UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
ESCUELA DE POSTGRADO Y EDUCACIÓN CONTINUA
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

SEGMENTACIÓN DE HIPERINTENSIDADES DE MATERIA BLANCA EN BASE A MODELOS DE *VISION TRANSFORMER*

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIA DE DATOS
MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO

CRISTIAN ALEXIS MUÑOZ BUSTAMANTE

PROFESORA GUÍA:
Nancy Hitschfeld Kahler

PROFESORES CO-GUÍA:
Mauricio Cerda Villablanca
Pablo Estévez Valencia

COMISIÓN:
Cecilia Okuma Ponce

Este trabajo ha sido parcialmente financiado por:
FONDECYT 1221696, ID20I10371, NAM21I0031, EQM 210020
y con el apoyo del supercomputador
Patagón de la Universidad Austral de Chile (FONDEQUIP EQM180042).

SANTIAGO DE CHILE
2024

RESUMEN DE TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIA
DE DATOS Y
MEMORIA PARA OPTAR AL TÍTULO
DE INGENIERO CIVIL ELÉCTRICO
POR: CRISTIAN ALEXIS MUÑOZ BUSTAMANTE
FECHA: 2024
PROF. GUÍA: NANCY HITSCHFELD KAHLER

SEGMENTACIÓN DE HIPERINTENSIDADES DE MATERIA BLANCA EN BASE A MODELOS DE *VISION TRANSFORMER*

Ante la creciente edad promedio de la población mundial, enfermedades neurodegenerativas como el Alzheimer comienzan a ser un foco de preocupación en el mundo. En la búsqueda de estrategias para el diagnóstico temprano se han explorado técnicas no invasivas como la resonancia magnética para el diagnóstico temprano. La resonancia magnética permite detectar lesiones en la sustancia blanca como las Hiperintensidades T2/FLAIR en Sustancia Blanca (WMH). Dichas lesiones se han reportado en pacientes con demencia y a su vez son predictoras de problemas cognitivos, derrames y demencia.

Actualmente, el análisis clínico de las lesiones en sustancia blanca es evaluado subjetivamente y por lo cual, se han planteado nuevos métodos para obtener evaluaciones cuantitativas de las lesiones, recurriendo a algoritmos de inteligencia computacional tales como aprendizaje automático y últimamente, aprendizaje profundo.

Estudios recientes en la tarea de segmentación general han demostrado el creciente rendimiento de modelos basados en Vision Transformer (ViT), lo cual motiva la hipótesis de que arquitecturas basadas en ViT que incorporen adaptaciones de arquitecturas U-Net para el problema de segmentación de WMH, en conjunto con variables de la literatura en neurociencia, permiten mejorar los resultados actuales para la segmentación automática de hiperintensidades en resonancias magnéticas cerebrales.

Para validar esta hipótesis se utiliza la base de datos internacional de segmentación de lesiones en sustancia blanca (*White Matter Hyperintensities Segmentation Challenge*) compuesta por 170 pacientes de múltiples países y utilizando la división original de datos con 60 pacientes para entrenamiento de 110 pacientes de prueba.

Con los experimentos computacionales realizados en esta tesis se logra observar que el desempeño de los nuevos modelos basados en ViT no logran sobrepasar al actual estado del arte en la tarea de segmentación de lesiones en sustancia blanca. Sin embargo, las ideas propuestas en esta tesis logran mejorar el rendimiento de los modelos base basados en ViT en 4 de 5 métricas evaluadas (Dice Score, H95, Recall y F1 score). Por otro lado, se observa un tiempo de inferencia casi 10 veces menos entre el modelo implementado y el actual estado del arte.

Para quienes han estado presente,
incluso sin estar.

Gracias

Agradecimientos

Quiero agradecer a mi familia quienes siempre han estado presente y dispuestos a conversar cuando estuve en los momentos más difíciles. Agradecer a mis amigos Alex, Rodrigo, Daniel, Pablo y Lukas, con quienes nos acompañamos largas noches de juegos para distraernos de la vida y una pandemia pasajera.

Un agradecimiento a quienes me acompañaron en mi vida universitaria, Pao, Pau, Cami, Nico, Rojo y Joaquín. Destacando a los dos últimos payasos quienes siempre estuvimos juntos desde el primer día, "estudiando", comiendo y carreteando en la universidad para luego llegar a la noche y seguir juntos jugando hasta las 3-4 AM. Joaquín amigo, espero estés viendo esto desde algún lado, lo logré.

Gracias a mi amigo de colegio, universidad y la vida Cristián quien fue el viejo gruñón (ya no tanto) que siempre estuvo conmigo estos años. Agradecer a Sofía quien me acompañó, apoyó y levantó cuando estaba en los momentos más difíciles.

Quisiera agradecer a mis compañeros en el Scian-Lab, mi buen amigo el Doctor alias "Carlos Navarro", Martín, Jorge, Yoya, Karina, Gonzalo, Malcom, Francisca, Cota, Roberto y Steffen (no me olvido de Mauricio, pero el va por separado), gracias a ustedes es tan agradable ir al laboratorio todos los días y han inspirado mi pasión por la academia.

Un día de invierno de pandemia recuerdo escribir un correo a un profesor que seguía una linea de investigación que siempre me ha apasionado, este correo solo tenia mis ganas de trabajar, sin experiencia, sin los cursos y conocimientos necesarios. Fuera de mis expectativas recibí una respuesta y comencé a trabajar. Hoy, ya casi tres años desde ese momento sigo pensando que ha sido una de las mejores decisiones de mi vida, muchas gracias al profesor Mauricio Cerda.

Un agradecimiento a los profesores Nancy Hitschfeld y Pablo Estévez por su ayuda, disponibilidad y comentarios para sacar esta tesis adelante.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Capítulo 1

Introducción

1.1. Motivación y definición del problema

La edad promedio de la población chilena se hace cada vez mayor, al igual que la mayoría de los países desarrollados. Un importante desafío es cómo mantener dicha población en buena condición física y mental [1]. En particular, la demencia por Alzheimer va a afectar al 10% de la población de 65 años o más y a un 47% de la población de más de 85 años, sin que exista una cura, en consecuencia, la organización mundial de la salud recomienda como estrategia en estas patologías comenzar con un diagnóstico temprano y promover el bienestar físico, cognitivo, y general [2]. En este contexto, las resonancias magnéticas son técnicas de imagenología que ofrecen información que puede servir como biomarcador temprano no invasivo. Sin embargo, la calificación de esta información se realiza en general cualitativamente.

Las Imágenes de Resonancias Magnéticas (IRM o en inglés MRI) permiten detectar lesiones en la sustancia blanca del cerebro como las hiperintensidades de señal en secuencias ponderadas en T2 y FLAIR. Se ha reportado en la literatura que las WMH, con presunto origen vascular, se encuentran comúnmente en pacientes con demencia. Por otro lado, las apariciones de estas WMH en el cerebro se correlacionan con problemas cognitivos, aumento del riesgo de derrames y mayor riesgo de demencia [3].

Para evaluar las WMH, los médicos radiólogos utilizan habitualmente escalas subjetivas, la más conocidallamada Fazekas [4] cuyos valores son 0, 1, 2, 3 que significan respectivamente: Sin lesión, Baja, Moderada, Alta. Catalogando así el riesgo potencial que posee el paciente de una manera semicuantitativa dependiente del operador. Por otro lado, en lo que respecta a la segmentación de WMH, el especialista debe observar detalladamente la secuencia de imágenes 3D de MRI e identificar cada región como una posible hiperintensidad. Como alternativa a esta evaluación se ha buscado cuantificar las WMH mediante algoritmos de *Machine Learning* (ML) segmentando cada una de las lesiones. Por ejemplo, los métodos basados en procesamiento de imágenes (e.g. filtros y umbrales), aprendizaje automático (*random forest*, *clustering*) [5] y, más recientemente, aprendizaje profundo (U-Net) como el actual método del estado del arte *"White matter hyperintensities segmentation using the ensemble u-net with multi-scale highlighting foregrounds"* o por el nombre de su modelo PGS [6].

1.2. Hipótesis:

Las recientes arquitecturas basadas en ViT que incorporen adaptaciones de arquitecturas U-Net, permiten mejorar los resultados actuales para la segmentación automática de WMH, considerando las métricas de la competencia *White Matter Hyperintensities Segmentation Challenge* (WMHSC).

1.3. Objetivos

Objetivo general: Proponer un modelo de segmentación para lesiones en sustancia blanca en resonancias magnéticas utilizando inteligencia computacional, específicamente, arquitecturas basadas en ViT, adaptadas a modelos U-Net en el estado del arte de esta tarea con el fin de superar o igualar los resultados obtenidos por los ganadores de la competencia internacional WMHSC.
Objetivos específicos:

1. Desarrollar un modelo de segmentación propio basado en Vision Transformers.
2. Incorporar nuevas variables en base a la literatura existente en neurología.
3. Comparar y evaluar resultados obtenidos por algoritmos en el estado del arte con modelos propios.

1.4. Estructura del documento

Este documento de tesis se organiza en 4 capítulos continuando de esta introducción. Iniciando con los antecedentes necesarios para comprender el proceso de adquisición de una Imagen de Resonancia Magnética (IRM) con un contexto clínico de las hiperintensidades en sustancia blanca, además, se abarca el estudio de técnicas en inteligencia computacional y como han evolucionado en la tarea de segmentación automática, pasando por métodos simples como la segmentación por umbrales a métodos más actuales basados en redes convolucionales y/o ViT. En el capítulo 3 se propone un nuevo algoritmo capaz de procesar la información en múltiples capas de profundidad denominado *Resized Multi-Output Segmentation* (ReMOS) en conjunto con la descripción del dataset utilizado y el preprocesamiento realizado. En el capítulo 4 se presentan los resultados de ReMOS variando los parámetros y comparando con el estado del arte, los cuales son discutidos y analizados. Finalmente, en el capítulo 5 se concluye el documento resumiendo el trabajo realizado, resultados principales y trabajos futuros.

Capítulo 2

Antecedentes

Este capítulo detalla antecedentes de la física de adquisición, biológicos y computacionales de las imágenes por resonancias magnética y WMH.

2.1. Imágenes por Resonancia Magnética

Las IRM son una técnica no invasiva y no ionizante de obtención de imágenes médicas basándose en poderosos imanes, radiofrecuencias (RF) y computadores, la cual se basa en el principio de resonancia magnética nuclear para crear visualizaciones de los núcleos de hidrógeno y otros en el interior del cuerpo. Estas visualizaciones permiten analizar distintas partes del cuerpo, específicamente tejidos blandos, órganos y huesos.

Por como se describe por su nombre, las imágenes por resonancia magnética integran 3 elementos [7]:

- Campo Magnético (**M**). El centro del equipo está rodeado de imanes, los cuales al rotar generan un campo magnético muy potente (medido en la unidad de flujo magnético *Tesla* [T] y que puede variar desde 0.3 a 7 [T]) que permite alinear los protones en los tejidos del cuerpo.
- Resonancia (**R**). Pulsos de radiofrecuencia (8 a 130 MHz) son aplicados al cuerpo, el cual permite redireccionar los polos de los protones organizados por el campo magnético y que estos emitan en respuesta a su relajación ondas de radiofrecuencia.
- Imagen (**I**). La energía disipada por los átomos al relajarse son capturados y visualizados en un computador donde se procesan por secciones o como imágenes en 3 dimensiones.

2.1.1. Secuencias ponderadas en T1, T2 & FLAIR

Existen 3 fases de adquisición de imágenes: T1, Densidad de Protones (PD), T2, las cuales están dadas por el tiempo dentro del ciclo de emisión (ver Figura 2.1). Las imágenes ponderadas T1 se obtienen dentro de la magnetización longitudinal (paralelo al campo magnético) y donde el mayor contraste entre los distintos tipos de tejidos en el cuerpo se da en un Tiempo de Repetición (TR) y Tiempo de Eco (TE) corto (ver Figura 2.1, 0 - 1300 ms), estas imágenes poseen un tiempo de adquisición más rápido y permiten identificar (y/o segmentar) más fácilmente tejidos de alta dureza como huesos y materia grasa. En este mismo ciclo longitudinal se pueden adquirir las imágenes PD, que si bien están presentes en ambos ciclos de magnetización poseen mejor contraste al final de la magnetización longitudinal, las imágenes PD se

caracterizan por tener un mayor contraste en TR largo y TE corto (ver Figura 2.1, 1300 - 2200 ms). Este tipo de imágenes se utiliza en resonancias cerebrales debido a su capacidad de distinguir líquidos de otros tejidos, sin embargo, su uso se ha visto remplazado por las imágenes *Fluid Attenuated Inversion Recovery* (FLAIR) las cuales se obtienen al extender el tiempo de adquisición en el período de magnetización transversal (perpendicular al campo magnético) que se tiene comúnmente para una imagen ponderada T2.

Las imágenes T2 como se mencionó se obtienen en el período de magnetización transversal y el mayor contraste entre distintos tipos de tejidos se obtiene en un TR y TE largos (ver Figura 2.1, 2200+ ms), este tipo de imágenes se visualizan como un inverso del tipo T1, sin embargo, posee la capacidad de que los valores de los contrastes entre cada material sea aún más resaltado.

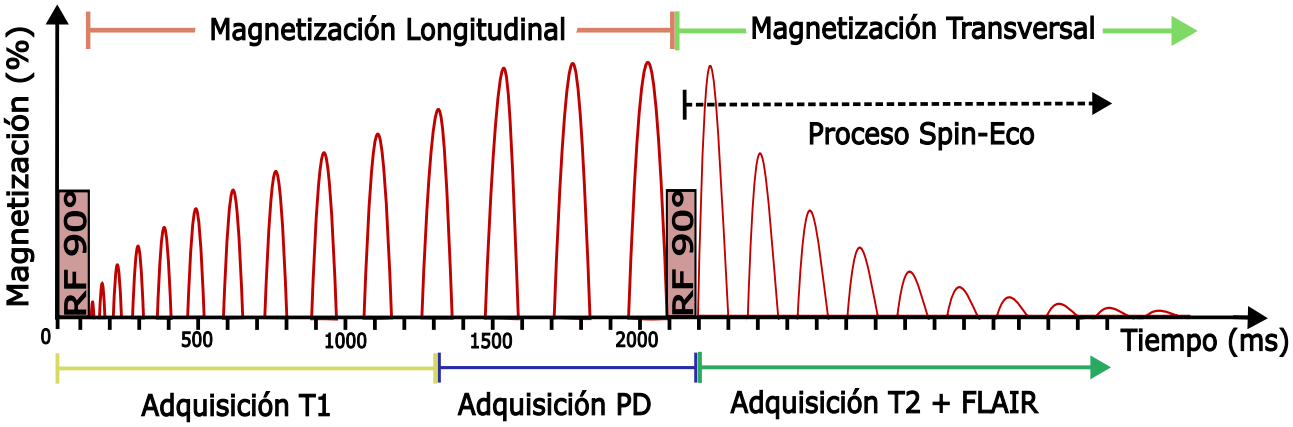


Figura 2.1: Fases longitudinal y transversal de un ciclo de emisión en una Resonancia Magnética

La técnica de adquisición spin-eco para resonancias magnéticas (ver Figura 2.2) consiste en la emisión de un pulso de RF en 90° seguido de un pulso RF 180° tras un tiempo τ, a continuación la señal (*Eco*) es capturada en un Tiempo de Eco (TE) que generalmente es dado por 2τ desde la primera emisión. Esta secuencia de emisión posee un Tiempo de Repetición (TR) que determina la frecuencia del ciclo de adquisición.

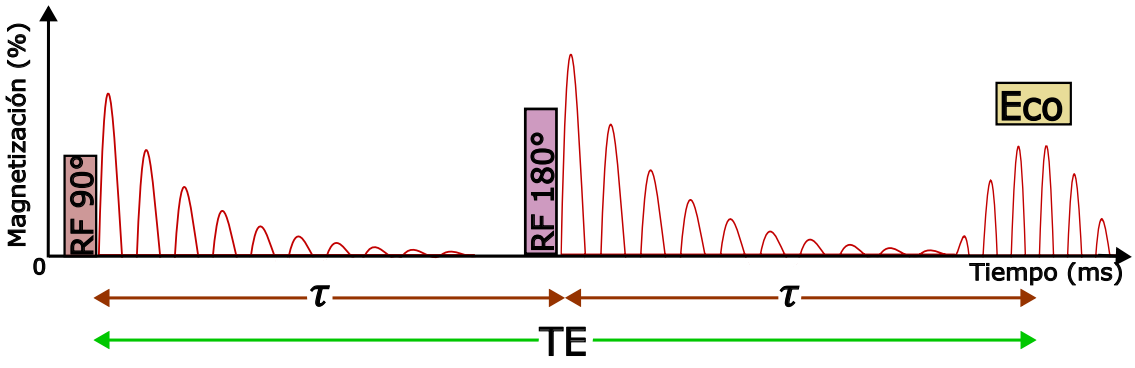


Figura 2.2: Técnica de adquisición spin-eco en Resonancias Magnéticas

2.1.2. Hiperintensidades en Sutancia Blanca en secuancias T2/FLAIR: relevancia clínica

Estudios en el campo de la Neurología han permitido correlacionar las WMH con distintas patologías. Por ejemplo, Prins & Scheltens [8] indican que una alta prevalencia de estas lesiones en pacientes con microangiopatía (enfermedad de vasos pequeños), lo cual a su vez se relaciona con el deterioro cognitivo y demencia. Tradicionalmente en la práctica clínica se evalúan con escalas de medición cualitativas, siendo la más usada FAZEKAS [4]. Por otro lado, existen estudios recientes que se han aproximado a un análisis cuantitativo de las lesiones [9] los cuales segmentan con herramientas computacionales las regiones de interes y las correspondientes lesiones permitiéndoles asociar la posición con la severidad de éstas.

2.2. Segmentación de Imágenes

La segmentación en su versión más simple, es una tarea fundamental en el campo de la visión computacional que implica asignar una etiqueta o categoría a cada píxel en una imagen digital. A diferencia de la detección de objetos, donde se identifican los objetos en una imagen, la segmentación se centra en comprender la estructura interna de estos objetos mediante la asignación de etiquetas a cada píxel que los compone. En particular, la segmentación semántica asigna a cada píxel en una imagen a una categoría específica, como "automóvil", "edificio", "persona", "árbol", etc.

2.2.1. Métricas de evaluación

Con el fin de medir la calidad de los modelos de segmentación, existen métricas que permiten evaluar qué tan bien un modelo puede asignar etiquetas semánticas a cada píxel en una imagen. Específicamente, en esta sección se mencionarán las métricas claves para la evaluación de segmentación de WMH proporcionadas por la competencia internacional *White Matter Hyperintesities Segmentation Challenge* (WMHSC) [10], la cual fue implementada con el fin de fomentar el desarrollo de tecnologías de segmentación automática para lesiones en sustancia blanca en resonancias magnéticas, proporcionando 170 pacientes con sus respectivas segmentaciones manuales (ver Sección 3.2). Las métricas utilizadas son:

- **Dice (Coeficiente de similitud Dice o Dice score)**. El coeficiente de similitud de Dice es una métrica que se utiliza para medir la similitud entre dos conjuntos, en este caso, dos segmentaciones. Se calcula como el doble de la intersección entre los dos conjuntos dividido por la suma de sus tamaños,

$$\text{Dice} = \frac{2 \times \text{Intersection}}{\text{Prediction} + \text{Ground Truth}}. \tag{2.1}$$

  Un valor de Dice 1 indica una superposición perfecta y un valor de Dice 0 indica que la predicción no se intercepta con la segmentación manual. Los rangos aceptables para modelos de segmentación son bastante variables y dependientes de la tarea, en caso de segmentación binaria fluctúan entre 0.7 a 0.8 (ver Capítulo 4).

- **Distancia Hausdorff, modificado al percentil 95 (HD95)**. La distancia de Hausdorff mide la distancia máxima entre dos conjuntos, para esto se considera todos los puntos de un conjunto respecto al siguiente conjunto de puntos más cercano. La versión modificada de la distancia de Hausdorff utiliza el percentil 95 para dar una medida más robusta, ya que se evita que valores atípicos tengan un impacto excesivo,

$$\text{Distancia Hausdorff} = \text{Modificado al percentil 95 de} \left(\max_{p_1 \in P_1} \left(\min_{p_2 \in P_2} \|p_1 - p_2\|_2 \right) \right), \quad (2.2)$$

donde P_1 y P_2 son los puntos que representan las dos segmentaciones. Un valor bajo de distancia de Hausdorff indica una mejor concordancia entre las dos segmentaciones, siendo 0 la superposición perfecta de los conjuntos con la mínima distancia.

- **Diferencia Volumétrica Promedio, en porcentaje (AVD)**. Esta métrica mide la diferencia promedio en volumen entre la segmentación predicha y la segmentación de referencia, expresada como un porcentaje del volumen de referencia.

$$\text{Average Volume Difference} = \frac{|\text{Prediction Volume} - \text{Ground Truth Volume}|}{\text{Ground Truth Volume}} \times 100\%. \quad (2.3)$$

Un valor cercano al 0% indica una buena concordancia en términos de volumen, considerando como 0 la superposición volumétrica perfecta entre los dos conjuntos.

- **Sensibilidad por Lesión, en porcentaje (recall)**. La sensibilidad, también conocida como tasa de verdaderos positivos o *recall*, mide la capacidad del modelo para detectar correctamente una lesión específica en comparación con la segmentación de referencia. Se expresa como un porcentaje y cuantifica qué proporción de la lesión de referencia se detectó correctamente.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \times 100\%. \quad (2.4)$$

- **F1-score por Lesión**. El F1-Score es una métrica que combina precisión y sensibilidad. Esta métrica proporciona una medida equilibrada del rendimiento en términos de detección y precisión de una lesión específica.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}. \quad (2.5)$$

Un valor cercano a 1 indica un buen rendimiento en la segmentación de la lesión, considerando 1 como la segmentación perfecta y 0 el conjunto vacío entre la segmentación y la predicción.

Las métricas utilizadas en WMHSC están específicamente diseñadas para la tarea de segmentación de lesiones en sustancia blanca (exceptuando Dice), esto indica que no existe referencia para otros modelos o tareas existentes. Los valores referenciales se encuentran en el Capítulo 4 y en la página oficial de la competencia.

2.2.2. Segmentación basada en umbrales globales

Los métodos de segmentación basados en umbrales globales son una categoría de técnicas que utilizan un valor de umbral único para separar los píxeles de una imagen en dos grupos: uno que cumple cierta condición con respecto al umbral (por ejemplo, píxeles más claros) y otro que no la cumple (por ejemplo, píxeles más oscuros). Comúnmente dentro de estos métodos se pueden encontrar técnicas con **umbral fijo**, es decir, seleccionar un valor de umbral y clasificar directamente cada píxel. Existen métodos optimizados de selección de umbral como **Otsu** [11], el cual es un algoritmo que busca de manera automática el umbral óptimo para la segmentación. Se calcula un umbral minimizando la varianza intra-clase y maximizando la varianza entre clases.

2.2.3. Segmentación basada en Redes Neuronales Convolucionales

Las *Convolutional Neural Networks* (CNNs) son un tipo de arquitectura de red neuronal profunda diseñada específicamente para procesar imágenes 2D o 3D. La característica clave de las CNN es la capa de convolución, que aplica filtros a pequeñas regiones de la imagen para detectar características locales, como bordes, texturas y patrones en general, que se infieren desde los datos. Los métodos basados en CNN han permitido desarrollar la visión por computadora y la segmentación de imágenes al permitir el aprendizaje automático de características en lugar de depender de características diseñadas manualmente. Ronneberger, Fischer,& Brox (2015) [12] introdujeron un nuevo tipo de arquitectura llamado U-Net el cual se compone de 2 partes claves; el *encoder* o etapa de compresión que mediante capas convolucionales y de *pooling* comprimen la imagen original, luego, esta información comprimida entra al *decoder* o capa de expansión la que a través de métodos de *upscaling* y otras convolucionales devuelven la imagen a su tamaño original. Este proceso se denominó U-Net debido a la forma de U que se genera entre el *encoder* y el *decoder*, además, es una de las arquitecturas de CNN que lideró muchos años la tarea de segmentación debido a su capacidad de manejar y regular distintas dimensiones de información. Por ejemplo, arquitecturas como APCNet [13] alcanzan un IoU promedio de 45.38% en el dataset ADE20K [14]. Sin embargo, estudios recientes han mostrado que en otras tareas de segmentación en imágenes médicas, los modelos a base de Vision Transformers (ViT [15]), particularmente DEIT-S [16] presentan un mejor desempeño [17] en IoU (*Intersection over Union*), alcanzando una mejora de 0.043 y 0.002 para las bases de datos ISIC2018 (cáncer de piel) [18][19] y CSAW-S (cáncer de mama) [20] respectivamente.

2.2.4. Arquitectura Swin-Transformers

La arquitectura Transformer se hizo famosa por su éxito en tareas de procesamiento de lenguaje natural, como la traducción automática y el procesamiento de texto [21]. Se basa en

mecanismos de atención, que permiten que la red se centre en partes específicas de la entrada, lo que resulta en un procesamiento más efectivo de secuencias de datos. Dado el éxito de los Transformers en el procesamiento de lenguaje natural, los investigadores comenzaron a explorar cómo aplicarlos a tareas de visión por computadora. Esto condujo al desarrollo de arquitecturas conocidas como *Vision Transformers* (ViTs), que son extensiones de los Transformers para el procesamiento de imágenes. Los ViTs han demostrado ventajas en tareas de segmentación de imágenes, especialmente en la captura de relaciones de largo alcance y la comprensión de contextos globales en imágenes. Esto los hace efectivos en escenarios en los que las CNN pueden tener dificultades. Algunos ejemplos de estos se ven con aplicaciones del modelo inicial ViT [15] adaptado a segmentación, continuando con arquitecturas que toman esta idea base y la desarrollan a tareas más específicas como DETR (DEtection TRansformer) [22], el cual se basa en arquitecturas Transformer para detectar y segmentar objetos de manera eficiente y finalmente modelos como Swin Transformer [23] que se basan en una arquitectura de jerarquía de bloques similar al concepto detrás de modelos U-Net con el fin de capturar con mayor detalle las características de la imagen.

Swin Transformer posee 3 características clave: i) División jerárquica de la imagen (ver Figura 2.3 (a)), la cual consiste en disminuir la dimensión de los parches que se introducen al modelo a medida que avanza en las etapas. ii) La selección móvil del conjunto de parches para realizar auto-atención conocida como Ventana móvil (ver Figura 2.3 (b)), para esto se selecciona un conjunto de parches (ventana) para realizar auto-atención, este conjunto de ventanas se desplaza diagonalmente rellenando los bordes que exceden la dimensión de la imagen con ceros (*padding*). iii) Finalmente, se utiliza el método de ventana móvil o *Shifted Windowing Multi-head Self Attention* (SW-MSA) alternando con una selección de ventanas regular (WMSA) como se puede ver en la Figura 2.3 (c) que demuestra como se conectan 2 bloques seguidos de Swin Transformers.

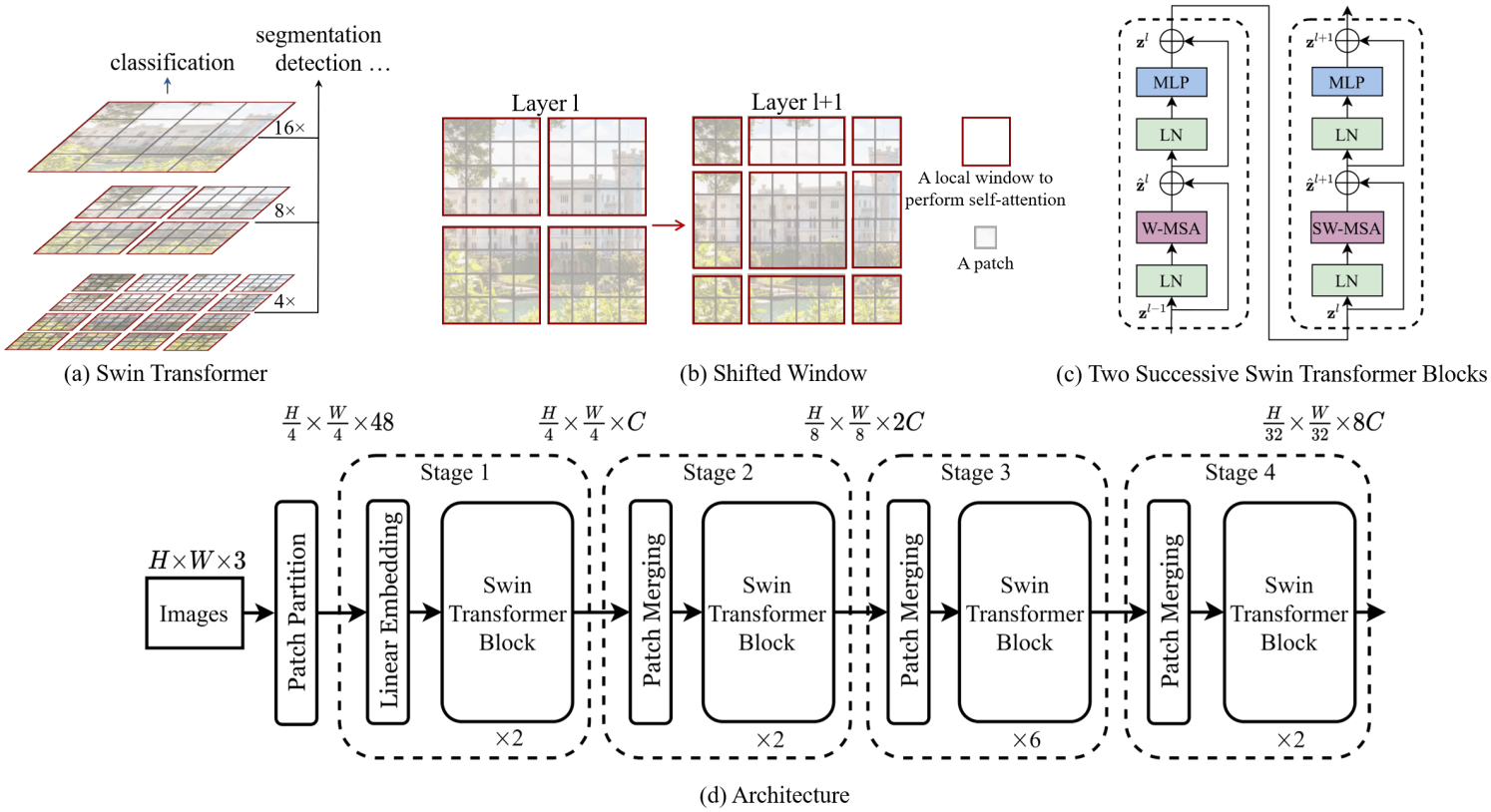


Figura 2.3: Arquitectura Swin-Transformer. a) Flujo de extracción de características en cada capa para la tarea de segmentación y de extracción única para la tarea de clasificación. b) Dos pasos consecutivos de selección de ventanas de atención utilizando la técnica *Shifted Window*. c) Dos bloques consecutivos de Swin Transformer demostrando el cambio entre W-MSA y SW-MSA. d) Estructura general de la arquitectura *Swin-Transformer*.

La arquitectura de Swin Transformer se representa en la Figura 2.3 (d), la cual se compone de los cuatro bloques principales de Swin Transformer regulados por una capa de normalización de los parches. Esta arquitectura muestra la configuración utilizada para la tarea de **clasificación,** que como se puede observar en (a) representa la unión de los resultados de cada bloque y la cual logra un 87.3% de top-1 accuracy para la base de datos Imagenet-1k [24]. Sin embargo, para la tarea de **segmentación** esta unión representa una pérdida de información sobre la representación jerárquica utilizada por Swin Transformer.

2.2.5. Segmentación con Swin-Transformers

La implementación original de Swin transformer posee la capacidad de clasificar por sí mismo, sin embargo, para tareas más complejas como es la segmentación se requiere de un cabezal (*decoder*) el cual permite procesar la información en múltiples dimensiones. En específico, utiliza Swin Transformer como encoder para crear cuatro representaciones de distinta dimensión, las cuales se entregan al decoder (UPerNet [25]) para crear la segmentación (ver Figura 2.4).

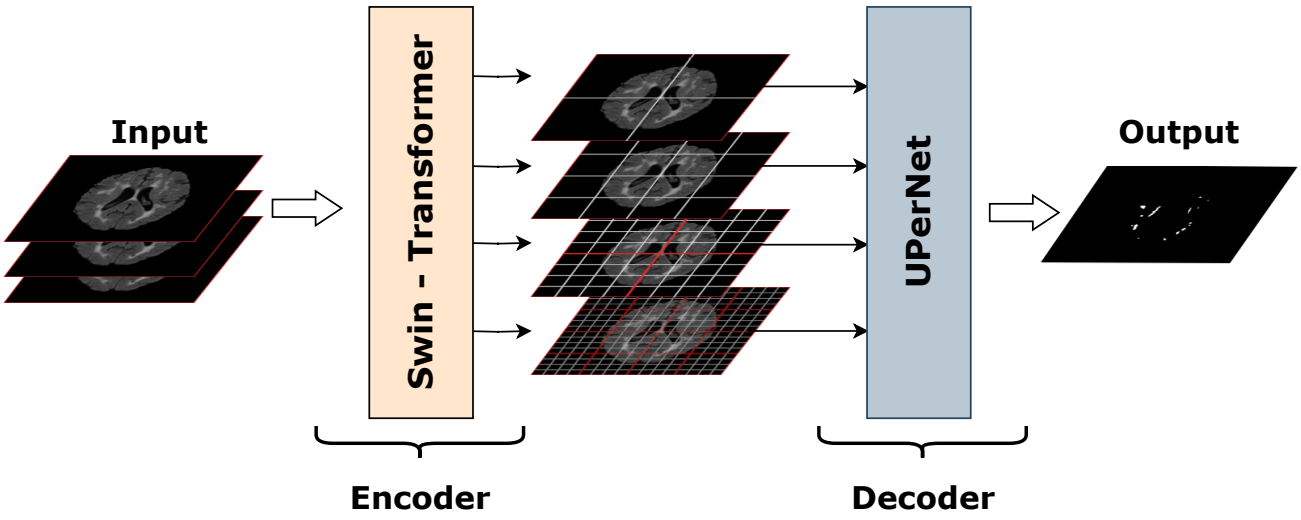


Figura 2.4: Diagrama general Swin - UPerNet. La arquitectura Swin Transformer utilizada como encoder entrega 4 imágenes en distintas resoluciones, las cuales son procesadas por el decoder (UPerNet) para lograr la segmentación.

2.2.5.1. Encoder

Como se ha mencionado anteriormente, Swin Transformer posee la capacidad de adaptar su arquitectura basándose en la tarea a realizar. En la Figura 2.5 se puede observar Swin adaptado para funcionar como encoder en la tarea de segmentación, para esto la imagen inicial debe estar conformada por 3 dimensiones, ancho, largo y canal, en donde este último representa la distribución RGB (*Red, Green & Blue*) de una imagen a color. Luego, esta imagen ingresa a la arquitectura Transformer mediante la división en parches de 1/4 de la dimensión original y además disminuyendo a medida que aumentan las etapas (conocida como jerarquía de bloques), sin embargo, como se trata de una tarea de segmentación no se utiliza el resultado del último bloque, sino que se utiliza la representación obtenida por cada una de las etapas. Resultando así en un encoder capaz de entregar 4 representaciones en distinta dimensiones de la imagen original.

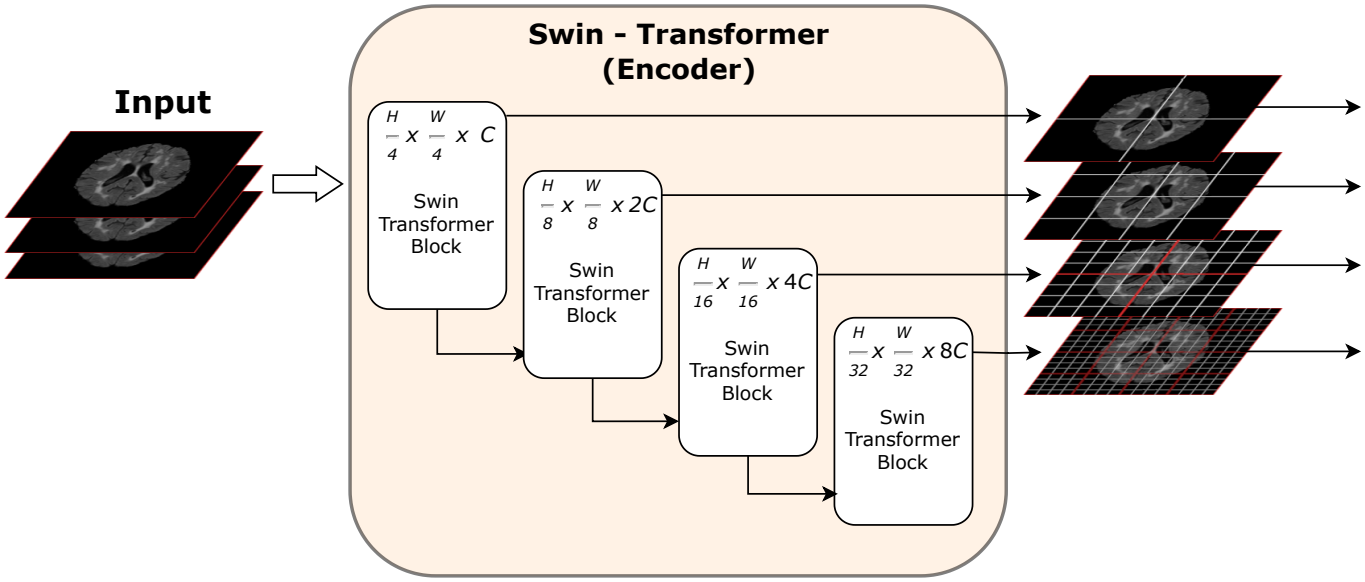


Figura 2.5: Diagrama Swin Transformer utilizado como Encoder

2.2.5.2. Decoder

Continuando con las 4 representaciones obtenidas por el encoder, el autor utiliza como decoder a UPerNet [25], la cual es una arquitectura diseñada para comprender y segmentar imágenes en distintas dimensiones (ver Figura 2.6), para esto a cada representación se le aplica una convolución de 1x1 (ponderación por un peso entrenable) exceptuando la imagen con mayor cantidad de parches que pasa por un bloque *Pyramid Pooling Module* (PPM) [26]. El bloque PPM distribuye la imagen en distintas dimensiones, distribuyendo en distintas composiciones de la misma imagen, luego, estas imágenes pasan por un bloque de convolución que permite unir los resultados para cada representación y redimensionar al tamaño original.

Los resultados obtenidos por los 3 bloques de convolución en UPerNet y el resultado del bloque PPM distribuyen su información mediante un arreglo piramidal (lo cual se ve representado en la Figura 2.6 como las flechas que conectan cada bloque de resultado interno de UPerNet). El arreglo piramidal consiste en redimensionar la imagen más pequeña y concatenar con la imagen de siguiente tamaño, permitiendo así a cada bloque poseer la información codificada del resultado anterior. Finalmente, se juntan todos los resultados obtenidos mediante la concatenación en el bloque *Fuse*, el cual a su vez realiza una convolución 1x1 para obtener la segmentación final.

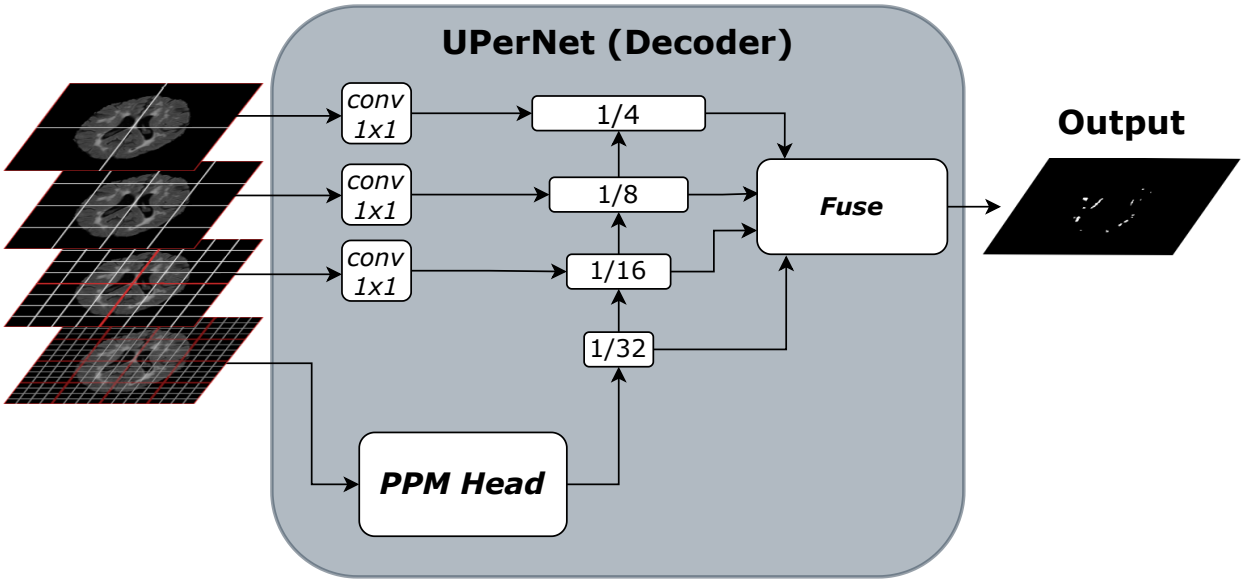


Figura 2.6: Diagrama UperNet utilizado como Decoder

La arquitectura Swin Transformer + UPerNet utilizada por los autores logró alcanzar un 62.8% de mIoU (Intersección sobre unión promedio) en la tarea de segmentación semántica utilizando el conjunto de prueba de la base de datos ADE20K [14]. Cabe destacar que la tarea de segmentación semántica es más compleja que una clasificación global de una imagen y, por lo tanto, las métricas utilizadas (como IoU) presentan un mayor desafío que el 87.3% de top-1 accuracy obtenido para la base de dato Imagenet-1K[24].

2.2.6. Segformer

Segformer [27] es una arquitectura especializada en segmentación, ya que esta arquitectura incluye los 2 bloques principales necesarios para esta tarea (encoder y decoder) y no presenta adaptaciones para otro tipo de tarea (ver Figura 2.7). Segformer une el concepto de modelos de *Vision Transformer* (ViT) como encoder para obtener una representación de múltiples dimensiones de una imagen, estas imágenes se unen mediante una capa convolucional para obtener la segmentación final (decoder). Esta arquitectura presenta un principio muy similar a Swin Transformer pero con una estructura más simple.

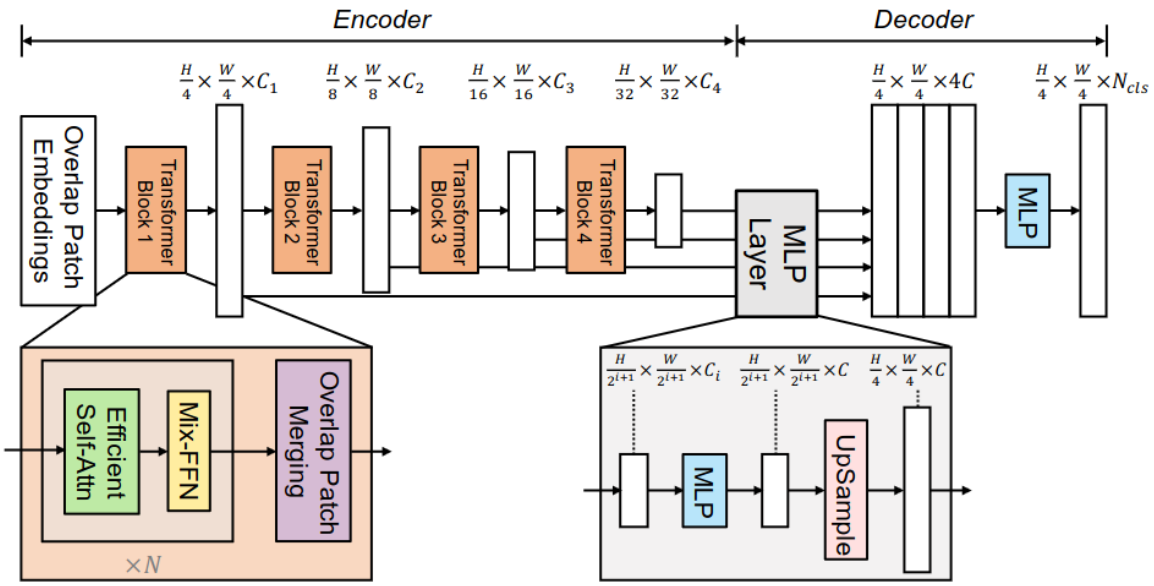


Figura 2.7: Arquitectura Segformer tomado de [27]. Compuesto por un encoder de 4 bloques secuenciales de Transformer con *Efficient self-attention* y con dimensión de parches reducida tras cada bloque. Bloque decoder con inicio de una capa MultiLayer Perceptron (MLP) seguido de bloques convolucionales.

Por como se puede ver en la Figura 2.7 el encoder se compone por 4 etapas de bloques de transformer y que a diferencia de Swin presenta métodos más sencillos de auto-atención y selección de parches. Estas 4 etapas entregan 4 representaciones de distinta dimensión al decoder. Cada representación es procesada por una etapa de Perceptrones multicapa (MLP Layer) para luego ser concatenada y procesada para obtener la segmentación final.

2.3. Estado del Arte en segmentación de lesiones en Materia Blanca

2.3.1. Segmentación basada en umbrales globales

Una multitud de métodos de segmentación de lesiones han sido propuestos durante las últimas décadas, utilizando distintos enfoques. En particular, un grupo de métodos propone tratar la segmentación de lesiones como una detección de anormalidades, por ejemplo, utilizando un registro de imágenes, como es el caso de *Matlab/SPM-LST(LGA)* [5]. Schmidt et al. [5] compara el nuevo paciente con un atlas previo de cerebros sanos (conocimiento previo) y en base de la desviación que ocurre en cada tejido entre las dos imágenes se procede a entregar un primer estimado de lesiones. Esta primera estimación de lesiones, que en general suele subestimar su volumen, se expande basándose en un crecimiento probabilístico dado por un umbral previamente seleccionado (parámetro del usuario).

Esta herramienta es ampliamente reconocida y utilizada por radiólogos y neurocientificos, ya que, es de fácil acceso y posee un Dice score promedio de 0.7531 pero con un resultado que varía dependiendo de los parámetros ajustados por el usuario, llegando a un mínimo de 0.4658 hasta un máximo de 0.9253 en la misma métrica en una base de datos propia de 70

pacientes.

2.3.2. Segmentación basada en Redes Convolucionales

Al ser una de las arquitecturas más comunes en la segmentación de imágenes médicas, los métodos basados en redes convolucionales (específicamente U-Net) se pudo observar una presencia mayoritaria en la competencia internacional *White Matter Hypterintesities Segmentation Challenge* [10], la cual existe con el fin de generar algoritmos que puedan detectar y segmentar precisamente las lesiones en sustancia blanca. En particular, el ganador PGS [6] se basa en un modelo U-Net 2D adaptado a nuevas características de las MRI, por ejemplo, la inclusión de nuevas imágenes (T1) concatenadas a las comúnmente utilizadas para esta tarea (FLAIR) (ver Input en Figura 2.8). PGS también incluye supervisión profunda (Deep Supervision) que consiste en permitir a PGS aprender en distintas etapas de la arquitectura previas al output, esto se puede identificar en las distintas salidas en la profundidad de los niveles en la Figura 2.8. Finalmente, para cada una de estas etapas de supervisión profunda, la imagen de salida es un resultado escalado de la imagen original, es decir, que por cada etapa de aprendizaje que se profundiza se le presenta un resultado aumentado que le entrega una mayor relevancia a las lesiones pequeñas.

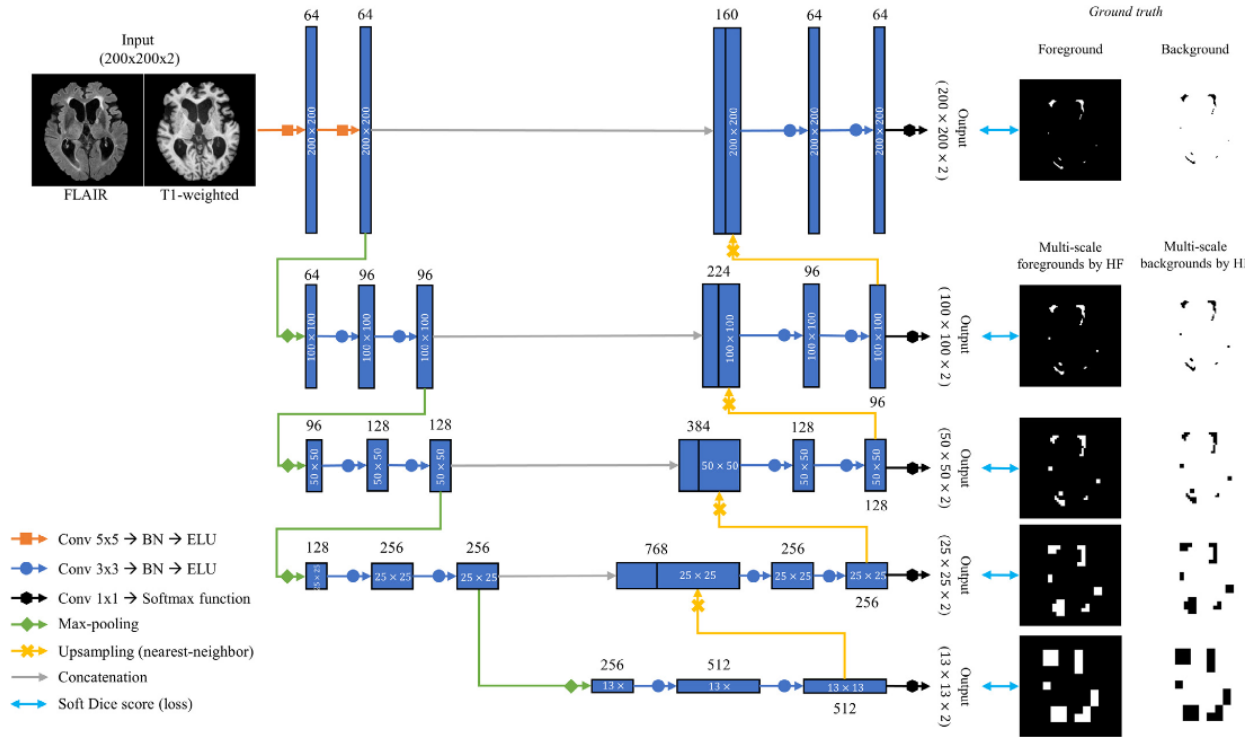


Figura 2.8: Modelo PGS tomado de [6]. UNet2D diseñada para la tarea de segmentación de lesiones en sustancia blanca, utilizando técnicas de *Deep Supervision, Highlight Foreground y Multi input.*

Con un Dice Score 0.8107, HD95 5.68mm, AVD 16.74%, recall 0.75 y F1-score 0.75, PGS se posiciona como el actual (desde 2022) primer lugar en la competencia WMHSC.

Capítulo 3

Algoritmo Propuesto y Metodología de evaluación

Como se explicó en el capítulo 2, se han utilizado técnicas de supervisión profunda para mejorar la segmentación de WMH. Por otra parte, los métodos como Swin-Transformers han aparecido como alternativas de gran rendimiento en tareas de segmentación generales. Estos nuevos modelos basados en *Transformer* se pueden descomponer en dos partes, el codificador o *BackBone* el cual se encarga de procesar y extraer la mayor cantidad de información posible de las imágenes, continuando con el decodificador o *Decode Head* encargado de representar y/o clasificar la información del Backbone para la tarea específica.

3.1. Método propuesto

Como unión de todas las ideas se propone *REsized Multi Output Segmentation* (REMOS), esta arquitectura de aprendizaje profundo aplicado a los diferentes niveles de representación del input que se recibe, es decir, cada uno de los canales (con distinta dimensión) posee dos opciones de aprendizaje: el inicial que proviene del bloque de fusión original de Segformer y el segundo que proviene de una proyección de la segmentación original basada en la técnica de *Highlight Foreground* de PGS (2.3.2). Este nuevo modelo posee 3 características claves.

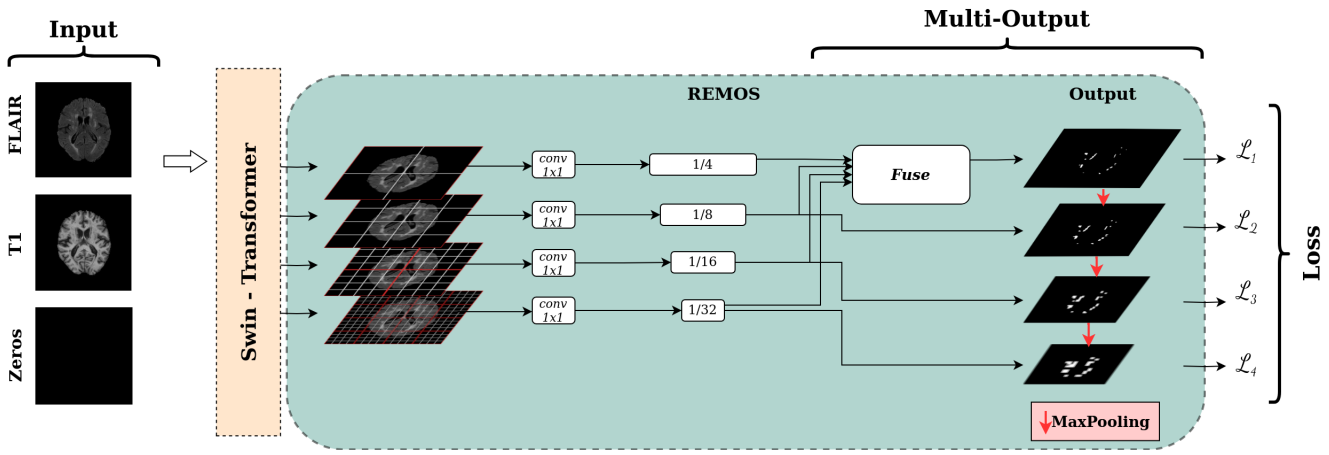


Figura 3.1: REMOS Head. Arquitectura compuesta por el método Swin Transformer como encoder y REMOS como decoder, esta nueva arquitectura como conjunto utiliza las técnicas de *Multi input, Multi output* redimensionado (previamente introducido como *Highlight Foreground*), y *Deep Supervision*

3.1.1. Input

Las imágenes de entrada se disponen para entrenar con 3 canales, donde cada canal se utiliza para alojar un tipo de imagen, ya sea *Fluid Attenuated Inversion Recovery* (FLAIR), T1, ceros u otra composición. Específicamente, se realizaron pruebas con: (1) FLAIR, T1, T1/T2, (2) FLAIR, T1, ceros y finalmente (3) FLAIR en los 3 canales.

3.1.2. Multi Output

Con el fin de que la red pueda aprender en profundidad sobre las imágenes, se le permite tener hasta 3 fases de aprendizaje extra. Cada una de estas etapas se componen de una imagen escalada de la segmentación original aplicando la técnica de *maxpooling* que consiste en seleccionar el mayor valor en una sección de la imagen (en este caso 2x2 píxeles), lo cual al ser una segmentación manual con valores binarios permite expresar con mayor importancia las lesiones más pequeñas (como se aplica en la técnica de PGS).

Sin embargo, como se puede observar en la Figura 3.1 (Multi-output) no existe una evolución progresiva del cálculo de la segmentación (como se observa en la sección decoder de las arquitecturas U-Net y en específico PGS) en donde el resultado de un bloque se transfiere directamente al siguiente, si no que los resultados de cada imagen son obtenidos en paralelo, por lo tanto, se aplica el aprendizaje profundo a cada una de estos resultados y que luego es concatenado en el bloque *fuse*.

3.1.3. Función de Pérdida

Para poder compatibilizar la existencia de múltiples outputs o aprendizajes se propone una función de pérdida adaptable por pesos del usuario. Esta función de pérdida se compone como una suma ponderada de cada una de las pérdidas generadas por las segmentaciones, es decir,

$$\frac{1}{n}\sum_{i=0}^{n}\omega_i\mathscr{L}_i,$$

donde n es el número de salidas (hiperparámetro), $\mathscr{L}$ la función de pérdida de la salida y ω el peso asociado dado por el usuario. En específico, la función de pérdida está dada por

$$\mathscr{L} = Log(Cosh(1 - DiceCoef)), \tag{3.1}$$

la cual es una función diseñada para tareas de segmentación binaria [28] en el ámbito clínico, ya que permite tener una mejor representación en caso de clases altamente desbalanceadas como es la tarea de segmentación de cráneos (*NBFS Skull-stripping dataset* [29]), donde Eq. 3.1 supera a otros 8 tipos de funciones de pérdida con respecto al coeficiente Dice.

3.1.4. Modelo Base

Para definir el proceso de evaluación y selección de la arquitectura es necesario definir una base de comparación, para esto se presentan las evaluaciones iniciales de los modelos Swin Segformer, Swin Upernet y los resultados de PGS obtenidos al ser replicado (ver Tabla 4.1).

Tanto el encoder (Swin transformer) como el decoder (Segformer) se pueden variar. En este trabajo, debido al tiempo, se evaluará cambios solo en el decoder. Estudios iniciales de los modelos base definen como se comparan las arquitecturas basadas en transformer. Para esta tarea en específico se prueba Swin Transformer con 2 cabezales distintos: Segformer y UPerNet (Original).

3.2. Dataset

Los datos utilizados provienen de la competencia internacional de segmentación de hiperintensidades en sustancia blanca [10] (*WMHSC*) los cuales consisten en 170 pacientes de múltiples países, centros, resonadores y dimensiones (Ver Tabla 3.1), cada lesión en las resonancias fue analizada y segmentada por 2 observadores siguiendo el procedimiento *STandards for ReportIng Vascular changes on nEuroimaging* (STRIVE), en donde el primer observador (O1) realizó una segmentación manual de los bordes de cada hiperintensidad los cuales fueron exhaustivamente revisados por el observador 2 (O2), por otro lado, las segmentaciones para el conjunto de entrenamiento fueron revisadas una segunda vez por otros 2 observadores (O3 y O4) siguiendo el mismo protocolo.

Tabla 3.1: Datos WMHSC [10]

Instituto	Escáner	Cantidad	Dimensión
UMC Ultrecht	3T Philips Achieva	50	230x230x48
NUHS Singapore	3T Siemens TrioTim	50	232x256x48
VU Amsterdam	3T GE Signa HDxt	50	132x256x103
	1.5T GE Signa HDxt	10	128x256x103
	3T Philips Ungenuity	10	321x240x83

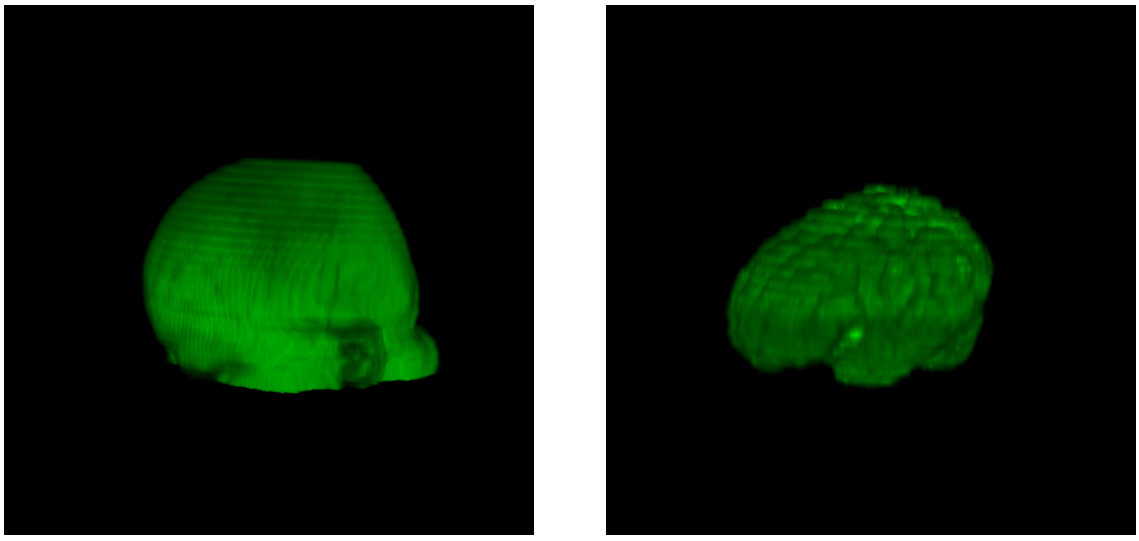
La selección oficial del conjunto de entrenamiento consiste en 20 pacientes de UMC Ultrecht, 20 pacientes de NUHS Singapore y 20 pacientes de VU Amsterdam (3T GE Signa HDxt), siendo los 110 restantes utilizados como conjunto de prueba.

Cada carpeta de paciente posee 3 tipos de secuencias, T1, FLAIR y 3DT1, las cuales a su vez fueron pre-procesados por los mismos creadores de la base de datos para redistribuir la

intensidad de la señal basándose en las características del resonador, cabe destacar que esta normalización de la señal es una operación intrínseca de cada resonador y es un procedimiento que se realiza comúnmente por clínicos al momento de visualizar las resonancias magnéticas.

3.3. Preprocesamiento

Una vez obtenidos los datos de la competencia con la redistribución intrínseca de cada resonador para regular el campo magnético detectado se procede a programar un proceso de normalización de las imagenes, para ello y con el fin de obtener información más precisa y detallada de las zonas de interés (sustancia blanca) se extrae la corteza cerebral de las resonancias magnéticas utilizando la herramienta ROBEX [30]. Una vez extraída la corteza cerebral se utiliza una normalización gaussiana entre el 2% y el 98% de los valores y se extrae imágenes sin información en el eje Z.



(a) Imagen original

(b) Imagen Post ROBEX

Figura 3.2: Extracción de corteza cerebral con ROBEX

Al ser una base de datos limitada en la cantidad de pacientes, se utiliza técnicas de aumento de datos que no generen distorsión en los cerebros, esto implica realizar cambios morfológicos a la imagen total y no corromper a nivel de píxel. Los métodos utilizados fueron desplazamiento aleatorio de 15% en el eje X e Y, rotación de 15°, reflejar aleatorio en eje X e Y y finalmente recorte (Shear) aleatorio 10%. Todos estos métodos fueron aplicados con una probabilidad de 50% y generados al momento de ingresar al entrenamiento. Del total de imágenes de entrenamiento (2610 post procesamiento) se generan un total de 37390 nuevas imágenes. En el proceso de selección de los métodos de aumento de datos se evaluó diferentes composiciones, iniciando con aumentos simples como rotación y reflejar en ejes X e Y, continuando con la evaluación para el conjunto de rotación, reflejar, shear y translación. Cada método fue variado entre los rangos 10 a 25 grados (rotación), solo reflejo horizontal y 5 a 20 de magnitud (shear), seleccionando finalmente la configuración con el mejor resultado.

3.4. Librerías y repositorio

Este trabajo se realizó utilizando el lenguaje de programación Python 3.10. Los modelos base y protocolo de entrenamiento y evaluación se basaron en el repositorio OpenMMLab/MMSegmentation [31], el cual consiste en un conjunto de modelos en el estado del arte para segmentación y un proceso de entrenamiento optimizado para el uso de CPU y/o GPU.

Los modelos evaluados fueron entrenados en 80000 ciclos con un tamaño de batch de 36 ejemplos, utilizando optimizador AdamW, $Lr = 0.0025$, $weightdecay = 0.01$, $betas = (0.9, 0.999)$. $Lineal\ Warmup = 2000$ ciclos y continuando $CosineLRDecay$. Los parámetros fueron optimizados en los rangos 0.01 a 0.0003 (Lr), 0.01 a 0.002 (weight decay), 20000 a 320000 (ciclos), 2 a 36 (batch size), *Cosine* y *Poly* (LrDecay) y 2000 a 8000 (Lineal Warmup) y se seleccionaron los parámetros con mejores resultados.

```
_base_ = [
    '../_base_/models/uper_remos_wmh.py', '../_base_/datasets/whmdataset.py',
    '../_base_/default_runtime.py', '../_base_/schedules/schedule_80k.py'
]

model = dict(
    decode_head=dict(
        remos_weight=[0.1, 0.1, 0.1, 0.7])
    )
# AdamW optimizer, no weight decay for position embedding & layer norm
# in backbone
optim_wrapper = dict(
    _delete_=True,
    type='OptimWrapper',
    optimizer=dict(
        type='AdamW', lr=0.0001, weight_decay=0.01, betas=(0.9, 0.999)),
    paramwise_cfg=dict(
        custom_keys={
            'backbone': dict(#lr_mult=0.1,
                             decay_mult=1.0
                             ),
            'absolute_pos_embed': dict(decay_mult=1.0),
            'relative_position_bias_table': dict(decay_mult=0.),
            'norm': dict(decay_mult=0.)
        }))

param_scheduler = [
    dict(
        type='LinearLR', start_factor=1e-3, by_epoch=False, begin=0, end=6000),
    dict(
        type='PolyLR',
        eta_min=0.0,
        power=1.0,
        begin=4000,
        end=80000,
        by_epoch=False,
    )
]
```

```
39
40 # By default, models are trained on 8 GPUs with 2 images per GPU
41 train_dataloader = dict(batch_size=36)
42 val_dataloader = dict(batch_size=1)
43 test_dataloader = val_dataloader
```

Código 3.1: Config-Uper-Remos.py

En el Código 3.1 se puede observar un ejemplo de definición de parámetros, carga de modelo ya definido y ejecución de los procesos de entrenamiento y test. En específico, en la lista **_base_** se selecciona el modelo creado (*uper_remos_wmh.py*), el preprocesamiento y aumento de datos en *whmdataset.py*, método de evaluación y parámetros de ejecución *default_runtime.py* y parámetros de entrenamiento *schedule_80k.py*. Los parámetros seleccionados son luego sobreescritos para adaptarse a la prueba realizada (en este caso variando los pesos del decoder) y cambiando detalles de entrenamiento (función de optimización y regulación de tasa de aprendizaje). El preprocesamiento, métodos de aumento de datos, modelos y otras optimizaciones a los procesos se pueden encontrar almacenados en el repositorio Github https://github.com/CCMunozB/mmsegmentation_whm .

Capítulo 4

Resultados y Discusión

En este capítulo se evaluará el desempeño del modelo propuesto ReMOS, se plantea un modelo base, el cual se optimizará variando manualmente sus parámetros de entrada, función de pérdida y distribución de pesos, con el fin de comparar su rendimiento con el actual estado del arte. A continuación, se realiza un estudio de los resultados obtenidos con la arquitectura optimizada para comprender e identificar patrones, finalmente, se muestran ejemplos visuales de los patrones previamente detectados e identificando posibles razones de su comportamiento.

En el siguiente segmento de discusión se planea un análisis más detallado de los resultados obtenidos y se realiza una comparación de ReMOS con los métodos actuales en la tarea de segmentación de lesiones en sustancia blanca.

4.1. Resultados

Los resultados obtenidos presentan una variación de parámetros conforme a las distintas implementaciones que se plantean en el modelo propuesto, las métricas fueron calculadas utilizando el código de evaluación oficial de la competencia WMHSC y el conjunto de pacientes de prueba oficial.

4.1.1. Modelo Base

Los Modelos Base se conforman por las arquitecturas basadas en *Vision Transformer* que fueron estudiadas en este trabajo. Inicialmente, se exploró modelos especializados en segmentación, para esto se evaluó el desempeño de Swin Transformer en conjunto con UPerNet y que al ser comparado con el estado del arte (PGS) se observa un mejor desempeño en 2 de las 5 métricas propuestas (ver Tabla 4.1).

Tabla 4.1: Comparación modelos base y PGS

Modelo	DSC	H95 (mm)	AVD (%)	Recall	F1
Swin Segformer	0.728	6.36	**17.21**	0.62	0.68
Swin Upernet	0.750	**6.13**	17.79	0.67	0.71
PGS (baseline)	**0.794**	8.33	20.62	**0.80**	**0.78**

Continuando con la exploración se determina que la capacidad de Swin + Upernet es buena, sin embargo, no lo suficiente para satisfacer la hipótesis, en este ámbito se explora la

variación del decoder, incorporando como nueva arquitectura para segmentar a Segformer, la cual como se mencionó previamente en la Sección 2.2.6, es una arquitectura simplificada de la implementada por Swin + Upernet y lo cual ve reflejado con los resultados obtenidos de la Tabla 4.1.

4.1.2. ReMOS

La aplicación de ReMOS a los modelos previamente mencionados se realiza con los parámetros base de 3 *pooling* (3 ReMOS), es decir, la salida original más 3 salidas reescaladas y el peso de cada una de estas es distribuido de equitativamente (0.25 para cada función de pérdida LogCosh DiceLoss). Los resultados se pueden observar en la tabla 4.2.

Tabla 4.2: Comparación ReMOS base y PGS

Modelo	DSC	H95 (mm)	AVD (%)	Recall	F1
Seg Remos	0.712	7.28	**16.79**	0.57	0.66
UPer Remos	0.746	**5.72**	17.50	0.66	0.72
PGS (baseline)	**0.794**	8.33	20.62	**0.80**	**0.78**

Por el desempeño obtenido para UPerNet y Segformer al aplicar ReMOS se selecciona UPer ReMOS como el mejor candidato para iterar en la fase de optimización.

4.1.3. ReMOS optimizado

El conjunto de parámetros que se pueden ajustar en los modelos de inteligencia computacional con muy amplios, sin embargo, para este método en particular se centran en el tipo de entrada, función de pérdida y los pesos asociados a cada capa de ReMOS.

Tipo de entrada

Como se menciona en 3.1.1 se obtiene los resultados para el conjunto base de ReMOS UPerNet (FLAIR, T1, 0), incluyendo variables de neurociencia (FLAIR, T1, T1/T2) y utilizando solo FLAIR (FLAIR, FLAIR, FLAIR). Esta comparación se puede observar en la tabla 4.3.

Tabla 4.3: Comparación Remos variación Input y PGS

Modelo	DSC	H95 (mm)	AVD (%)	Recall	F1
UPer Remos	0.746	5.72	17.50	0.66	0.72
UPer Remos (t1/t2)	0.745	6.24	19.02	0.65	0.71
UPer Remos (flair)	0.748	**5.61**	**17.23**	0.65	0.70
PGS (baseline)	**0.794**	8.33	20.62	**0.80**	**0.78**

Función de pérdida

En la función de pérdida base utilizada se puede ver un desempeño similar entre Diceloss y Logcosh diceloss para el modelo UPer-ReMOS (ver tabla 4.4), sin embargo, Logcosh diceloss obtiene mejores resultados en 3 de 5 métricas.

Tabla 4.4: Comparación Función de pérdida.

Func. Pérdida	DSC	H95 (mm)	AVD (%)	Recall	F1
Uper Remos (DiceLoss)	0.753	5.96	18.07	0.66	0.71
UPer Remos (LogCosh DiceLoss)	0.746	**5.72**	**17.50**	0.66	0.72
PGS (baseline)	**0.794**	8.33	20.62	**0.80**	**0.78**

Pesos en ReMOS

Los pesos utilizados para cada etapa de ReMOS son un parámetro ajustable, estos parámetros siguen la estructura de [ReMOS 1], [ReMOS 2], [ReMOS 3], [Output] y son distribuidos dando prioridad a la salida original de la arquitectura, por como se puede observar en la tabla 4.5 el peso que le da mayor relevancia al output final (0.250, 0.125, 0.0625, 0.5625) posee los mejores resultados en H95 y Recall.

Tabla 4.5: Comparación en pesos de ReMOS

UPer Remos (Pesos)	DSC	H95 (mm)	AVD (%)	Recall	F1
0.250, 0.250, 0.250, 0.250	0.753	5.96	18.07	0.66	0.71
0.125, 0.125, 0.125, 0.625	0.755	6.13	19.45	0.67	0.71
0.100, 0.100, 0.100, 0.700	0.757	5.87	**17.99**	0.67	0.71
0.250, 0.125, 0.0625, 0.5625	0.756	**5.70**	18.96	0.68	0.71
PGS (baseline)	**0.794**	8.33	20.62	**0.80**	**0.78**

UPer ReMOS optimizado

Finalmente, se entrena un modelo utilizando UPer-ReMOS, función de pérdida LogCoshDiceLoss, pesos [0.250, 0.125, 0.0625, 0.5625] y el tipo de input FLAIR.

Tabla 4.6: Comparación con diferentes inputs para selección final

Modelo	DSC	H95 (mm)	AVD (%)	Recall	F1
UPer remos optimizado	0.764	**5.56**	**18.63**	0.68	0.73
PGS (baseline)	**0.794**	8.33	20.62	**0.80**	**0.78**

4.2. Validación Cruzada

Con el fin de comprender la influencia de la distribución en los datos para el conjunto de prueba se realiza una prueba de validación cruzada, la cual consiste en variar los pacientes que se utilizan para entrenamiento y prueba, manteniendo la misma distribución que los conjuntos originales (110 pacientes de prueba y 60 de entrenamiento) para un total de 10 pruebas. Cabe señalar que este estudio no es comparable a modelos en la competencia de WMHSC, ya que, en esta competencia, se ha predefinido los pacientes para entrenamiento y prueba, y, por lo tanto, al variar la distribución de pacientes no es clara la comparación entre cada modelo.

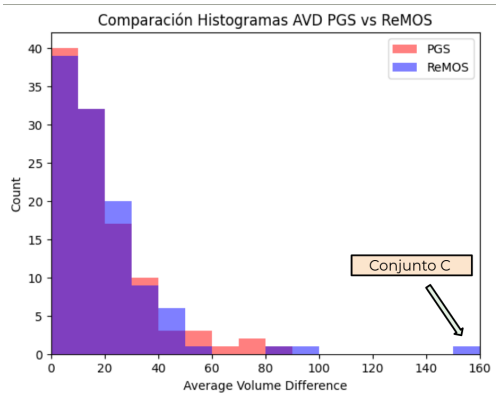
Tabla 4.7: Prueba de Validación cruzada con un total de 10 conjuntos.

	Dice	**H95**	**AVD**	**Recall**	**F1**
0	0.766	5.009	15.266	0.676	0.725
1	0.763	5.667	13.119	0.639	0.715
2	0.777	5.192	12.083	0.657	0.729
3	0.779	4.732	12.983	0.661	0.725
4	0.766	5.791	13.150	0.645	0.720
5	0.777	5.339	13.436	0.655	0.721
6	0.771	5.591	14.425	0.669	0.726
7	0.770	5.005	13.581	0.667	0.724
8	0.765	5.575	13.910	0.662	0.732
9	0.770	4.988	11.808	0.674	0.727
Mean	0.770	5.289	13.376	0.661	0.724
Std	0.006	0.356	1.022	0.012	0.005

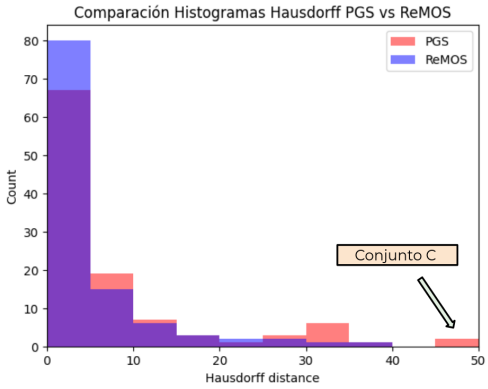
En la tabla 4.7 se presenta los resultados de las 10 pruebas de validación cruzada. Al comparar con el conjunto de prueba oficial de la competencia de WMHSC (Tabla 4.6) se observa que las diferencias observadas en la partición fija en las 5 métricas no parecen explicarse por la partición de los datos. Por ejemplo, en la Tabla 4.6 Uper remos optimizado tiene un DSC de -0.03 respecto a línea base (PGS). En la Tabla 4.6, vemos que el mejor resultado de DSC llega a 0.779, es decir todavía inferior a la línea base.

4.3. Comparación ReMOS vs PGS

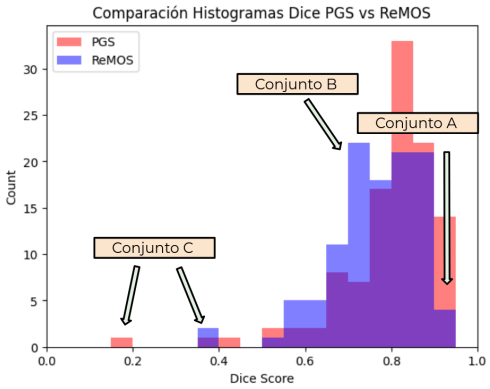
Al comparar los resultados de ReMOS optimizado y PGS se puede observar un mejor rendimiento en H95 y AVD. Esto se ve representado en la figura 4.1 (a) y (b) con un histograma concentrado en 0 y menos esparcido para AVD y H95.



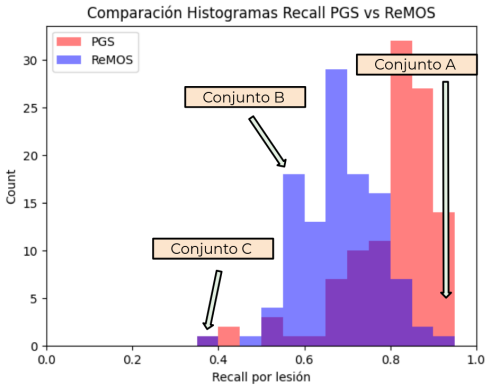
(a) Histograma de AVD



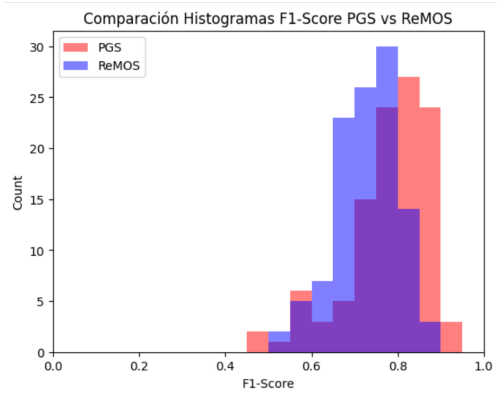
(b) Histograma de H95



(c) Histograma de Dice score



(d) Histograma de Recall



(e) Histograma de F1 score

Figura 4.1: Comparación por histograma en métricas de AVD (a), H95 (b), DSC (c), Recall (d) y F1 score (e). En estos histogramas se identificaron 3 conjuntos de comportamiento de los resultados. Conjunto A, compuesto por resultados de ReMOS que superan a PGS. Conjunto B, compuesto por resultados con sobresegmentación. Conjunto C, compuesto por errores y outliers.

En las métricas de DSC, Recall y F1 score se pueden observar una distribución claramente

desplazada con respecto a PGS, en particular se observa 3 grupos de comportamiento, los cuales se denominan como conjunto a, conjunto b y conjunto c.

4.4. Visualización de resultados

4.4.1. Conjunto A

El primer conjunto se desprende de los mejores resultados obtenidos por ReMOS. Como se puede observar en la Figura 4.2 tanto el paciente 26 como el paciente 119 poseen una gran área afectada por lesiones en sustancia blanca, tanto en la zona periventricular como en sectores alejados del centro. Sin embargo, para el paciente 119 se observa como PGS no logra procesar la parte delantera del cerebro perjudicando sus resultados.

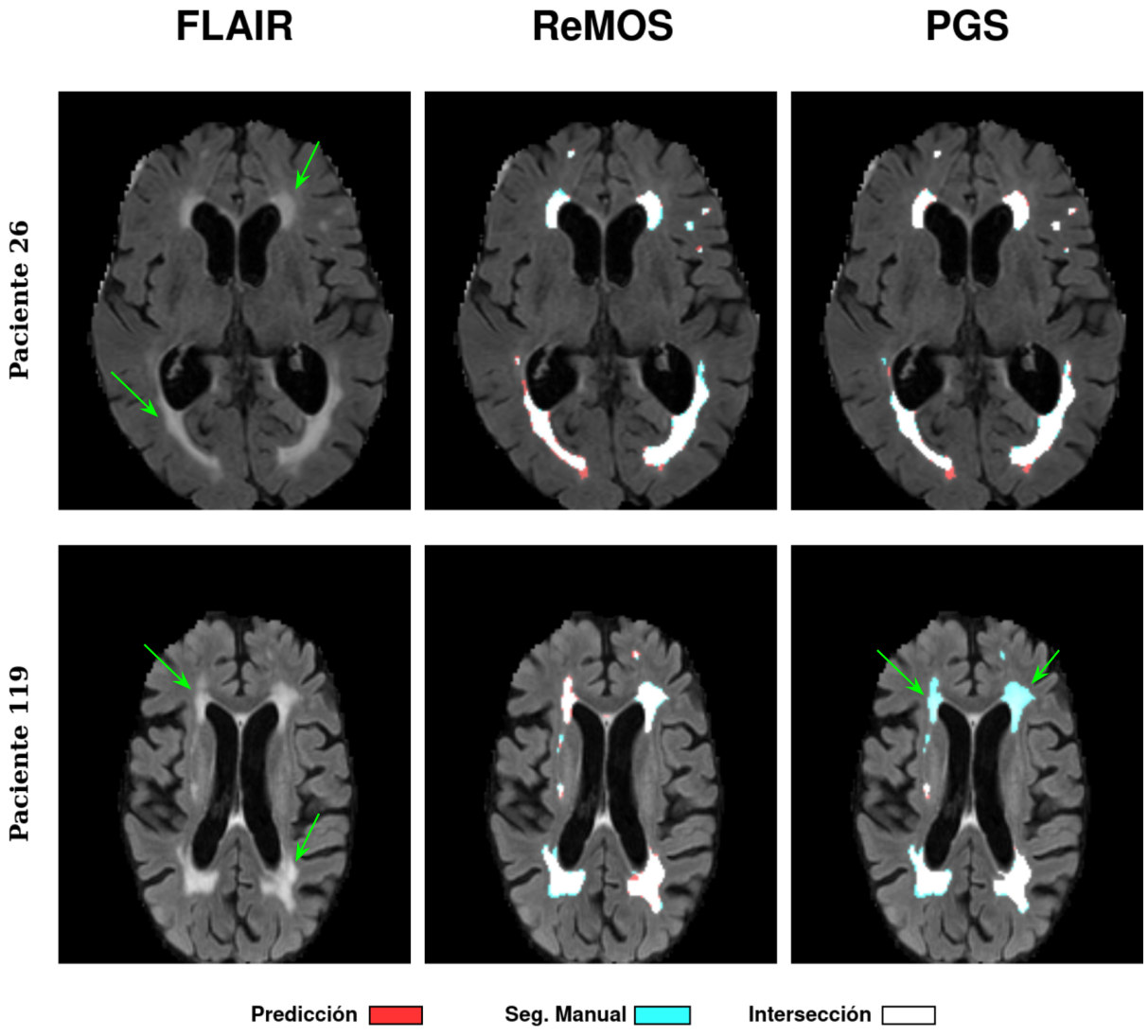


Figura 4.2: Ejemplos con buena segmentación en ReMOS. Se observa en la columna FLAIR lesiones grandes y totalmente segmentadas por el personal clínico (flecha verde FLAIR), por otro lado, se observa errores en la segmentación por parte de PGS al no detectar el sector superior de la imagen (flecha verde PGS).

4.4.2. Conjunto B

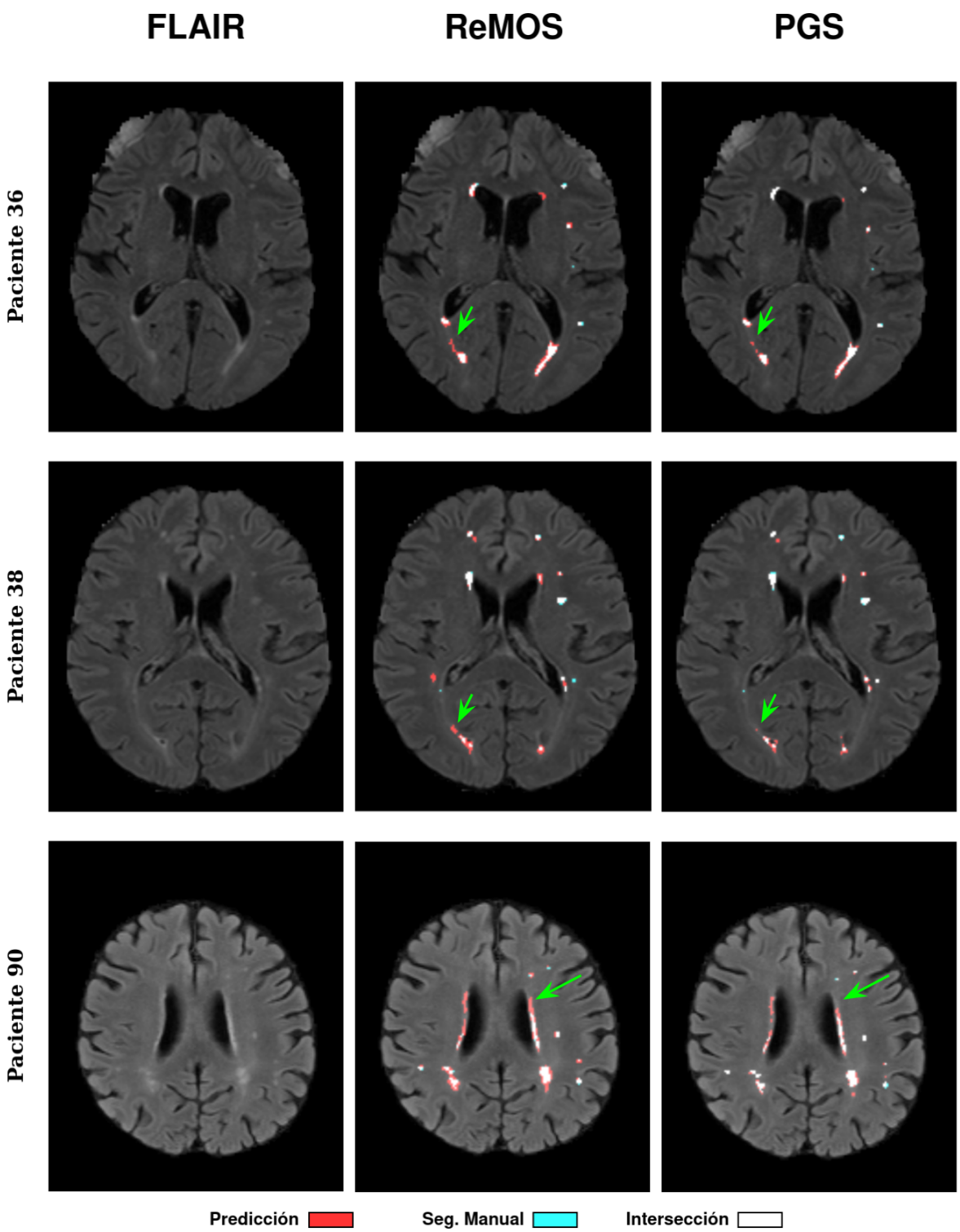


Figura 4.3: Ejemplos sobre segmentación en ReMOS y PGS. Al comparar la segmentación realizada por ReMOS y PGS se observa una mayor área segmentada por ReMOS para una misma lesión (flecha verde)

Ejemplos dentro del conjunto B se pueden observar en la Figura 4.3 donde los pacientes 36, 38 y 90 presentan lesiones en la zona periventricular (bordes del centro negro interior) la cuales poseen una tendencia a ser sobreestimadas en ambos modelos. Es decir, la segmentación predicha es mayor a la realmente segmentada por los observadores. Esta sobre segmentación es considerablemente mayor en el caso de ReMOS como se observa en los pacientes 38 y 90.

4.4.3. Conjunto C

El tercer conjunto explorado se compone de imágenes que se escapan de la norma, para esto se observa que tanto PGS como ReMOS no logran predecir correctamente las lesiones segmentadas (ver Figura 4.4). Estas resonancias poseen 3 características que las diferencias de un ejemplo normal.

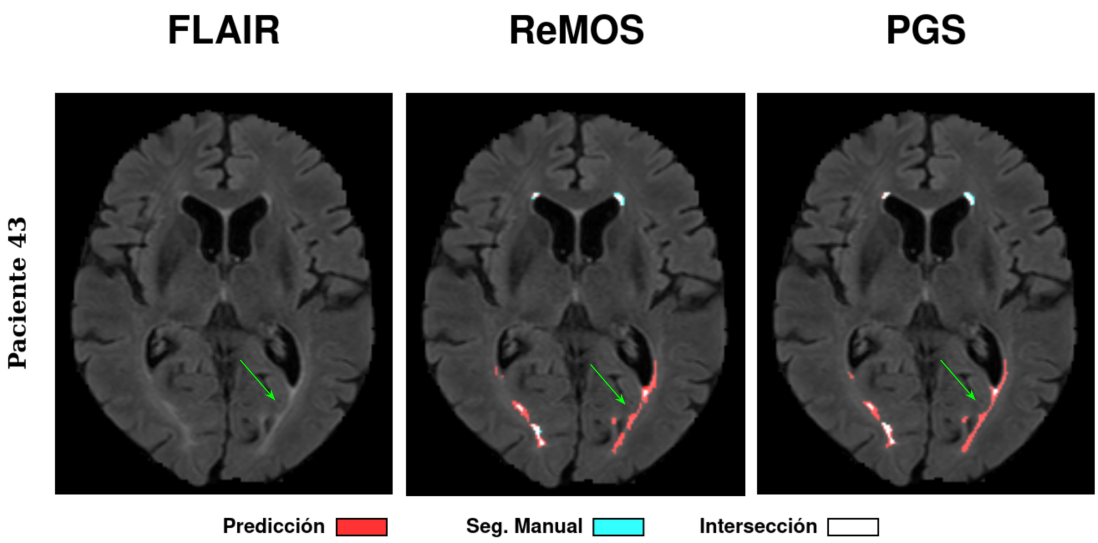
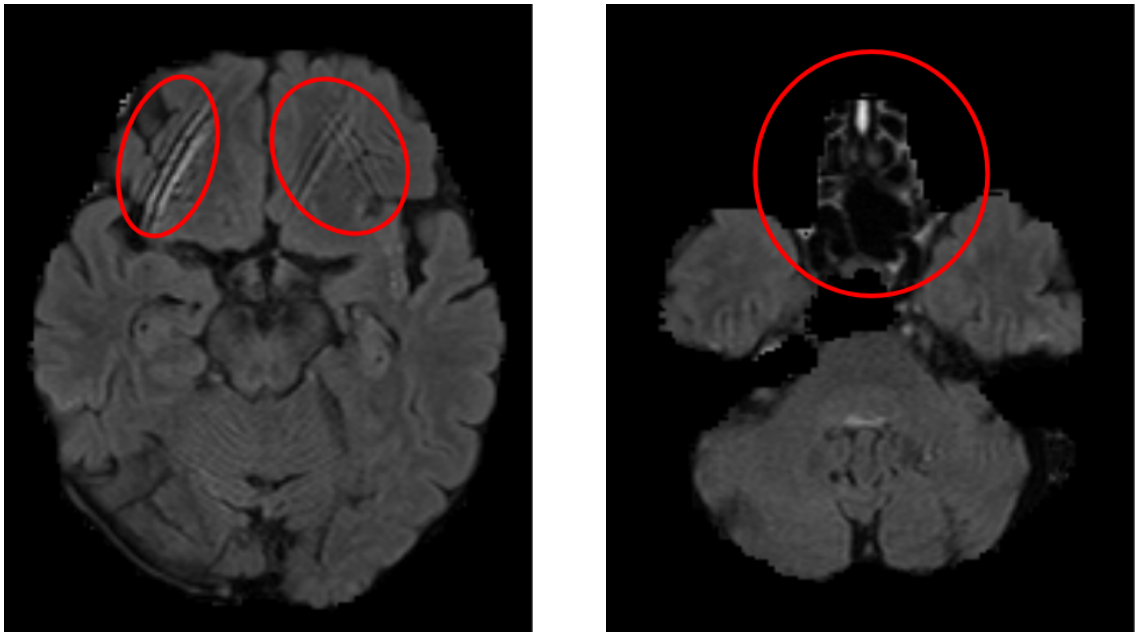


Figura 4.4: Comparación de Outliers PGS vs ReMOS. En este conjunto se observa una sobresegmentación de la lesión tanto en ReMOS como en PGS (flecha verde).

1. Lesiones sobre segmentadas. En particular existe la presencia de lesiones sobre segmentadas, sin embargo, en este conjunto las lesiones predichas poseen una diferencia significativa en área con la segmentación manual, tanto para ReMOS como para PGS (ver Figura 4.4).

2. Ruido magnético. En el proceso de adquisición de una resonancia magnética existen variables que no se pueden controlar, una de ellas viene dado por el movimiento de la persona o ruido electromagnético como se presenta en esta imagen y que generalmente se ve como oleadas dentro de la imagen de resonancia magnética (ver Figura 4.5 (a)).

3. Error ROBEX. ROBEX es un método de extracción cerebral que se basa en atlas previamente obtenidos con una gran recolección de cerebros, sin embargo, ante fallas inesperadas (como las nombradas anteriormente) puede presentar fallos al momento de realizar la extracción cerebral, en particular, se puede observar problemas al extraer parte de boca y cara del paciente en la Figura 4.5 (b)).



(a) Ruido en campo magnético

(b) Error en extracción de cerebro (ROBEX)

Figura 4.5: Ejemplo outliers (Paciente 43). (a) Ruido en el campo magnético observable como rasgaduras en la adquisición. (b) Error en la extracción de la corteza cerebral por el programa ROBEX.

4.5. Discusión

Previamente, se evaluó el desempeño de modelos ViTs aplicados a la tarea de segmentación de lesiones en sustancia blanca, lo cual mejora el rendimiento en las métricas de H95 y AVD, sin embargo, no es suficiente para alcanzar o igualar el rendimiento de PGS. Al aplicar ReMOS en los modelos basados en ViT se mejora el rendimiento para UPerNET en cuanto a las métricas previamente mejoradas, pero se entiende que esta es una arquitectura base y que posee una gran cantidad de parámetros que pueden afectar el resultado de la segmentación.

4.5.1. Optimización ReMOS

Al optimizar los parámetros de ReMOS se busca abarcar todas las variables dentro de la nueva implementación que puedan afectar el rendimiento, es decir, evitar el estudio de parámetros fuera de la arquitectura propuesta, ya que la cantidad de pruebas es muy extensa para un solo estudio. Para esto se realiza un estudio manual del rendimiento del modelo respecto al tipo de entrada, función de pérdida y los pesos asociados a cada capa de ReMOS. Al variar el tipo de entrada que recibe el modelo se observa que no existe un cambio significativo al usar (FLAIR, T1, 0) y solo FLAIR, por otro lado, añadir una nueva variable de la combinación de estos (T1/T2) añade información que perjudica el rendimiento de la segmentación.

La función de pérdida es una herramienta que permite al modelo aprender la tarea objetivo. En este estudio se considera: i) la función dice de pérdida, enfocada a maximizar la intersección entre la segmentación predicha con el objetivo, y ii) la función Logcosh diceloss la cual es una nueva función de pérdida con un objetivo equivalente que Dice, sin embargo,

considera el desequilibrio de clases típico en la segmentación binaria en imágenes médicas. Lo anterior se refleja en los resultados de la tabla 4.4, en donde Logcosh Diceloss supera a Diceloss en 3 métricas e igualando en recall.

Cada una de las salidas de ReMOS posee un peso asociado que determina su relevancia a la función de pérdida global, en la tabla 4.5 se puede observar una leve diferencia entre una distribución equitativa (0.25 en todas la capas) con respecto a los pesos que dan una mayor relevancia a la segmentación final (peso 4) indicando que para el modelo es más facil lograr una buena predicción mirando con más detalle al objetivo original. Dentro de los 3 pesos que poseen mayor impacto en la segmentación final se destaca el conjunto número 4 (0.250, 0.125, 0.0625, 0.5625) con la mayor influencia en el objetivo principal y obtenido mejores resultados en H95 y recall.

Finalmente, en el estudio de parámetros se determinó la selección de input FLAIR, función de pérdida Logcosh Diceloss y con un peso asociado de (0.250, 0.125, 0.0625, 0.5625) para la prueba final de comparación. La comparación final supera a PGS en 2 de 5 métricas (H95 y AVD), y mejora el rendimiento del modelo base UPerNET en 4 de 5 métricas.

4.5.2. Sobre segmentación

Con el fin de comprender el comportamiento de ReMOS se realiza una comparación de cada paciente, los resultados principales indican 2 tipos de comportamientos que diferencia ReMOS de PGS. El primer conjunto se compone por una subdivisión en las métricas de recall y Dice, en particular el desempeño en estas métricas indica una baja en la intersección entre el resultado predicho y la segmentación manual. Sin embargo, la diferencia volumétrica (AVD) es bastante alta, es decir que la alternativa posible a un mal desempeño en estas métricas implica una sobre segmentación del área de las lesiones. Por otro lado, el análisis indica un conjunto de imágenes con valores anormales (outliers), si bien la magnitud de cada uno de estos es distinto para cada modelo se encuentra un solapamiento entre los pacientes anormales tanto PGS como para ReMOS indicando que este comportamiento no es característico de cada arquitectura.

Confirmando los resultados obtenidos en el análisis en la figura 4.3, se presentan 3 ejemplos de sobre segmentación en ReMOS y se comparan con los resultados obtenidos por PGS. La sobre segmentación de los modelos se da principalmente en sectores cercanos al centro (líquido cefalorraquídeo) y que poseen una intensidad parecida a las lesiones vasculares, sin embargo, ReMOS posee una tendencia a extender esta sobre segmentación debido a la influencia de las capas de *multi-output*. Esta área periventricular es reconocida por un alto índice de lesiones vasculares e hiperintensidades naturales (sin causa), por lo anterior, es común para radiólogos obviar estas lesiones o darle menor relevancia al momento de segmentar.

4.5.3. Errores externos

Con respecto a las imágenes con valores fuera de la norma, se establece que existen 2 características, más allá de una sobre segmentación de las lesiones. La primera característica se compone por el ruido magnético generado por imperfecciones al momento de adquirir las imágenes de MRI. La segunda característica da referencia a la extracción incompleta o errónea del cerebro por ROBEX. Particularmente, estos componentes no presentan un problema para los modelos, ya que, se encuentran presentes en otros pacientes. Sin embargo, la presencia de los 2 en un mismo paciente impide a ambos modelos identificar las lesiones como corresponde.

Estos dos factores se encuentran dentro de un conjunto de comportamiento anómalo, ya sea los errores en la extracción de las imágenes (ruido magnético) o preprocesamiento del cerebro (ROBEX) no se considera como una variable solucionable dentro de este trabajo.

4.5.4. Costo computacional

Entre de las ventajas de implementar nuevas herramientas se encuentran la velocidad de procesamiento y el costo computacional. En este trabajo se evaluó el desempeño temporal de 3 algoritmos, PGS, ReMOS y la actual herramienta de uso clínico SPM/LSTM (LGA) *Matlab*. En primer lugar SPM/LSTM posee un tiempo análisis de 0.1 pacientes por minuto debido a su comportamiento iterativo para llegar a su solución. PGS por su parte es una herramienta basada en U-Net/pytorch, sin embargo, no se encuentra diseñado para optimizar el tiempo y recursos logrando tiempo análisis de 1.3 pacientes por minuto, finalmente, ReMOS al ser diseñado en una estructura optimizada puede procesar 11 pacientes por minuto, mejorando considerablemente el tiempo análisis de lesiones en sustancia blanca.

4.5.5. Etiquetado disponible

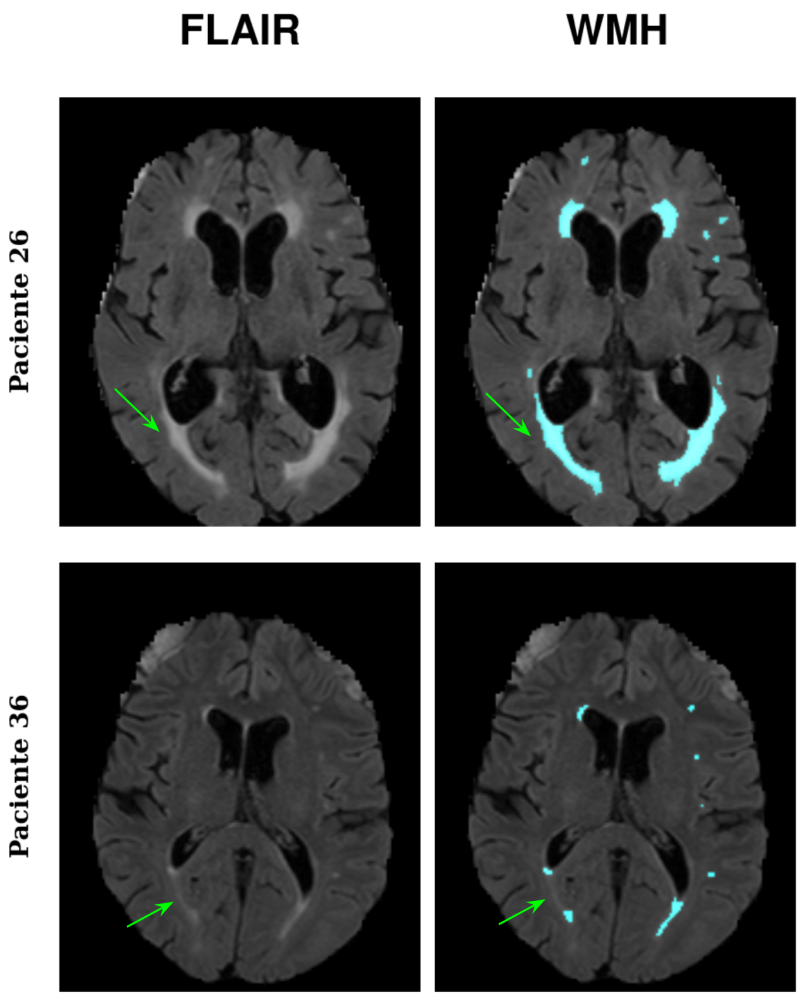


Figura 4.6: Diferencias en etiquetado para distintos pacientes. En el paciente 26 se observa toda el área con presencia de hiperintensidad identificada como lesión (flecha verde), por otro lado, el paciente 36 solo presenta una fracción del área con hiperintensidad identificada como lesión (flecha verde).

Dentro de las pruebas realizadas se pudo observar diferencias en el etiquetado disponible, en la Figura 4.6 se encuentra el ejemplo de una buena segmentación realizada por ReMOS (paciente 26) y una sobre segmentación (paciente 36). En esta comparación se observa una segmentación completa de la sección periventricular para el paciente 26, sin embargo, en la misma zona para el paciente 36 se ve que no se marca por completo como una lesión. Estas diferencias implican la existencia de diferencias en la precisión de la segmentación manual por parte de los observadores.

Capítulo 5

Conclusión

En esta tesis se propuso un modelo basado en ViT que incorpora ideas de arquitecturas U-Net para resaltar la segmentación de objetos pequeños como es el *deep supervision*, aplicado a la segmentación de lesiones de sustancia blanca. Este modelo se denominó *Resized Multi Output Segmentation o ReMOS* el cual logra un tiempo de análisis de 11 pacientes por minuto en comparación de un método basado en U-Net (PGS, actual estado del arte) con un tiempo de análisis de 1.3 pacientes por minuto. Por otro lado, ReMOS posee un mejor rendimiento que los modelos base de ViT aplicados al mismo problema de segmentación (Swin Transformer + UPerNET o Segformer) en 4 de las 5 métricas propuestas (DSC, H95, Recall y F1), lo cual implica que la técnica de supervisión profunda provee una mejora a las arquitecturas basadas en transformer para la tarea de segmentación semántica.

Al realizar una evaluación del modelo ReMOS y comparar las métricas Dice, Hausdorf (H95), diferencia volumétrica (AVD), recall y f1-score siguiendo la metodología de la competencia internacional de segmentación de lesiones en sustancia blanca [10] se observa que ReMOS logra superar a PGS en 2.27 mm en H95 y 1.99% en AVD. Sin embargo, posee un peor rendimiento en coeficiente dice por 0.3, recall en 0.12 y f1-score en 0.05. En vista de lo anterior, se determina que la arquitectura propuesta ReMOS no logra superar al actual estado del arte en 3 de las 5 métricas propuestas y en consecuencia se rechaza la hipótesis de qué recientes arquitecturas basadas en ViT que incorporen adaptaciones de arquitecturas U-Net para el problema de segmentación de WMH, en conjunto con variables de la literatura en neurociencia permiten mejorar los resultados actuales para la segmentación automática de hiperintensidades en resonancias magnéticas cerebrales. Sin embargo, esto no descarta la posibilidad de que otras configuraciones o modelos basados en ViT posean mejor rendimiento que modelos basados en CNNs para esta tarea. En esta línea, una hipótesis que podría explicar la diferencia de ReMOS con PGS es que se requiere de una mayor exploración de parámetros, pero al ser esta costosa computacionalmente (10 horas de entrenamiento por modelo con GPU NVIDIA GeForce RTX 4090) se hace compleja de realizar de manera extensiva.

Como trabajo futuro se plantea la optimización del modelo ReMOS abarcando distintas aristas no evaluadas en esta tesis tales como: variar la arquitectura del encoder (Swin transformer), utilizar métodos de penalización para restringir la sobre segmentación de las lesiones, plantear nuevas ideas de transferencia de aprendizaje al variar entre un modelo preentrenado con ImageNet a uno utilizando tareas con resonancias magnéticas, y evaluar ReMOS utilizando validación cruzada.

Otro punto a evaluar es la concordancia de la segmentación manual en las resonancias

de la competencia, lo que implica el consultar a personal experto en el área de la radiología con el fin de adquirir una tercera opinioń de la segmentación obtenida y la referencias en la competencia. Finalmente, esta tesis posee un enfoque algoritmico, sin embargo, un trabajo futuro de interes es la implementación de un software para el uso clínico y científico.

Bibliografía

[1] Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., Ballard, C., Banerjee, S., Burns, A., Cohen-Mansfield, J., *et al.*, "Dementia prevention, intervention, and care," The Lancet, vol. 390, no. 10113, pp. 2673–2734, 2017.

[2] "Dementia.", https://www.who.int/news-room/fact-sheets/detail/dementia. Accessed on 20.09.2023.

[3] Wardlaw, J. M., Valdés Hernández, M. C., y Muñoz-Maniega, S., "What are white matter hyperintensities made of? relevance to vascular cognitive impairment," Journal of the American Heart Association, vol. 4, no. 6, p. e001140, 2015.

[4] Fazekas, F., Barkhof, F., Wahlund, L., Pantoni, L., Erkinjuntti, T., Scheltens, P., y Schmidt, R., "Ct and mri rating of white matter lesions," Cerebrovascular diseases, vol. 13, no. Suppl. 2, pp. 31–36, 2002.

[5] Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V. J., Zimmer, C., *et al.*, "An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis," Neuroimage, vol. 59, no. 4, pp. 3774–3783, 2012.

[6] Park, G., Hong, J., Duffy, B. A., Lee, J.-M., y Kim, H., "White matter hyperintensities segmentation using the ensemble u-net with multi-scale highlighting foregrounds," Neuroimage, vol. 237, p. 118140, 2021.

[7] Freire, M. J., "Fundamentos físicos de las imágenes médicas: Resonancia magnética.", https://personal.us.es/alberto/ffisim/material/Resonancia.pdf. Accessed on 20.09.2023.

[8] Prins, N. D. y Scheltens, P., "White matter hyperintensities, cognitive impairment and dementia: an update," Nature Reviews Neurology, vol. 11, no. 3, pp. 157–165, 2015.

[9] Medel, V., Vidal, V., Gonzalez-Gomez, R., Vergara, R., Wainstein, G., Orellana, P., Ibanez, A., Shine, J. M., Delgado, C., Delano, P. H., *et al.*, "Cortical white matter hyperintensities are associated with locus coeruleus atrophy in elder subjects," en Alzheimer's Association International Conference, ALZ, 2023.

[10] Kuijf, H. J., Biesbroek, J. M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M. J., Casamitjana, A., Collins, D. L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Lladó, X., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B.-Y., Park, H., Park, S. H., Pezold, S., Puybareau, E., Rittner, L., Sudre, C. H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Chen, C., van der Flier, W., Barkhof, F., Viergever, M. A., y Biessels, G. J., "Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge," IEEE Transactions on Medical Imaging, vol. 38, no. 11,

pp. 2556–2568, 2019, doi:10.1109/TMI.2019.2905770.

[11] Liu, D. y Yu, J., "Otsu method and k-means," en 2009 Ninth International Conference on Hybrid Intelligent Systems, vol. 1, pp. 344–349, 2009, doi:10.1109/HIS.2009.74.

[12] Ronneberger, O., Fischer, P., y Brox, T., "U-net: Convolutional networks for biomedical image segmentation," en Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241, Springer, 2015.

[13] He, J., Deng, Z., Zhou, L., Wang, Y., y Qiao, Y., "Adaptive pyramid context network for semantic segmentation," en Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[14] Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., y Torralba, A., "Semantic understanding of scenes through the ade20k dataset," International Journal of Computer Vision, vol. 127, pp. 302–321, 2019.

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[16] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., y Jégou, H., "Training data-efficient image transformers & distillation through attention," en International conference on machine learning, pp. 10347–10357, PMLR, 2021.

[17] Matsoukas, C., Haslum, J. F., Söderberg, M., y Smith, K., "Is it time to replace cnns with transformers for medical images?," arXiv preprint arXiv:2108.09038, 2021.

[18] Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., y Halpern, A., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," en 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172, 2018, doi:10.1109/ISBI.2018.8363547.

[19] Tschandl, P., Rosendahl, C., y Kittler, H., "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific data, vol. 5, no. 1, pp. 1–9, 2018.

[20] Matsoukas, C., Hernandez, A. B., Liu, Y., Dembrower, K., Miranda, G., Konuk, E., Haslum, J. F., Zouzos, A., Lindholm, P., Strand, F., *et al.*, "Adding seemingly uninformative labels helps in low data regimes," en International Conference on Machine Learning, pp. 6775–6784, PMLR, 2020.

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, $\mathcal{L}$., y Polosukhin, I., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[22] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., y Zagoruyko, S., "End-to-end object detection with transformers," en European conference on computer vision, pp. 213–229, Springer, 2020.

[23] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., y Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," en Proceedings of the

IEEE/CVF international conference on computer vision, pp. 10012–10022, 2021.

[24] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., y Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," en 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, Ieee, 2009.

[25] Xiao, T., Liu, Y., Zhou, B., Jiang, Y., y Sun, J., "Unified perceptual parsing for scene understanding," 2018.

[26] Zhao, H., Shi, J., Qi, X., Wang, X., y Jia, J., "Pyramid scene parsing network," en Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890, 2017.

[27] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., y Luo, P., "Segformer: Simple and efficient design for semantic segmentation with transformers," Advances in Neural Information Processing Systems, vol. 34, pp. 12077–12090, 2021.

[28] Jadon, S., "A survey of loss functions for semantic segmentation," en 2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB), pp. 1–7, IEEE, 2020.

[29] Puccio, B., Pooley, J. P., Pellman, J. S., Taverna, E. C., y Craddock, R. C., "The preprocessed connectomes project repository of manually corrected skull-stripped t1-weighted anatomical mri data," Gigascience, vol. 5, no. 1, pp. s13742–016, 2016.

[30] Iglesias, J. E., Liu, C.-Y., Thompson, P. M., y Tu, Z., "Robust brain extraction across datasets and comparison with publicly available methods," IEEE Transactions on Medical Imaging, vol. 30, no. 9, pp. 1617–1634, 2011, doi:10.1109/TMI.2011.2138152.

[31] MMSegmentation Contributors, "OpenMMLab Semantic Segmentation Toolbox and Benchmark," 2020, https://github.com/open-mmlab/mmsegmentation.