UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Sistemas Informáticos y Computación

Fusión Multimodal de Datos Satelitales para la Segmentación de Imágenes

Trabajo Fin de Máster

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

AUTOR/A: Obrador Reina, Miquel

Tutor/a: Paredes Palacios, Roberto

Cotutor/a: Albiol Colomer, Alberto

CURSO ACADÉMICO: 2023/2024

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

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

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València

Fusión Multimodal de Datos Satelitales para la Segmentación de Imágenes

TRABAJO FIN DE MÁSTER

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

Autor: Miquel Obrador Reina

Tutor: Roberto Paredes Palacios
Alberto Albiol Colomer

Curso 2023-2024

Resum

Aquest treball investiga tècniques avançades de segmentació d'imatges satel·litàries mitjançant la fusió multimodal de dades de diferents satèl·lits, específicament Sentinel-1 i Sentinel-2.

Sentinel-1 ofereix imatges RADAR amb dues bandes, capaces de capturar imatges independentment de les condicions meteorològiques, mentre que Sentinel-2 proporciona imatges d'alta resolució en tretze bandes, incloent RGB, cosa que permet una exploració més detallada per a la segmentació. En aquest estudi, es comparará el rendiment dels models utilitzant únicament imatges de Sentinel-1 o Sentinel-2, així com diferents estratègies de fusió: fusió primerenca (early fusion) on totes les imatges s'introdueixen en un únic model; fusió intermèdia (intermediate fusion), en què es combinen característiques dels dos satèl·lits dins de l'arquitectura del model; i fusió tardana (late fusion), que integra les sortides de models independents per a cada conjunt de dades satel·litàries, combinant-les al final per obtenir la predicció final.

Es faran servir sistemes de segmentació basats en xarxes neuronals convolucionals (CNNs), amb arquitectures adaptades per integrar la fusió multimodal. Aquest estudi té com a objectiu avaluar com la fusió de dades multimodal pot millorar la precisió i l'eficiència en la segmentació d'imatges satelitals.

Paraules clau: Imatges de Satèl·lit, Segmentació en Imatges Satelitals, Fusió Multimodal de Dades, Xarxes Neuronals

Resumen

Este trabajo investiga técnicas avanzadas de segmentación de imágenes satelitales mediante la fusión multimodal de datos de diferentes satélites, específicamente Sentinel-1 y Sentinel-2.

Sentinel-1 ofrece imágenes RADAR con dos bandas, capaces de capturar imágenes independientemente de las condiciones meteorológicas, mientras que Sentinel-2 proporciona imágenes de alta resolución en trece bandas, incluyendo RGB, lo que permite una exploración más detallada para la segmentación. En este estudio, se comparará el rendimiento de los modelos utilizando únicamente imágenes de Sentinel-1 o Sentinel-2, así como diferentes estrategias de fusión: fusión temprana (early fusion), donde todas las imágenes se introducen en un único modelo; fusión intermedia (intermediate fusion), en la que se combinan características de los dos satélites dentro de la arquitectura del modelo; y fusión tardía (late fusion), que integra las salidas de modelos independientes para cada conjunto de datos satelitales, combinándolas al final para obtener la predicción final.

Se emplearán sistemas de segmentación basados en redes neuronales convolucionales (CNNs), con arquitecturas adaptadas para integrar la fusión multimodal. Este estudio tiene como objetivo evaluar cómo la fusión de datos multimodal puede mejorar la precisión y eficiencia en la segmentación de imágenes satelitales.

Palabras clave: Imágenes de Satélite, Segmentación en Imágenes Satelitales, Fusión Multimodal de Datos, Redes Neuronales

Abstract

This work investigates advanced satellite image segmentation techniques by multimodal fusion of data from different satellites, specifically Sentinel-1 and Sentinel-2.

Sentinel-1 provides RADAR imagery with two bands, capable of capturing images regardless of weather conditions, while Sentinel-2 provides high-resolution imagery in thirteen bands, including RGB, allowing for more detailed scanning for segmentation. In this study, model performance will be compared using only Sentinel-1 or Sentinel-2 imagery, as well as different fusion strategies: early fusion, where all images are fed into a single model; intermediate fusion, where features from the two satellites are combined within the model architecture; and late fusion, which integrates independent model outputs for each satellite dataset, combining them at the end to obtain the final prediction.

Segmentation systems based on convolutional neural networks (CNNs) will be used, with architectures adapted to integrate multimodal fusion. This study aims to evaluate how multimodal data fusion can improve the accuracy and efficiency of satellite image segmentation.

Key words: Satellite Imagery, Segmentation in Satellite Imagery, Multimodal Data Fusion, Neural Networks

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

En este primer capítulo se expone la motivación y se introduce el problema de Segmentación en Imágenes Satelitales. Se explicará el funcionamiento de los distintos satélites y las características de sus imágenes. Finalmente, se exponen los objetivos del Trabajo de Fin de Máster y su estructura.

1.1 Motivación

La segmentación de imágenes satelitales es una herramienta crucial en múltiples aplicaciones, desde la monitorización ambiental y la gestión de recursos naturales hasta la planificación urbana y la respuesta a desastres naturales [16], entre otras. Sin embargo, la precisión y eficiencia de estas segmentaciones pueden verse afectadas significativamente por las limitaciones inherentes a los datos obtenidos de un único tipo de sensor satelital.

Los satélites de imagen óptica, como Landsat y Sentinel-2, capturan la luz que entra en sus sensores y ofrecen imágenes de alta resolución y detalle en las bandas visibles, ultravioleta e infrarrojas del espectro electromagnético (Figura 1.1). Sin embargo, estos satélites se ven afectados negativamente por condiciones meteorológicas adversas, como las nubes, que reflejan y absorben la luz, impidiendo que llegue al sensor. Además, la oscuridad también es un obstáculo, ya que la falta de luz impide la captura de imágenes.

Por otro lado, existen satélites con tecnología SAR (Radar de apertura sintética), como el Sentinel-1, que emiten ondas de microondas (con una longitud de aproximadamente 5 cm) y las reciben de vuelta. Debido a su gran longitud de onda, estas pueden atravesar las nubes, lo que permite la captura de imágenes independientemente de las condiciones meteorológicas. Además, al ser emitidas por el satélite, no se ven afectadas por las condiciones de luz, permitiendo la captura de imágenes tanto de día como de noche. Sin embargo, las imágenes de radar, aunque proporcionan una gran ventaja al poder operar en cualquier condición meteorológica y de iluminación, suelen tener una resolución espacial más baja y carecen de la riqueza espectral que ofrecen los satélites ópticos. Esto se traduce en una menor capacidad para distinguir entre diferentes tipos de superficies y materiales, ya que las imágenes SAR están principalmente basadas en la intensidad de la señal reflejada y la geometría de las superficies, en lugar de la variabilidad espectral detallada. Además, interpretar las imágenes de radar puede ser más complejo debido a los efectos de la geometría y la rugosidad de los objetos en la señal recibida, lo que puede dificultar la segmentación precisa de ciertos elementos del terreno.

En nuestro proyecto, nos proponemos distinguir las zonas con presencia de agua y desbordamientos, conocidas como *Water Bodies*, del resto de los terrenos. Por lo tanto, nos interesa segmentar las áreas de agua píxel por píxel de otros tipos de superficies. Adicio-

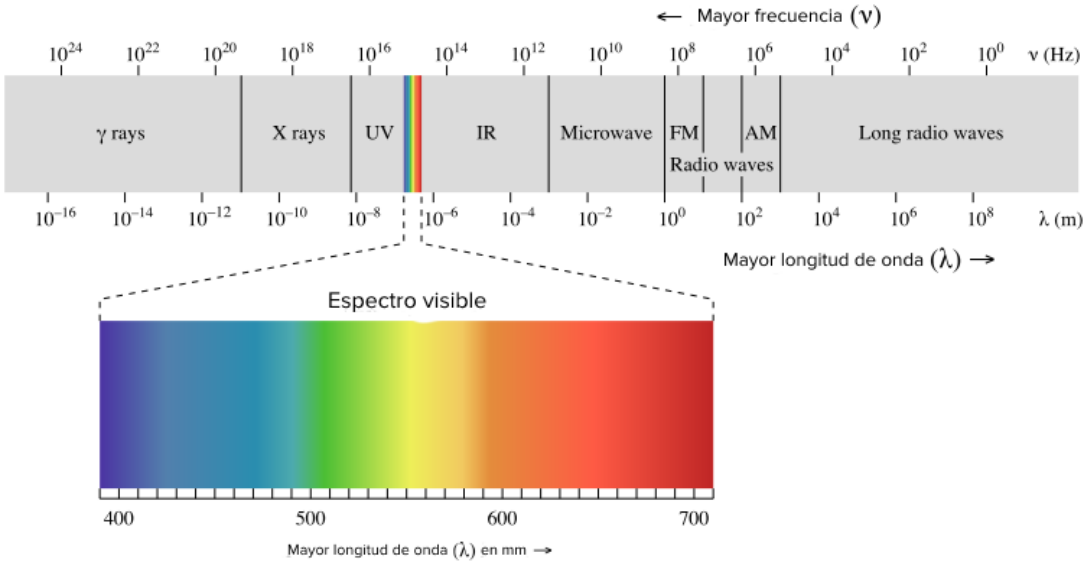

Figura 1.1: Espectro electromagnético

nalmente, esto es especialmente relevante en condiciones meteorológicas adversas, como en imágenes completamente o parcialmente nubladas, así como en condiciones de nula luminosidad.

A pesar de los avances en la segmentación de imágenes, aún existen desafíos significativos debido a la complejidad y variabilidad de las imágenes satelitales. Actualmente, los satélites ópticos parecen brindar los mejores resultados para la detección de cuerpos de agua. Sin embargo, cuando la imagen está nublada o no hay luz, es imposible realizar una segmentación precisa. Por lo tanto, es necesario esperar otra imagen de la misma zona (que suele tener un tiempo de revisita de 5 días) y confiar en que esté lo suficientemente despejada para una segmentación adecuada. Alternativamente, se puede implementar un sistema complejo que evalúe la calidad de la imagen y decida si es mejor utilizar los datos de satélite óptico o recurrir a datos de radar, que ofrecen una segmentación menos precisa pero generalmente aceptable. Sin embargo, montar este sistema es costoso tanto en términos económicos como computacionales, y aún así puede no ofrecer resultados óptimos. Tal vez, la fusión multimodal podría ser la solución a ambos problemas de una forma más sencilla y eficiente.

La fusión multimodal de datos de diferentes satélites ha demostrado ser una estrategia efectiva para mejorar la precisión y eficiencia en la segmentación, ya que permite combinar información complementaria de diferentes fuentes [18].

La combinación de distintos tipos de datos a través de técnicas avanzadas de fusión multimodal tiene el potencial de mejorar significativamente la calidad y robustez de las segmentaciones de imágenes satelitales. En particular, se espera que esta fusión permita una mayor precisión en la identificación y clasificación de diferentes características del terreno, mejorando así la toma de decisiones en las aplicaciones mencionadas.

La motivación de este estudio radica en la necesidad de mejorar las técnicas de segmentación de imágenes satelitales. La capacidad de obtener segmentaciones más precisas y eficientes puede tener un impacto significativo en áreas como la agricultura de precisión, la gestión de recursos hídricos, la conservación de la biodiversidad y la respuesta rápida a desastres naturales, contribuyendo así a un uso más sostenible y eficaz de los recursos naturales y a la mejora de la resiliencia frente a eventos adversos.

1.2 Agradecimientos

Quiero expresar mi más profundo agradecimiento a ValgrAI por el apoyo y la confianza depositada en mí durante todo el proceso de realización de este Máster. Su contribución ha sido fundamental para poder concentrarme plenamente en mis estudios y en la elaboración de este trabajo. Además, las oportunidades brindadas por ValgrAI, tanto en el ámbito académico como profesional, han sido invaluables para mi desarrollo y crecimiento en el campo de la inteligencia artificial. Su compromiso con la formación y el impulso de nuevos talentos es inspirador, y me siento muy afortunado de haber contado con su respaldo en esta etapa tan importante de mi carrera.

1.3 Sentinel-1 (S1) y Sentinel-2 (S2)

Sentinel-1 y Sentinel-2 son dos misiones satelitales operadas por la Agencia Espacial Europea (ESA) como parte del programa Copernicus, diseñado para ofrecer datos y servicios de observación de la Tierra de manera continua y confiable. Ambos satélites proporcionan información crucial para una amplia gama de aplicaciones ambientales y de seguridad, aunque emplean tecnologías de sensores diferentes y complementarias.

1.3.1. Sentinel-1

Sentinel-1 es una misión de radar de apertura sintética (SAR) compuesta por dos satélites, Sentinel-1A y Sentinel-1B, lanzados en 2014 y 2016, respectivamente. Estos satélites operan en la banda C y están diseñados para proporcionar datos de radar en todo tipo de condiciones meteorológicas, día y noche. Las imágenes SAR son especialmente útiles para monitorear cambios en la superficie terrestre, detectar movimientos del suelo y cartografiar zonas inundadas, entre otras aplicaciones.

Características de las Imágenes de Sentinel-1

1. Resolución: Sentinel-1 ofrece imágenes con resoluciones que varían entre 5 y 40 metros, dependiendo del modo de operación.

2. Modos de Operación: Los satélites Sentinel-1 pueden operar en varios modos, incluyendo:

   - **Interferometric Wide (IW):** Modo principal para observaciones terrestres, con una resolución espacial de 5 x 20 metros.
   - **Extra-Wide Swath (EW):** Utilizado principalmente para el monitoreo de áreas costeras y marítimas, con una resolución espacial de 20 x 40 metros.
   - **Stripmap (SM):** Proporciona la mayor resolución (hasta 5 metros), utilizado para aplicaciones que requieren detalles finos.

3. Bandas: Sentinel-1 utiliza la banda C (5.405 GHz), proporcionando dos polarizaciones (VV y VH) en el modo IW, lo que permite obtener información detallada sobre la estructura y propiedades de la superficie terrestre. Las polarizaciones en el contexto de la imagen de radar de apertura sintética (SAR) se refieren a la orientación de la onda electromagnética transmitida y recibida. A continuación, se explican las dos polarizaciones principales utilizadas por Sentinel-1:

- **VV (Vertical-Vertical) Polarización**: En la polarización VV, la señal de radar se transmite verticalmente y se recibe también verticalmente. Esto significa que la antena del radar emite una onda electromagnética con polarización vertical y luego recibe la reflexión de esa onda con la misma polarización vertical.
- **VH (Vertical-Horizontal) Polarización**: En la polarización VH, la señal de radar se transmite verticalmente pero se recibe horizontalmente. Esto implica que la antena del radar emite una onda electromagnética con polarización vertical y luego recibe la reflexión de esa onda con polarización horizontal.

Cada banda de polarización puede resaltar diferentes entidades de la superficie terrestre. Por ejemplo, la polarización VV puede ser más efectiva para detectar superficies reflectantes como agua, mientras que la polarización VH puede ser más útil para detectar vegetación o estructuras urbanas.

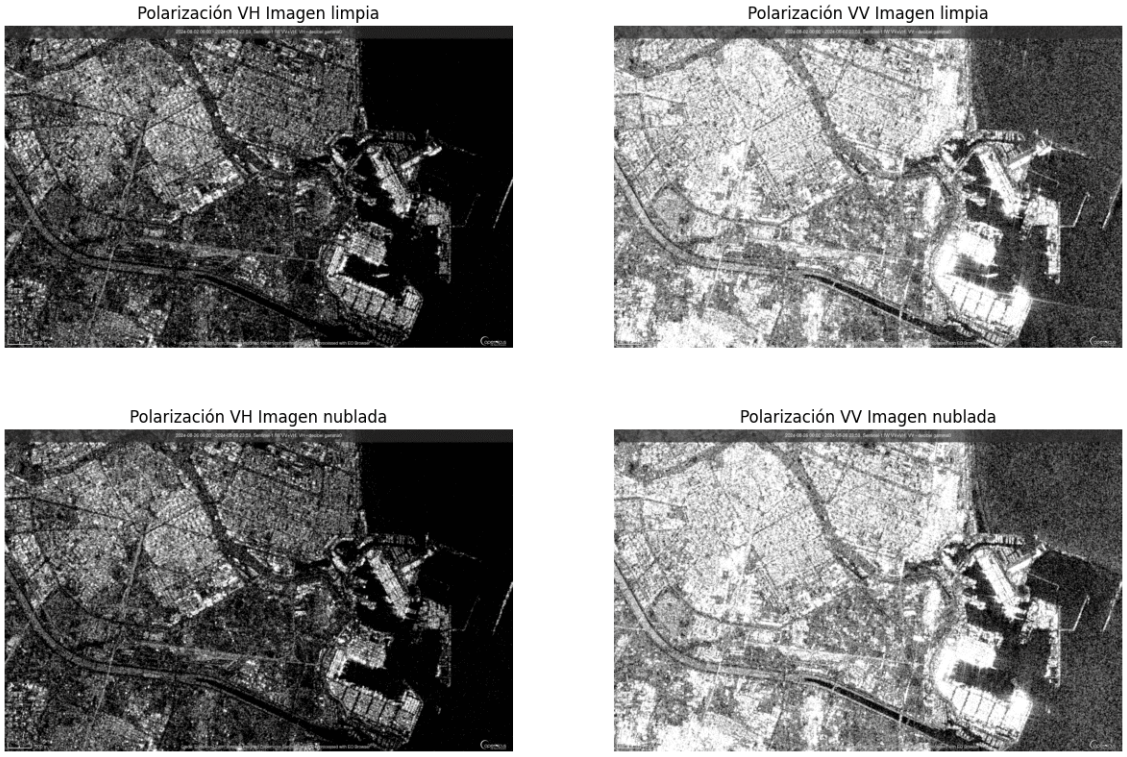

Figura 1.2: Imágenes Sentinel-1

En la Figura 1.2, se muestra una imagen capturada por Sentinel-1, con sus dos polarizaciones en modo de operación IW, correspondiente a la ciudad de Valencia. En la primera fila, el día era claro y despejado, mientras que en la segunda fila, el día era nublado. Sin embargo, los datos parecen no verse afectados por la presencia de nubes. Por otro lado, las imágenes de radar son difíciles de interpretar, ya que aunque podemos intuir qué es cada cosa en la imagen, es difícil distinguir detalles finos.

1.3.2. Sentinel-2

Sentinel-2 es una misión óptica compuesta por dos satélites, Sentinel-2A y Sentinel-2B, lanzados en 2015 y 2017, respectivamente. En 2024 se lanzo un tercero, Seninel-2C, aunque no disponemos de sus datos para este trabajo. Estos satélites están equipados con el instrumento Multispectral Instrument (MSI), que captura imágenes en 13 bandas

espectrales, abarcando desde el visible hasta el infrarrojo de onda corta (SWIR). Sentinel-2 proporciona datos de alta resolución espacial y temporal, ideales para aplicaciones en agricultura, silvicultura, manejo de recursos hídricos y monitoreo ambiental.

Características de las Imágenes de Sentinel-2

1. Resolución: Sentinel-2 ofrece imágenes con resoluciones espaciales de 10, 20 y 60 metros, dependiendo de la banda espectral.

2. Bandas Espectrales: Sentinel-2 cuenta con 13 bandas espectrales, cada una diseñada para aplicaciones específicas:

   - **Bandas de 10 metros:** Bandas 2 (Azul), 3 (Verde), 4 (Rojo) y 8 (Infrarrojo cercano - NIR), estas bandas cubren el espectro visible y el infrarrojo cercano. Estas bandas de alta resolución son ideales para aplicaciones que requieren detalles finos, como el monitoreo de cambios en áreas urbanas, la evaluación de la salud de cultivos y la detección de características geográficas específicas.
   - **Bandas de 20 metros:** Bandas 5 (Red Edge 1), 6 (Red Edge 2), 7 (Red Edge 3), 8A (NIR estrecho), 11 (SWIR 1) y 12 (SWIR 2), estas bandas cubren el infrarrojo cercano y el infrarrojo de onda corta. Estas bandas son adecuadas para análisis que requieren una buena resolución pero no necesitan los detalles más finos. Por ejemplo, se pueden usar para monitorear la salud de los bosques, detectar incendios forestales y evaluar la calidad del agua en ríos y lagos.
   - **Bandas de 60 metros:** Bandas 1 (Aerosoles costeros), 9 (Vapor de agua) y 10 (Cirrus). Estas bandas cubren el ultravioleta, el azul y el infrarrojo de onda corta. La banda 1 se utiliza para la detección de aerosoles, la banda 9 para la detección de nubes y la banda 10 para la detección de vapor de agua. Estas bandas son útiles para aplicaciones que requieren una visión más amplia y no necesitan una resolución tan alta. Por ejemplo, se pueden usar para monitorear la calidad del aire, detectar nubes y evaluar la humedad

3. Cobertura y Frecuencia: Sentinel-2 ofrece una cobertura global con un tiempo de revisita de 5 días, asegurando datos frecuentes y actualizados para monitorear cambios rápidos en la superficie terrestre.

En la Figura 1.3, al igual que en el ejemplo del Sentinel-1, podemos ver una imagen de la ciudad de Valencia capturada por Sentinel-2. En este caso, solo se muestra el espectro visible, utilizando las bandas 2, 3 y 4, que corresponden a azul, verde y rojo, respectivamente. Estas bandas son las más amigables para el ojo humano y ofrecen la mayor resolución. Estas imágenes son significativamente diferentes de las obtenidas por Sentinel-1, ya que la gran resolución y la riqueza espectral (que se incrementa aún más al utilizar más bandas) no se pueden comparar con las tipo RADAR del otro satélite.

Sin embargo, podemos observar un gran problema asociado con estos tipos de datos: las nubes. A medida que crecen en tamaño u opacidad, bloquean parcial o completamente gran parte de la imagen, haciéndola casi inservible. Este efecto afecta a todas las bandas de manera similar como podemos observar en la Figura 1.4. Además, cuando las nubes son del tipo niebla, la imagen no se ve nítida, dificultando la tarea de igual forma.

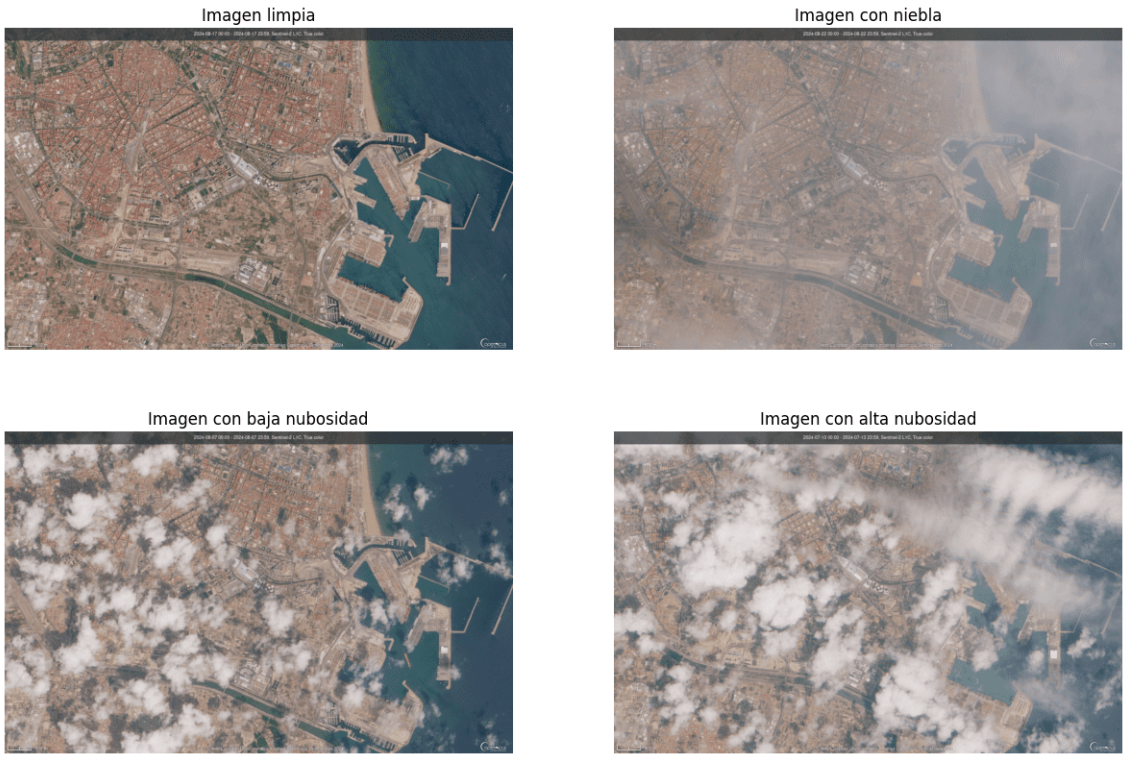

Figura 1.3: Imágenes (RGB) Sentinel-2

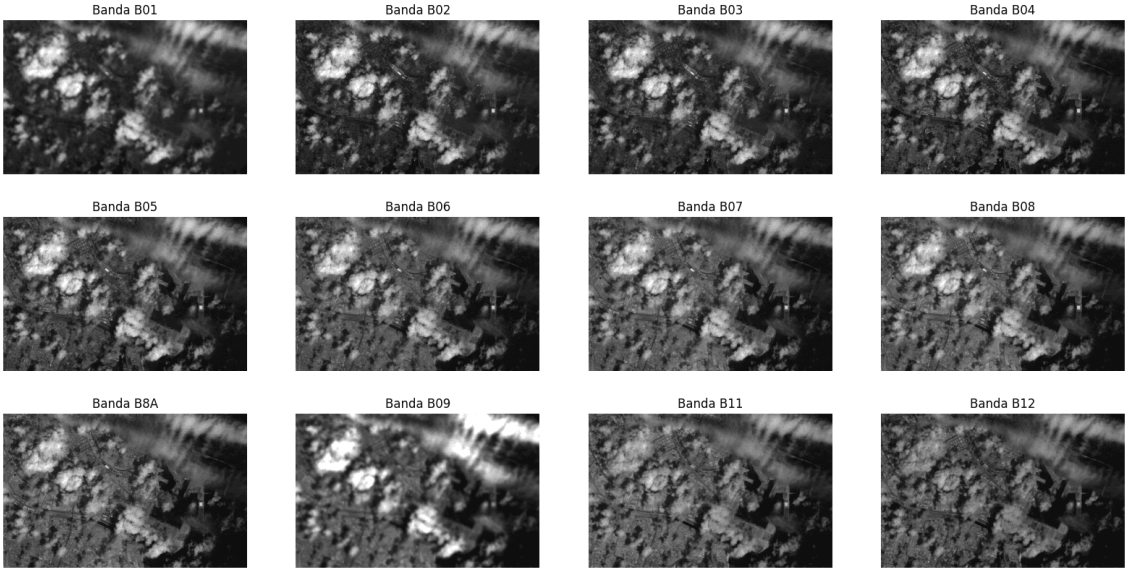

Figura 1.4: Efecto de las nubes en las distintas bandas del Sentinel-2

1.3.3. Comparación y Sinergia entre Sentinel-1 y Sentinel-2

Sentinel-1 y Sentinel-2 proporcionan datos complementarios que, al fusionarse, ofrecen una visión más completa y precisa de la superficie terrestre. La sinergia entre estos satélites se basa en sus diferentes capacidades de captación de datos:

1. **Condiciones de Captura**: Sentinel-1 puede adquirir imágenes independientemente de las condiciones meteorológicas y de iluminación, mientras que Sentinel-2 proporciona datos detallados en condiciones de buen tiempo. Si no hay luz o las condi-

ciones meteorológicas son desfavorables (nubes, tormenta o niebla) las zonas afectadas de las imágenes no proporcionan ninguna información útil.

2. **Resolución y Detalle**: Sentinel-2, con sus múltiples bandas espectrales de alta resolución, permite una segmentación detallada de características específicas del terreno, como la vegetación, el suelo y el agua. Sentinel-1, aunque con una resolución espacial menor, es crucial para detectar cambios estructurales y movimientos en la superficie terrestre.

3. **Aplicaciones**: La combinación de datos SAR y ópticos permite mejorar significativamente la precisión en la identificación y clasificación de diferentes características del terreno. Esto es especialmente útil en aplicaciones como la detección de cambios, el monitoreo de desastres naturales, la agricultura de precisión y la gestión de recursos naturales.

En resumen, la fusión de datos de Sentinel-1 y Sentinel-2 ofrece una herramienta poderosa para la observación de la Tierra, aprovechando la complementariedad de sus sensores para superar las limitaciones individuales y proporcionar información más robusta y precisa para diversas aplicaciones ambientales y de gestión de recursos.

1.4 Fuente de Datos

Una de las principales ventajas de estos satélites es que proporciona acceso gratuito y abierto a los datos de observación de la Tierra, lo que permite a investigadores, instituciones y empresas de todo el mundo utilizar esta valiosa información para diversas aplicaciones.

Los datos tanto de S1 y S2 están disponibles a través de varias plataformas y servicios proporcionados por la ESA. Por ejemplo **Copernicus Open Access Hub**, es la principal plataforma de distribución de datos Sentinel, donde los usuarios pueden buscar y descargar imágenes SAR de Sentinel-1. La plataforma está disponible en: https://dataspace.copernicus.eu.

Aún así encontrar un conjunto de datos grande con imágenes del Sentinel-1 y Sentinel-2 de la misma zona en franjas horarias similares etiquetado con máscaras de agua en condiciones meteorológicas adversas no ha sido posible.

El conjunto de datos que hemos decidido utilizar es parte del 2020 IEEE GRSS DATA FUSION CONTEST [6] organizado por el Comité Técnico de Análisis de Imágenes y Fusión de Datos (IADF TC) de la Sociedad Geofísica y de Sensores Remotos (GRSS) de IEEE y la Universidad Técnica de Munich. El objetivo del concurso era promover la investigación en el mapeo de cobertura terrestre a gran escala utilizando datos multimodales satelitales de alta resolución, esto es clasificar cada píxel de una imágen en 10 clases (Bosque, Matorral, Sabana, Pastizal, Humedal, Tierra de cultivo, Urbano, Nieve, Yermo y Agua), en nuestro caso al interesarnos sobretodo la segmentación de agua, solo nos quedaremos con la clase 10 (Agua) y el problema será una clasificación binaria a nivel de pixel.

El conjunto de datos proviene del SEN12MS dataset [21], este subconjunto para la competición contiene unos 6000 samples alrededor de todo el globo, cada sample es un triplet de imágen de S1, imágen de S2 y el label o máscara objetivo con la clase correspondiente a cada píxel (recordemos que solo nos quedamos con la clase agua, por lo tanto, cada pixel será una clasificación binaria de 1 para agua y 0 para no agua), cada imágen tiene una resolución nativa de 10-20m por píxel i con 256x256 píxeles.

Todas las imágenes de este conjunto de datos presentan una baja cantidad de luz, como se puede observar en la Figura 1.5, la cual ya ha sido preprocesada aumentando la exposición. Esto es ideal, ya que nos permite evaluar el sistema en condiciones lumínicas desfavorables. Sin embargo, todos los *samples* están completamente libres de nubes, por lo que se añadirán de manera artificial mediante un proceso algorítmico, que se describirá en detalle en el capítulo de Metodología.

Figura 1.5: *Sample* de nuestro conjunto de datos de Sentinel-2

1.5 Objetivos

Este Trabajo de Fin de Máster se centra en explorar y comparar diferentes estrategias de fusión multimodal y sistemas de segmentación basados en redes neuronales convolucionales (CNNs) para mejorar el proceso de segmentación de imágenes satelitales. Por lo que en este apartado se presentan los principales objetivos.

El primer objetivo será desarrollar un sistema que combine datos de satélites ópticos (Sentinel-2) y de radar (Sentinel-1) para mejorar la precisión y robustez en la segmentación de imágenes satelitales, especialmente en condiciones adversas como nubosidad y falta de luz. A la vez que se obtiene un sistema único que integre los datos de ambos sensores de manera eficiente, reduciendo la carga computacional asociada con el uso de múltiples sistemas separados.

El segundo objetivo será evaluar y comparar diferentes métodos de fusión de datos, como la *Early Fusion* y la *Intermediate Fusion* al igual que explorar diferentes arquitecturas de redes neuronales como U-Net o DeepLab, y evaluar su desempeño en la segmentación de agua en imágenes satelitales.

Como último objetivo, se propone una comparativa entre los modelos diseñados en el tercer objetivo y el Estado del Arte actual en la segmentación de agua en imágenes satelitales.

1.6 Estructura de la memoria

La memoria se estructura del siguiente modo:

- **Capítulo 1, Introducción**: En este capítulo se presenta la motivación y los objetivos del proyecto, explicando la importancia de la segmentación de imágenes satelitales y los desafíos asociados a esta tarea. Además, se introduce el problema específico abordado y los satélites utilizados, detallando sus características y capacidades.
- **Capítulo 2, Estado del Arte**: Se revisan las principales técnicas de segmentación de imágenes satelitales, con un enfoque tanto en las arquitecturas de redes neuronales convolucionales para segmentación semántica genérica, como U-Net, así como en arquitecturas específicas para la segmentación de cuerpos de agua, como DeepWaterMapV2 y WatNet.
- **Capítulo 3, Metodología**: Este capítulo describe en detalle la preparación de los datos, incluyendo el filtrado y la generación de nubes artificiales para simular condiciones reales desfavorables. Se presentan las arquitecturas de redes neuronales propuestas, así como las técnicas de fusión de datos utilizadas, como la fusión temprana y la fusión intermedia.
- **Capítulo 4, Experimentos y Resultados**: En este capítulo se evalúan las arquitecturas propuestas mediante diversos experimentos. Se comparan los resultados de los modelos utilizando diferentes estrategias de fusión de datos y se analiza su desempeño bajo condiciones de imágenes distintas, proporcionando una comparativa detallada de las métricas de precisión.
- **Capítulo 5, Conclusiones y Trabajos Futuros**: Finalmente, se resumen los hallazgos más importantes del estudio, destacando las contribuciones del trabajo a la mejora de la segmentación de imágenes satelitales mediante la fusión multimodal. También se proponen posibles direcciones para investigaciones futuras que podrían ampliar y mejorar los resultados obtenidos

CAPÍTULO 2
Estado del arte

En este segundo capítulo se presenta el Estado del Arte en este ámbito, en el cual podemos encontrar diversas aproximaciones o métodos de segmentación, todos basados en *Deep Learning* y en Fully Convolutional Networks (FCN) [22]. En una primera sección se revisarán métodos utilizados en segmentación semántica en general y luego veremos como estos se pueden modificar o adaptar para la segmentación de agua. Cabe destacar que todos estos métodos suelen utilizar un único tipo sensor, normalmente el Sentinel-2 o similares.

2.1 Segmentación semántica

2.1.1. U-Net

U-Net es una arquitectura de red neuronal profunda que ha demostrado ser altamente efectiva para tareas de segmentación semántica. Fue propuesta inicialmente por Ronneberger, Fischer y Brox en 2015 [19] para segmentación de imágenes biomédicas, pero su diseño ha sido adaptado con éxito para una amplia gama de aplicaciones, incluyendo la segmentación de agua en imágenes satelitales.

La U-Net se caracteriza por una estructura en forma de "U" (Figura 2.1), que consiste en dos partes principales: el *encoder* (contracción) y el *decoder* (expansión). El *encoder* está compuesto por una serie de capas de convolución y pooling que reducen progresivamente la resolución espacial de la imagen de entrada mientras aumentan la profundidad de las características aprendidas.

El *decoder*, por otro lado, está compuesto por capas de convolución y upsampling, estas últimas se encargan de incrementan la resolución espacial de la representación aprendida, normalmente al doble (*Up-conv 2x2*). Esto se puede implementar de distintas formas, en nuestro caso, utilizamos capas de *Upsampling2d* de interpolación bilineal con un factor de escalado de 2, seguido de una convolución 3x3 tradicional. De igual forma se pueden utilizar convoluciones traspuestas para hacer el escalado y la convolución en una única capa, no hay evidencia de que unas implementaciones funcionen mejor que otras. Otra característica distintiva de U-Net es el uso de conexiones de skip entre las capas correspondientes del *encoder* y el *decoder*. Estas conexiones permiten que la información de alta resolución de las capas tempranas del *encoder* se combine con las características de alta abstracción en el *decoder*, lo que mejora significativamente la precisión de la segmentación.

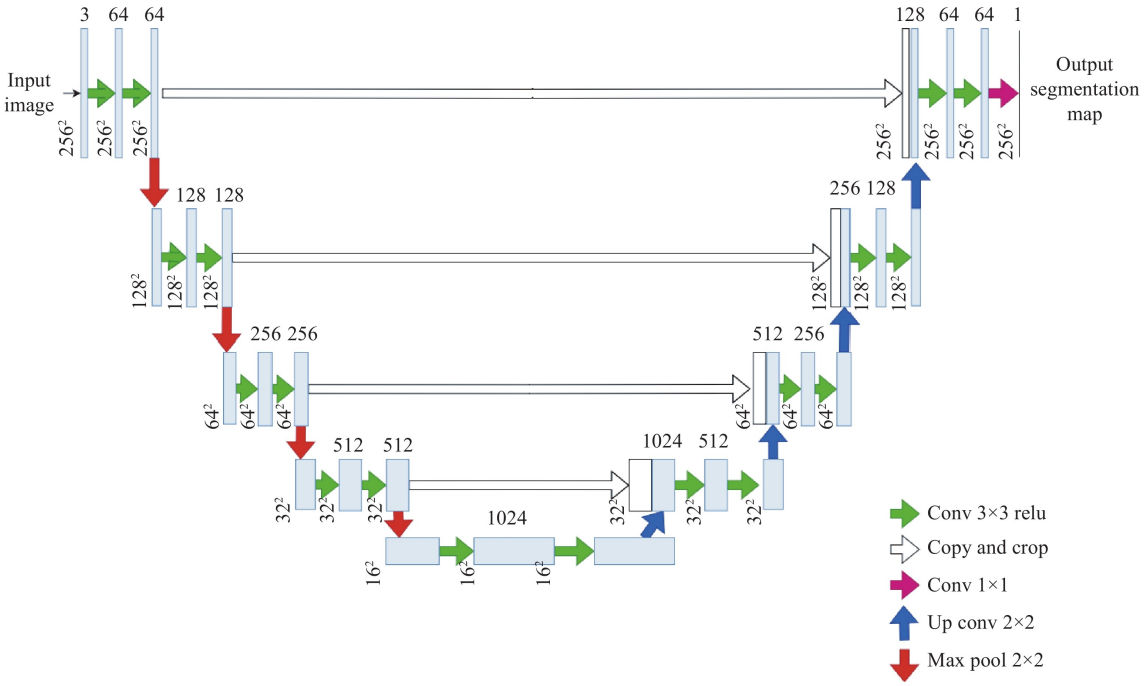

Figura 2.1: Arquitectura U-Net

Normalmente después de cada convolución se aplica una normalización por lotes (*Batch Normalization*) [9], seguido de una capa de activación no lineal con la función ReLU. Este proceso permite a la red capturar información contextual de alto nivel.

La función de pérdida utilizada en U-Net suele ser una combinación de cross-entropy y otras métricas, como el coeficiente de Dice, que son adecuadas para evaluar la precisión en la segmentación de objetos.

El uso de U-Net en la segmentación de agua ha demostrado ser efectivo [5], permitiendo identificar con precisión los cuerpos de agua en diversas condiciones y resoluciones espaciales. Además, la capacidad de U-Net para combinar información de diferentes resoluciones mediante las conexiones de skip resulta particularmente útil en el contexto de imágenes satelitales, donde las características de interés pueden variar considerablemente en tamaño y forma.

En resumen, U-Net representa una herramienta poderosa y versátil para la segmentación semántica, y sus adaptaciones específicas han permitido su aplicación exitosa en la identificación de cuerpos de agua, proporcionando resultados precisos y confiables en el análisis de imágenes satelitales.

2.1.2. ResNet como *encoder*

Para mejorar la capacidad de extracción de características de U-Net, una estrategia efectiva es sustituir la parte del *encoder* con una arquitectura de red neural convolucional pre-entrenada, como la ResNet (*Residual Network*) [7]. ResNet, propuesta por He, Zhang, Ren y Sun en 2015, introdujo el concepto de *residual learning* para facilitar el entrenamiento de redes neuronales muy profundas. La idea principal detrás de ResNet es el uso de bloques residuales, que permiten que las capas aprendan una función residual con respecto a las entradas originales. Esto aborda el problema del desvanecimiento del gradiente en redes muy profundas, el cual al calcularse de atrás hacia delante, no consigue llegar a las primeras capas de la red.

ResNet está compuesta por múltiples bloques residuales (Figura 2.2). Cada bloque residual incluye una o más capas de convolución y una conexión de atajo (skip connection) que salta una o más capas y luego se suma al resultado de estas capas. Esta arquitectura permite entrenar redes mucho más profundas sin los problemas de *vanishing gradients* y la degradación del rendimiento.

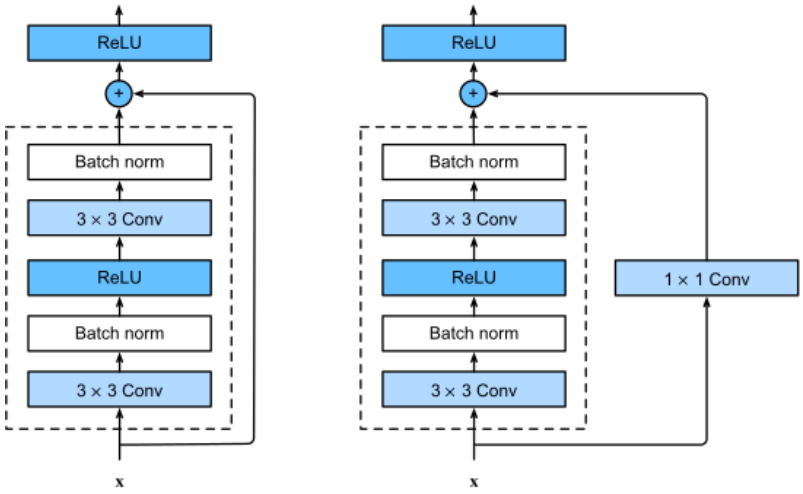

Figura 2.2: Bloque Residual

Los modelos ResNet se presentan en varias profundidades, como ResNet-18, ResNet-34, ResNet-50, ResNet-101 y ResNet-152, donde el número indica la cantidad de capas.

Una práctica común al utilizar ResNet como *encoder* en U-Net es emplear una versión pre-entrenada en el con *ImageNet*. *ImageNet* es un gran conjunto de datos que contiene millones de imágenes etiquetadas en miles de categorías, lo que permite que las redes pre-entrenadas aprendan características de bajo y alto nivel que son útiles para una amplia gama de tareas de visión por computadora.

Al incorporar una ResNet pre-entrenada como *encoder*, se obtienen varios beneficios:

- **Mejor extracción de características**: En la U-Net original, la arquitectura del *encoder*, encargada de la extracción de características, consiste en convoluciones simples colocadas una detrás de otra. Al usar una ResNet, la capacidad de extracción de características aumenta notablemente, ya que las ResNet obtienen mucho mejores resultados en la clasificación de *ImageNet* en comparación con arquitecturas planas. Esto se debe a su diseño de bloques residuales, que permiten una mejor representación y captura de características complejas en las imágenes.

- **Transfer Learning**: El uso de pesos pre-entrenados permite aprovechar características aprendidas de *ImageNet*, facilitando la extracción de patrones robustos desde las primeras etapas del entrenamiento.

- **Convergencia más rápida**: Dado que la red ya ha aprendido a extraer características útiles, el modelo converge más rápidamente comparado con entrenar una U-Net desde cero.

- **Mejor desempeño con menos datos**: Utilizar una ResNet pre-entrenada puede ser especialmente ventajoso cuando se dispone de un conjunto de datos limitado para la tarea específica de segmentación de agua, ya que la red ya tiene un conocimiento previo significativo.

En conclusión, reemplazar el *encoder* de U-Net con una ResNet pre-entrenada proporciona una mejora significativa en la extracción de características, acelerando el entrenamiento y mejorando el desempeño general del modelo en la tarea de segmentación semántica.

2.2 Segmentación de agua

2.2.1. DeepWaterMapV2

DeepWaterMapV2 [11] es una arquitectura de red neuronal completamente convolucional (FCN) propuesta por Isikdogan, Bovik y Passalacqua (2019) para la segmentación del agua en imágenes de satélite. Este modelo es un intento de abordar el problema de la cobertura de nubes en las imágenes de satélite ópticos.

La arquitectura de DeepWaterMapV2 (Figura 2.3) consiste en una estructura *encoder-decoder* con conexiones residuales. El *encoder* sigue una arquitectura de red residual (ResNet) [7], que incluye múltiples bloques de downscaling. Cada unidad de downscaling consta de una convolucion de kernel 5x5 con stride 2, reduciendo el tamaño de la imagen a la mitad, seguido de una convolución 3x3 y una conexion residual.

Al final del *encoder* y antes del *decoder* la unidad de bottleneck consiste de 2 convoluciones 3x3 con conexión residual, este bloque no cambia la resolución espacial de la imagen.

En el *decoder* se añaden conexiones de salto entre las capas *encoder* para preservar la información espacial y mejorar los resultados de la segmentación. Estos mapas de características se suman en la upscaling unit y utilizando una transformación depth-to-space se re-organizan los píxeles para pasar de *WxHxC* a *2Wx2HxC/4*. Luego esta transformación es seguida de dos convoluciones 3x3 con conexión residual.

Todas las capas convolucionales van seguidas de *Batch Normalization* y una activación ReLU a excepción de la primera y última capa. La capa de entrada es una capa convolucional 1 × 1 que actúa como una capa de compresión lineal por canales. La capa de salida también es una capa convolucional de 1 × 1, pero tiene una activación sigmoide que emite los valores que corresponden a las probabilidad de que cada píxel de la entrada sea un píxel de agua.

Realmente, esta arquitectura es similar a combinar la U-Net [19] mencionada anteriormente sustituyendo la parte *encoder* por una ResNet [7].

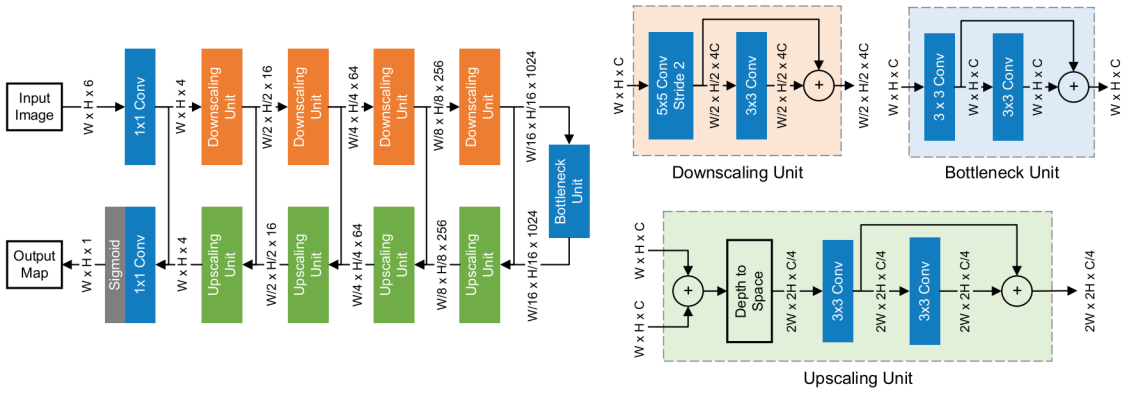

Figura 2.3: Arquitectura DeepWaterMapV2

La entrada del modelo DeepWaterMapV2 es una imagen de 6 canales. Para el entrenamiento de la red se han utilizado imágenes del Landsat-7, pero realmente no requiere que las imágenes de entrada sean Landsat-7, Landsat-8 o Sentinel-2. Solo se espera que las bandas o canales de entrada coincidan aproximadamente con las bandas Landsat:

- B2: Blue
- B3: Green
- B4: Red
- B5: Near Infrared (NIR)
- B6: Shortwave Infrared 1 (SWIR1)
- B7: Shortwave Infrared 2 (SWIR2)

La salida es una máscara binaria que indica la presencia de agua en la imagen. El modelo se entrena usando una combinación de max-pooled adaptative loss [10], que sirve para centrar la atención en esos píxeles con la pérdida más alta usando su relación espacial y la focal loss [14] ayuda a mitigar el desbalanceo de las clases fijándose en los ejemplos más difíciles.

Se ha demostrado que DeepWaterMapV2 ofrece resultados de estado de arte en segmentación de agua en varios conjuntos de datos de imágenes de satélite, como Landsat y Sentinel-2.

En resumen, DeepWaterMapV2 es una potente arquitectura CNN para la segmentación de agua en imágenes de satélite. Su estructura *encoder-decoder* con conexiones de salto y conexiones residuales con bottlenecks permiten al modelo captar características tanto locales como globales, mientras que su función de pérdida equilibra el desequilibrio de clases entre los píxeles de agua y los que no lo son.

2.2.2. WatNet

El modelo WatNet [15], propuesto por Luo, Tong y Hu en 2021, al igual que DeepWaterMapV2, es una arquitectura completamente convolucional *encoder-decoder* para la segmentación automática de aguas superficiales a partir de imágenes de satélite multiespectrales.

En la Figura 2.4 podemos observar que el *encoder* en este caso es una red convolucional pre-entrenada, la MobileNetV2 [20], esta es una red neuronal convolucional diseñada para dispositivos móviles y aplicaciones con recursos limitados.

MobileNetV2 utiliza bloques residuales invertidos, los cuales difieren de los tradicionales en que primero expanden el número de canales, aplican convoluciones de profundidad, y luego reducen los canales de nuevo (Ver Figura 2.5). Cada bloque consta de tres capas principales: una convolución 1x1 para expandir la dimensión de características, una convolución depthwise separable 3x3 para capturar las relaciones espaciales, y otra convolución 1x1 para reducir la dimensión de características. Además, incorpora conexiones residuales al igual que la ResNet, haciendo así posible el entrenamiento de la redes muy profundas.

Esta arquitectura optimiza el balance entre precisión y eficiencia computacional, haciendo posible ejecutar modelos de visión por computadora en dispositivos con capacidad de procesamiento limitada u obtener una gran velocidad de procesamiento, esto

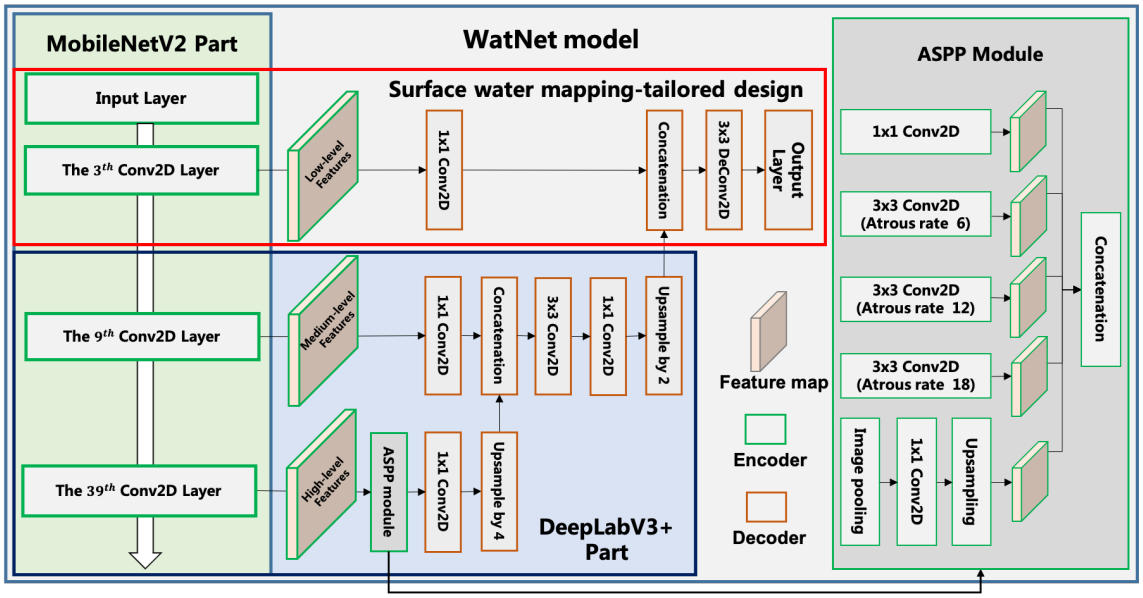

Figura 2.4: WatNet

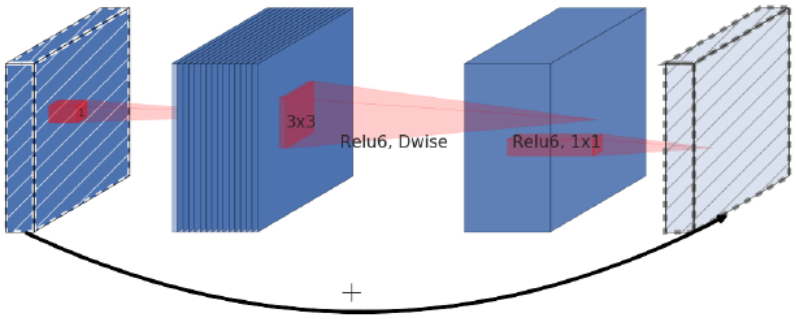

Figura 2.5: Bloque MobileNetV2

junto a la elección del *decoder* que se explicará a continuación hacen que esta arquitectura sea muy rápida y liviana.

En el *decoder* se encuentra una versión de DeepLabV3+ [1] ligeramente modificada. DeepLabV3+ (Ver Figura 2.6) es una arquitectura o método para segmentación de imágenes que utiliza convoluciones dilatadas para capturar información contextual en múltiples escalas, lo que permite segmentar objetos de diferentes tamaños con precisión. El *encoder*, en este caso MobileNetV2, desempeña un papel crucial extrayendo características de alto nivel. Estas características son posteriormente procesadas por el módulo de Atrous Spatial Pyramid Pooling (ASPP), el cual utiliza múltiples convoluciones dilatadas con diferentes tasas de dilatación, esto permite capturar características en diversas escalas.

A continuación, el módulo de decodificación refina estas características mediante la combinación de información de resoluciones altas con características de nivel intermedio (mapa de la capa 9 de MobileNetV2) y resoluciones bajas con características de alto nivel (mapa de la capa 39) permitiendo generar mapas de segmentación más precisos y detallados. En el caso de WatNet se añade un mapa adicional de bajo nivel y muy alta resolución espacial, el de la capa 3, para aumentar aún mas la precisión en la segmentación. Esta estructura permite alcanzar un rendimiento similar a la U-Net en segmentación semántica pero de forma más eficiente, manteniendo un equilibrio entre precisión y eficiencia computacional.

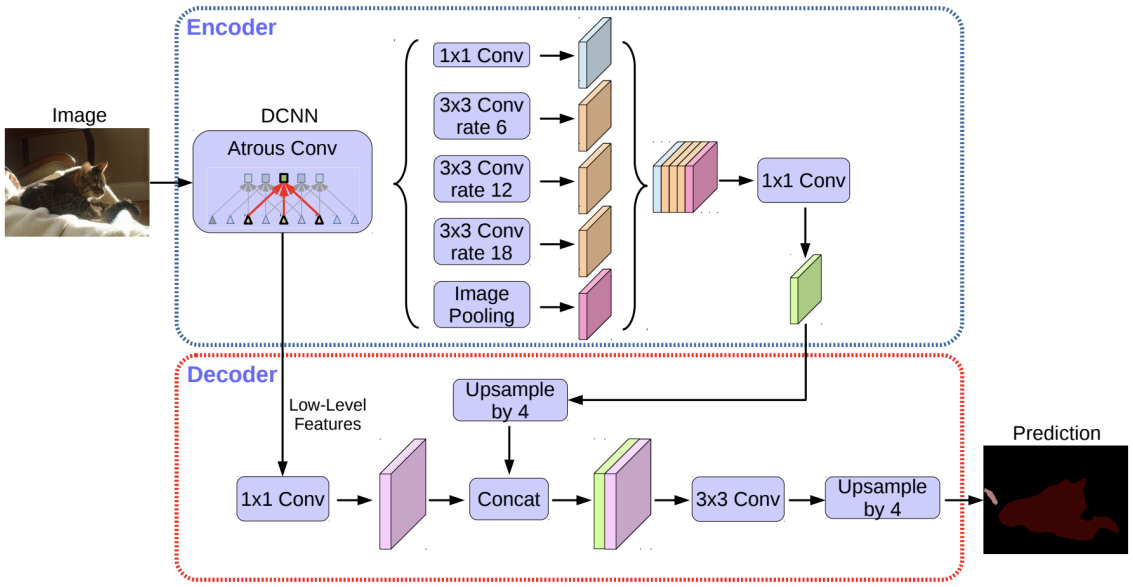

Figura 2.6: Arquitectura DeepLabV3+

Al igual que DeepWaterMapV2, en la capa de salida se debe aplicar una activación sigmoide para emitir los valores que corresponden a la probabilidad de agua en cada píxel. La entrada de este modelo, también es una imágen de 6 canales, y ha sido entrenado con imágenes de Sentinel-2 con las siguientes bandas, las cuales son las mismas que en DeepWaterMapV2, pero utilizando el satelite Landsat:

- B2: Blue
- B3: Green
- B4: Red
- B8: Near Infrared (NIR)
- B11: Shortwave Infrared 1 (SWIR1)
- B12: Shortwave Infrared 2 (SWIR2)

La salida al igual que el modelo anterior es una máscara binaria que indica la probabilidad de agua en cada píxel de la imagen. La red se entrena utilizado la entropía binaria como función de pérdida para cada píxel de cada imagen del *batch*.

En resumen, WatNet se presenta como una solución eficaz y eficiente para la segmentación de aguas superficiales en imágenes satelitales multiespectrales, superando en ocasiones a DeepWaterMapV2 en precisión. Integrando la MobileNetV2 como *encoder* y una versión modificada de DeepLabV3+ como *decoder*, esta arquitectura logra un balance óptimo entre precisión y eficiencia computacional obteniendo mapas de segmentación detallados y precisos.

Además, la inclusión de un mapa de características adicional de alta resolución espacial contribuye a mejorar aún más la exactitud de la segmentación. Este enfoque demuestra ser altamente adecuado para su implementación en dispositivos con recursos limitados, manteniendo una alta velocidad de procesamiento sin sacrificar la calidad de los resultados.

2.3 Resumen estado del arte

En este capítulo, se ha presentado un estado del arte en el ámbito de la segmentación semántica y la segmentación de agua utilizando arquitecturas de redes neuronales profundas, específicamente basadas en Fully Convolutional Networks (FCN).

Se han revisado diversas aproximaciones y métodos, todas las arquitecturas revisadas se centran en arquitecturas convolucionales *encoder-decoder* similares a las usadas en segmentación semántica general.

Sin embargo, todas ellas parecen usar un solo tipo de sensor, lo que nos puede servir para sembrar un modelo de referencia (*Baseline*) e intentar mejorarlo.

La utilización de estas arquitecturas ha demostrado ser efectiva en la segmentación de agua en imágenes satelitales, y su adaptabilidad en distintas tareas y precisión las convierten en herramientas valiosas para futuras investigaciones y aplicaciones en este campo.

CAPÍTULO 3
Metodología

En este capítulo se presenta la Metodología, donde se describe de forma detallada el procedimiento llevado a cabo para realizar el trabajo.

3.1 Preparación de los datos

Como hemos mencionado anteriormente, el conjunto de datos utilizado proviene de la competición 2020 IEEE GRSS DATA FUSION CONTEST [6]. Este conjunto de datos consta de 6114 imágenes para entrenamiento y 986 imágenes para validación. Dado que, al ser una competición el conjunto de prueba es desconocido, evaluaremos nuestros resultados en el conjunto de validación, asumiendo que ha sido cuidadosamente seleccionado para representar la población de datos.

3.1.1. Filtrado de imágenes

En primer lugar, se aplicó un filtrado a las imágenes del conjunto de datos. Cada píxel en las imágenes puede pertenecer a 10 clases diferentes, correspondientes a la cobertura terrestre. Se eliminaron las imágenes que no contenían la clase *agua* y se redujeron las demás clases a una máscara binaria de 1 y 0, donde 1 indica la presencia de agua y 0 indica su ausencia. Este filtrado resultó en 4508 muestras en el conjunto de entrenamiento y 799 en el conjunto de validación.

3.1.2. Generación nubes artificiales

Para simular un entorno más realista, se agregaron nubes artificiales a las imágenes utilizando la herramienta Satellite Cloud Generator (SCG) [2]. Esta herramienta permite crear nubes y sombras artificiales en imágenes mediante ruido estructural (ruido Perlin). Debido a la complejidad computacional del algoritmo, no fue posible agregar nubes de forma aleatoria durante el entrenamiento. En su lugar, se agregaron nubes a todas las imágenes del conjunto de entrenamiento y validación, obteniendo pares de imágenes limpias y con nubes para el entrenamiento y la evaluación. Este proceso solo se aplicó a imágenes del Sentinel-2, ya que las imágenes del Sentinel-1, al ser de radar, no se ven afectadas por las nubes.

La herramienta SCG puede generar diferentes tipos de nubes utilizando distintos hiperparámetros. Afortunadamente, hay configuraciones preestablecidas para los tipos de nubes más comunes (Ver Figura 3.1). La herramienta puede actuar sobre todas las bandas de las imágenes del Sentinel-2, excepto la banda 10, que se recomienda eliminar ya que

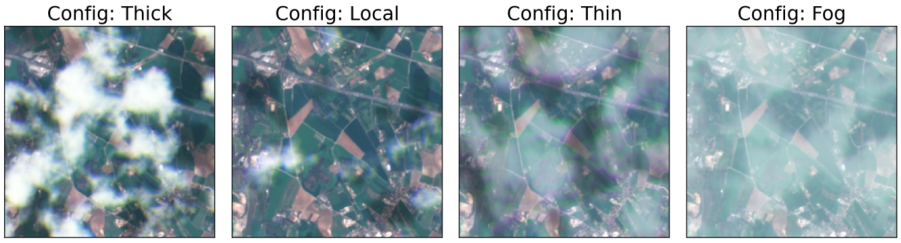

Figura 3.1: Nubes generadas con SCG

responde principalmente a los reflejos de los cirros en la parte superior de la atmósfera y tiene una estructura muy distinta a las otras bandas, como podemos ver en las Figuras 3.2 y 3.3.

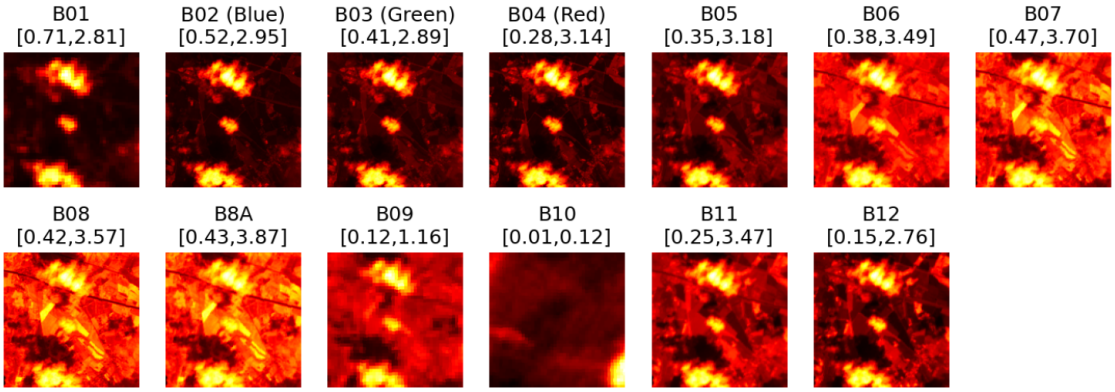

Figura 3.2: Ejemplo de contenido de bandas multiespectrales individuales en cada canal para una imagen nublada de Sentinel-2.

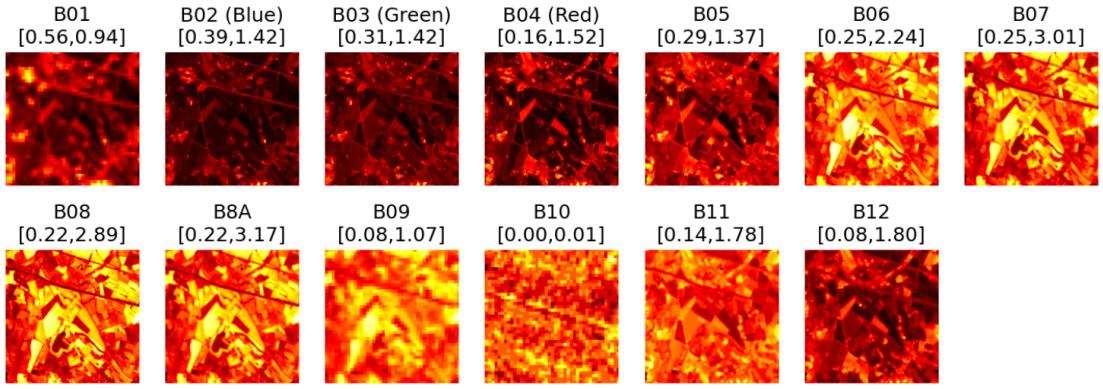

Figura 3.3: Ejemplo de contenido de bandas multiespectrales individuales en cada canal para una imagen Sentinel-2 sin nubes.

Este efecto no está modelizado actualmente por el simulador de nubes y de ahí la exclusión. Las dos figuras contienen visualizaciones de bandas individuales del Sentinel-2 para una muestra nublada y otra despejada. En ambos casos, todas las bandas excepto la Banda-10 (SWIR-Cirrus) parecen estar altamente correlacionadas. En la imagen nublada, las bandas tienden a contener una presencia similar de la nube, mientras que en la Banda 10 este objeto parece ausente. Del mismo modo, la Banda 10 en la imagen clara parece detectar una estructura bastante diferente a la de las otras bandas.

Aun así el artículo demuestra que las nubes generadas artificialmente tienen un impacto similar al de las nubes reales en modelos neuronales. A continuación en las Figuras

3.4, 3.5 y 3.6, podemos observar ejemplos de nubes creadas en nuestro conjuto de datos con distintas configuraciones del generador de nubes.

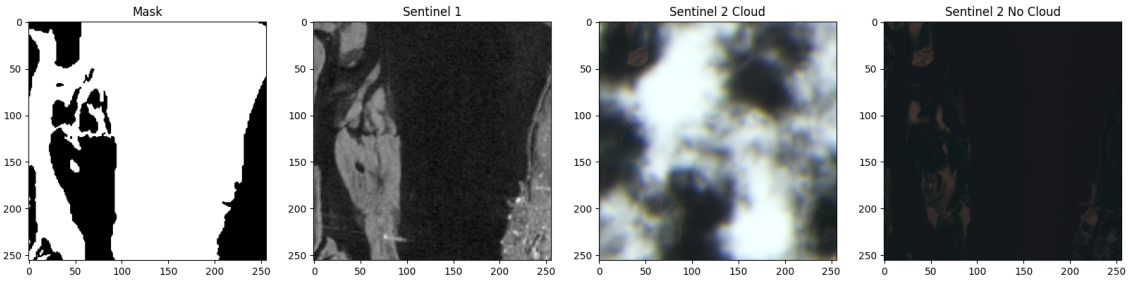

Figura 3.4: Ejemplo de nubes grandes

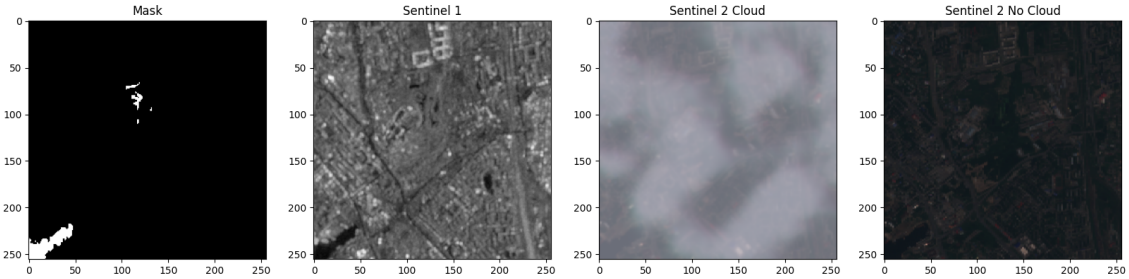

Figura 3.5: Ejemplo de nubes tipo neblina

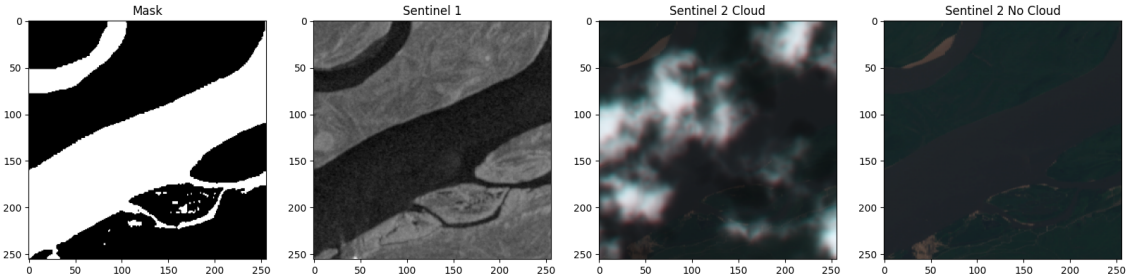

Figura 3.6: Ejemplo de nubes pequeñas

3.2 Arquitecturas propuestas

En esta sección explicamos las distintas arquitecturas propuestas así como las distintas técnicas de fusión de datos utilizadas en cada una de ellas.

3.2.1. *Baseline* un solo sensor

En primer lugar, nos propusimos evaluar el desempeño de arquitecturas similares a las mencionadas en el estado del arte utilizando nuestros propios datos. Para ello, implementamos las dos redes referenciadas en el estado del arte: DeepWaterMapV2 y WatNet. Posteriormente, compararemos sus resultados, y la red que obtenga los mejores se establecerá como nuestro modelo *Baseline*, el cual servirá como base para los experimentos posteriores.

La primera arquitectura neuronal que implementaremos como posible modelo *Baseline* será muy similar a la anteriormente mencionada DeepWaterMapV2. Dado que no

tenemos un conjunto de datos muy grande, utilizaremos un *encoder* pre-entrenado con Imagenet [4]. La arquitectura elegida para este es la ResNet-50, ya que nos brinda un buen balance entre velocidad computacional y precisión. El *decoder* será una U-Net, que al igual que DeepWaterMapV2 contiene conexiones de salto entre las distintas partes del *decoder* y *encoder*, permitiendo así una segmentación fina y precisa.

En la Figura 3.7 podemos ver un diagrama de esta arquitectura.

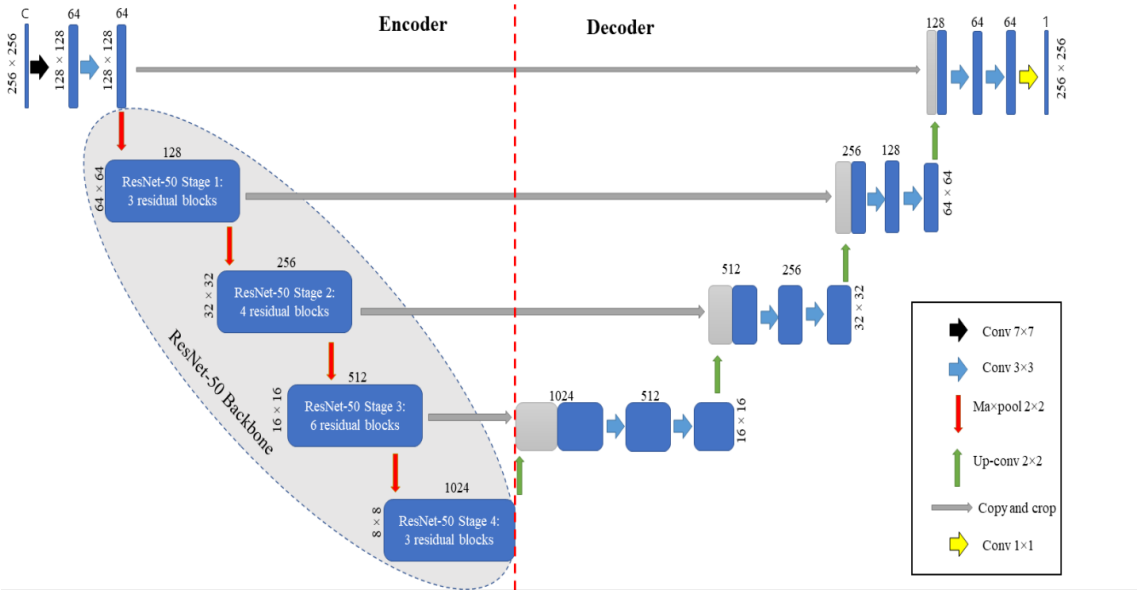

Figura 3.7: Diagrama de una U-Net con Res-Net50 como *encoder* [17]

La segunda opción como modelo de referencia será WatNet. En este caso, se ha seguido fielmente la arquitectura original presentada en la Figura 2.4. El *encoder* utilizado es MobileNetV2, y se ha añadido la extracción de características a alta resolución.

Las arquitecturas neuronales mencionadas anteriormente utilizan imágenes ópticas (Sentinel-2, Landsat) con distintas bandas para realizar la segmentación. Por lo tanto, la elección del modelo de referencia se llevará a cabo de esta forma, utilizando todas las capas del Sentinel-2.

Posteriormente, una vez elegido el modelo de referencia, se comparará su rendimiento con distintos sensores y bandas. Aquí buscamos comprobar el impacto, las ventajas y las desventajas de utilizar imágenes ópticas frente a imágenes de radar, como las del Sentinel-1.

3.2.2. *Single Encoder U-Net* y *Early Fusion*

En esta sección se describe la arquitectura utilizada para la fusión temprana (*Early Fusion*) de las imágenes de Sentinel-1 y Sentinel-2. La fusión temprana implica combinar los canales de las imágenes de radar y ópticas antes de ingresar a la red neuronal.

La arquitectura de la red será exactamente igual que la mencionada anteriormente en la *Baseline*, siendo una red tipo *U-Net*, sustituyendo la parte del *encoder* por una ResNet-50 pre-entrenada con Imagenet.

En este caso, se modifica para aceptar una entrada combinada de los 2 canales de Sentinel-1 y los 12 canales de Sentinel-2, resultando en una entrada de 14 canales. Esto se logra concatenando los tensores tridimensionales en la dimensión canal, resultando en una matriz de 14x256x256. Esta matriz ingresa directamente a la arquitectura U-Net, que procesa la información combinada de manera integral y se espera que sea capaz de

captar las relaciones y características específicas entre las distintas bandas y obtener una segmentación más precisa. En la Figura 3.8 podemos ver un diagrama de estar fusión y la arquitectura

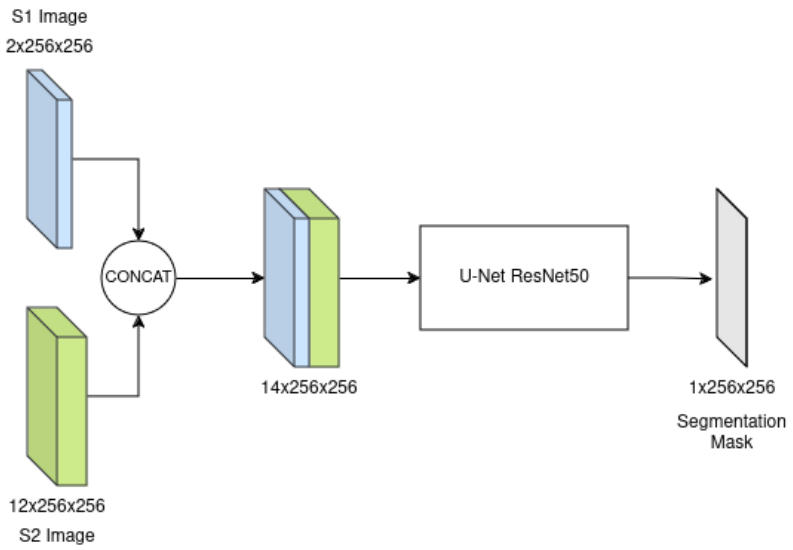

Figura 3.8: Diagrama *Early Fusion*

Las ventajas de la fusión temprana frente a la fusión intermedia o la tardía son que al combinar los canales temprano, se reduce la complejidad del modelo, ya que no se necesitan múltiples ramas de codificadores separados como veremos en la siguiente sección. De la misma forma, es más eficiente en términos de recursos computacionales, ya que se procesa una sola entrada combinada en lugar de múltiples entradas separadas.

Sin embargo, al combinar toda la información al principio, la red puede tener dificultades para capturar características específicas de cada tipo de imagen en las etapas más tempranas de la red.

La fusión temprana puede ser una opción viable y vale la pena evaluar su rendimiento para combinar imágenes de radar y ópticas, aunque puede tener limitaciones a la hora de capturar correctamente las características específicas de cada tipo de imagen.

3.2.3. *Dual Encoder U-Net* y *Intermediate Fusion*

La siguiente arquitectura que proponemos se basa en procesar las imágenes de S1 y S2 por separado en *encoder* distintos para luego realizar una fusión de las características. A continuación se describen en detalle los componentes de esta arquitectura, la cual podemos observar en la Figura 3.9.

En el *Encoder*, el modelo utiliza dos ramas separadas de codificadores basados en ResNet-50, una para cada tipo de imagen, Sentinel-1 con 2 canales y Sentinel-2 con 12 canales, cada rama procesa las imágenes de su respectivo satélite. Como arquitectura de ambos *encoder* decidimos seguir con ResNet-50.

Las características extraídas por los codificadores ResNet-50 se fusionan en varios puntos a lo largo de la red. Está fusión se realiza para combinar la información relevante de ambas fuentes de imagen (radar y óptica), permitiendo que el modelo aproveche las fortalezas complementarias de los diferentes tipos de datos. Las fusiones (F1 a F5) deben asegurar que el modelo pueda integrar información tanto de la estructura superficial como de la textura, lo que es crucial para la identificación precisa de cuerpos de agua.

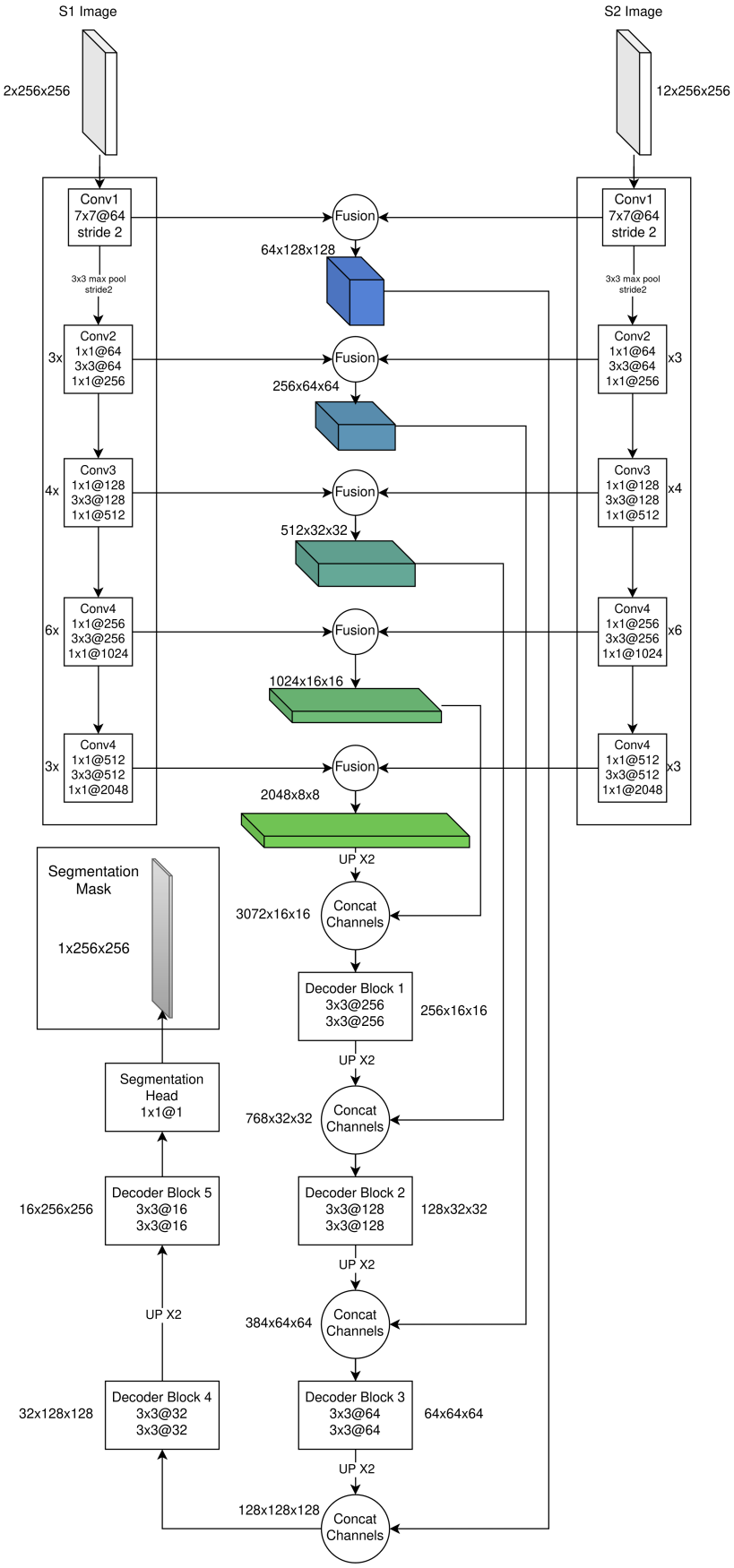

Figura 3.9: Arquitectura *Dual Encoder U-Net* y *Intermediate Fusion*

Por lo tanto, es crucial utilizar una función de fusión adecuada, en nuestro caso, decidimos utilizar una fusión modificada a nivel de canal propuesta por Yimian Dai et al. en Attentional Feature Fusion [3].

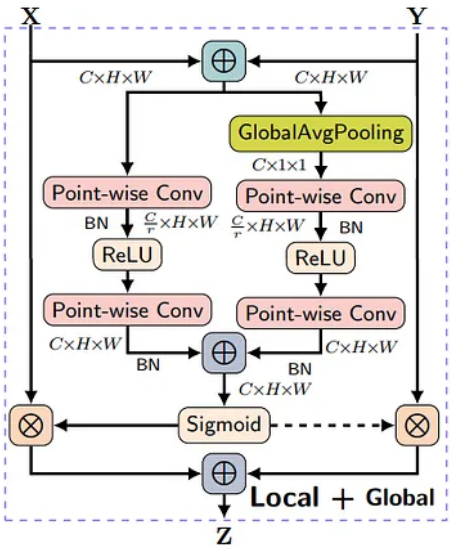

Figura 3.10: Función de fusión a nivel canal

En la Figura 3.10 podemos observar la función de fusión que combina características globales y locas de dos entradas X e Y (mapas de características de los *encoder* a distinto nivel) para producir un mapa fusionado Z.

1. **Suma de las entradas**: En primer lugar los mapas X e Y, los cuales son del mismo tamaño, se suman y se divide en 2 ramas, una va a procesar las características locales (izquierda) mientras que otra se va a encargar de las globales (derecha).

2. **Procesado de las características locales**: Para procesar las características a nivel de canal primero se aplica una convolución puntual de 1x1 con un ratio de reducción *r*, la idea de esta reducción proviene del artículo de Squeeze-and-Excitation Networks [8], este parámetro se encarga de mejorar la representación capturando las características más relevantes y de mejorar la eficiencia reduciendo el número de operaciones necesarias, un buen valor teniendo en cuenta precisión y eficiencia suele ser 16. Esta transformación va seguida de una operación de *Batch Normalization* y una activación *ReLU*. Finalmente otra convolución de 1x1 es aplicada volviendo al número de canales original seguido de una *Batch Normalization* final.

3. **Procesado de las características globales**: El procesado de las características globales es exactamente igual que el de las locales, con la única diferencia que antes de enviar el mapa por las transformaciones se pasa por una capa de *GlobalAvgPooling*, la cual reduce la dimensionalidad espacial de HxW a 1x1, centrándose solo en el promedio global de cada canal, lo que ayuda a capturar características contextuales globales.

4. **Coeficientes de atención**: Finalmente los mapas locales y globales preprocesado se vuelven a sumar, el tensor de Cx1x1 se replica para que coincida con las dimensiones del tensor CxHxW y se puedan sumar. A continuación se aplica una activación sigmoide a las características refinadas para escalarlas entre 0 y 1. Estas actúan como coeficientes de atención que ponderan las características locales originales.

5. **Fusión final**: Finalmente las características originales son multiplicadas elemento a elemento por los coeficiente de atención obtenidos (*AttScores*), permitiendo que

ciertas características sean realzadas mientras que otras se atenúan. En el caso de X se multiplica por los *AttScores* e Y se multiplica por (1 - *AttScores*) como indica la linea discontinua. El resultado de esta multiplicación se suma y se obtiene el resultado final Z.

El resultado final Z es una combinación de las características locales y globales, lo que puede mejorar la capacidad del modelo para capturar tanto detalles finos como contextos más amplios en tareas como la clasificación y la segmentación de imágenes.

3.2.4. *Dual U-Net* y *Late Fusion*

La última arquitectura propuesta, denominada *Dual U-Net* con *Late Fusion*, se centra en procesar las imágenes de los sensores Sentinel-1 y Sentinel-2 mediante redes U-Net separadas, fusionando los resultados en una etapa tardía para la segmentación final. Esta técnica busca maximizar las capacidades específicas de cada sensor, permitiendo que cada red U-Net se especialice en la representación de sus respectivas entradas antes de la combinación final. La Figura 3.11 muestra un diagrama de esta arquitectura.

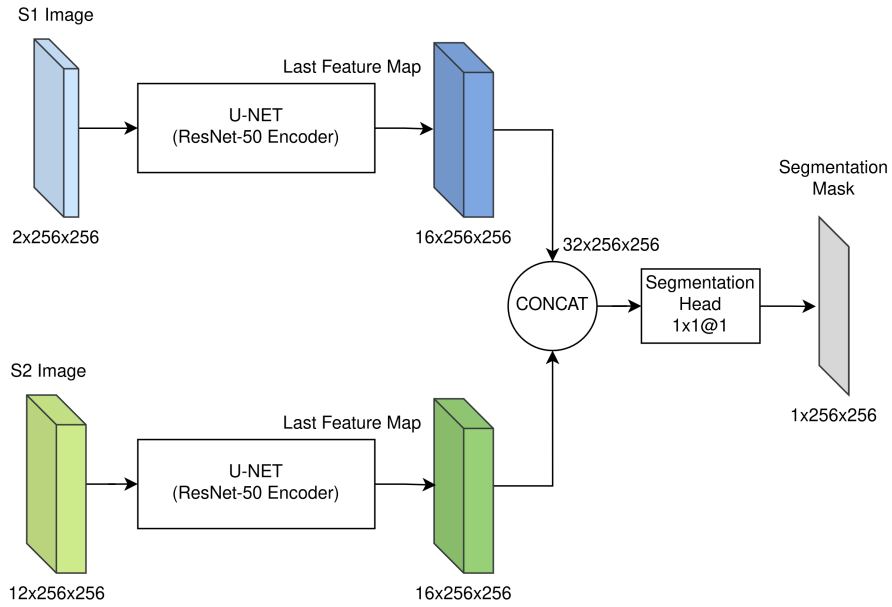

Figura 3.11: Diagrama *Late Fusion*

El modelo consta de dos redes U-Net independientes: una dedicada al procesamiento de las imágenes de Sentinel-1 y otra a las imágenes de Sentinel-2. Similar a las arquitecturas anteriores, cada U-Net se compone de un *encoder* basado en ResNet-50 pre-entrenado con Imagenet, seguido por un *decoder* que reconstruye las características de entrada para generar un mapa de segmentación de tamaño completo.

La fusión tardía se realiza después del procesamiento completo de cada rama U-Net. Las salidas de ambas redes, que representan mapas de características a nivel de píxel, se concatenan para combinar la información proveniente de ambos sensores. Posteriormente, esta combinación pasa por una cabeza de segmentación que consiste en una capa convolucional de 1x1, cuyo objetivo es reducir la dimensionalidad y producir el mapa de segmentación final. Se espera que esta capa ajuste los pesos para integrar las características combinadas y genere una predicción más precisa, gracias a la utilización diferenciada y especializada de los datos de cada sensor.

Este enfoque permite que cada red capture de manera óptima y especializada las características específicas de sus respectivas entradas antes de la fusión, maximizando así el aprovechamiento de las fortalezas inherentes de los datos de radar y ópticos. Sin embargo, es también la arquitectura más compleja, ya que esencialmente consiste en la implementación de dos modelos de segmentación por separado.

3.3 Métricas de evaluación

Para evaluar el rendimiento de nuestra red neuronal en la tarea de segmentación de agua, se utilizaron varias métricas comúnmente empleadas en problemas de segmentación de imágenes. A continuación, se describen las métricas utilizadas:

La Accuracy es una métrica que evalúa la proporción de píxeles correctamente clasificados. Se calcula como:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{3.1}$$

donde TN es el número de píxeles verdaderos negativos (píxeles que no pertenecen a la clase y no son clasificados como si lo hicieran), TP es el número de píxeles verdaderos positivos (píxeles que pertenecen a la clase y son correctamente clasificados), FP es el número de píxeles falsos positivos (píxeles que no pertenecen a la clase pero son clasificados como si lo hicieran) y FN es el número de píxeles falsos negativos (píxeles que pertenecen a la clase pero no son clasificados como si lo hicieran).

Un valor de Accuracy cercano a 1 indica una buena precisión en la clasificación de píxeles. Aunque esta métrica sea bastante utilizada en este ámbito puede no ser la más adecuada ya que es muy susceptible al desbalanceo de las clases, ya que puede ser engañosa ya que una clasificación que siempre predice la clase mayoritaria puede tener una Accuracy alta, pero no ser muy precisa en la práctica. Por lo tanto, utilizaremos otras métricas como F1 score o la *Mean IoU*.

El F1 score es una métrica que combina la *Precision* y el *Recall* para evaluar el rendimiento de la segmentación. Se calcula como:

$$F1Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3.2}$$

donde *Precision* es el número de píxeles verdaderos positivos dividido por el número total de píxeles positivos (TP + FP), y *Recall* es el número de píxeles verdaderos positivos dividido por el número total de píxeles que pertenecen a la clase (TP + FN). Un valor de F1 score cercano a 1 indica una buena precisión y *Recall* en la segmentación.

La métrica *Mean IoU* se utiliza para evaluar la precisión de la segmentación. Se calcula como la media de la intersección sobre la unión (IoU) entre la predicción y la etiqueta de verdad para cada imagen. La IoU se calcula como:

$$IoU = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3.3}$$

Las métricas se calculan para cada imagen individualmente (con los TP, FP, TN y FN obtenidos para cada píxel) y posteriormente se hace la media de todos estos valores para obtener la evaluación de todo el conjunto de datos.

3.4 Condiciones de Entrenamiento

La reproducibilidad de los resultados en investigaciones de aprendizaje automático y procesamiento de imágenes es un aspecto fundamental para validar la eficacia y fiabilidad de los métodos propuestos. En este sentido, es esencial detallar las condiciones bajo las cuales se entrenaron los modelos utilizados en este estudio, incluyendo la descripción del hardware, los parámetros de entrenamiento, el número de epochs, el optimizador seleccionado, y otras configuraciones relevantes.

3.4.1. Hardware Utilizado

El entrenamiento de los modelos se realizó utilizando un equipo con las siguientes especificaciones:

- **Procesador**: 13th Gen Intel(R) Core(TM) i7-13z
- **Memoria RAM**: 16GB
- **Targeta Gráfica**: NVIDIA GeForce RTX 4060 Max-Q
- **Sistema Operativo**: Fedora Linux 40

3.4.2. Optimizador y Función de Pérdida

El optimizador utilizado para el entrenamiento del modelo fue el optimizador ADAM [13], que se ha demostrado eficaz en una variedad de tareas de aprendizaje profundo. Las hiperparámetros del optimizador ADAM se configuraron de la siguiente manera:

- **Tasa de Aprendizaje**: 0.0001
- **Beta 1**: 0.9
- **Beta 2**: 0.999

La funcion de péridida utilizada fue la Dice Loss [23], que es común en tareas de segmentación.

3.4.3. *Epochs* y *Bach Size*

Al utilizar modelos pre-entrenados y contar con una limitada cantidad de datos, no es necesario entrenar el modelo durante un gran número de épocas. En nuestro caso, 20 épocas parecieron ser suficientes. Esto se debe a que guardamos el modelo cuando obteníamos los mejores resultados de validación, y a partir de las épocas 11-13, comenzamos a observar sobreajuste en los datos de entrenamiento.

El tamaño de lote (*batch size*) elegido para el entrenamiento se determinó en función de los recursos computacionales disponibles. Para los modelos simples, se utilizó un tamaño de lote de 16, mientras que para los modelos más complejos, que incluyen dos encoders, se redujo a 8. Esta reducción fue necesaria para evitar sobrepasar la capacidad de la memoria de vídeo disponible, asegurando que el entrenamiento se pudiera realizar de manera eficiente y sin interrupciones.

CAPÍTULO 4
Experimentos y Resultados

En este capítulo se realizan los experimentos siguiendo la Metodología descrita anteriormente. Se entrenaran y evaluaran cada una de las arquitecturas propuestas anteriormente, variando el conjunto de entrenamiento y validación con imágenes con y sin nubes.

4.1 *Baseline*

4.1.1. Elección de modelo *Baseline*

Como hemos mencionado anteriormente, primero se realizó la elección modelo de referencia entre DeepWaterMapV2 modificado y WatNet. Los resultados obtenidos con ambas arquitecturas se muestran en la Tabla 4.1.

Tabla 4.1: Resultados eleccion *baseline*

Datos	Arquitectura	Accuracy	F1-Score	MeanIoU
S2	DeepWaterMapV2 (MOD)	**0.9962**	**0.7976**	**0.7521**
S2	WatNet	0.9942	0.7281	0.6852

DeepWaterMapV2 modificado obtuvo una *Accuracy* de 0.9962, un *F1-Score* de 0.7976 y un *MeanIoU* de 0.7521, lo cual indica un rendimiento superior tanto en precisión general como en la capacidad de segmentar correctamente los cuerpos de agua. En comparación, WatNet alcanzó una *Accuracy* de 0.9942, con un *F1-Score* de 0.7281 y un *MeanIoU* de 0.6852, valores significativamente menores que reflejan una menor capacidad para identificar correctamente los píxeles de agua.

DeepWaterMapV2 modificada claramente se destaca como la mejor opción para establecer el modelo *baseline*, gracias a sus métricas superiores en segmentación de agua en imágenes.

Sin embargo, al considerar el tiempo de procesamiento por imagen, la Figura 4.1 WatNet tiene una clara ventaja en términos de eficiencia computacional al estar utilizando una arquitectura más ligera. DeepWaterMapV2 modificado tarda aproximadamente 5 milisegundos por imagen, mientras que WatNet reduce este tiempo a cerca de 3 milisegundos.

Esta situación podría agravarse aún más al utilizar equipos con menor capacidad de procesamiento, especialmente aquellos que carecen de GPU. En entornos donde los recursos computacionales son limitados, como dispositivos de bajo costo o sistemas em-

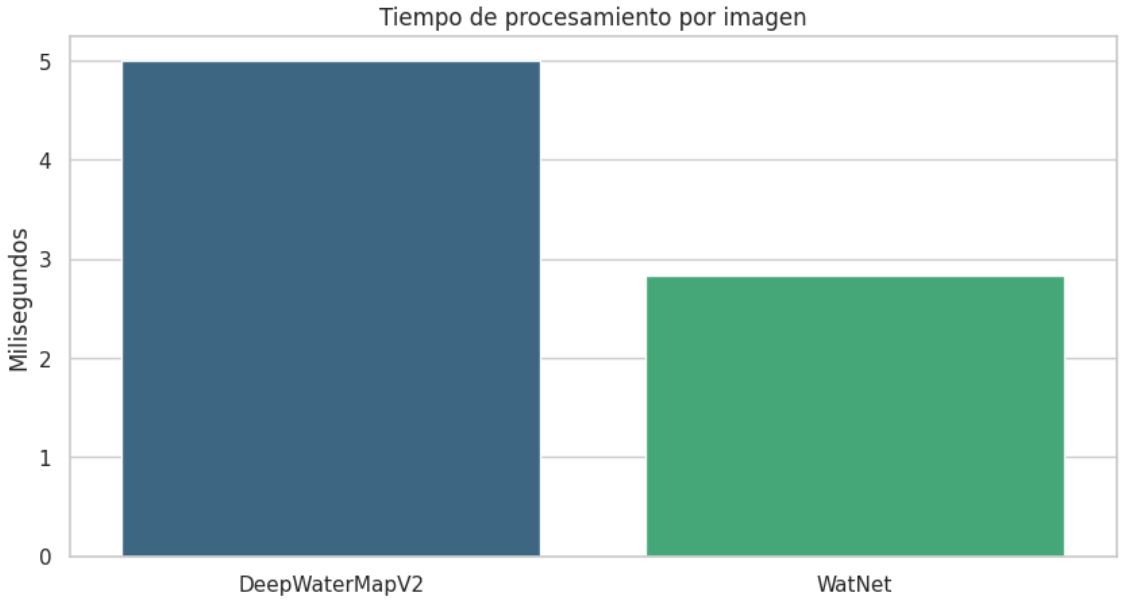

Figura 4.1: Comparativa tiempo de inferencia

bebidos, el tiempo de inferencia puede aumentar significativamente para modelos más complejos y pesados. WatNet seguirá siendo una opción viable para estos escenarios.

Afortunadamente, este no es nuestro caso, ya que contamos con equipos de alta capacidad y acceso a GPU para acelerar el procesamiento de los modelos. Dado que no enfrentamos limitaciones de hardware significativas, podemos priorizar la precisión y la calidad de segmentación sobre la velocidad de inferencia. Por lo tanto, continuaremos utilizando DeepWaterMapV2 modificado como nuestro sistema de referencia *Baseline*.

4.1.2. Comparativa de sensores

Una vez elegido el modelo que se va a utilizar como referencia (DeepWaterMapV2 modificado) se comparará su rendimiento con distintos sensores individualmente.

Los resultados obtenidos del modelo de referencia con distintos sensores se pueden observar en la Tabla 4.2.

Tabla 4.2: Resultados modelo *baseline*

Datos Train	Datos Val	Accuracy	F1-Score	MeanIoU
S1	S1	0.9928	0.7884	0.7367
S2 (RGB)	S2 (RGB)	0.9931	0.6398	0.6142
S2 (NO NUBES)	S2 (NO NUBES)	**0.9962**	**0.7976**	**0.7521**
S2 (NO NUBES)	S2 (NUBES)	0.6607	0.2105	0.1532
S2 (NUBES)	S2 (NUBES)	0.9876	0.6867	0.6403

El uso de datos de radar S1 proporciona un rendimiento general aceptable, con una *Accuracy* muy alta. Sin embargo, tanto el *F1-Score* como el *MeanIoU* son ligeramente inferiores, con valores de 0.7884 y 0.7367, respectivamente. Esto sugiere que, aunque el modelo es generalmente preciso, podría haber un desbalance entre las clases o una menor capacidad para identificar correctamente todos los píxeles de agua (TP) en comparación con los falsos positivos (FP) y falsos negativos (FN). Este comportamiento es compren-

sible, dado que los datos de radar pueden ser más complejos de interpretar para ciertas características del terreno en comparación con los datos ópticos.

Al utilizar únicamente las bandas RGB de Sentinel-2, se observa nuevamente una alta *Accuracy*, pero tanto el *F1-Score* como el *MeanIoU* disminuyen notablemente, alcanzando valores de 0.6398 y 0.6142, respectivamente. Esto indica que las bandas RGB, si bien útiles, no capturan toda la información necesaria para una segmentación precisa de cuerpos de agua. Es probable que esto se deba a la ausencia de otras bandas espectrales disponibles en Sentinel-2, como el infrarrojo cercano (NIR), que es clave para distinguir el agua de otras superficies.

Cuando se emplean todas las bandas de Sentinel-2, el rendimiento del modelo mejora significativamente en todas las métricas. La *Accuracy* aumenta a 0.9962, y tanto el *F1-Score* como el *MeanIoU* muestran mejoras considerables, con valores de 0.7976 y 0.7521, respectivamente. Esto refuerza la idea de que la inclusión de bandas adicionales, más allá del espectro visible, proporciona información crucial que mejora la capacidad del modelo para segmentar correctamente las imágenes. En particular, estas bandas adicionales permiten una mejor discriminación entre el agua y otras superficies, lo que incrementa tanto la precisión como la consistencia de las predicciones.

En cuanto a los resultados obtenidos al incorporar nubes en las imágenes de Sentinel-2, cuando el modelo no ha sido entrenado con imágenes nubladas, los resultados se deterioran significativamente. El *F1-Score* cae a 0.2105 y el *MeanIoU* a 0.1532, lo que indica que el modelo no está preparado para manejar este tipo de datos, lo cual es comprensible dada la falta de ejemplos durante el entrenamiento.

Por otro lado, cuando el modelo ha sido entrenado con imágenes que contienen nubes, su rendimiento mejora notablemente frente a estas condiciones. Aunque los resultados siguen siendo inferiores a los obtenidos con imágenes limpias o con datos de radar en condiciones nubladas, se alcanzan métricas aceptables, con un *F1-Score* de 0.6867 y un *MeanIoU* de 0.6403. Esto sugiere que, aunque las nubes complican la tarea al ocultar información clave, la riqueza espectral de los datos de Sentinel-2 ayuda al modelo a mitigar parcialmente estos efectos y a obtener resultados útiles.

En resumen, el análisis de los resultados indica que la alta resolución y la riqueza espectral de todas las bandas de Sentinel-2 proporcionan un rendimiento superior en comparación con el uso exclusivo de datos de radar. Esto subraya la importancia de combinar la información radar para enfrentar condiciones adversas con los datos ópticos para mejorar la precisión en la segmentación de cuerpos de agua.

4.2 *Single Encoder U-Net y Early Fusion*

La Tabla 4.3 presenta los resultados obtenidos al entrenar y evaluar el modelo *Early Fusion*, que combina información de los datos de radar de Sentinel-1 (S1) y los datos ópticos de Sentinel-2 (S2). Este enfoque se compara con los resultados del modelo *baseline*, donde se utilizan diferentes combinaciones de datos, incluidas variaciones con y sin nubes, así como la aplicación de técnicas de aumento de datos (DA, por sus siglas en inglés), como rotaciones parciales o completas, vertical u horizontalmente e incluso translaciones.

Inicialmente, al entrenar el modelo *Early Fusion* utilizando datos de Sentinel-1 y Sentinel-2 sin nubes, se logra una *Accuracy* de 0.9959, con un *F1-Score* de 0.8059 y un *MeanIoU* de 0.7596. Estos resultados son ligeramente superiores a los obtenidos con el modelo *baseline* que emplea todas las bandas de Sentinel-2, lo que refuerza la idea de que la combinación de datos de radar y ópticos puede ser útil para una mejor segmentación. Al incorporar técnicas de aumento de datos (DA) durante el entrenamiento con datos sin nubes,

Tabla 4.3: Resultados modelo *Early Fusion*

Datos Train	Datos Val	Accuracy	F1-Score	MeanIoU
S1 + S2 (NO NUBES)	S1 + S2 (NO NUBES)	0.9959	0.8059	0.7596
S1 + S2 (NO NUBES) + DA	S1 + S2 (NO NUBES)	**0.9960**	**0.8247**	**0.7754**
S1 + S2 (NO NUBES) + DA	S1 + S2 (NUBES)	0.6659	0.2372	0.1754
S1 + S2 (NUBES) + DA	S1 + S2 (NO NUBES)	0.9842	0.7720	0.7284
S1 + S2 (NUBES) + DA	S1 + S2 (NUBES)	0.9928	0.7245	0.6825

el rendimiento del modelo mejora aún más. Se observa un incremento en el *F1-Score* y en el *MeanIoU*, alcanzando valores de 0.8247 y 0.7754, respectivamente. El aumento de datos ayuda a mejorar la robustez del modelo al proporcionar una mayor diversidad de ejemplos de entrenamiento, lo que facilita la generalización y precisión en la tarea de segmentación.

Sin embargo, cuando este modelo entrenado con datos sin nubes y se evalúa con imágenes que contienen nubes, el rendimiento se ve gravemente afectado. La *Accuracy* cae a 0.6659, con un *F1-Score* de solo 0.2372 y un *MeanIoU* de 0.1754. Este patrón es similar al observado en el modelo *baseline*, indicando que la presencia de nubes continúa siendo un desafío significativo, incluso cuando se utilizan tanto datos de radar como ópticos.

Por otro lado, cuando el modelo se entrena con datos que incluyen nubes y se aplican técnicas de aumento de datos, los resultados mejoran notablemente en presencia de nubes, pero se mantiene una caída en las métricas en comparación con las condiciones sin nubes. En particular, cuando se evalúa con datos sin nubes, la *Accuracy* es de 0.9842, con un *F1-Score* de 0.7720 y un *MeanIoU* de 0.7284. Esto sugiere que, aunque el modelo es más robusto frente a la variabilidad introducida por las nubes, sigue existiendo una brecha de rendimiento comparado con condiciones ideales.

Finalmente, al evaluar el modelo entrenado con nubes y aumento de datos en imágenes que también contienen nubes, se observa un *F1-Score* de 0.7245 y un *MeanIoU* de 0.6825. Aunque estas cifras son inferiores a las obtenidas sin nubes, representan una mejora respecto a las condiciones sin entrenamiento previo con nubes, lo que indica que el modelo ha aprendido a manejar parcialmente la complejidad añadida por la presencia de nubes, beneficiándose de la riqueza espectral proporcionada por la combinación de datos radar y ópticos.

En resumen, los resultados del modelo *Early Fusion* subrayan la importancia de utilizar datos multisensoriales (radar y óptico) para la segmentación de cuerpos de agua, especialmente en condiciones ideales. Además, la aplicación de técnicas de aumento de datos demuestra ser una estrategia eficaz para mejorar la robustez del modelo. Sin embargo, la presencia de nubes sigue siendo un desafío considerable, aunque mitigado parcialmente mediante el entrenamiento con datos nublados y la estrategia de fusión de datos.

4.3 *Dual Encoder U-Net* e *Intermediate Fusion*

La Tabla 4.4 muestra los resultados del modelo *Intermediate Fusion*, que incorpora la fusión de características a través de una arquitectura de *Dual Encoder U-Net* en distintas etapas del modelo utilizando la función de fusión explicada en el Capítulo 3.

Cuando donde el modelo se entrena con datos sin nubes y con técnicas de aumento de datos (DA), se obtiene una *Accuracy* de 0.9955, un *F1-Score* de 0.8206, y un *MeanIoU* de

Tabla 4.4: Resultados modelo *Intermediate Fusion*

Datos Train	Datos Val	Accuracy	F1-Score	MeanIoU
S1 + S2 (NO NUBES) + DA	S1 + S2 (NO NUBES)	0.9955	0.8206	0.7707
S1 + S2 (NO NUBES) + DA	S1 + S2 (NUBES)	0.7913	0.4265	0.3478
S1 + S2 (NUBES) + DA	S1 + S2 (NO NUBES)	**0.9946**	**0.8251**	**0.7707**
S1 + S2 (NUBES) + DA	S1 + S2 (NUBES)	0.9920	0.7949	0.7398

0.7707. Estos valores son muy competitivos y están en línea con los resultados obtenidos con el modelo *Early Fusion*. La alta precisión y los sólidos valores de *F1-Score* y *MeanIoU* sugieren que la atención por canal es eficaz para extraer y combinar las características más relevantes de los datos radar y ópticos, lo que mejora la capacidad del modelo para distinguir entre cuerpos de agua y otras clases.

Al evaluar este modelo entrenado con datos sin nubes en un conjunto de validación con nubes, el rendimiento sufre un deterioro significativo, como era de esperar. La *Accuracy* desciende a 0.7913, mientras que el *F1-Score* y el *MeanIoU* caen a 0.4265 y 0.3478, respectivamente.

En contraste, cuando el modelo es entrenado con datos que incluyen nubes y con técnicas de aumento de datos, se observa un rendimiento sólido tanto en imágenes con nubes como sin ellas. Al evaluar el modelo con datos sin nubes, se obtiene la mejor *Accuracy* (0.9946), acompañada de un *F1-Score* de 0.8251 y un *MeanIoU* de 0.7707, lo que subraya la robustez del modelo en condiciones ideales.

Además, en la evaluación con datos que contienen nubes, el modelo alcanza una *Accuracy* de 0.9920, con un *F1-Score* de 0.7949 y un *MeanIoU* de 0.7398. Aunque estos valores son ligeramente inferiores a los obtenidos en condiciones sin nubes, representan una mejora notable respecto a modelos anteriores con datos nublados. Esto demuestra que el modelo *Channel Attention*, cuando se entrena adecuadamente, es capaz de manejar la variabilidad introducida por las nubes, aprovechando la capacidad de atención para enfocarse en las características más relevantes de los datos multisensoriales.

El modelo con *Intermediate Fusion* muestra un rendimiento sobresaliente en la segmentación de cuerpos de agua, especialmente cuando se entrena con una combinación de datos con y sin nubes. La atención por canal mejora la capacidad del modelo para discriminar entre diferentes clases, incluso en condiciones desfavorables.

4.4 *Dual U-Net* y *Late Fusion*

Finalmente el la Tabla 4.5 podemos ver los resultados del último modelo de *Late Fusion* con una arquitectura *Dual U-Net*.

Tabla 4.5: Resultados modelo *Late Fusion*

Datos Train	Datos Val	Accuracy	F1-Score	MeanIoU
S1 + S2 (NO NUBES) + DA	S1 + S2 (NO NUBES)	**0.9961**	**0.8238**	**0.7766**
S1 + S2 (NO NUBES) + DA	S1 + S2 (NUBES)	0.7046	0.3126	0.2388
S1 + S2 (NUBES) + DA	S1 + S2 (NO NUBES)	0.9955	0.8099	0.7608
S1 + S2 (NUBES) + DA	S1 + S2 (NUBES)	0.9944	0.7954	0.7422

Este modelo, entrenado con datos de Sentinel-1 y Sentinel-2 (sin nubes) con técnicas de aumento de datos (DA), alcanzó una *Accuracy* de 0.9961, un *F1-Score* de 0.8238 y un

MeanIoU de 0.7766 cuando fue evaluado con imágenes limpias. Estos resultados son los mejores obtenidos en condiciones sin nubes, superando ligeramente a todos los otros modelos en términos de precisión y consistencia en la segmentación, situándose como el mejor modelo para imágenes no limpias.

Como en los otros casos cuando el mismo modelo se evalúa en condiciones desfavorables sus resultados se ven muy deteriorados, obteniendo una *Accuracy* de 0.7046, un *F1-Score* de 0.3126 y un *MeanIoU* de 0.2388.

Cuando el modelo se entrenó con imágenes nubladas y se evaluó con datos limpios, los resultados fueron también bastante robustos, con una *Accuracy* de 0.9955, un *F1-Score* de 0.8099 y un *MeanIoU* de 0.7608. Esto indica que el modelo mantiene un buen rendimiento, aunque ligeramente inferior al mejor escenario de entrenamiento sin nubes y viéndose superado en pequeña medida por el modelo de *Intermediate Fusion*.

Finalmente, cuando tanto el entrenamiento como la validación incluyeron imágenes con nubes, el modelo logró una *Accuracy* de 0.9944, un *F1-Score* de 0.7954 y un *MeanIoU* de 0.7422, mostrando que es capaz de manejar mejor las condiciones desafiantes en comparación a otros modelos y situándose otra vez en primer lugar en esta situación en concreto.

La arquitectura *Dual U-Net* con *Late Fusion* ofrece un excelente rendimiento en imágenes limpias y muestra una mejora en la capacidad de generalización frente a imágenes nubladas.

4.5 Benchmark generales

En este apartado se compara el rendimiento de diferentes modelos de segmentación en imágenes limpias, es decir, sin la presencia de nubes (entrenados sin nubes y entrenados con nubes) y en imágenes nubladas.

4.5.1. Imágenes limpias

Entrenamiento con imágenes limpias

En la Tabla 4.6 se pueden observar los resultados de los distintos modelos frente a imágenes limpias, entrenados a su vez con imágenes libres de nubes.

Tabla 4.6: Resultados modelos en imágenes limpias

Arquitectura	Datos usados	Accuracy	F1-Score	MeanIoU
Baseline	S1	0.9928	0.7884	0.7367
Baseline	S2 (RGB)	0.9931	0.6398	0.6142
Baseline	S2	0.9962	0.7976	0.7521
Early Fusion	S1 + S2	0.9959	0.8059	0.7596
Early Fusion	S1 + S2 + DA	0.9960	0.8247	0.7754
Intermediate Fusion	S1 + S2 + DA	0.9955	0.8206	0.7707
Late Fusion	S1 + S2 + DA	**0.9961**	**0.8238**	**0.7766**

En términos generales, todos los modelos presentan resultados aceptables. El modelo *Baseline*, el más simple, entrenado con todas las bandas de Sentinel-2 (S2), alcanza una *Accuracy* de 0.9962, un *F1-Score* de 0.7976 y un *MeanIoU* de 0.7521. Aunque estos resul-

tados son muy buenos, los modelos que combinan datos de radar de Sentinel-1 (S1) y ópticos de Sentinel-2 (S2) muestran un rendimiento superior.

El modelo *Early Fusion*, al combinar ambos tipos de datos con aumento de datos (DA), alcanza una *Accuracy* de 0.9960, un *F1-Score* de 0.8247 y un *MeanIoU* de 0.7754, consolidándose como una de las opciones más robustas.

El modelo *Intermediate Fusion*, también con DA, sigue de cerca con métricas de *F1-Score* de 0.8206 y *MeanIoU* de 0.7707, presentando un rendimiento comparable pero ligeramente inferior al *Early Fusion*.

Finalmente, el modelo *Late Fusion*, que emplea una arquitectura *Dual U-Net* con DA, se destaca como el mejor modelo al obtener la mayor *Accuracy* de 0.9961, un *F1-Score* de 0.8238 y un *MeanIoU* de 0.7766. Esto lo posiciona como la opción más robusta y precisa para la segmentación en imágenes limpias, superando a los demás en términos de precisión y consistencia.

En conclusión, para imágenes limpias, el modelo *Late Fusion* es la elección preferida, debido a su capacidad para integrar eficientemente la información de múltiples fuentes y maximizar las métricas clave de segmentación. Aunque todos los modelos basado en fusión se proclaman como opciones sólidas y mejores a utilizar un solo sensor.

Entrenamiento con imágenes nubladas

Para presentar un resultado más realista, especialmente si el objetivo es utilizar el mismo modelo para imágenes con y sin nubes, es crucial entrenar el modelo para la tarea más desafiante, es decir, las imágenes con nubes. Esta evaluación es clave para entender cómo los modelos, entrenados en condiciones más desafiantes, se desempeñan en situaciones ideales, comparándolos con los resultados anteriores obtenidos en imágenes limpias.

Tabla 4.7: Resultados modelos en imágenes limpias entrenamiento con nubes

Arquitectura	Datos usados	Accuracy	F1-Score	MeanIoU
Baseline	S1	0.9928	0.7884	0.7367
Early Fusion	S1 + S2 + DA	0.9842	0.7720	0.7284
Intermediate Fusion	S1 + S2 + DA	**0.9946**	**0.8251**	**0.7707**
Late Fusion	S1 + S2 + DA	0.9955	0.8099	0.7608

En la Tabla 4.7 se evalúa el rendimiento de los modelos en imágenes limpias después de haber sido entrenados con datos nublados.

El modelo que utiliza datos de radar (S1), al no haber sido entrenado con datos ópticos, no ve afectado su rendimiento por la presencia de nubes, manteniendo resultados consistentes con los obtenidos en la tabla anterior.

En el modelo de *Early Fusion*, se observa una ligera disminución en el rendimiento en imágenes limpias en comparación con los resultados anteriores. La *Accuracy* disminuye a 0.9842, el *F1-Score* a 0.7720, y el *MeanIoU* a 0.7284. Aunque este modelo fue entrenado con datos nublados, parece que esta adaptación ha introducido cierta pérdida de precisión en imágenes limpias.

Por otro lado, el modelo *Intermediate Fusion* muestra un excelente rendimiento en imágenes limpias, incluso después de ser entrenado con datos nublados. Con una *Accuracy* de 0.9946, un *F1-Score* de 0.8251, y un *MeanIoU* de 0.7707, supera al modelo *Early Fusion* en todas las métricas y se acerca significativamente a los mejores resultados obtenidos

anteriormente en condiciones ideales. Esto sugiere que la fusión intermedia con atención es capaz de mantener una alta precisión, incluso cuando el modelo es entrenado en condiciones más desafiantes.

Finalmente, el modelo *Late Fusion* muestra un rendimiento robusto pero ligeramente inferior en comparación con su desempeño en imágenes limpias, obteniendo una *Accuracy* de 0.9955, un *F1-Score* de 0.8099 y un *MeanIoU* de 0.7608. Aunque estos resultados son sólidos, son un poco menores a los del *Intermediate Fusion*, lo cual podría deberse a la complejidad adicional introducida por la fusión tardía, que podría ser más sensible a la variabilidad en los datos de entrenamiento con nubes.

En general, entrenar con imágenes nubladas parece tener un impacto en la capacidad de generalización de los modelos cuando se evalúan en condiciones ideales (imágenes limpias). Sin embargo, tanto el *Intermediate Fusion* como el *Late Fusion* demuestran ser opciones viables para manejar ambos escenarios, con el *Intermediate Fusion* destacándose como el modelo más equilibrado y robusto en la tarea de segmentación de cuerpos de agua en imágenes con y sin nubes. Esto subraya la importancia de seleccionar modelos con una buena capacidad de adaptación, especialmente para aplicaciones en entornos complejos y variables.

4.5.2. Imágenes nubladas

Finalmente se evalúan los modelos en imágenes nubladas, que representan un escenario más desafiante debido a la interferencia visual de las nubes. Esta evaluación es crucial para determinar la capacidad de los modelos de manejar condiciones adversas y ofrecer una segmentación precisa bajo estas circunstancias. En la Tabla 4.8 podemos ver los resultados.

Tabla 4.8: Resultados modelos en imágenes nubladas

Arquitectura	Datos usados	Accuracy	F1-Score	MeanIoU
Baseline	S1	0.9928	0.7884	0.7367
Baseline	S2	0.9876	0.6867	0.6403
Early Fusion	S1 + S2 + DA	0.9928	0.7245	0.6825
Intermediate Fusion	S1 + S2 + DA	0.9920	0.7949	0.7398
Late Fusion	S1 + S2 + DA	**0.9944**	**0.7954**	**0.7422**

Los datos de radar (S1), como anteriormente al no verse afectado por la nubes mantiene sus resultados. En contraste, el *Baseline* entrenado con datos ópticos de Sentinel-2 (S2) muestra una disminución significativa en el rendimiento, evidenciando las limitaciones de los datos ópticos bajo condiciones nubladas.

El modelo *Early Fusion*, que combina datos radar y ópticos con técnicas de aumento de datos (DA), logra mejorar el rendimiento respecto al *Baseline* basado en S2, pero aún se queda atrás comparado con el modelo basado exclusivamente en S1. Sus métricas muestran que, aunque el enfoque de fusión temprana ofrece cierta mejora, aún no logra mitigar completamente los efectos adversos de las nubes.

El modelo *Intermediate Fusion* muestra una notable capacidad para manejar imágenes nubladas, obteniendo una *Accuracy* de 0.9920, un *F1-Score* de 0.7949 y un *MeanIoU* de 0.7398, lo que demuestra una integración más efectiva de la información complementaria de los datos radar y ópticos. Sin embargo, el modelo *Late Fusion* se destaca como el mejor en este escenario adverso, alcanzando las métricas más altas con una *Accuracy* de 0.9944, un *F1-Score* de 0.7954 y un *MeanIoU* de 0.7422. Esto sugiere que la fusión tardía

maneja de manera óptima la combinación de ambas fuentes de datos, proporcionando la segmentación más precisa y robusta en presencia de nubes.

Mientras que los datos de radar ofrecen un rendimiento constante en condiciones nubladas, los modelos de fusión, especialmente el *Late Fusion*, sobresalen al integrar de manera efectiva tanto datos radar como ópticos, convirtiéndose en la mejor opción para segmentación en escenarios con nubes, debido a su superior capacidad de adaptación y precisión.

4.5.3. Resumen general

Los resultados de las evaluaciones muestran que el modelo *Late Fusion* es la mejor opción para segmentación en imágenes nubladas, mientras que el modelo *Intermediate Fusion* sobresale en imágenes limpias. Esto resalta la importancia de adaptar la elección del modelo a las condiciones específicas del entorno. Si se espera que en una zona predominen los días nublados, como en climas tropicales o estaciones lluviosas, la *Late Fusion* se perfila como la opción más robusta debido a su capacidad para manejar eficazmente la presencia de nubes al combinar los datos de radar y ópticos. En cambio, en regiones o estaciones con mayor prevalencia de días claros, la *Intermediate Fusion* podría ser preferible, ya que maximiza la precisión y eficiencia en condiciones ideales de observación. Por tanto, la elección entre ambos modelos debe considerar la variabilidad meteorológica del área de interés para optimizar los resultados de segmentación de cuerpos de agua en imágenes satelitales.

CAPÍTULO 5
Conclusiones y Trabajos Futuros

En conclusión, en este trabajo se ha investigado la aplicación de técnicas avanzadas de fusión multimodal de datos para la segmentación de imágenes satelitales, empleando datos provenientes de los satélites Sentinel-1 y Sentinel-2. La principal contribución de este estudio ha sido la evaluación del rendimiento de diferentes estrategias de fusión, y cómo estas pueden mejorar la precisión y eficiencia en la segmentación de imágenes en diversas condiciones atmosféricas y de iluminación.

En primer lugar, los resultados obtenidos demuestran que la fusión multimodal de datos es una estrategia eficaz para superar las limitaciones inherentes a los datos de un único tipo de sensor. En particular, la combinación de imágenes de radar de Sentinel-1, que son independientes de las condiciones meteorológicas, con las imágenes ópticas de alta resolución de Sentinel-2, permite una segmentación más robusta y precisa, especialmente en escenarios complejos como áreas nubladas o con baja iluminación. Este hallazgo refuerza la hipótesis inicial de que la fusión de diferentes modalidades de datos puede complementar la información, mejorando así el desempeño de los modelos de segmentación.

Además, la comparación entre las distintas estrategias de fusión (temprana, intermedia y tardía) reveló diferencias significativas en términos de rendimiento. Tanto la fusión intermedia como la fusión tardía, que emplean modelos especializados para cada tipo de datos antes de combinarlos, han mostrado ser particularmente efectivas, logrando un equilibrio óptimo entre precisión y eficiencia computacional. Estos enfoque permite aprovechar al máximo las características distintivas de cada conjunto de datos, lo que se traduce en una mejora notable en la calidad de las segmentaciones obtenidas.

Por otro lado, las pruebas realizadas con imágenes nubladas y sin nubes han puesto en evidencia la importancia de la preparación y preprocesamiento de los datos. La generación de nubes artificiales y la simulación de diferentes condiciones meteorológicas durante el entrenamiento de los modelos resultaron ser técnicas cruciales para aumentar la resiliencia de los sistemas de segmentación frente a variaciones inesperadas en los datos de entrada. Este enfoque no solo mejora la precisión en la segmentación, sino que también incrementa la robustez de los modelos ante escenarios adversos, lo que es esencial en aplicaciones críticas como la gestión de desastres naturales.

Los resultados de los benchmarks generales han confirmado que las arquitecturas basadas en U-Net, en combinación con redes preentrenadas como ResNet50, proporcionan un rendimiento sólido en la tarea de segmentación de imágenes satelitales. Sin embargo, es importante destacar que la elección del modelo adecuado depende en gran medida del tipo de datos y las condiciones específicas del entorno en el que se va a aplicar. En este sentido, la flexibilidad y adaptabilidad de las arquitecturas propuestas son características clave que deben ser consideradas en futuros desarrollos.

En conclusión, el estudio realizado no solo ha permitido confirmar la eficacia de la fusión multimodal de datos en la mejora de la segmentación de imágenes satelitales, sino que también ha identificado áreas clave para futuras investigaciones. Entre estas, se destacan la exploración de nuevas arquitecturas de fusión que puedan integrarse más profundamente con modelos basados en transformers, y la evaluación de técnicas de fusión aplicadas a otros tipos de datos satelitales, como los proporcionados por satélites de próxima generación con capacidades mejoradas.

Finalmente, es esencial considerar la aplicabilidad práctica de los resultados obtenidos en contextos reales. La implementación de estos sistemas en escenarios operativos, como la monitorización ambiental, la gestión de recursos naturales y la respuesta ante desastres, puede contribuir significativamente a la mejora de las decisiones basadas en datos. En este sentido, la fusión multimodal de datos satelitales se presenta como una herramienta poderosa para abordar los desafíos complejos y dinámicos que caracterizan estas aplicaciones, subrayando la importancia de seguir investigando y perfeccionando estas técnicas en futuros estudios.

5.1 Trabajos Futuros

Los resultados obtenidos en este estudio abren una serie de posibles direcciones para investigaciones futuras, tanto en términos de perfeccionamiento de las técnicas utilizadas como en la exploración de nuevas aplicaciones y enfoques. A continuación, se presentan algunas de las líneas de trabajo más prometedoras que podrían ser exploradas para avanzar en el campo de la segmentación de imágenes satelitales mediante fusión multimodal de datos.

1. **Optimización de modelos basados en Transformers**: Aunque este estudio se ha centrado principalmente en arquitecturas basadas en U-Net y ResNet50, la creciente popularidad de los modelos basados en Transformers en el ámbito de la visión por computadora sugiere que estos podrían ofrecer mejoras significativas en la tarea de segmentación. Una línea de investigación futura podría explorar la integración de modelos Transformers con las estrategias de fusión multimodal, evaluando su capacidad para manejar la complejidad y la diversidad de los datos satelitales. Específicamente, sería interesante investigar cómo estos modelos pueden ser adaptados para procesar y fusionar datos de radar y ópticos de manera más eficiente. Incluso recientemente, con la creación de modelos tipo transformer preentrenados con datos satelitales, como el modelo Prithvi-100M [12], desarrollado por IBM y la NASA.

2. **Desarrollo de técnicas avanzadas de fusión**: Este trabajo ha identificado la fusión intermedia como una estrategia particularmente efectiva, pero existe un amplio margen para el desarrollo de técnicas de fusión más avanzadas. Investigaciones futuras podrían enfocarse en la creación de métodos de fusión que no solo combinen información de diferentes modalidades, sino que también tengan en cuenta la temporalidad de los datos. Esto podría permitir la creación de modelos que sean capaces de captar patrones temporales y estacionales en los datos satelitales, lo que es crucial para aplicaciones como la monitorización del cambio climático o la gestión agrícola.

3. **Ampliación a otros tipos de datos satelitales**: Si bien este trabajo se ha centrado en los datos proporcionados por los satélites Sentinel-1 y Sentinel-2, futuros estudios podrían explorar la aplicación de las técnicas desarrolladas a otros conjuntos

de datos satelitales. La inclusión de datos hiperespectrales o información de elevación DEM (Digital Elevation Model), por ejemplo, podría ofrecer una riqueza de información que mejore aún más la precisión y aplicabilidad de los modelos de segmentación.

4. **Aplicación en escenarios realistas**: Aunque el presente estudio ha demostrado el potencial de las técnicas de fusión multimodal en un entorno controlado, futuros trabajos deberían enfocarse en la validación y optimización de estos modelos en escenarios operativos reales. Esto implica no solo adaptar los modelos para que sean más robustos y eficientes en entornos de producción, sino también trabajar en la integración de estas técnicas en sistemas de monitoreo y gestión de desastres, donde la rapidez y la precisión de la segmentación de imágenes pueden tener un impacto directo en la toma de decisiones.

En resumen, las posibilidades para futuras investigaciones en este campo son vastas y variadas. Las técnicas de fusión multimodal y la segmentación de imágenes satelitales continúan evolucionando, y los avances tecnológicos prometen abrir nuevas fronteras en este campo. La continuación de este trabajo podría contribuir significativamente al desarrollo de sistemas más robustos y precisos con un impacto positivo en una amplia gama de aplicaciones críticas.

Bibliografía

[1] Liang-Chieh Chen et al. *Rethinking Atrous Convolution for Semantic Image Segmentation*. 2017. arXiv: 1706.05587 [cs.CV]. URL: https://arxiv.org/abs/1706.05587.

[2] Mikolaj Czerkawski et al. «SatelliteCloudGenerator: Controllable Cloud and Shadow Synthesis for Multi-Spectral Optical Satellite Images». En: *Remote Sensing* 15.17 (2023). ISSN: 2072-4292. DOI: 10.3390/rs15174138. URL: https://www.mdpi.com/2072-4292/15/17/4138.

[3] Yimian Dai et al. *Attentional Feature Fusion*. 2020. arXiv: 2009.14082 [cs.CV]. URL: https://arxiv.org/abs/2009.14082.

[4] Jia Deng et al. «Imagenet: A large-scale hierarchical image database». En: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, págs. 248-255.

[5] Dmytro Filatov y Ghulam Nabi Ahmad Hassan Yar. *Forest and Water Bodies Segmentation Through Satellite Images Using U-Net*. 2022. arXiv: 2207.11222 [cs.CV]. URL: https://arxiv.org/abs/2207.11222.

[6] Michael Schmitt; Lloyd Hughes; Pedram Ghamisi; Naoto Yokoya; Ronny Hänsch. *2020 IEEE GRSS Data Fusion Contest*. 2019. DOI: 10.21227/rha7-m332. URL: https://dx.doi.org/10.21227/rha7-m332.

[7] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV]. URL: https://arxiv.org/abs/1512.03385.

[8] Jie Hu et al. *Squeeze-and-Excitation Networks*. 2019. arXiv: 1709.01507 [cs.CV]. URL: https://arxiv.org/abs/1709.01507.

[9] Sergey Ioffe y Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG]. URL: https://arxiv.org/abs/1502.03167.

[10] Furkan Isikdogan, Alan Bovik y Paola Passalacqua. «Learning a River Network Extractor Using an Adaptive Loss Function». En: *IEEE Geoscience and Remote Sensing Letters* 15.6 (2018), págs. 813-817. DOI: 10.1109/LGRS.2018.2811754.

[11] Leo F. Isikdogan, Alan Bovik y Paola Passalacqua. «Seeing Through the Clouds With DeepWaterMap». En: *IEEE Geoscience and Remote Sensing Letters* 17.10 (2020), págs. 1662-1666. DOI: 10.1109/LGRS.2019.2953261.

[12] Johannes Jakubik et al. *Prithvi-100M*. Ago. de 2023. DOI: 10.57967/hf/0952.

[13] Diederik P. Kingma y Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[14] Tsung-Yi Lin et al. *Focal Loss for Dense Object Detection*. 2018. arXiv: 1708.02002 [cs.CV]. URL: https://arxiv.org/abs/1708.02002.

[15] Xin Luo, Xiaohua Tong y Zhongwen Hu. «An applicable and automatic method for earth surface water mapping based on multispectral images». En: *International Journal of Applied Earth Observation and Geoinformation* 103 (2021), pág. 102472. ISSN: 1569-8432. DOI: https://doi.org/10.1016/j.jag.2021.102472. URL: https://www.sciencedirect.com/science/article/pii/S0303243421001793.

[16] Pankaj Malik et al. «Satellite Image Segmentation Using Neural Networks: A Comprehensive Review». En: *International Journal of Enhanced Research in Educational Development* Vol. 11 (nov. de 2023), págs. 2320-8708.

[17] Elias Manos et al. «Convolutional Neural Networks for Automated Built Infrastructure Detection in the Arctic Using Sub-Meter Spatial Resolution Satellite Imagery». En: *Remote Sensing* 14 (jun. de 2022), pág. 2719. DOI: 10.3390/rs14112719.

[18] Cesar Augusto Valbuena Calderón Nicola Clerici y Juan Manuel Posada. «Fusion of Sentinel-1A and Sentinel-2A data for land cover mapping: a case study in the lower Magdalena region, Colombia». En: *Journal of Maps* 13.2 (2017), págs. 718-726. DOI: 10.1080/17445647.2017.1372316. eprint: https://doi.org/10.1080/17445647.2017.1372316. URL: https://doi.org/10.1080/17445647.2017.1372316.

[19] Olaf Ronneberger, Philipp Fischer y Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597.

[20] Mark Sandler et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2019. arXiv: 1801.04381 [cs.CV]. URL: https://arxiv.org/abs/1801.04381.

[21] Michael Schmitt et al. *SEN12MS – A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion*. 2019. arXiv: 1906.07789 [cs.CV]. URL: https://arxiv.org/abs/1906.07789.

[22] Evan Shelhamer, Jonathan Long y Trevor Darrell. «Fully Convolutional Networks for Semantic Segmentation». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (2017), págs. 640-651. DOI: 10.1109/TPAMI.2016.2572683.

[23] Carole H. Sudre et al. «Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations». En: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer International Publishing, 2017, págs. 240-248. ISBN: 9783319675589. DOI: 10.1007/978-3-319-67558-9_28. URL: http://dx.doi.org/10.1007/978-3-319-67558-9_28.

APÉNDICE A

Relación del proyecto con los Objetivos de Desarrollo Sostenible

El trabajo descrito, que involucra segmentación de imágenes satelitales mediante la fusión multimodal de datos tiene una conexión directa con varios de los Objetivos de Desarrollo Sostenible (ODS) establecidos por las Naciones Unidas. Estas metas globales buscan abordar desafíos clave como la pobreza, la desigualdad, el cambio climático y la degradación ambiental, entre otros. A continuación, se detalla cómo las tecnologías y metodologías desarrolladas en este estudio contribuyen a algunos de estos objetivos:

- ODS 9 - Industria, Innovación e Infraestructura: Este objetivo busca promover la construcción de infraestructuras resilientes, el fomento de la innovación y el fomento de la adopción de tecnologías limpias y sostenibles. El desarrollo de tecnologías avanzadas para la segmentación de imágenes satelitales y la fusión multimodal de datos representa un avance significativo en el campo de la teledetección y la observación de la Tierra. Estas innovaciones no solo mejoran la capacidad de monitorear y gestionar recursos naturales, sino que también impulsan el desarrollo de nuevas aplicaciones en diversos sectores, desde la agricultura hasta la gestión de desastres.
- ODS 13 - Acción por el clima: Este objetivo se centra en tomar medidas urgentes para combatir el cambio climático y sus impactos. La segmentación precisa de imágenes satelitales, especialmente en combinación con datos multimodales, permite un monitoreo más eficiente y detallado de fenómenos relacionados con el cambio climático, como la deforestación, el deshielo de los glaciares, y la desertificación. La capacidad de segmentar imágenes en condiciones meteorológicas adversas, gracias a la fusión de datos de radar y ópticos, mejora la precisión de los sistemas de alerta temprana y contribuye a una mejor planificación y respuesta ante desastres naturales, lo cual es esencial para mitigar los impactos del cambio climático.
- ODS 14 - Vida Submarina: Aunque el foco de este trabajo ha sido la segmentación de imágenes terrestres, las técnicas desarrolladas también pueden adaptarse para la monitorización de ambientes marinos y costeros. Esto es relevante para el ODS 14, que busca conservar y utilizar de manera sostenible los océanos, mares y recursos marinos. El monitoreo costero, incluyendo la identificación de áreas de erosión, la evaluación de la salud de los ecosistemas marinos, y la gestión de zonas de pesca, se beneficia enormemente de las tecnologías avanzadas de segmentación y fusión de datos.
- ODS 15 - Vida de Ecosistemas Terrestres: Este objetivo se centra en proteger, restaurar y promover el uso sostenible de los ecosistemas terrestres, gestionando los bosques de manera sostenible, combatiendo la desertificación y deteniendo la pérdida

de biodiversidad. La tecnología desarrollada en este trabajo permite la monitorización constante y detallada de los ecosistemas terrestres, facilitando la identificación de cambios en el uso del suelo, la extensión de áreas forestales y la salud de los ecosistemas. Estas capacidades son fundamentales para la conservación de la biodiversidad y la gestión sostenible de los recursos naturales.

Es importante tener en cuenta que la aplicación de la inteligencia artificial y el aprendizaje por refuerzo debe llevarse a cabo de manera ética y responsable, teniendo en cuenta los impactos sociales, económicos y ambientales. El desarrollo de políticas y marcos regulatorios sólidos también es crucial para garantizar que estas tecnologías se utilicen de manera sostenible y beneficiosa para la sociedad en su conjunto.

Tabla A.1: Impacto del trabajo en los Objetivos de Desarrollo Sostenible (ODS)

ODS	**Nivel de Impacto**
ODS 1 - Fin de la pobreza	Nulo
ODS 2 - Hambre cero	Bajo a Medio
ODS 3 - Salud y bienestar	Bajo
ODS 4 - Educación de calidad	Nulo
ODS 5 - Igualdad de género	Nulo
ODS 6 - Agua limpia y saneamiento	Medio a Alto
ODS 7 - Energía asequible y no contaminante	Medio a Alto
ODS 8 - Trabajo decente y crecimiento económico	Nulo
ODS 9 - Industria, Innovación e Infraestructura	Alto
ODS 10 - Reducción de las desigualdades	Nulo
ODS 11 - Ciudades y comunidades sostenibles	Medio a Alto
ODS 12 - Producción y consumo responsables	Nulo
ODS 13 - Acción por el clima	Alto
ODS 14 - Vida submarina	Medio a Alto
ODS 15 - Vida de ecosistemas terrestres	Medio a Alto
ODS 16 - Paz, justicia e instituciones sólidas	Nulo
ODS 17 - Alianzas para lograr los objetivos	Nulo