2D	86.076	81.637	93.076	87.512	89.270	74.623	68.514	90.726	89.990	94.499	85.592
UM_Bot_3D	84.052	87.700	91.286	88.272	90.320	79.214	76.192	90.315	89.593	94.082	87.103
UM_Enc_3D	83.873	86.913	91.056	81.861	88.418	83.409	73.337	88.487	86.838	92.711	85.690

Cuadro 6.9: Resultados de las arquitecturas U-Mamba 'Bot' y 'Enc' entrenadas desde cero (NSD ↑)

Modelo	MM-WHS							ACDC			NSD
	MVI	AI	VI	AD	VD	AA	AP	VD	MVI	VI	
UM_Bot_2D	77.676	63.901	71.594	55.980	63.087	69.539	51.505	98.290	99.480	99.412	75.046
UM_Enc_2D	79.077	63.245	77.395	64.075	64.825	69.600	54.362	98.048	99.426	99.471	76.952
UM_Bot_3D	78.917	70.575	76.019	66.046	69.970	72.912	64.494	98.223	99.523	99.465	79.614
UM_Enc_3D	78.082	66.771	75.878	54.991	64.974	77.453	60.662	97.141	98.894	98.426	77.327

En el cuadro 6.8 en la evaluación usando DSC, la variante `UM_Enc_2D` demuestra ser particularmente efectiva, superando a `UM_Bot_2D` en la mayoría de las etiquetas, destacando en la segmentación del VI (93.076) y la AD (87.512) en el conjunto MM-WHS con altos valores que reflejan una segmentación precisa de estas estructuras cardíacas grandes. Similarmente, en el cuadro 6.9 para NSD, `UM_Enc_2D` muestra excelentes capacidades para capturar los contornos precisos del VI, indicando su habilidad para detallar finalmente las fronteras anatómicas necesarias para aplicaciones clínicas críticas. Las variantes 3D, como `UM_Bot_3D`, también se destacan, especialmente en la precisión de los contornos de estructuras como el VD (98.223) y el MVI (99.523) en ambos conjuntos, con `UM_Bot_3D` mostrando superioridad en la captura de contornos tridimensionales que son fundamentales para un diagnóstico y planificación quirúrgica precisos.

Sin embargo, todas las variantes enfrentan desafíos en la segmentación de la AP, como se refleja en los bajos valores tanto en DSC como en NSD. Esto sugiere que las estructuras más pequeñas o anatómicamente complejas aún representan un obstáculo significativo para los modelos actuales, destacando una área crucial para futuras mejoras y desarrollos. Tomando de nuevo en cuenta una evaluación equilibrada que considera la métrica DSC como la NSD, los tres mejores modelos de esta arquitectura son `UM_Enc_2D` (81.272), `UM_Enc_3D` (81.509) y `UM_Bot_3D` (83.359). Finalmente, echando la vista atrás, los tres mejores modelos obtenidos han sido `LMS_AdamW_1e-05` (82.602), `UM_Bot_3D` (83.359) y `MS_AdamW_1e-06` (89.639).

Como último comentario sobre los resultados de esta arquitectura, el modelo que intentó abarcar todos los conjuntos de datos no ha proporcionado resultados satisfactorios. Este enfoque, aunque ambicioso en su intento de generalizar a través de múltiples fuentes de datos

con una única arquitectura, enfrentó dificultades significativas para adaptarse a las variaciones y especificaciones inherentes a cada conjunto de datos. Este hallazgo destaca la necesidad de ajustar más específicamente las arquitecturas o de considerar estrategias de entrenamiento diferentes para manejar mejor la heterogeneidad entre diferentes tipos de muestras médicas.

6.3 SAMed

A continuación, se presentan los resultados de SAMed, la tercera y última arquitectura examinada en este TFM. SAMed representa una evolución significativa en términos de rendimiento dentro de la arquitecturas estudiadas, especialmente en su variante SAMed_h, la cual, sin necesidad de ajustes adicionales, alcanza un rendimiento superior al de su predecesor SAMed.

A pesar de que el tamaño del modelo en su versión `vit_h` es mucho más grande comparado con la versión `vit_b`, el punto de control de LoRA para SAMed_h solo muestra un incremento modesto. Esto implica costes de despliegue y almacenamiento casi equivalentes. Es por ello que, además con el contexto de la industria, donde frecuentemente se prefieren modelos más grandes y de mayor rendimiento por sus ventajas, se prefiere realizar los experimentos únicamente con este modelo más grande.

El siguiente cuadro, al igual que para las anteriores arquitecturas, muestra la diferencia en el número de parámetros entre ambas variantes de la arquitectura `vit` en su versión `vit_b` y `vit_h`.

Cuadro 6.10: Número de parámetros de las variantes 'vit' en la arquitectura SAMed

Modelo	# Parámetros
SAMed	92,052,148
SAMed_h	638,474,453

Protocolo de entrenamiento y entorno experimental

El preprocesamiento de datos se enfoca en preparar las imágenes de una manera que mejore la capacidad del modelo para aprender características relevantes sin ser afectado por variaciones innecesarias o artefactos en los datos. Inicialmente, las intensidades de las imágenes se normalizan, estandarizando los rasgos de intensidad de las imágenes a través del conjunto de datos. Esta normalización asegura que el modelo no sea sesgado hacia características particulares de ciertos conjuntos de datos debido a diferencias en la captura o el procesamiento de imagen.

En segundo lugar, las imágenes son recortadas y remuestreadas a un tamaño estándar 224×224 para asegurar que todas las entradas al modelo tengan las mismas dimensiones. El ajuste del tamaño también se lleva a cabo para las máscaras de segmentación, utilizando técnicas que preservan las etiquetas de clase sin introducir distorsiones significativas.

Tras esto, los datos preparados se almacenan en un formato adecuado para su uso durante el entrenamiento, manteniendo la asociación entre las imágenes y sus etiquetas de segmentación correspondientes. Esta organización facilita el acceso eficiente durante el proceso de entrenamiento y asegura que se mantenga la correspondencia correcta entre las imágenes y sus máscaras.

Finalmente, cada imagen y su correspondiente máscara se someten a una serie de transformaciones antes de la inferencia por la red para asegurar la uniformidad de los datos. Esto incluye rotaciones y volteos aleatorios, lo que ayuda a aumentar la generalidad del modelo al exponerlo a diferentes orientaciones de los mismos objetos. Este tipo de aumento de datos es vital para entrenar modelos robustos que pueden interpretar correctamente imágenes médicas.

El proceso de entrenamiento de los modelos se basa en una configuración detallada de hiperparámetros que juega un papel crucial en la optimización del aprendizaje y la eficiencia. La configuración incluye la definición del número de épocas, tasas de aprendizaje, elección de optimizadores y ajuste dinámico de las tasas de aprendizaje a largo del proceso.

El entrenamiento está diseñado para ejecutarse a través de un máximo de 50 épocas, pero incluye un mecanismo de detención anticipada que interrumpe el proceso si se cumplen ciertas condiciones, como la convergencia o el estancamiento del aprendizaje, para evitar el sobreentrenamiento. La tasa de aprendizaje inicial está configurada en 5e-03, 5e-04 y 5e-05 y puede ser ajustada durante el entrenamiento para responder a la dinámica del aprendizaje. Este incluye un período de calentamiento donde la tasa de aprendizaje se incrementa gradualmente desde un valor inicialmente más bajo, permitiendo un inicio suave del entrenamiento.

En cuanto a los optimizadores, se emplea AdamW y SGD dependiendo de la configuración específica. AdamW es favorecido por su manejo eficiente del decaimiento de los pesos y su adaptabilidad en escenarios de DL, configurado con un *weight decay* de 0.1. El SGD, usado con un *momentum* de 0.9 y un *weight decay* de 1e-04, es la otra opción viable que favorece la convergencia estable en espacios de solución complejos. Para la función de pérdida, se utiliza una combinación de entropía cruzada y DiceLoss, donde esta última ayuda a contrarrestar las deficiencias de la primera en situaciones de desbalance de clases. La pérdida total es una combinación ponderada de estas dos métricas, con un peso predominante en la DiceLoss, enfatizando la importancia de la precisión en la segmentación sobre la clasificación general.

Por último, se incorporan técnicas para acelerar el entrenamiento y mejorar la eficiencia del modelo. Esto incluye el uso de precisión mixta cuando está habilitado, lo que permite una reducción significativa en el consumo de memoria y un incremento de la velocidad de entrenamiento al utilizar cálculos de menor precisión donde es posible. También se contempla la compilación del modelo para optimizar su rendimiento en un *hardware* específico y, se activa el uso de TF32, que ofrece un equilibrio óptimo entre rendimiento y precisión.

Resultados

En los siguientes cuadros se presentan los resultados obtenidos de los modelos de la arquitectura SAMed, donde cada modelo es, de nuevo, identificado y diferenciado principalmente por la combinación del optimizador utilizado y la tasa de aprendizaje aplicada, además del prefijo correspondiente a la arquitectura.

De nuevo, con la intención de optimizar el uso de recursos y maximizar el potencial de los datos disponibles, se considera conveniente emplear la variante de mayor éxito en un "super" conjunto de datos que consolide todas las muestras de todos los conjuntos disponibles. Esta táctica, en teoría, facilita una mejora significativa en el aprendizaje y la generalización del modelo elegido, siguiendo las pautas y configuraciones explicitadas en el cuadro 6.1. Los cuadros subsiguientes presentan los resultados obtenidos en el conjunto de test para cada variante de la arquitectura con ambos conjuntos de datos:

Cuadro 6.11: Resultados de las variantes entrenadas a partir de la arquitectura SAMed_h (DSC ↑)

Modelo	MM-WHS							ACDC			DSC
	MVI	AI	VI	AD	VD	AA	AP	VD	MVI	VI	
Sed_SGD_5e-03	77.119	80.915	87.609	74.392	80.880	78.249	65.111	82.515	76.346	84.548	78.768
Sed_SGD_5e-04	61.692	51.185	81.841	14.782	58.474	57.093	31.513	68.522	65.518	77.433	56.805
Sed_SGD_5e-05	41.315	36.837	60.351	18.968	18.300	0.000	0.003	58.608	58.322	69.312	36.202
Sed_AdamW_5e-03	86.126	87.140	92.536	85.098	88.850	81.252	73.601	0.000	0.000	0.000	59.461
Sed_AdamW_5e-04	84.508	83.802	92.459	84.537	87.908	79.192	70.893	90.883	88.185	92.979	85.535
Sed_AdamW_5e-05	79.315	82.317	89.517	77.264	79.643	68.518	64.904	88.858	85.042	91.309	80.669

Cuadro 6.12: Resultados de las variantes entrenadas a partir de la arquitectura SAMed_h (NSD ↑)

Modelo	MM-WHS							ACDC			NSD
	MVI	AI	VI	AD	VD	AA	AP	VD	MVI	VI	
Sed_SGD_5e-03	62.887	47.551	58.461	36.706	46.509	58.925	43.189	94.331	93.924	92.777	63.526
Sed_SGD_5e-04	41.496	22.691	40.197	15.136	21.895	32.475	15.512	83.016	86.283	86.601	44.530
Sed_SGD_5e-05	26.973	10.169	19.688	6.759	10.614	0.000	0.030	72.309	82.130	78.156	30.683
Sed_AdamW_5e-03	79.546	66.077	76.548	56.587	64.347	71.104	59.803	0.000	0.000	0.000	47.401
Sed_AdamW_5e-04	75.918	59.932	73.986	55.066	60.738	67.131	52.232	98.752	99.544	99.324	74.262
Sed_AdamW_5e-05	67.096	53.980	67.741	39.991	48.304	58.241	44.807	97.734	98.392	98.425	67.471

Los modelos SAMed, analizados mediante las métricas DSC y NSD, exhiben un desempeño variante que refleja tanto fortalezas como las limitaciones de cada configuración, especialmente éstas últimas, bajo diferentes parámetros de entrenamiento. El cuadro 6.11 revela que los modelos entrenados con AdamW, especialmente en las configuraciones `AdamW_5e-04` y `AdamW_5e-05`, muestran un rendimiento superior en la mayoría de las etiquetas. El modelo `AdamW_5e-04` se destaca notablemente, alcanzando los valores más altos en etiquetas críticas como el VI (92.979) y el VD (90.883) en el conjunto de datos ACDC. En cuanto a la métrica NSD en el cuadro 6.12, `AdamW_5e-04` y `AdamW_5e-05` nuevamente sobresalen, proporcionando los valores más altos en casi todas las etiquetas y, particularmente, `AdamW_5e-04` demuestra una alta precisión en las segmentaciones del VD en el conjunto ACDC y el VI en el conjunto

MM-WHS. Por otro lado, los modelos entrenados con SGD, como `SGD_5e-03` también muestran un buen rendimiento, aunque son generalmente superados por las variantes de AdamW.

Las etiquetas como la AP y la AA en MM-WHS muestran valores más bajos, sugiriendo, de nuevo, que dichas estructuras suponen un desafío mayor que otras más grandes. Estos resultados subrayan la importancia de seleccionar cuidadosamente los hiperparámetros y las configuraciones del modelo para optimizar tanto la precisión general de la clasificación como la exactitud de los contornos en la segmentación médica. Entre los modelos SAMed, después de considerar igual de importantes ambas métricas, los tres mejores modelos son `SGD_5e-03` (71.147), `AdamW_5e-05` (74.070) y `AdamW_5e-04` (79.899). Estos modelos han demostrado un rendimiento superior frente al resto, destacando la decisión del optimizador AdamW frente a SGD en los entrenamientos. Finalmente, teniendo en cuenta todos los modelos entrenados con sus resultados, han sido `LMS_AdamW_1e-05` (82.602), `UM_Bot_3D` (83.359) y `MS_AdamW_1e-06` (89.639). Como último comentario sobre los resultados de la arquitectura SAMed, el modelo SAMed_h con la mejor configuración obtenida ha conseguido resultados relativamente mejores, con puntuaciones de 83.498 en DSC y 73.379 en NSD. Este rendimiento, aunque destacable dentro de su serie, aún no supera a otros modelos más especializados, como el configurado con AdamW_5e-04 y entrenado exclusivamente con los conjuntos MM-WHS y ACDC.

7 Resultados y conclusiones

El trabajo realizado ha permitido analizar exhaustivamente diversas arquitecturas de modelos de segmentación aplicadas a imágenes médicas, específicamente enfocadas en la segmentación de estructuras cardíacas. A lo largo del proyecto, se han entrenado y evaluado múltiples modelos, destacando las configuraciones optimizadas que ofrecen un mejor rendimiento según las métricas DSC y NSD. El análisis detallado de los resultados ha permitido establecer un *ranking* objetivo de las diferentes arquitecturas, identificando las características que contribuyen al éxito de ciertos modelos en la segmentación precisa de estructuras anatómicas críticas.

Cuadro 7.1: Ranking de todos los modelos entrenados

	Modelo	DSC	NSD	Total
1	**MS_AdamW_1e-06***	**91.778**	**87.499**	**89.639**
2	**UM_Bot_3D**	**87.103**	**79.614**	**83.359**
3	**LMS_AdamW_1e-05***	**84.765**	**80.438**	**82.602**
4	LMS_CLMR_5e-05*	84.794	80.238	82.516
5	LMS_CLMR_1e-05*	84.615	79.769	82.192
6	LMS_AdamW_1e-06*	84.465	79.653	82.059
7	**S_AdamW_1e-06**	**86.501**	**76.724**	**81.613**
8	UM_Enc_3D	85.690	77.327	81.509
9	LMS_AdamW_5e-05*	84.107	78.587	81.347
10	UM_Enc_2D	85.592	76.952	81.272
11	LMS_SGD_1e-05*	83.426	78.009	80.718
12	LMS_Adam_1e-06*	83.572	77.399	80.486
13	(MS) Checkpoint base	81.953	78.942	80.448
14	LMS_SGD_5e-05*	83.309	77.085	80.197
15	**Sed_AdamW_5e-04**	**85.535**	**74.262**	**79.899**
16	LMS_RMSprop_1e-06*	83.070	76.642	79.856
17	UM_Bot_2D	83.869	75.046	79.458
18	LMS_Adam_1e-05*	82.652	75.435	79.044
19	LMS_CLMR_1e-06*	81.927	75.695	78.811
20	Sed_AdamW_5e-04*	83.498	73.379	78.439
21	LMS_RMSprop_1e-05*	81.928	74.175	78.052
22	(S) Checkpoint base	80.689	72.720	76.705
23	LMS_SGD_1e-06*	80.082	73.096	76.589
24	Sed_AdamW_5e-05	80.669	67.471	74.070
25	(LMS) Checkpoint base	74.203	72.810	73.507

Continúa en la siguiente página

Cuadro 7.1: Ranking de todos los modelos entrenados (*continuación*)

	Modelo	DSC	NSD	Total
26	Sed_SGD_5e-03	78.768	63.526	71.147
27	LMS_RMSprop_5e-05*	76.118	65.328	70.723
28	LMS_Adam_5e-05*	75.059	63.310	69.185
29	MS_AdamW_1e-05*	70.793	59.714	65.254
30	S_AdamW_1e-05	72.883	57.135	65.009
31	Sed_AdamW_5e-03	59.461	47.401	53.431
32	Sed_SGD_5e-04	56.805	44.530	50.668
33	Sed_SGD_5e-05	36.202	30.683	33.443
34	S_AdamW_1e-04	0.000	0.000	0.000
-	MS_AdamW_1e-04*	0.000	0.000	0.000
-	UM_Bot_3D*	0.000	0.000	0.000

* Entrenado con todos los conjuntos de datos

El *ranking* de todos los modelos entrenados ofrece una visión completa que trasciende la simple ubicación de cada modelo en una escala de rendimiento; es crucial también para determinar la posición relativa de cada arquitectura en función de su éxito en las pruebas realizadas. Esta clasificación no solo destaca cuáles modelos y arquitecturas sobresalen, sino que también facilita la identificación de aquellas que pueden requerir refinamientos adicionales. Además, este *ranking* es instrumental para seleccionar el modelo más prometedor de cada arquitectura, con el objetivo de someterlo a pruebas más detalladas y visuales utilizando diversas muestras del conjunto de test. Esta estrategia no solo proporciona una verificación adicional del rendimiento teórico de los modelos en condiciones controladas, sino que también ofrece una oportunidad para evaluar su efectividad en aplicaciones prácticas, permitiendo observaciones más profundas y directas de su capacidad para manejar datos reales en un entorno clínico.

Figura 7.1: Algunos ejemplos de las predicciones realizadas con los modelos

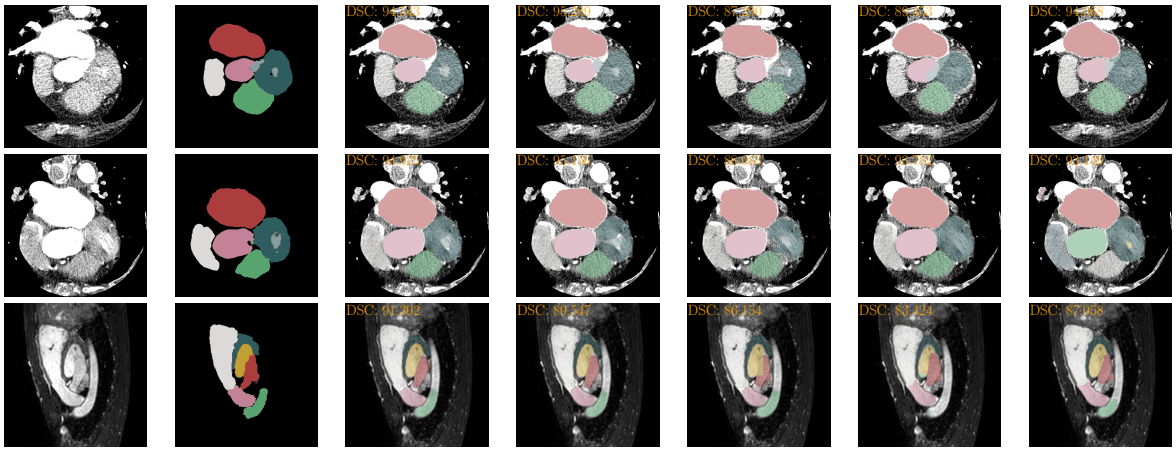
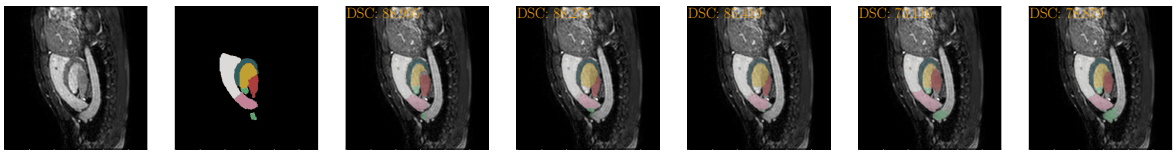


Figura 7.1: Algunos ejemplos de las predicciones realizadas con los modelos (*continuación*)



Las imágenes adjuntadas en la figura 7.1 del conjunto de datos MM-WHS presentan un interesante conjunto de resultados de las arquitecturas trabajadas en términos de DSC. Para la primera serie de imágenes, la arquitectura MedSAM (95.289) muestra un excelente rendimiento, superando ligeramente a LiteMedSAM (94.543) y siendo seguida de cerca por U-Mamba (94.068). Tras estos resultados, las arquitecturas SAM (87.290) y SAMed (89.453) se alejan bastante, lo que indica que no manejan tan bien los detalles de grano fino como las otras arquitecturas. Del mismo modo, para la segunda serie de imágenes, los resultados son consistentemente altos en todas las arquitecturas excepto en SAM (88.683), con MedSAM (95.156) y LiteMedSAM (94.925) mostrando nuevamente un rendimiento superior frente al resto, seguidas muy de cerca de U-Mamba (93.149). Estos altos valores para ambas series demuestran que estas arquitecturas están bien adaptadas para trabajar con imágenes de TC.

En cambio, la evaluación de las imágenes de RM (tercera y cuarta serie) ofrecen una perspectiva diferente. En la tercera serie LiteMedSAM (91.302) supera ampliamente a MedSAM (89.547) pero SAM (86.154) y SAMed (83.424) siguen mostrando un rendimiento más bajo con respecto al resto. Para la cuarta y última serie, el cambio más significativo aparece en U-Mamba (78.879), dado que para las otras muestras siempre se situaba cerca de las dos mejores arquitecturas, para esta muestra no ha sucedido lo mismo. Como último detalle, las muestras de imágenes de RM en comparación con las imágenes de TC muestran un rendimiento más bajo, posiblemente debido a diferencias en la calidad de la imagen o en las características intrínsecas del tejido que son más difíciles de capturar con estas arquitecturas.

Por lo general, los resultados obtenidos en la parte de arriba del *ranking* son realmente buenos comparados con el estado del arte actual. Al analizar más detalladamente, es evidente que obtener contexto de varios conjuntos de datos mejora notablemente la actuación de los modelos. Esto se refleja claramente en el *ranking*, donde la mayoría de los modelos mejor clasificados han sido entrenados utilizando múltiples conjuntos de datos. Además, tanto el uso del optimizador AdamW como la implementación de tasas de aprendizaje bajas han sido claves para alcanzar una buena convergencia, en comparación con tasas ligeramente más altas o el uso de SGD. Finalmente, aunque los modelos desarrollados han demostrado ser efectivos, aún existen áreas para futuras investigaciones, como la exploración de nuevas configuraciones de hiperparámetros, la optimización de la eficiencia computacional y la adaptación de los modelos a otros tipos de imágenes médicas. Estos esfuerzos contribuirán a mejorar aún más la precisión y aplicabilidad de los modelos en el campo de la medicina, consolidando su relevancia en el diagnóstico y tratamiento de enfermedades cardiovasculares.

7.1 Trabajo futuro

Se plantean varias líneas de investigación y desarrollo que podrían extender y enriquecer las capacidades del presente proyecto. Estas propuestas buscan no solo mejorar la precisión de los modelos actuales sino también ampliar su aplicabilidad en contextos clínicos, proporcionando herramientas más robustas y versátiles para la toma de decisiones médicas.

Una dirección prometedora para futuras investigaciones es la posibilidad de extraer biomarcadores cuantitativos, como el volumen de las cavidades cardíacas, directamente de las imágenes procesadas por los modelos de segmentación. Este tipo de información es fundamental para evaluar la función cardíaca y podría ser crucial para la monitorización y diagnóstico de condiciones preexistentes o emergentes en los pacientes. La precisión en la medición de estos volúmenes permitiría no solo entender mejor la patología cardíaca del paciente sino también ajustar los tratamientos de manera más efectiva.

Otro ámbito de interés es el desarrollo de modelos capaces de detectar y segmentar calcificaciones dentro de los vasos sanguíneos o tejidos cardíacos. La detección de calcificaciones es un indicador vital de enfermedades cardiovasculares, incluyendo la arteriosclerosis, que puede llevar a complicaciones severas como infartos o insuficiencia cardíaca. Incorporar esta capacidad en los modelos no solo mejoraría la evaluación del riesgo cardiovascular sino que también facilitaría la planificación de intervenciones quirúrgicas, asegurando que los pacientes reciban el cuidado preventivo necesario antes de procedimientos de alto riesgo como los trasplantes de órganos.

Finalmente, sería de gran valor desarrollar modelos que no solo describan el estado actual del corazón, sino que también predigan el riesgo de complicaciones futuras, como el fallo cardíaco, especialmente en el contexto de cirugías mayores como los trasplantes de hígado, basándose en los biomarcadores extraídos e información externa como edad, sexo, altura, peso, entre otros. Recordar que este trabajo nace de un proyecto con el objetivo de evaluar la función cardíaca preoperatoria para prever complicaciones dentro de la cirugía, lo que resalta la importancia de predecir posibles desenlaces negativos basados en la evaluación cardíaca previa a la cirugía. La implementación de modelos predictivos podría transformar significativamente la manera en que se preparan y gestionan estos pacientes, optimizando los resultados y minimizando los riesgos asociados con el trasplante.

Estas áreas de trabajo futuro no solo expanden el alcance técnico y clínico del proyecto actual sino que también subrayan la relevancia de la tecnología de segmentación y análisis de imágenes médicas en la mejora continua de los cuidados de salud.

Bibliografía

Amazon. (2023). *¿qué es la visión artificial?* Descargado de https://aws.amazon.com/es/what-is/computer-vision/

Ayuso Lera, A. (2022). *Segmentación automática de imágenes de resonancia magnética cerebral mediante redes neuronales convolucionales.* Universidad de Valladolid. Descargado de https://uvadoc.uva.es/handle/10324/53670

Azad, R., Arimond, R., Aghdam, E. K., Kazerouni, A., y Merhof, D. (2023). *Dae-former: Dual attention-guided efficient transformer for medical image segmentation.* Descargado de https://arxiv.org/abs/2212.13504

Bayer. (2002). *El científico que descubrió los rayos x.* Descargado de https://www.bayer.com/es/es/blog/espana-el-cientifico-que-descubrio-los-rayos-x#:~:text=El%208%20de%20noviembre%20se,en%201895%2C%20los%20rayos%20X.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., y cols. (2018, Nov). Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, *37*(11), 2514–2525. doi: 10.1109/TMI.2018.2837502

Campello, V. M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P. M., … Lekadir, K. (2021). Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge. *IEEE Transactions on Medical Imaging*, *40*(12), 3543-3554. doi: 10.1109/TMI.2021.3090082

Carballo Pacheco, J. J. (2022). *Clasificación de imágenes médicas con técnicas de deep learning.* Universidad de Extremadura. Descargado de http://hdl.handle.net/10662/16482

Carpentry, D. (2024). *Thresholding - image processing with python.* Descargado de https://datacarpentry.org/image-processing/07-thresholding

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., y Yuille, A. L. (2017). *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.*

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., y Adam, H. (2018). *Encoder-decoder with atrous separable convolution for semantic image segmentation.* Descargado de https://arxiv.org/abs/1802.02611

contributors, N. (s.f.). *nilearn.* Descargado de https://github.com/nilearn/nilearn doi: https://doi.org/10.5281/zenodo.8397156

de Ingeniería del Conocimiento, I. (2023). *La medicina del futuro en bnew.* Descargado de https://www.iic.uam.es/lasalud/la-medicina-del-futuro-en-bnew/

Demush, R. (2019). *Una breve historia de la visión artificial (y las redes neuronales convolucionales).* Descargado de https://hackernoon.com/es/una-breve-historia-de-la-vision-por-computadora-y-las-redes-neuronales-convolucionales-8fe8aacc79f3

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., y Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. En *2009 ieee conference on computer vision and pattern recognition.*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.*

Dumoulin, V., y Visin, F. (2016, mar). A guide to convolution arithmetic for deep learning. *ArXiv e-prints.*

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., … Kikinis, R. (2012, Nov). 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, *30*(9), 1323-1341.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics.*

García Ferrando, G. A. (2017). *Deep learning en segmentación de imagen médica.* Universidad Politécnica de Valencia. Descargado de http://hdl.handle.net/10251/86225

Girshick, R., Donahue, J., Darrell, T., y Malik, J. (2014). *Rich feature hierarchies for accurate object detection and semantic segmentation.*

Gu, A., y Dao, T. (2024). *Mamba: Linear-time sequence modeling with selective state spaces.* Descargado de https://arxiv.org/abs/2312.00752

Gu, A., Goel, K., y Ré, C. (2022). *Efficiently modeling long sequences with structured state spaces.* Descargado de https://arxiv.org/abs/2111.00396

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., y Xu, D. (2022). *Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images.* Descargado de https://arxiv.org/abs/2201.01266

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., … Xu, D. (2021). *Unetr: Transformers for 3d medical image segmentation.* Descargado de https://arxiv.org/abs/2103.10504

He, K., Gkioxari, G., Dollár, P., y Girshick, R. (2018). *Mask r-cnn.*

Himmel, D. P., y Peasner, D. (1974). *A large-scale optical character recognition system simulation.* Descargado de https://informs-sim.org/wsc74papers/1974_0024.pdf

Hubel, D., y Wiesel, T. (1959). *Receptive fields of single neurones in the cat's striate cortex.* Descargado de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363130/pdf/jphysiol01298-0128.pdf

IBM. (2024a). *What is computer vision?* Descargado de https://www.ibm.com/topics/computer-vision

IBM. (2024b). *What is image segmentation?* Descargado de https://www.ibm.com/topics/image-segmentation

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., … Maier-Hein, K. H. (2018). *nnu-net: Self-adapting framework for u-net-based medical image segmentation.* Descargado de https://arxiv.org/abs/1809.10486

Kiefer, J., y Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, *23*, 462-466. Descargado de https://api.semanticscholar.org/CorpusID:122078986

Kingma, D. P., y Ba, J. (2017). *Adam: A method for stochastic optimization.* Descargado de https://arxiv.org/abs/1412.6980

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., … Girshick, R. (2023). *Segment anything.*

Krizhevsky, A., Sutskever, I., y Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. En *Advances in neural information processing systems.* Descargado de https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Li, X., Ding, H., Yuan, H., Zhang, W., Pang, J., Cheng, G., … Loy, C. C. (2023). *Transformer-based visual segmentation: A survey.*

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., y Belongie, S. (2017). *Feature pyramid networks for object detection.*

Long, J., Shelhamer, E., y Darrell, T. (2015). *Fully convolutional networks for semantic segmentation.*

Loshchilov, I., y Hutter, F. (2019). *Decoupled weight decay regularization.* Descargado de https://arxiv.org/abs/1711.05101

Ma, J. (2020). *Segmentation loss odyssey.* Descargado de https://arxiv.org/abs/2005.13449

Ma, J., He, Y., Li, F., Han, L., You, C., y Wang, B. (2024). *Segment anything in medical images.* Descargado de https://arxiv.org/abs/2304.12306 doi: https://doi.org/{docm.doi}

Ma, J., Li, F., y Wang, B. (2024). *U-mamba: Enhancing long-range dependency for biomedical image segmentation.* Descargado de https://arxiv.org/abs/2401.04722

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., … Jäger, P. F. (2024, febrero). Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, *21*(2), 195–212. Descargado de http://dx.doi.org/10.1038/s41592-023-02151-z doi: 10.1038/s41592-023-02151-z

Marhuenda Tendero, L. J. (2023). *Detección de comportamientos compatibles con enfermedades a largo plazo.* Universidad de Alicante. Descargado de http://hdl.handle.net/10045/135378

Mortazi, A., Cicek, V., Keles, E., y Bagci, U. (2023). *Selecting the best optimizers for deep learning based medical image segmentation.* Descargado de https://arxiv.org/abs/2302.02289

Myronenko, A. (2018). *3d mri brain tumor segmentation using autoencoder regularization.* Descargado de https://arxiv.org/abs/1810.11654

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*. doi: 10.1109/TSMC.1979.4310076

Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., y Murphy, R. (2020). Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *arXiv preprint arXiv:2012.02951*.

Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., y Murphy, R. R. (2021). Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, *9*, 89644-89654. doi: 10.1109/ACCESS.2021.3090981

Redmon, J., Divvala, S., Girshick, R., y Farhadi, A. (2016). *You only look once: Unified, real-time object detection.*

Ronneberger, O., Fischer, P., y Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation.*

Ruder, S. (2017). *An overview of gradient descent optimization algorithms.* Descargado de https://arxiv.org/abs/1609.04747

Seidlitz, S., Sellner, J., Odenthal, J., Özdemir, B., Studier-Fischer, A., Knödler, S., … Maier-Hein, L. (2022, agosto). Robust deep learning-based semantic organ segmentation in hyperspectral images. *Medical Image Analysis*, *80*, 102488. Descargado de http://dx.doi.org/10.1016/j.media.2022.102488 doi: 10.1016/j.media.2022.102488

Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., … Cardoso, M. J. (2019a). *A large annotated medical image dataset for the development and evaluation of segmentation algorithms.* Descargado de https://arxiv.org/abs/1902.09063

Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., … Cardoso, M. J. (2019b). *A large annotated medical image dataset for the development and evaluation of segmentation algorithms.*

Tang, Y., Yang, D., Li, W., Roth, H., Landman, B., Xu, D., … Hatamizadeh, A. (2022). *Self-supervised pre-training of swin transformers for 3d medical image analysis.* Descargado de `https://arxiv.org/abs/2111.14791`

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). *Attention is all you need.*

Wikipedia. (2024a). *Esperanza de vida — wikipedia, la enciclopedia libre.* Descargado de `https://es.wikipedia.org/w/index.php?title=Esperanza_de_vida&oldid=158844978`

Wikipedia. (2024b). *Historia de la medicina — wikipedia, la enciclopedia libre.* Descargado de `https://es.wikipedia.org/w/index.php?title=Historia_de_la_medicina&oldid=158313022`

Wikipedia. (2024c). *Microscopio — wikipedia, la enciclopedia libre.* Descargado de `https://es.wikipedia.org/w/index.php?title=Microscopio&oldid=158642990`

Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., y Yuan, L. (2022). Tinyvit: Fast pretraining distillation for small vision transformers. En *European conference on computer vision (eccv).*

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., … Darrell, T. (2020). *Bdd100k: A diverse driving dataset for heterogeneous multitask learning.*

Zhang, K., y Liu, D. (2023). *Customized segment anything model for medical image segmentation.* Descargado de `https://arxiv.org/abs/2304.13785`

Zhuang, X. (2013). Challenges and methodologies of fully automatic whole heart segmentation: A review. *Journal of Healthcare Engineering*, *4*(3), 371–407.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M. P., … others (2019). Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, *58*, 101537.

Zhuang, X., Rhode, K., Razavi, R., Hawkes, D. J., y Ourselin, S. (2010). A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI. *IEEE Transactions on Medical Imaging*, *29*(9), 1612–1625.

Zhuang, X., y Shen, J. (2016). Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Medical Image Analysis*, *31*, 77–87.

Lista de Acrónimos y Abreviaturas

AA	Arteria Aorta.
ACDC	Automatic Cardiac Diagnosis Challenge.
AD	Aurícula Derecha.
AI	Aurícula Izquierda.
AP	Arteria Pulmonar.
CLIP	Contrastive Language-Image Pre-Training.
CNN	Convolutional Neural Network.
CPU	Central Processing Unit.
DCNN	Deep Convolutional Neural Networks.
DL	Deep Learning.
DSC	Dice Similarity Coefficient.
EPI	Echo-Planar Imaging.
FCCRF	Fully Connected Conditional Random Fields.
FCN	Fully Convolutional Network.
FIS	Fondos de Investigaciones Sanitarias.
FPN	Feature Pyramid Network.
GELU	Gaussian Error Linear Unit.
GPU	Graphics Processing Unit.
IA	Inteligencia Artificial.
ICR	Reconocimiento Inteligente de Caracteres.
INE	Instituto Nacional de Estadística.
IoU	Intersection over Union.
LASC	Left Atrial Segmentation Challenge.
LoRA	Low-Rank Adaptation.
MAE	Masked Autoencoders.
MedSAM	Segment Anything in Medical Images.
MICCAI	Medical Image Computing and Computer Assisted Intervention Society.
ML	Machine Learning.
MLP	Multilayer Perceptron.
MM-WHS	Multi-Modality Whole Heart Segmentation.
MONAI	Medical Open Network for Artificial Intelligence.
MSD	Medical Segmentation Decathlon.
MVI	Miocardio del Ventrículo Izquierdo.
NifTI	Neuroimaging Informatics Technology Initiative.
NLP	Procesamiento del Lenguaje Natural.

NSD	Normalized Surface Distance.
OCR	Reconocimiento Óptico de Caracteres.
R-CNN	Region-based Convolutional Neural Network.
RELU	Rectified Linear Unit.
RGB	Red-Green-Blue.
RM	Resonancia Magnética.
RNN	Recurrent Neural Network.
S4	Structured State Space Sequence Model.
SAM	Segment Anything.
SETR	Segmentation Transformer.
SLURM	Simple Linux Utility for Resource Management.
SOP	Procedimiento Operativo Estándar.
SSM	State Space Sequence Model.
TC	Tomografía Computarizada.
TFM	Trabajo de Fin de Máster.
TL	Transfer Learning.
VD	Ventrículo Derecho.
VI	Ventrículo Izquierdo.
ViT	Vision Transformer.
VPC	Visión Por Computador.
YOLO	You Only Look Once.