PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

FUSIÓN DE DATOS PARA SEGMENTACIÓN SEMÁNTICA EN APLICACIONES URBANAS DE TELEDETECCIÓN AÉREA USANDO ALGORITMOS DE APRENDIZAJE PROFUNDO

Tesis para optar el grado de Magíster en Informática con mención en Ciencias de la Computación

AUTOR

Miguel Angel Chicchón Apaza

ASESOR

Dr. Ivan Anselmo Sipiran Mendoza

LIMA – PERÚ

2018

Agradecimientos

Agradezco a mi familia por su apoyo incondicional. A los profesores de la Maestría en informática de la PUCP por las experiencias transmitidas durante las clases. Al Grupo de Investigación en Tecnologías Exponenciales de la Universidad de Lima por el apoyo brindado durante la realización de este trabajo.

Resumen

La creciente urbanización requiere un mapeo y monitoreo preciso del sistema urbano para planificar futuros desarrollos. La teledetección permite obtener información de la superficie de la Tierra y a partir de esta comprender el proceso de urbanización. Esta información hoy en dia puede ser obtenida en forma masiva utilizando vehículos aéreos no tripulados. Esta información puede ser variada incluyendo imágenes ópticas rgb, multiespectrales y modelos digitales de superficie, generandose la necesida de contar con técnicas de fusión multisensorial eficientes y efectivas para explotarlas completamente.

La segmentación semántica en teledetección urbana permite la interpretación automática de los datos y es útil en tareas como el mapeo de la cobertura terrestre y la planificación urbana.

Actualmente, el aprendizaje profundo se ha vuelto de interés en Visión por computador y Teledetección, existiendo diferentes estudios de la aplicación de variantes de redes neuronales convolucionales (CNN) en segmentación semántica.

En el presente trabajo de tesis se investiga la utilización de métodos de fusión de datos basado en algoritmos de aprendizaje profundo para la segmentación semántica en aplicaciones urbanas de teledetección.

Palabras clave: Teledetección, fusión de datos, segmentación semántica, aprendizaje profundo.

Abstract

The growing urbanization requires a precise mapping and monitoring of the urban system to plan future developments. Remote sensing allows us to obtain information about the surface of the Earth and from this we can understand the urbanization process. This information today can be obtained in bulk using unmanned aerial vehicles. This information can be varied including rgb, multispectral optical images and digital surface models, generating the need to have efficient and effective multisensory fusion techniques to fully exploit them.

The semantic segmentation in urban remote sensing allows the automatic interpretation of the data and is useful in tasks such as land cover mapping and urban planning.

Currently, deep learning has become of interest in computer vision and remote sensing, there are different studies of the application of variants of convolutional neural networks (CNN) in semantic segmentation.

In the present thesis work we investigate the use of data fusion methods based on deep learning algorithms for semantic segmentation in urban remote sensing applications.

Keywords: Remote sensing, data fusión, semantic segmentation, deeplearning

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Generalidades

1.1 Definición del problema

Según informe de la ONU, en el 2014 aproximadamente el 54% de la población mundial vivía en las regiones urbanas y se espera que la tendencia aumente al 66% para el 2050. La ONU concluye que la gestión de las áreas urbanas es uno de los desafíos clave del presente siglo [2]. Esta creciente urbanización requiere un mapeo y monitoreo preciso del sistema urbano para planificar futuros desarrollos. En este sentido la teledetección permite obtener información de la superficie de la Tierra y a partir de esta comprender el proceso de urbanización [3]. Tradicionalmente esta información es obtenida de imágenes satelitales o fotografías aéreas, pero actualmente la aparición de vehículos aéreos no tripulados permite acceder a imágenes de alta resolución a un costo razonable. Un paso esencial para comprender una imagen de teledetección de alta resolución es realizar una segmentación semántica, que consiste en etiquetar cada píxel de la imagen con la categoría semántica del objeto al que pertenece.

A las imágenes ópticas de alta resolución se le puede sumar datos multiespectrales, que pueden discriminar diversos materiales en función de las características de reflectancia espectral, así como datos tridimensionales a partir de procesamiento de imágenes o sistemas LiDAR (un acrónimo del inglés, Light Detection and Ranging o Laser Imaging Detection and Ranging), que aportan características geométricas. La combinación de estas múltiples fuentes de datos puede complementarse entre sí, obteniéndose un mejor rendimiento en tareas como la segmentación semántica. Por lo

expuesto, se necesitan técnicas de fusión multisensorial eficientes y efectivas para explotar completamente estas modalidades de datos complementarios [4].

En los últimos años, el aprendizaje profundo se ha vuelto de interés en Visión por computador y Teledetección [5]. Variantes de redes neuronales convolucionales (CNN) ahora se consideran como el estado del arte en segmentación semántica con mayor estudio en visión por computador. La aplicación de estos algoritmos en Teledetección continúa siendo un desafío debido a diferencias como la adquisición de datos a vista de pájaro, por lo que la perspectiva se ve significativamente alterada con respecto a los conjuntos de datos usuales de visión por computador. Los objetos se encuentran dentro de un plano 2D, lo que hace que el ángulo de visión sea consistente, pero reduce el número de sugerencias relacionadas con la profundidad, como las sombras proyectadas. Además, cada píxel en las imágenes de Teledetección tiene un significado semántico [6].

1.2 Objetivo general

Desarrollar un método de fusión de datos basado en algoritmos de aprendizaje profundo para la segmentación semántica en aplicaciones urbanas de teledetección. Los datos son obtenidos de cámaras RGB, cámaras multiespectrales y sistemas LiDAR transportados en vehículos aéreos.

1.3 Objetivos específicos

- Revisar el estado del arte de técnicas de fusión de datos aplicadas en problemas de segmentación semántica.
- Constituir los conjuntos de datos considerando los mecanismos de medición de la calidad de la segmentación semántica a realizar.
- Implementar algoritmos de fusión de datos basados en aprendizaje profundo para la segmentación semántica aplicada a la determinación de cobertura de suelo en entornos urbanos.
- Evaluar el desempeño de las implementaciones de las técnicas de fusión de datos para segmentación semántica aplicadas al conjunto de datos recolectado.

1.4 Resultados esperados

- Documentación del Estado del Arte.
- Conjunto de imágenes y nubes de puntos obtenidos a partir del estado del arte y fuentes propias.
- Algoritmos de fusión de datos basados en aprendizaje profundo para segmentación semántica implementados en lenguaje Python.
- Métricas de evaluación en segmentación semántica y resultados de la comparación de los algoritmos implementados.

1.5 Justificación

La fusión de datos permite obtener mejores resultados en tareas de segmentación semántica. La segmentación semántica en teledetección urbana permite la interpretación automática de los datos, siendo esta una tarea de primordial importancia para una amplia gama de aplicaciones prácticas, como el mapeo de la cobertura terrestre (análisis urbano, agricultura de precisión) [3][7], planificación urbana y monitoreo del tráfico [8][9], detección de daños (por ejemplo, en desastres naturales como inundaciones, huracanes, terremotos, derrames de petróleo en los mares). Además, las adquisiciones repetidas de una escena en diferentes momentos permiten monitorear los recursos naturales y variables ambientales (fenología de la vegetación, capa de nieve), efectos antropológicos (expansión urbana, deforestación), cambios climáticos (desertificación, erosión costera) entre otros [10].

La segmentación semántica en teledetección, a pesar del esfuerzo realizado, aun es una tarea desafiante debido a la apariencia muy heterogénea de objetos como edificios, calles, árboles y automóviles en datos de muy alta resolución, produciendo una gran variación dentro de la clase mientras que la varianza entre clases es baja [11].

Los métodos basados en redes neuronales convolucionales, CNN del acrónimo en inglés, han tenido un gran éxito últimamente en una amplia gama de tareas de visión artificial y teledetección. Su utilización en extracción de características ha logrado mejores resultados que los métodos que usan características hechas a mano y representan el estado del arte en segmentación semántica [12]. Existen muchos modelos preentrenados con imágenes RGB que pueden ser utilizados en tareas de segmentación semántica con datos adicionales como imágenes multiespectrales, índices de vegetación, modelos digitales de superficie y nube de puntos. En la revisión del estado del arte hay antecedentes acerca de la mejora en la precisión de la segmentación semántica con estas modalidades adicionales [7], pero la fusión no es un problema trivial. La adaptación de los métodos RGB con pesos preentrenados para su utilización en fusión de múltiples modalidades de datos requiere que las modalidades adicionales se manejen por separado o las arquitecturas de la red se modifiquen, dificultando de esta manera el uso de los

pesos preentrenados. Por lo tanto, es motivante investigar cómo se pueden fusionar modalidades adicionales en la red, conservando los beneficios de los métodos de segmentación semántica para imágenes RGB.

1.6 Límites

La segmentación semántica se puede realizar tanto en imágenes (2D) como nube de puntos (3D) [10]. Sin embargo, la presente tesis se limita a la segmentación semántica de imágenes, utilizando la nube de puntos como una entrada del algoritmo de fusión de datos a implementar, con la finalidad de evaluar mejoras en los resultados.

La teledetección recopila información a diferentes altitudes utilizando principalmente satélites y aviones [13]. En la presente tesis se tiene un gran interés en datos de alta resolución tomados a baja altitud por vehiculos aereos no tripulados. Sin embargo, debido a la falta de conjuntos de datos etiquetados y considerando una referencia estándar para la evaluación de los algoritmos a implementar se selecciona, a partir de la revisión del estado del arte, el conjunto de datos de referencia Postdam-Vaihingen de la Sociedad Internacional de Fotogrametría y Teledetección (ISPRS, por sus siglas en inglés).

Aunque existen múltiples objetivos de clase en segmentación semántica de imágenes en Teledetección con aplicación a cobertura de suelo en áreas urbanas. La presente tesis se limita a la segmentación de las clases del conjunto de datos seleccionado.

El conjunto de datos esta limitado a imágenes aereas de tipo RGB y multiespectral, y modelos digitales de superficie. En la presente tesis en la etapa de preprocesamiento se considera la generación de índices de vegetación, como el ndvi, y se explora la generación de un nuevo conjunto de datos a partir del conjunto de datos 2D y 3D de la ciudad de Vaihingen proporcionado separadamente por el ISPRS. Finalmente, solo se ha considerado la implementación de métodos de extremo a extremo del aprendizaje profundo, en consecuencia, los posibles métodos de postprocesamiento han sido omitidos.

1.7 Aportes

El método planteado para la segmentación semántica 2D a partir de multiples fuentes de datos de alta resolución es un aporte en el campo de la Teledetección y plantea un abanico de oportunidades para futuras aplicaciones en este campo. Se demuestra la efectividad de las redes neuronales convolucionales profundas para fusionar datos en tareas de segmentación semántica, siendo este un punto de partida para posteriores desarrollos como la segmentación semántica 3D.

1.8 Esquema

El presente documento consta de 7 capítulos: El capítulo 1 busca definir el problema y el enfoque adoptado en esta tesis para darle solución. El Capítulo 2 presenta la teoría sobre teledetección, fusión de datos, los métodos de segmentación semántica y las redes neuronales convolucionales. El capítulo 3 describe los trabajos previos que busquen segmentar imágenes fusionando múltiples fuentes de datos antes y después de la aparición de las redes neuronales convolucionales. El capítulo 4 discute los métodos utilizados en la tesis. El Capítulo 5 describe los experimentos realizados, considerando los conjuntos de datos, las configuraciones de red y la metodología de evaluación. Tambien se enumeran los resultados cuantitativos y se ilustran los resultados cualitativos con imágenes. El capítulo 6 contiene las conclusiones de los experimentos y la discusión sobre el trabajo futuro.

Capítulo 2

Marco Conceptual

2.1 Teledetección

La teledetección es el arte y la ciencia de obtener información de la superficie o el subsuelo de la Tierra sin necesidad de estar en contacto con ella. Esto se logra detectando y registrando energía emitida o reflejada para luego procesar, analizar e interpretar la información recuperada y así poder tomar decisiones [11].

Existen tecnologías de teledetección basadas en el aire, en el espacio, en la tierra y en el mar, que utilizan una gran cantidad de sensores a bordo de diferentes plataformas. Dichos sensores están diseñados para observar la energía electromagnética, acústica, ultrasónica, sísmica y magnética para el monitoreo ambiental y la observación de la Tierra. Los sensores para teledetección se pueden clasificar en pasivos y activos.

Los sensores pasivos miden la distribución de energía que está disponible naturalmente a través de procesos de transferencia radiativa. Estos sensores pueden detectar la reflectancia (luz visible), la emisión (el infrarrojo y el infrarrojo térmico), y/o porciones de microondas del espectro electromagnético utilizando diferentes tipos de radiómetros y espectrómetros. Estos sensores pueden generar imágenes pancromáticas, RGB, infrarrojas, hiperespectrales y multiespectrales.

Los sensores activos envían su propia energía para la iluminación. Esto significa que los sensores pueden emitir su propia radiación dirigida hacia el objetivo de interés. La radiación reflejada de ese objetivo de interés es detectada y medida por el sensor. Un

sensor activo en sistemas de teledetección puede ser un radar, un láser o un LiDAR (del inglés, Light Detection and Ranging o Laser Imaging Detection and Ranging).

En los sistemas de teledetección, las medidas por lo general se cuantifican y se convierten en una imagen digital, donde cada píxel tiene un valor discreto en unidades de número digital (DN) asociado a resoluciones espaciales, espectrales, radiométricas y/o temporales [11]. Estas imágenes pueden contener deficiencias como presencia de ruido debido múltiples factores como la vibración anormal de los sistemas de observación. Para tratar estos defectos se realiza procedimientos de procesamiento adicionales con el objetivo primordial de producir una mejor imagen y así ayudar en la visualización o extracción de información.

Se han desarrollado diversas técnicas de procesamiento de imágenes para ayudar en la interpretación o la extracción de información de imágenes obtenidas por teledetección. Entre las técnicas de preprocesamiento se puede mencionar correcciones atmosféricas, radiométricas y geométricas, transformaciones geométricas, remuestreo, generación de mosaicos y relleno de espacios. Además, para mejorar aún más la propiedad o calidad de las imágenes en teledetección se aplican generalmente cuatro técnicas avanzadas de procesamiento de imágenes: mejoramiento, restauración, transformación y segmentación [11].

2.2 Teledetección aérea

La teledetección tradicionalmente se ha asociado con satélites o aeronaves tripuladas con un conjunto de sensores aerotransportados. En la última década, se han producido desarrollos y mejoras en plataformas no tripuladas y en las tecnologías de detección instaladas a bordo de dichas plataformas, brindando excelentes oportunidades para aplicaciones de teledetección. Además de la versatilidad, flexibilidad, fácil planificación y seguridad, estas plataformas pueden volar a baja altura y a baja velocidad, permitiendo adquirir datos espaciales y temporales de alta resolución, lo que representa ventajas importantes contra las plataformas convencionales que se han utilizado ampliamente a lo largo de los años [13].

Entre la terminología utilizada para referirse a las plataformas aéreas no tripuladas tenemos [14]:

- UAV: vehículo aéreo no tripulado.
- Drone: denominación de ámbito militar.
- RPA: Aeronave pilotada a distancia. Según la normativa peruana [14]: “un RPA es una aeronave pilotada por un piloto remoto, emplazado en una estación de piloto remoto ubicada fuera de la aeronave quien monitorea la aeronave en todo momento y tiene responsabilidad directa de la conducción segura de la aeronave durante todo su vuelo. Una RPA puede poseer varios tipos de tecnología de piloto

automático, pero, en todo momento, el piloto remoto puede intervenir en la gestión del vuelo".

- RPAS: Sistema de aeronave pilotada a distancia, conjunto de elementos configurables integrado por una aeronave pilotada a distancia, sus estaciones de piloto remoto, los enlaces de mando y control y cualquier otro elemento de sistema que pueda requerirse en cualquier punto durante la operación de vuelo.

- UAS: Sistema de aeronave no tripulada, se define como el conjunto aeronave, enlace de comunicaciones y estación de tierra que permiten la operación sin piloto a bordo.

En teledetección, aunque el equipo de UAS es necesario para capturar información, que luego es manejada (procesada, analizada o almacenada), comúnmente se usa el término "UAV" [14]. Los UAV se pueden clasificar en dos tipos: aviones de ala fija y ala giratoria. Cada tipo es adecuado para aplicaciones y tareas específicas. Los aviones de ala fija permiten cubrir grandes extensiones de terreno a altas velocidades de vuelo. Los aviones de ala giratoria permites velocidades de vuelo lentas o incluso estacionarias. Ofrecen una excelente maniobrabilidad y no requieren una pista para el despegue y el aterrizaje.

Figura 1. (a) avión de ala fija - modelo eBee de SenseFly. (b) avión de ala giratoria – modelo cuadricóptero phantom 4 de DJI **[15]**

Para realizar las misiones de teledetección los UAV se encuentran equipados con diferentes instrumentos y sensores, tales como, sistemas de posicionamiento global (GPS), sensores de navegación inercial (INS), sensores de altitud (AS), sensores anticolisión basados en ultrasonido, cámaras o escáneres laser 3D. Estos elementos junto con estrategias de control permiten una navegación autónoma de rutas previamente planificada, con la capacidad de tomar decisiones autónomas, como en el caso de evitar obstáculos. Otras tareas a tener en cuenta dentro de la planificación de la ruta son las estrategias para el posicionamiento, el aterrizaje y el despegue, que se pueden dar incluso en condiciones adversas por presencia de viento, además de la logística necesaria para cubrir las extensas áreas de la misión, como la recarga de la batería [13].

Los límites de carga útil a bordo de UAV representan una desventaja en el uso de sensores y genera nuevos desafíos en la industria. La teledetección requiere que cada imagen obtenida mediante cámaras (ópticas, térmicas, multiespectrales o hiperespectrales) estén asociadas con la correspondiente ubicación GPS, altitud y orientación del UAV, con el objetivo de obtener productos geométricos como mapas 3D, imágenes georreferenciadas y ortofotos [13].

Los sensores y los instrumentos para teledetección están en continuo mejoramiento y desarrollo. En [13] se describe detalladamente las aplicaciones y las tecnologías utilizadas actualmente en teledetección basada en UAV. Entre las tecnologías utilizadas con éxito podemos citar:

- Cámaras en el espectro visible: son ampliamente usadas y su rango de operación es de longitudes de onda de 390 nm a 700 nm. Se requiere el uso de sistemas que brinden alta estabilidad con un peso y consumo de energía adecuado. La necesidad de cubrir grandes áreas de visión permitió el desarrollo de sistemas de visión omnidireccional y lentes de ojo de pez. Aplicaciones que se pueden realizar son la prevención, detección y monitoreo de incendios, vigilancia, monitoreo de líneas eléctricas o corredores de tuberías.

- Cámaras en el espectro Infrarrojo termal: Un sensor térmico infrarrojo detecta la energía radiante emitida por objetos con temperaturas superiores al cero absoluto. Dicha radiación infrarroja se asume una función de la longitud de onda y la temperatura. Las longitudes de onda de las bandas espectrales varían 0.78 μm hasta 3 μm (infrarrojo cercano), 3 μm hasta 50 μm (media de infrarrojo), y 50 μm a 1000 μm (infrarrojo lejano). Las cámaras infrarrojas térmicas son dispositivos capaces de operar en condiciones climáticas adversas o con poca iluminación, incluidas las observaciones durante la noche.

- LiDAR: Los dispositivos de detección y alineación de luz (Light Detection and Ranging, en inglés) son utilizados para medir distancias hacia objetivos explorando la escena con la luz, emitida por un láser. La adaptación de estos sistemas en UAV lo ha convertida en una gran opción para la vigilancia o mapeo de estructuras naturales y artificiales.

- Cámaras multiespectral e hiperespectral: La diferencia entre ambos es el número de bandas espectrales y el rango de longitud de onda cubierto, incluido el espectro visible. Los sensores hiperespectrales se basan en el escaneo de líneas a través del movimiento estabilizado del UAV. Los sensores multiespectrales no escanean y, en general, proporcionan resoluciones de imagen inferiores en comparación con los sensores hiperespectral.

2.3 Fusión de datos

El Sub panel de Fusión de Datos de la Junta de Directores de Laboratorios (JDL) del Departamento de Defensa de los Estados Unidos [3], define a la fusión de datos como "un proceso multinivel y multiaspecto que trata del registro, detección, asociación, correlación y combinación de datos e información de múltiples fuentes para lograr una estimación refinada del estado y de la identidad, y evaluaciones completas oportunas de la situación, incluyendo amenazas y oportunidades"

JDL clasificó el proceso de fusión de datos en cinco niveles de procesamiento, una base de datos asociada y un bus de información que conecta los cinco componentes [11]:

- Nivel 0 - Preproceso de la fuente: se asigna datos a procesos adecuados y se realiza una prelectura de datos que mantiene información útil para procesos de alto nivel. Fusión a nivel de señal y píxel.

- Nivel 1 - Refinamiento del objeto: se transforma los datos procesados en estructuras de datos coherentes. Combina características extraídas de imágenes procesadas y refina la identificación de objetos y la clasificación.

- Nivel 2 - Evaluación de la situación: se relaciona los eventos con las situaciones probables observadas. Permite la interpretación los datos.
- Nivel 3 - Evaluación de amenazas: se evalúa el riesgo de eventos futuros mediante la evaluación de situaciones actuales.

- Nivel 4: Refinamiento del proceso: se supervisa el proceso de fusión de datos e identifica qué información adicional se requiere para mejorar el proceso de fusión de datos.

En [16] se da una definición más utilizada de la fusión de datos: "las técnicas de fusión de datos combinan datos de múltiples sensores e información relacionada de bases de datos asociadas para lograr una precisión mejorada e inferencias más específicas que las que se podrían lograr mediante el uso de un solo sensor".

En teledetección las fuentes de datos utilizadas para la fusión de datos pueden ser todos los tipos de información, como imágenes de sensores remotos de espacio o aire, puntos de datos discretos, archivos de vectores, imágenes de cámara, etc. Existen otros términos asociados a la fusión de datos como la fusión de sensores o la fusión de información. La fusión de la información se utiliza principalmente en la teoría de la información o la inteligencia artificial y apunta a fusionar información que no siempre puede ser representada por números reales. La fusión de sensores hace referencia a la mezcla de datos de sensores dispares para obtener más información de la que podría obtenerse con cada sensor individual y un caso particular es cuando las fuentes de datos se proporcionan en forma de imágenes, denominándose al proceso de fusión como fusión de imágenes [17].

En [5] se presenta una adaptación al contexto de teledetección del proceso de fusión de datos dado por la JDL. Siendo de principal interés el nivel 1 o refinamiento del objeto, el cual se centra en dos acciones centrales:

- La alineación de datos y la correlación datos/objetos forman conjuntamente en el contexto de teledetección la correspondencia espacial/temporal y el coregistro de diferentes datos sensoriales que muestran propiedades radiométricas, geométricas y de otro tipo potencialmente muy diferentes. Su objetivo es garantizar que las mediciones estén conectadas entre sí y con el objeto de interés respectivo.
- La estimación de atributos y estimación de identidad, constituyen la fusión, es decir, la explotación combinada de datos de medición alineados y correlacionados

en un marco de estimación. En la estimación de atributos se desean atributos que describan el objeto (por ejemplo, coordenadas tridimensionales, velocidades o parámetros tales como tamaño o área); mientras que, en la estimación de identidad, el objetivo es una identificación semántica del objeto (por ejemplo, reconocimiento del edificio o clasificación del uso de la tierra).

Las técnicas de fusión de datos se pueden agrupar en diferentes categorías en función de las relaciones entre las fuentes de datos de entrada [11] o la integración de los diferentes tipos de sensores [17]:

- La fusión competitiva ocurre cuando dos o más sensores proporcionan observaciones en el mismo lugar y grado de libertad.
- La fusión complementaria ocurre cuando un conjunto de fuentes de información ofrece información diferente, complementaria a la misma característica geométrica en diferentes grados de libertad.
- La fusión cooperativa ocurre cuando un sensor depende de la información de la observación previa de otro sensor.

En el contexto de la teledetección [11], la fusión de imágenes, tradicionalmente de procedencia satelital, corresponde a la integración de datos derivados de instrumentos en forma de imágenes con características multiespaciales, multitemporales o multiespectrales para generar una imagen fusionada con más información de interés para la observación de la Tierra y la supervisión del medio ambiente. En las últimas décadas, se desarrollaron una variedad de métodos y técnicas para la fusión de imágenes o datos. Estos métodos se pueden clasificar típicamente en las siguientes tres categorías [11][17][16]:

- Fusión a nivel de observación, los datos brutos se combinan directamente. En el caso de fusión de imágenes se refiere a la fusión de imágenes preprocesadas píxel por píxel después del registro preciso de la imagen. Un ejemplo típico es producir el índice de vegetación de diferencia normalizada (NDVI) mezclando información de dos imágenes medidas en el infrarrojo cercano y longitudes de onda rojas para crear una nueva imagen en la que las características de la cubierta vegetal se mejoran en gran medida [11].

- Fusión a nivel de características, implica una extracción preliminar de características representativas de los datos de cada sensor individual. Este proceso requiere la detección, segmentación y extracción de características distintivas de los objetos en diferentes dominios antes del esquema de fusión. Las características involucradas en la fusión de imágenes pueden ser radiométricas (por ejemplo, intensidades) y geométricas (por ejemplo, formas, tamaños, altura). El núcleo de la fusión de imágenes en el nivel de características es la selección y extracción precisa de características de múltiples imágenes recopiladas de sensores de sensores remotos multiespectrales, hiperespectrales o de microondas. Esta categoría se ha aplicado ampliamente en el mapeo del uso y cobertura del suelo, como en el mapeo forestal [11].

- Fusión a nivel de decisión, se usa solo después de que ya se haya logrado una primera determinación de los atributos de interés del objetivo por cada sensor individual. Se refiere a la combinación de las características extraídas mediante el uso de reglas de decisión externas para reforzar interpretaciones y dilucidar los objetivos observados con múltiples aspectos. En fusión de imágenes se establecen reglas de decisión para las características extraídas de cada imagen individual antes de fusionar directamente las decisiones finales combinando diferentes extracciones e interpretaciones de características a través de uno o más operadores matemáticos para la toma de decisiones. Un ejemplo es crear un mapa de uso y cobertura del suelo mediante el uso de un árbol binario simple con los umbrales adecuados para los valores de reflectancia en diferentes longitudes de onda. Luego se detectan y extraen distintas clases de uso y cobertura de suelo desde imágenes de teledetección multiespectral, que pueden considerarse características distintivas de decisión sobre el área de interés hacia la fusión del clasificador final a través de un árbol de decisión para la fusión a nivel de decisión [11].

2.4 Segmentación semántica

En visión por computadora, la segmentación de imágenes es el proceso de partición de una imagen digital en múltiples segmentos (conjuntos de píxeles). El objetivo de la segmentación es simplificar y/o cambiar la representación de una imagen en algo que sea más significativo y más fácil de analizar [19]. La segmentación semántica en cambio implica etiquetar cada píxel de una imagen, o un punto en una nube de puntos, con su correspondiente etiqueta semántica [18], es decir, reconocer o comprender lo que hay en la imagen en el nivel de píxel o punto. La segmentación semántica ha sido muy estudiada tanto en visión por computadora como en teledetección.

En visión por computadora, la segmentación semántica consiste en asignar una etiqueta semántica o clase a cada región coherente de una imagen. Esto se puede lograr utilizando modelos de predicción densa a nivel de píxel que puedan clasificar cada píxel de la imagen [5]. Una comprensión semántica del entorno facilita las tareas en robótica, como la navegación, localización y conducción autónoma [18].

En teledetección, la segmentación semántica se denomina generalmente clasificación de imágenes, pero debe considerarse como una tarea integral que combina los problemas tradicionales de reconocimiento, detección y segmentación de etiquetas múltiples en un único proceso [20]. La segmentación semántica de las imágenes no RGB tiene numerosas aplicaciones, como la clasificación de la cobertura terrestre, la clasificación de la vegetación y la planificación urbana [19]. En teledetección el resultado de la segmentación semántica puede responder a las dos preguntas siguientes:

- ¿Qué categorías de cobertura terrestre se observan en la imagen?

- ¿dónde aparecen?

2.5 Aprendizaje profundo

La inteligencia artificial es el campo que estudia la síntesis y el análisis de agentes computacionales que actúan de forma inteligente. Un agente computacional es un agente cuyas decisiones sobre sus acciones pueden explicarse en términos de cálculo [21]. El aprendizaje automático, en inglés machine learning, es un subconjunto de la inteligencia artificial en el campo de la informática que generalmente a partir de técnicas estadísticas da a las computadoras la capacidad de "aprender" con datos sin estar explícitamente programadas. El término aprender esta referido a mejorar progresivamente el rendimiento en una tarea específica [22]. A pesar de que la tecnología de aprendizaje automático comenzó a potenciar muchos aspectos de la sociedad moderna, las técnicas convencionales tenían una capacidad limitada para procesar datos naturales en su forma original. La construcción de estos sistemas requería de una ingeniería cuidadosa y una considerable experiencia en el dominio del problema o aplicación para diseñar un extractor de características que transforme los datos en bruto a una representación interna adecuada o vector de características desde el cual el subsistema de aprendizaje, generalmente un clasificador, podría detectar o clasificar patrones en la entrada [23].

Según LeCun [23], el aprendizaje profundo, en inglés deep learning, son un conjunto de algoritmos dentro del aprendizaje automático, en ingles machine learning, que están compuestos de múltiples capas de procesamiento para aprender representaciones de datos con múltiples niveles de abstracción. Dicho aprendizaje de representación se refiere a un conjunto de métodos que permite alimentar una máquina con datos sin procesar y descubrir automáticamente las representaciones necesarias para la detección o clasificación. De acuerdo con lo expuesto los métodos de aprendizaje profundo son métodos de representación-aprendizaje con múltiples niveles de representación, obtenidos por composición de módulos simples, pero no lineales que transforman la representación en un nivel (comenzando con la entrada en bruto) en una representación a un nivel más alto, algo más abstracto. Con la composición de tales transformaciones, se pueden aprender funciones muy complejas. Para las tareas de clasificación, las capas más altas de representación amplifican aspectos de la entrada que son importantes para la discriminación y suprimen las variaciones irrelevantes. Una arquitectura de aprendizaje profundo es una pila multicapa de módulos simples, todos (o la mayoría) de los cuales están sujetos al aprendizaje, y muchos de los cuales computan asignaciones de entrada-salida no lineales. Cada módulo en la pila transforma su entrada para aumentar tanto la selectividad como la invariancia de la representación.

2.6 Redes neuronales convolucionales

Las redes neuronales artificiales, ANN por sus siglas en inglés, fueron propuestas por primera vez en los años 40 por los profesores del MIT Warren McCullough y Walter Pitts. Inspirados por los avances en neurociencia, propusieron crear un sistema informático que reprodujera cómo funciona el cerebro. La idea central se basaba en un sistema informático que funcionaba como una red interconectada. Actualmente, el aprendizaje profundo es considerado el estudio contemporáneo de las redes neuronales. La principal diferencia es que las redes neuronales utilizadas en el aprendizaje profundo suelen tener un tamaño mucho mayor [24].

En la figura 2 se puede observar el esquema de una red neuronal artificial totalmente conectada con tres funciones de entrada, dos capas ocultas y dos salidas. También se muestra el modelo de una de las neuronas artificiales: los productos de las entradas, $\boldsymbol{x_j}$, y pesos, $\boldsymbol{w}$, se suma con un sesgo, $\mathbf{b_j}$. A la suma le sigue una función de activación, $\boldsymbol{\sigma}$, que realiza un mapeo no lineal, que es la salida de la neurona [25].

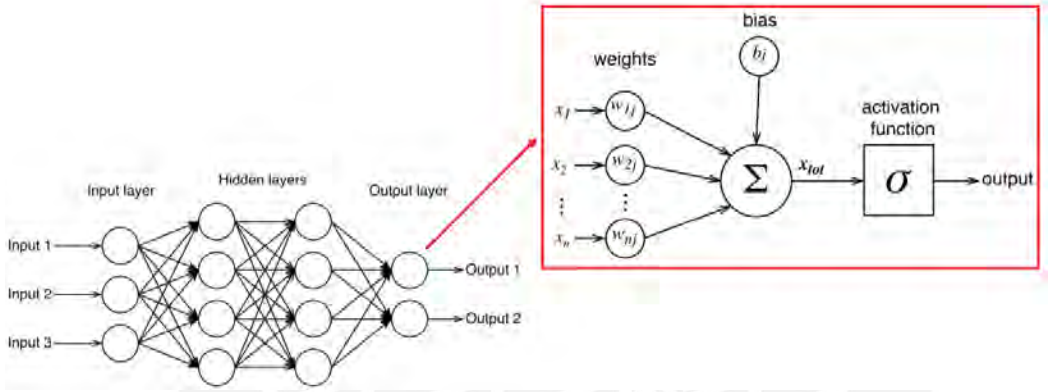

Figura 2. Esquema de una red neuronal artificial y modelo de una neurona **[25]**

Las redes neuronales convolucionales, CNN por sus siglas en inglés, son un tipo especializado de red neuronal para procesar datos que tienen una topología similar a una cuadrícula. Por ejemplo, en datos de series de tiempo, pueden considerarse como una cuadrícula 1-D tomando muestras a intervalos de tiempo regulares, y en datos de imágenes, se pueden considerar como una cuadrícula 2D de píxeles. Las redes convolucionales son simplemente redes neuronales que utilizan la operación matemática lineal de convolución en lugar de la multiplicación general de matrices en al menos una de sus capas [26].

Las CNN también son conocidas simplemente como redes convolucionales y es común encontrar en la bibliografía el término ConvNet para referirse a ellas. Las CNN tienen mejor desempeño sobre imagenes del mundo real en comparación con las clásicas ANN como los Perceptrones de múltiples capas (MLP) y esto se debe a que las CNN diferencias patrones en datos multidimensionales, es decir entienden el hecho de que los píxeles de imagen que están más cerca uno de otro está más relacionado que los píxeles que están más alejados. Las CNN organizan sus neuronas en tres dimensiones: ancho, altura y profundidad. Cada capa transforma su volumen de entrada 3D en un volumen de neuronas de salida 3D utilizando funciones de activación. Por ejemplo, en la figura 3, la capa de entrada roja contiene la imagen, por esta razón, su ancho y alto son las dimensiones de la imagen, y la profundidad seria tres, debido a los canales en rojo, verde y azul de la imagen [28].

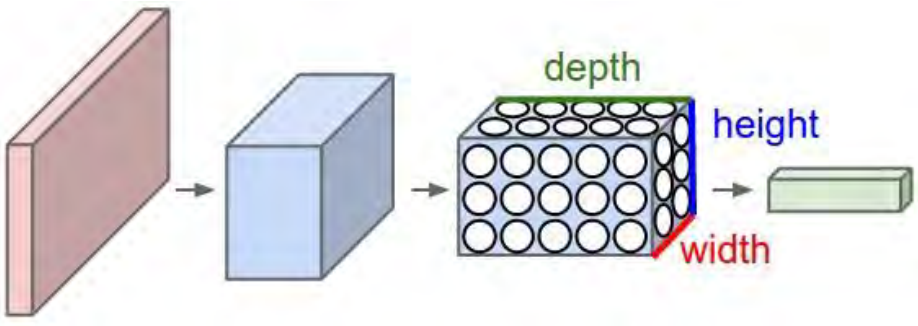

Figura 3. Esquema de una red neuronal artificial y modelo de una neurona **[27]**

Generalmente hay tres capas principales en una simple CNN: la capa de convolución (convolutional layer), la capa de agrupación (pooling layer) y la capa totalmente conectada (fully connected layer). Las capas de convolución filtran sus entradas de capa para encontrar características útiles dentro de esas entradas. En la parte superior de la figura 4 se muestra la operación de convolución 2-D en una imagen y su resultado. El núcleo de filtro tiene una profundidad que coincide con la profundidad de la entrada, 3 canales en este ejmplo. Aunque los tensores de entrada a una capa de convolución pueden tener cualquier número de canales. Es posible aplicar mas de un filtro con la finalidad de buscar diferentes patrones a la entrada. En la parte inferior de la figura 4 se observa la salida de la aplicación de 6 filtros convolucionales de 5x5x3 a una entrada de 32x32 pixeles y 3 canales [29].

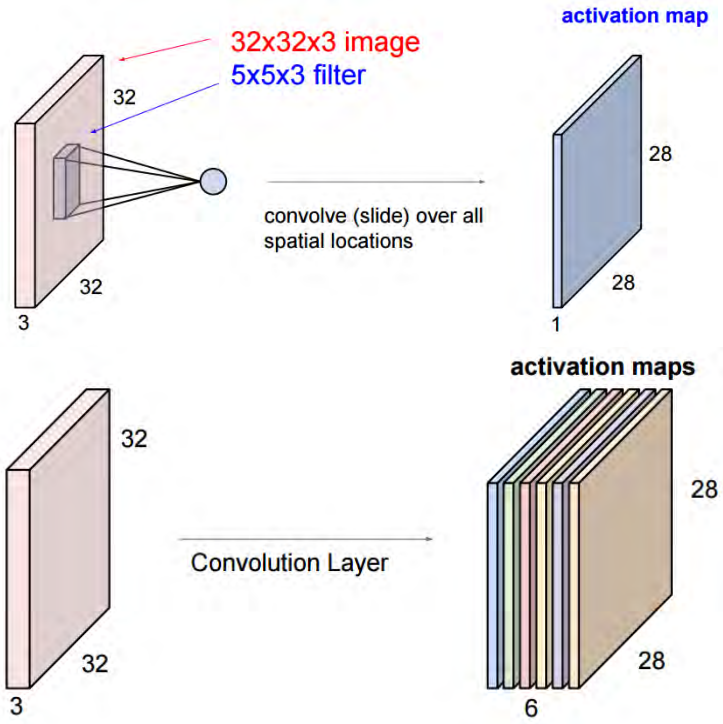

Figura 4. Aplicación de filtros convolucionales **[29]**

La capa de agrupación (pooling layer) se utiliza para reducir las dimensiones espaciales de los tensores de activación, pero no la profundidad del volumen, en una CNN. Una de sus grandes ventajas, que no tiene parámetros que aprender, es también su mayor desventaja, ya que la agrupación puede terminar simplemente desechando información importante. En la figura 5 se muestra el tipo más común de agrupación conocida como max-pooling layer, la cual consisten en deslizar una ventana, como una convolución normal, y luego en cada ubicación, establecer el mayor valor en la ventana como la salida [29]. Las neuronas de las capas totalmente conectada (fully connected layer) tienen conexiones completas con todas las activaciones en la capa anterior, como se ve en las redes neuronales regulares. Por lo tanto, sus activaciones se pueden calcular con una multiplicación de matrices seguida por un desplazamiento de sesgo [27].

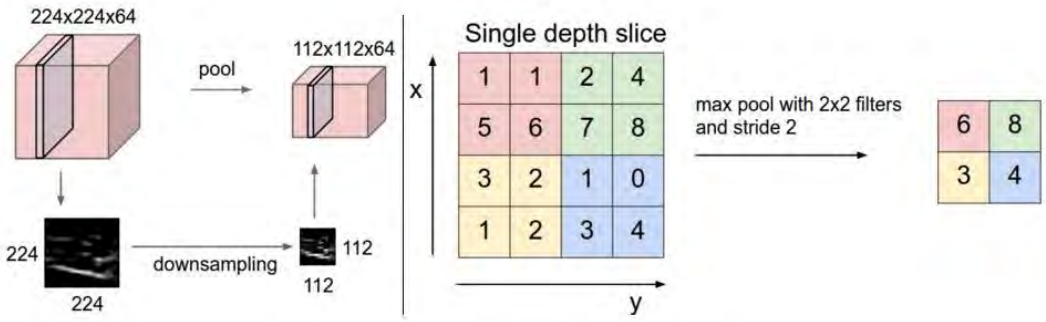

Figura 5. Esquema de una capa de agrupación máxima (max-pooling layer) **[29]**

Capítulo 3

Estado del Arte

3.1 Segmentación semántica

En visión por computadora, se ha desarrollado una variedad de algoritmos de segmentación semántica. En [30] [31] se presenta cuatro criterios diferentes para clasificar los algoritmos de segmentación según el tipo de datos en que operan y el tipo de segmentación que pueden producir:

- Clase permitida: La mayoría de los algoritmos funcionan con un conjunto fijo de clases, incluidos los clasificadores binarios que tienen primer plano vs fondo. Algunos algoritmos de segmentación son capaces de reconocer cuando no conocen una clase.

- Relación entre los píxeles: Algunos algoritmos pueden clasificar simultáneamente diferentes objetos en las mismas coordenadas de una sola imagen.

- Tipo de datos de entrada: Las imágenes pueden estar a color o en escala de grises, se puede excluir o incluir la profundidad, pueden ser de fuente simple, estéreo o incluso múltiples como en el caso del video.

- Modo de funcionamiento de los algoritmos: El estado de operación de la máquina clasificadora puede estar activa (robots) o pasiva (imagen recibida no puede ser influenciada). Los pasivos a su vez se pueden realizar de una manera completamente automática o trabajando en un modo interactivo.

Por lo general, la segmentación semántica se realiza con un clasificador entrenado con imágenes de tamaño fijo como entradas, para luego ser utilizado en imágenes de tamaño mayor a partir de regiones rectangulares llamadas ventanas deslizantes. Las redes neuronales pueden aplicar el enfoque de ventana deslizante de una manera muy eficiente manejando una red entrenada como una convolución y aplicando la convolución en la imagen completa. Existen otros enfoques como los campos aleatorios de Markov (MRF) y campos aleatorios condicionales (CRF) que toman la información de la imagen completa y la segmentan en un enfoque holístico [30].

Los enfoques aplicados en segmentación semántica se pueden dividir en aquellos que hacen uso del conocimiento del dominio, denominados enfoques tradicionales y aquellos basados en aprendizaje profundo [30].

3.1.1 Enfoques tradicionales

3.1.1.1 Características y métodos de preprocesamiento

En los enfoques tradicionales es muy importante la elección de las características tanto globales como locales. A continuación, se describen brevemente las principales características recogidas en [30]:

- Color de píxel: una imagen típica está en el espacio de color RGB, pero dependiendo del clasificador y del problema, otro espacio de color puede dar como resultado mejores segmentaciones. No se ha demostrado que ningún espacio de color sea superior a todos los demás en todos los contextos. Sin embargo, las opciones más comunes parecen ser RGB (simple y con gran soporte) y HSI (invariante a la iluminacion).

- Histograma de gradientes orientados -HoG, Histogram of oriented gradients-: este descriptor de características interpreta la imagen como una función discreta que mapea la posición (x, y) a un color. Transforma la imagen original en dos mapas de características de igual tamaño que representa el degradado. Estos mapas de características se dividen en parches y se calcula un histograma de las direcciones para cada parche.

- Transformación de característica en escala invariable -SIFT, Scale-invariant feature transform-: Este descriptor de características describe puntos clave en una imagen. Se toma un parche de imagen del tamaño 16x16 alrededor del punto clave. Este parche se divide en 16 partes distintas del tamaño 4x4. Para cada una de esas partes, un histograma de 8 orientaciones se calcula de forma similar a las características de HOG. Esto da como resultado un vector de características de 128 dimensiones para cada punto clave. SIFT es una característica global para una imagen completa.

- Bolsa de palabras visuales - BOV, Bag of visual Word-: También llamado bolsa de puntos clave, se basa en la cuantificación vectorial. Similar a las características

de HOG, las características de BOV son histogramas que cuentan el número de ocurrencias de ciertos patrones dentro de un parche de la imagen.

- Poselets: originalmente fueron utilizados para la estimación de la postura humana. Dependen de puntos clave extra añadidos, los cuales se pueden encontrar fácilmente para las clases de imágenes conocidas, como los humanos. Sin embargo, es difícil con otros objetos donde los anotadores humanos no conocen los puntos clave. Por consiguiente, los puntos clave deben elegirse para cada clase individual.

- Textons: se refiere a un bloque de visión mínimo. La literatura de visión por computadora no da una definición estricta para los texton, pero los detectores de bordes podrían ser un ejemplo. Se podría argumentar que las técnicas de aprendizaje profundo con Convolution Neuronal Networks (CNN) aprenden los textos en los primeros filtros.

Entre los métodos de prepocesamiento tenemos la reducción de la dimensionalidad, PCA por sus siglas en inglés, que permite reducir la resolución de la imagen de alta resolución a una variante de baja resolución. El problema con las imágenes de alta resolución es que tienen muchos píxeles que podrian generar más de un millón de características, y en el peor de los casos sin contener mucha más información, dificultando de esta manera el entrenamiento. Un problema de PCA es el hecho de que no distingue las diferentes clases. Esto significa que puede suceder que un conjunto perfectamente separable de vectores de características no se pueda separar en absoluto después de aplicar PCA [30].

3.1.1.2 Segmentación no supervisada

Los algoritmos de segmentación no supervisados no son semánticos, es decir detectan regiones consistentes o límites de región, pero se pueden usar en la segmentación supervisada como otra fuente de información o para refinar una segmentación. A continuación, se describen brevemente los principales algoritmos recpilados en [29]:

- Algoritmos de agrupamiento: se pueden aplicar directamente en los píxeles, cuando se proporciona un vector de características por píxel. Dos algoritmos de agrupación son k-means y el algoritmo mean-shift. El algoritmo k-means es un algoritmo de agrupación de propósito general que requiere como entrada la cantidad de clusters que se deben administrar. Inicialmente, coloca los k centroides aleatoriamente en el espacio de características. Luego, asigna cada punto de datos al centroide más cercano, mueve el centroide al centro del grupo y continúa el proceso hasta que se alcanza un criterio de detención. El algoritmo de cambio de media o mean-shift encuentra los centros del clúster inicializando los centroides en los puntos semilla al azar y desplazándolos iterativamente a la coordenada media dentro de un cierto rango, encontrando centros de clúster en las posiciones con una densidad local de puntos más alta.

- Segmentación de imágenes basada en grafos: estos algoritmos interpretan píxeles como vértices y un peso de borde es una medida de diferencia, como la diferencia de color. Elecciones de vecindarios de 4 (norte, este, sur oeste) u 8 (norte, noreste, este, sudeste, sur, suroeste, oeste, noroeste) son posibles para los bordes. Una forma de cortar los bordes es construir un árbol de expansión mínimo y eliminar los bordes por encima de un umbral. Este umbral puede ser constante, adaptado al gráfico o ajustado por el usuario. Después del paso de corte de borde, los componentes conectados son los segmentos.

- Algoritmo de camino aleatorio - Random Walks -: pertenece a los algoritmos de segmentación de imágenes basadas en grafos y generalmente funciona de la siguiente manera: los puntos de semillas se colocan en la imagen para los diferentes objetos en la imagen. De cada píxel individual, se calcula la probabilidad de alcanzar los diferentes puntos iniciales mediante una caminata aleatoria. Esto se hace tomando gradientes de imagen como las características de HOG. La clase del píxel es la clase de la cual se alcanzará un punto inicial con la probabilidad más alta. Al principio, este es un método de segmentación interactivo, pero puede extenderse para que no sea interactivo al utilizar otros resultados de métodos de segmentación como puntos iniciales.

- Modelos de contorno activo - ACMs, Active Contour Models-: Son algoritmos que segmentan las imágenes aproximadamente a lo largo de los bordes, pero también intentan encontrar un borde que sea liso. Esto se hace definiendo una llamada función de energía que se minimizará. Se describieron inicialmente en [31]. Los ACM pueden usarse para segmentar una imagen o para refinar la segmentación.

- Segmentación de línea divisoria de aguas – Wathershed-: este algoritmo toma una imagen en escala de grises y la interpreta como un mapa de altura. Los valores bajos son cuencas de captación y los valores más altos entre dos cuencas de captación vecinas es línea divisoria de aguas (wathershed). Las cuencas de captación deben contener lo que el desarrollador desea capturar. Esto implica que esas áreas deben estar oscuras en las imágenes en escala de grises. El algoritmo comienza a llenar las cuencas desde el punto más bajo. Cuando dos cuencas están conectadas, se encuentra una wathershed. El algoritmo se detiene cuando se alcanza el punto más alto [33].

3.1.1.3 Segmentación supervisada

- Bosques de decisión aleatoria - Random Decision Forest -: este tipo de clasificador aplica técnicas llamadas aprendizaje combinado (ensemble learning), donde se entrena a múltiples clasificadores y se usa una combinación de sus hipótesis. En este caso cada clasificador es un árbol de decisión. Un árbol de decisión es un árbol donde cada nodo interno usa una o más características para decidir en qué rama descender. Cada hoja es una clase. Una de las fortalezas de Random Decision Forests en comparación con muchos otros clasificadores como SVM y redes neuronales es que la escala de medida de las características (nominal, ordinal, intervalo, proporción) puede ser arbitraria. Otra ventaja de Random Decision Forests en

comparación con las SVM, por ejemplo, es la velocidad de entrenamiento y clasificación [34]. La aplicación a imágenes multiespectrales de teledetección se aplica en [35] para segmentar 4 clases (agua, campo, bosque y fuego), obtuviendose mejores resultados que SVM y redes neuronales.

- Máquinas de vectores de soporte – Support Vector Machine, SVM-: son una clase general de arquitectura de aprendizaje inspirada en la teoría del aprendizaje estadístico. Dado un dato de entrenamiento, el algoritmo de entrenamiento SVM obtiene el hiperplano de separación óptimo en términos de error de generalización. En [36] se realiza una segmentacion de imagenes multiespectrales con variantes de SVM con el objetivo de minimizar el número de puntos etiquetados necesarios para diseñar un clasificador.

3.1.1.4 Métodos probabilísticos

Aunque estos métodos son considerados en los enfoques tradicionales, también se están utilizando con éxito como complemento de los enfoques basados en redes neuronales artificiales.

- Campos aleatorios de Markov - MRF, Markov Random Fields-: son modelos gráficos probabilísticos no dirigidos. La idea general de MRF es asignar una variable aleatoria para cada característica y una variable aleatoria para cada píxel. Las variables aleatorias son condicionalmente independientes, dado su vecindario local. Estas independencias se pueden expresar con un grafo. En [36] se propone un modelo de fusión de múltiples capas para la segmentación adaptativa de imágenes de sensores remotos ópticos. El método realiza una segmentación de MRF de múltiples capas donde el etiquetado resultante se aplica para el entrenamiento automático de las capas individuales.

- Campos aleatorios condicionales - CRF, Conditional Random Fields-: Los CRF son MRF en los que todos los potenciales cliques, subgrafos totalmete conectados, están condicionados a las características de entrada. Esto significa que, en lugar de aprender la distribución P(y, x), la tarea se reformula para conocer la distribución P(y|x). Una consecuencia de esta reformulación es que los CRF necesitan muchos menos parámetros ya que la distribución de x no tiene que ser estimada. Otra ventaja de los CRF en comparación con los MRF es que no se debe realizar una suposición de distribución sobre x. En [38] se propone modelar el problema de segmentación mediante un CRF con capacitación discriminativa, empleando SVM para aprender los pesos de un conjunto informativo de descriptores de apariencia , mejorando la precisión promedio de la clase en un conjunto de datos que involucran imágenes satelitales urbanas de alta resolución.

3.1.1.5 Métodos de posprocesamiento

El posproceso refina una segmentación encontrada y elimina errores obvios. Por ejemplo, las operaciones morfológicas de apertura y cierre pueden eliminar el ruido. La operación de apertura es una dilatación seguida de una erosión que elimina pequeños segmentos. La operación de cierre es una erosión seguida de una dilatación que elimina pequeñas brechas en regiones que de otro modo estarían llenas. Otra forma de refinar la segmentación encontrada es ajustando la segmentación para que coincida con los bordes cercanos. Los modelos de contorno activo son otro ejemplo de un método de post-procesamiento [30].

3.1.2 Enfoques basados en aprendizaje profundo

La mayoría de las arquitecturas utilizadas se basan en los principios establecidos en [38], donde la aplicación de las redes completamente convolucionales, en inglés Fully Convolutional Networks (FCN), en problemas de segmentación semántica han demostrado lograr resultados impresionantes como en el conjunto de datos PASCAL VOC, muy popular para crear y evaluar algoritmos para clasificación de imágenes, detección de objetos y segmentación. La idea principal consiste en modificar la clasificación tradicional mediante las rede neuronales convolucionales, CNN, para que la salida no sea un vector de probabilidad sino un mapa de probabilidad. Por lo general una CNN estándar se utiliza como un codificador que extrae características, seguido de un decodificador que muestrea mapas de características a la resolución espacial original de la imagen de entrada. Finalmente, se obtiene un mapa de calor para cada clase.

A partir de FCN otras arquitecturas han obtenido buenos resultados en la segmentación semántica. Arquitecturas como DeepLab [39] y convoluciones dilatadas [41] han obtenido mejoras aumentando el campo de visión del codificador y eliminando las capas de acumulación para evitar cuellos de botella. Especificamente, en [41] la arquitectura se basa en el hecho de que las convoluciones dilatadas soportan la expansión exponencial del campo receptivo sin pérdida de resolución o cobertura. La predicción estructurada se ha investigado con modelos estructurados integrados tales como los campos aleatorios condicionales (CRF) dentro de la red profunda, CRFas-RNN [42][43]. La utilización de modelos graficos probabilísticos basados en CRF mejora la capacidad limitada de los algoritmos de aprendizaje profundo en el deliniado de objetos visuales.

Arquitecturas como ResNet [44] [45], redes neuronales recurrentes RRN [46] también proporcionaron nuevos conocimientos. En [44] se presenta con éxito un marco de aprendizaje residual para facilitar el entrenamiento de redes muy profundas, reformulando explícitamente las capas como funciones residuales de aprendizaje con referencia a las entradas de capa, en lugar de aprender funciones sin referencia. En [46] se incluye el uso de capas espacialmente recurrentes (ReNet) que capturan directamente los contextos globales y conducen a representaciones de características mejoradas. También se han aplicado autoencoders convolucionales [47] y se a investigado arquitecturas simétricas de codificador-decodificador como DeconvNet [48] y SegNet [49]. En [47] se presentas autoencoders apilados que integran propiedades discriminativas y generativas, proporcionando un enfoque unificado sin depender del muestreo durante el entrenamiento. DeconvNet adopta una red VGG de 16 capas en las capas convolucionales y en la red de deconvolución se compone por capas de desconvolución y desacoplamiento, que identifican las etiquetas de clase de píxeles y predicen las máscaras

de segmentación. Este método de segmentación identifica estructuras detalladas y maneja objetos en múltiples escalas de manera natural [48]. SegNet consiste en una red de codificador y una red de decodificador seguida de una capa de clasificación de píxeles. En el codificador se adopta una red VGG16 con 13 capas convolucionales y el decodificador asigna mapas de características del codificador de baja resolución a los mapas de características de resolución de entrada completa para la clasificación de píxeles. La novedad de SegNet radica en la manera en que el decodificador toma muestras de sus mapas de características de entrada de resolución más baja [49].

En Teledeccion, el aprendizaje profundo un campo de investigación muy activo. Uno de los primeros trabajos donde se ha utilizado con éxito CNN para clasificación y etiquetado denso de datos en aplicaciones de detección de camino es [50]. En [51] se ha demostrado que las características profundas basadas en CNN superan a los métodos tradicionales basados en extracción de características como Máquinas de Vector de Soporte para la clasificación de la cubierta terrestre. En el concurso Data Fusion 2015 [52] un modelo que utiliza aprendizaje profundo para la segmentación semántica superó a los métodos tradicionales [53], obtuviendo una precisión muy alta. En [54] se estableció que las redes profundas mejoran significativamente la línea base de SVM utilizado comúnmente en procesamiento de datos de teledetección [54]. Otro ejemplo de clasificación de datos de Teledetección utilizando un conjunto de CNN multiescalas es [55], y en [56] se ha mejorado resultados con la introducción de FCNs. Las arquitecturas completamente convolucionales aplicadas a imágenes de Teledetección al aprender a clasificar los píxeles (“Que”) permiten detectar diferentes tipos de cobertura terrestre y al predecir estructuras espaciales (“Donde"), pueden predecir las formas de los edificios o las curvas de los caminos [57].

3.2 Fusión de datos

En el contexto de la segmentación semántica, la fusión de datos de Teledetección generalmente se trabaja a un nivel de características y decisión.

3.2.1 Fusión a nivel de características

Extracción de características para las diferentes fuentes de entrada de datos seguida de fusión en el nivel de característica, este proceso puede incluir la concatenación y selección de características. Posteriormente, las características fusionadas se incorporarán a un esquema de entrenamiento supervisado para fines de clasificación [58].

3.2.2 Fusión a nivel de decisión

Utilización de diferentes métodos de procesamiento para cada fuente de entrada de datos y combinación las decisiones individuales del conjunto de clasificadores entrenados para obtener el resultado óptimo [59] [60] [61]. En [39] se propuso entrenar dos redes

neuronales separadas para cada fuente de entrada de datos y concatenar las características aprendidas en la última capa convolucional. En [7] se propuso una fusión a nivel de decisión combinando las salidas de probabilidad de dos clasificadores directamente de una manera fija. En cambio, en [4] se combino las salidas de los clasificadores con pesos de fusión a través de los CRF en los datos de entrenamiento y se aplicó redes neuronales totalmente convolucionales que aprenden a combinar información de capa gruesa con información de capa fina mientras usaban una CNN de resolución múltiple como extractor de características.

En [62] se comparó ambos procedimientos de fusión. Proponiendo una estrategia de promedios para la fusión a nivel de decisión y la corrección de características para la fusión a nivel de característica. Sus resultados mostraron que la corrección de características obtuvo ligeramente mejores resultados. Más adelante, en [7], se estudian sistemáticamente diferentes arquitecturas de red para la segmentación semántica de datos de teledetección multimodal de alta resolución y, más específicamente, se encuentra que la fusión tardía o nivel de decisión hace posible la recuperación de errores de datos ambiguos, mientras que la fusión temprana o nivel de características, permite un mejor aprendizaje de características conjuntas, pero a costa de una mayor sensibilidad a los datos faltantes.

Capítulo 4

Metodología

4.1 Introducción

La metodología se resume en los siguientes puntos:

- Recolección y selección de los datos: Se utilizará un conjunto de imágenes multimodales del ISPRS 2D Semantic Labeling Challenge. La segmentación semántica de las imágenes se realizará en seis clases.
- Entrenamiento del modelo: Para cumplir el objetivo de realizar una segmentación semántica de las imágenes, se entrenarán modelos de red neuronal basados en la arquitectura de red convolucional U-net.
- Evaluación del modelo: La evaluación se realizará principalmente en función a las reglas del ISPRS 2D Semantic Labeling Challenge y el estado del arte.

4.2 Modelos de línea base

En esta tesis dos estrategias de fusión de datos son aplicadas a partir de la línea base sobre la arquitectura de red profunda U-net.

La arquitectura U-net consiste en un camino de contracción (lado izquierdo) y un camino expansión (lado derecho). La estructura original se puede observar en la figura 6. El camino de contracción consiste en la aplicación repetida de dos convoluciones 3x3 cada una seguida por una unidad lineal rectificada (ReLU) y una operación de agrupación máxima de 2x2 con pasos de 2 para el submuestreo. En cada paso de submuestreo se duplica el número de canales de funciones. El camino de expansión consiste en un muestreo hacia arriba del mapa de características seguido de una deconvolución 2x2 que divide a la mitad número de canales de características, una concatenación con el mapa de características correspondiente recortado de la ruta de contratación, y dos convoluciones 3x3, cada una seguida por una ReLU. En la capa final, se utiliza una convolución 1x1 para asignar cada vector de características de 64 componentes al número deseado de clases. En total la red tiene 23 capas convolucionales [63].

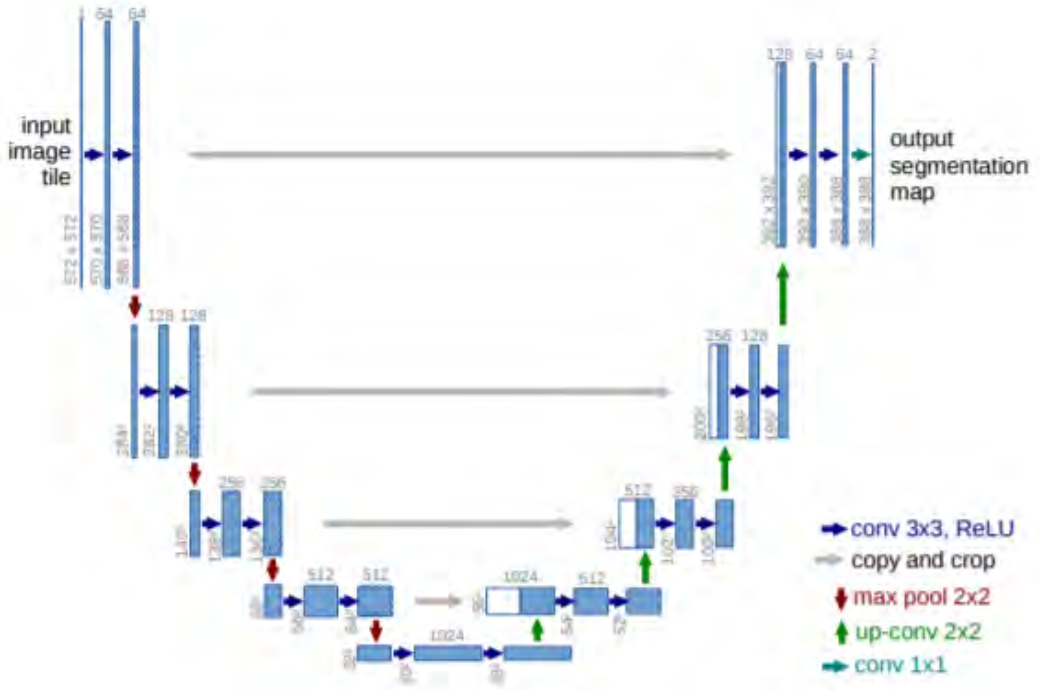

Figura 6. Arquitectura U-net [7]

En [63] la fusión de los datos se realiza mediante el apilamiento de canales y la Convolución fusionada tardía. El primer método implica la concatenación de múltiples modalidades de datos en canales y el aprendizaje de características combinadas desde el principio, mientras que el segundo método implica el aprendizaje individual para segmentar utilizando flujos separados, seguido de más representaciones fusionadas de aprendizaje.

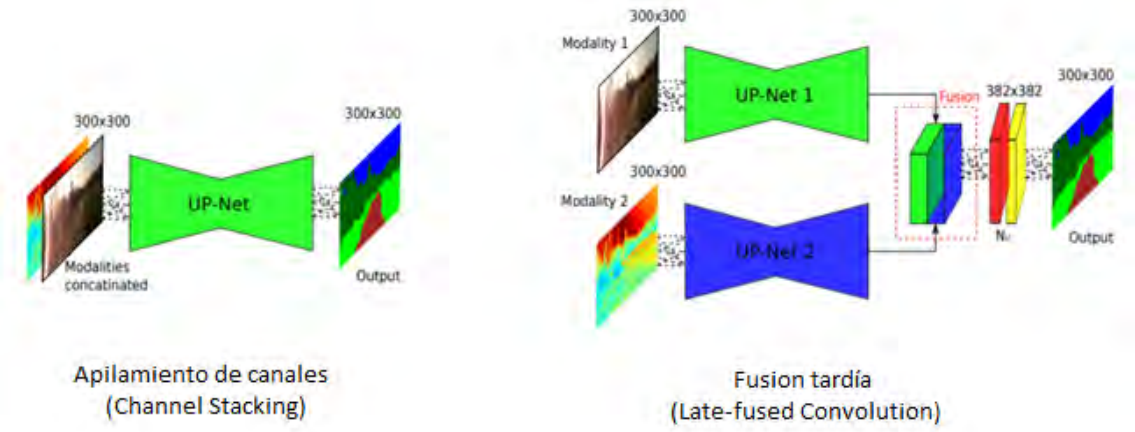

*Figura 7. Configuraciones de arquitectura de fusión profunda propuestas en [**64**]*

. En [6] se utiliza una arquitectura de red basadas en Segnet. La arquitectura se basa en una estructura de codificador-decodificador que produce una salida con la misma resolución que la entrada. El codificador está basado en las capas convolucionales de VGG-16. Se utiliza la fusión tardía mediante un módulo de corrección residual que consiste en una red neuronal convolucional residual que toma como entrada los últimos mapas de características de dos redes profundas. Cada red profunda es una red totalmente convolucional que ha sido entrenada por las imágenes RGB y la información auxiliar (NDVI y DSM).

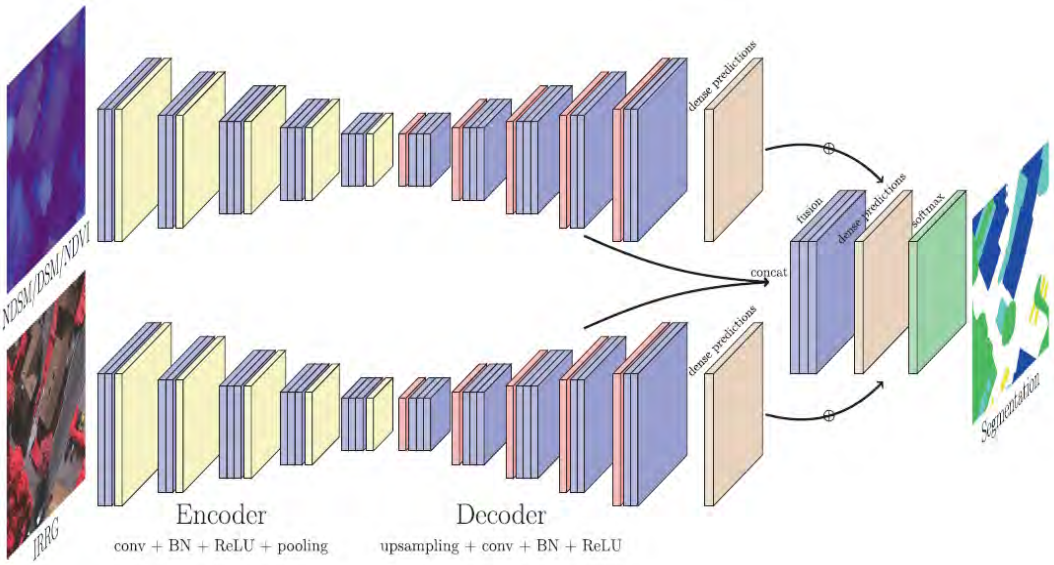

*Figura 8. Correccion residual para fusión tardía usando dos redes Segnet [**7**]*

4.3 Apilamiento de canales

El enfoque más simple para la fusión de datos investigado en esta tesis es apilar los 6 canales de entrada (RGB, IR, nDSM y NDVI) y utilizar la arquitectura U-net base. La figura 9 ilustra el enfoque de apilamiento de canales.

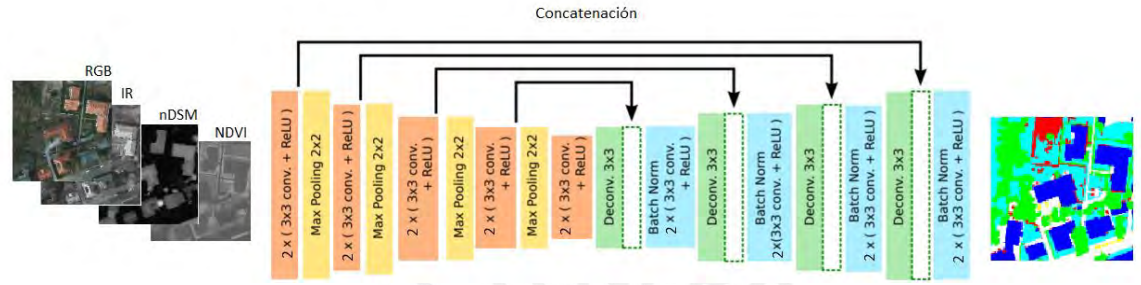

Figura 9.Arquitectura U-net con 6 canales de entrada

4.4 Fusión tardía

En este segundo enfoque, la fusión se da tardíamente mediante corrección residual [6], ver figura 10. Se agrega una red residual de 3 capas al final de las dos U-net. Se toma la entrada de los mapas de características anteriores y se aprende un término correctivo para aplicar a la predicción promediada.

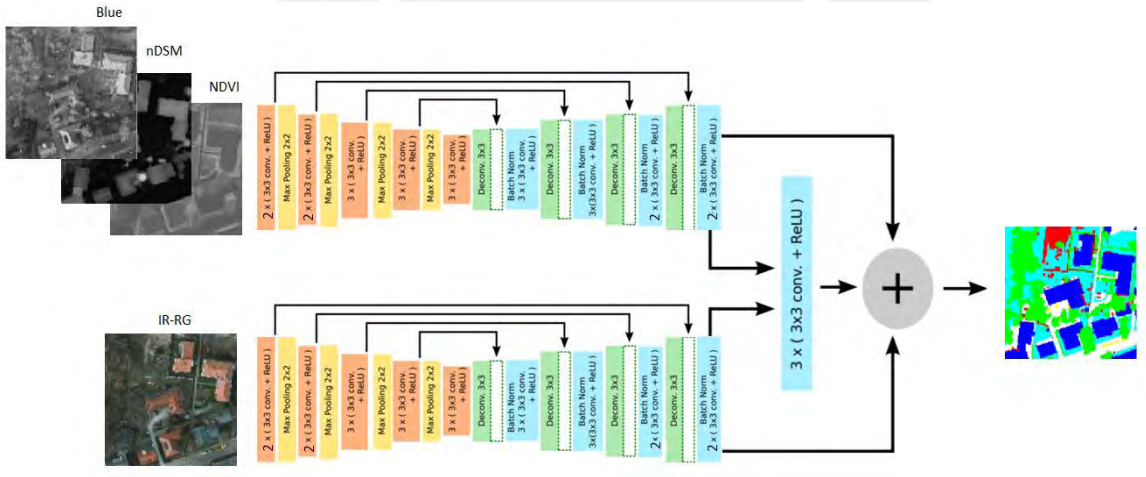

Figura 10. Arquitecturas U-net con 3 canales de entrada cada una.

4.5 Transferencia de aprendizaje

El enfoque de fusión tardía al tener arquitecturas con entradas de 3 canales permite utilizar arquitecturas pre-entrenadas. Mediante el enfoque de transferencia de aprendizaje se puede reutilizar modelos entrenados en grandes conjuntos de datos.

Basado en [1] se utilizara en el enfoque de fusión tardía una arquitectuta U-net con el encoder modificado en base al modelo pre-entrenado VGG16. En una de las ramas se utilizará los pesos obtenidos del entrenamiento en Imagenet mientras que en la otra rama se entrenará en base a los datos del ISPRS.

Capítulo 5

Experimentación y Resultados

5.1 Descripción de datos

El conjunto de datos para la experimentación es obtenido del Concurso de etiquetado semántico de ISPRS WG III/4. Este conjunto de datos de imágenes aerotransportadas consta de ortomosaicos y modelos de superficie obtenidos a partir de técnicas de coincidencia de imágenes. El área cubierta corresponde a escenarios urbanos de la ciudad de Potsdam en Alemania, la cual es una una ciudad histórica típica con grandes bloques de construcción, calles estrechas y una estructura de asentamiento densa.

El conjunto de datos de Potsdam consiste de 38 parches de 6000 x 6000 con 5cm de resolución, ver figura 11. Los datos se definen sobre la misma cuadricula y se organizaron en 4 grupos:

- Ortomosaicos RGBIR: archivos TIFF de 4 canales con resolución espectral de 8 bits.
- DSM normalizado: archivos JPG de un canal obtenidos a partir de los archivos DMS eliminando la altura de suelo en cada pixel.
- Etiquetas: también conocidos como ground-truth. Archivos TIFF con anotaciones de clase por cada pixel.
- Etiquetas sin borde: archivos TIFF con anotaciones de clase por cada pixel sin considerar bordes entre clases.

Figura 11. Mosaicos que conforman el área mapeada de Postdam **[12]**

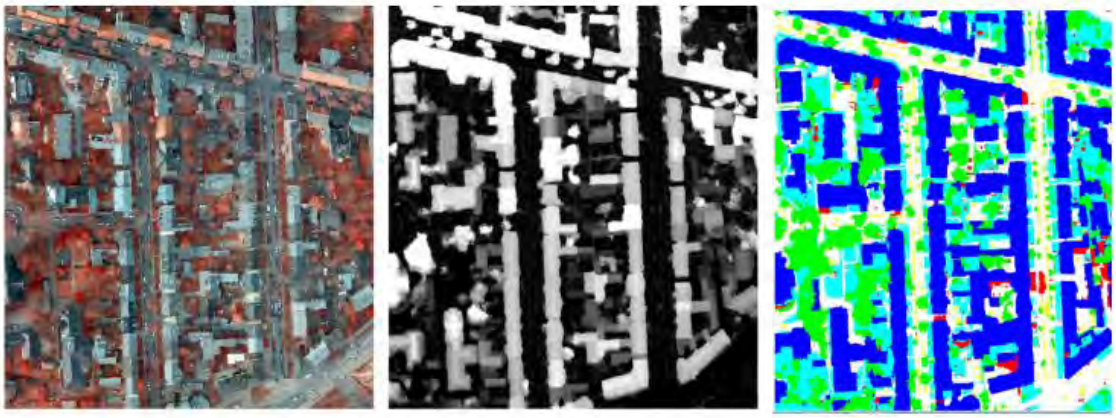

Figura 12. Ejemplos de datos: (a) Ortomosaico real, (b) DSM, y (c) ground truth **[12]**

5.1.1 Clases y anotaciones

El conjunto de datos se ha clasificado manualmente en seis de las clases de cobertura terrestre más comunes. Las clases definidas son:

- Superficies impermeables (Impervious - RGB: 255, 255, 255)
- Edificios (Building - RGB: 0, 0, 255)
- Vegetación baja (Low vegetation - RGB: 0, 255, 255)
- Árbol (Tree - RGB: 0, 255, 0)
- Autos (Car - RGB: 255, 255, 0)
- Desorden/fondo (Clutter - RGB: 255, 0, 0)

La clase de desorden/fondo incluye objetos que se ven muy diferentes las demás clases (por ejemplo, contenedores, canchas de tenis, piscinas) y que generalmente no son de interés en segmentación de escenas urbanas. En la figura 12(c) se puede observas la apariencia de las imágenes etiquetas a nivel de pixel de estas clases.

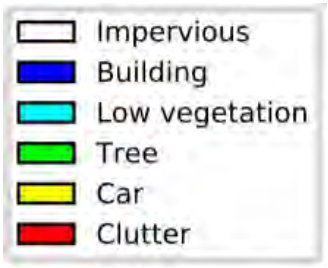

Figura 13. Las clases y sus correspodiente representación de color

5.1.2 División de los datos

Los datos de entrenamiento y validación se obtienen a partir de 24 ortomosaicos de 6000 x 6000 pixeles que tienen sus respectivas imágenes etiquetadas. Se utiliza un enfoque de validación cruzada de 3-fold para el entrenamiento y validación. Por limitaciones de hardware se utiliza ventanas de tamaño 256 x 256 con un solapamiento del 50% para el entrenamiento y ventanas de 2000 x 2000 para la validación.

Para los datos de pruebas se tienen 14 ortomosaicos con sus respectivas imágenes etiquetadas.

División	Imágenes utilizadas
Entrenamiento y validación	2-10, 3-10, 3-11, 3-12, 4-11, 4-12, 5- 10, 5-12, 6-10, 6-11, 6-12, 6-8, 6-9, 7-11, 7-12, 7-7, 7, 9, 2-11, 2-12, 4-10, 5-11, 6-7, 7-10, 7-8
Pruebas	2-13, 2-14, 3-13, 3-14, 4-13, 4-14, 4-15, 5-13, 5-14, 5-15, 6-13, 6-14, 6-15, 7-13

Tabla 1. Division de las imágenes del conjunto de datos.

5.1.3 NDVI y DSM

En el experimento se tiene como entradas imágenes RGB, pero además se busca trabajar con dos fuentes adicionales de datos (NDVI y DSM). En agricultura y forestación es común utilizar índices NDVI para realizar análisis sobre vegetación. Este se puede obtener a partir de información espectral de la banda roja e infraroja cercana de cámaras multiespectrales. El índice NDVI se encuentra entre [-1, 1] y se calcula mediante la ecuación 1.

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad (1)$$

Los modelos digitales de superficie (DSM) pueden ser obtenidos mediante fotogrametría o nube de puntos lidar. Esto brindan información de elevación de objetos sobre la superficie del suelo respecto al nivel del mar. Esta información se puede referenciar respecto al nivel del suelo tomando un modelo geodésico como WGS84 para obtener un DSM normalizado. De esta manera se obtiene información de altitud de las clases del experimento.

5.2 Configuraciones de red

En total, 2 diferentes configuraciones de red han sido entrenadas y evaluadas en la tesis con el objetivo de investigar cómo las modalidades de imagen adicionales individuales (ndvi y dsm) afectan el rendimiento de la clasificación. Las diferentes redes han sido entrenadas con un tamaño máximo de lote permitido por la memoria de la GPU. Las configuraciones de red se especifican en la Tabla 2.

Arquitectura	Configuraciones
U-net con apilamiento de canal	Total de parametros: 7,772,614 Parametros entrenables: 7,766,726
U-net con fusión tardia	Total de parametros: 55,852,242 Parametros entrenables: 41,126,440

Tabla 2. Configuraciones de red.

5.2.1 Parametros de implementación y entrenamiento

Todas las configuraciones de red se implementaron en redes neuronales usando el API Keras con soporte de GPU. Todas las sesiones de entrenamiento se ejecutaron en NVIDIA Tesla K80 con 24GB de memoria. Las redes se han entrenado para 100 épocas. Las capas convolucionales se han inicializado con el método de Xavier. La velocidad de aprendizaje inicial se estableció en 0,001 para ambas configuraciones y se redujo en un factor 10 cada 30 épocas. Solo el mejor modelo para cada configuración, con respecto a la pérdida de validación después de cada época, fue almacenado y utilizado para la prueba.

5.3 Evaluación de la metodología

La evaluación se realizará principalmente en función a las reglas del ISPRS 2D Semantic Labeling Challenge, las cuales se basan en el cálculo de matrices de confusión basadas en píxeles por mosaico y una matriz de confusión acumulada. A partir de esas matrices se derivan las métricas: recall, precisión y F1 score, y mediante la normalización

de la traza de la matriz de confusión se obtendrá la exactitud global (Overall Accuracy). Adicionalmente se utilizan las métricas aplicadas a segmentación semántica recopiladas en [39].

$$Pixel\ Accuracy\ (overall) = \frac{\sum_i n_{ii}}{\sum_i t_i}$$

$$Mean\ Accuracy = \frac{1}{nl}\sum_i \frac{n_{ii}}{t_i}$$

$$Mean\ IoU = \frac{1}{nl}\sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$$

$$Frequency\ Weighted\ IoU = \frac{1}{\sum_k t_k}\sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$$

Donde:

- nl: número de clases incluidas en la segmentación de la etiqueta (ground truth)
- nij: número de píxeles predecidos de la clase i que pertenecen a la clase j
- ti: número total de píxeles de la clase i en la segmentación de la etiqueta (ground truth)

5.4 Resultados Cuantitativos

Esta sección presenta los resultados cuantitativos de los experimentos basados en las imágenes de prueba anotadas. Se probaron 2 enfoques de fusión de datos basados en la arquitectura de red neuronal convolucional U-net.

Las pruebas iniciales se realizaron utilizando el enfoque de fusión de apilamiento de canales a la entrada de un modelo U-net. Se experimentó combinando los canales de entrada generandose 6 casos de prueba, detallados en la tabla 3. La segunda fase de pruebas se centra en el enfoque de fusión tardía sobre un modelo U-net con encoder basado en VGG 16.

En la tabla 4 se muestran los resultados promedios de las pruebas realizadas utilizando las métricas exactitud del conjunto (overall accuracy), exactitud promedio, F1 promedio, IoU promedio y IoU de frecuencia ponderada. Se puede observar en el enfoque por apilamiento de canales que al incrementar un canal con una fuente de datos adicional permite mejorar los resultados. En el caso de los canales infrarojo y ndvi se obtuvo resultados similares. En general los mejores resultados se obtuvieron con el enfoque de fusión tardia basada en U-net con encoder VGG16.dos se obtiene valores ligeramente mayores con la fusión tardía. En las tablas 5, 6 y 7 se pueden observar los resultados de los modelos por cada clase segmentada en función a las métricas exactitud, F1 e IoU respectivamente. Respecto a la métrica IoU se obtienen los mejores resultados con el

enfoque de fusión tardia en todas las clases excepto en vegetación baja. El mejor resultado se obtiene para edificaciones.

Casos de prueba	Descripción
U-net RGB	Modelo U-net básico con 3 canales de entrada (rojo,verde, azul).
U-net RGB-IR	Modelo U-net básico con 4 canales de entrada (rojo,verde, azul, infrarojo).
U-net RGB-NDVI	Modelo U-net básico con 4 canales de entrada (rojo,verde, azul, índice ndvi).
U-net RGB-nDSM	Modelo U-net básico con 4 canales de entrada (rojo,verde, azul, dsm normalizado).
U-net RGB-IR-nDSM- NDVI	Modelo U-net básico con 6 canales de entrada (rojo,verde, azul, infrarojo, dsm normalizado, índice ndvi).
U-net IR-RG	Modelo U-net básico con 3 canales de entrada (infrarojo, rojo,verde).
U-net VGG16 IR-RG	Modelo U-net con encoder VGG16 con 3 canales de entrada (infrarojo, rojo,verde).
U-net VGG16 Late Fusion	Modelo basado en dos U-net con encoder VGG16. El primer modelo con 3 canales de entrada (infrarojo, rojo,verde) y el segundo modelo con 3 canales de entrada (azul, dsm normalizado y ndvi)

Tabla 3. Casos de prueba realizados

.

Pruebas	Overall Accuracy	Mean Accuracy	Mean F1	Mean IoU	FW IoU
U-net RGB	0.8240	0.7530	0.7602	0.6444	0.7031
U-net RGB-IR	0.8460	0.7954	0.8084	0.6931	0.7338
U-net RGB-NDVI	0.8481	0.8013	0.8131	0.6986	0.7380
U-net RGB-nDSM	0.8284	0.7780	0.7894	0.6713	0.7062
U-net RGB-IR-nDSM- NDVI	0.8636	0.8023	0.8115	0.7081	0.7626
U-net IR-RG	0.8597	0.7913	0.8012	0.6986	0.7575
U-net VGG16 IR-RG	0.8320	0.8100	0.8204	0.7042	0.7135
U-net VGG16 Late Fusion	**0.8792**	**0.8412**	**0.8365**	**0.7439**	**0.7941**

Tabla 4. Resultados promedio.

	U-net RGB	U-net RGB-IR	U-net RGB-NDVI	U-net RGB-nDSM	U-net RGB-IR-nDSM-NDVI	U-net IR-RG	U-net VGG16 IR-RG	U-net VGG16 Late Fusion
Superficie imp.	0.8576	0.8835	0.8767	0.8833	0.9192	0.9111	0.9162	**0.9195**
Edificaciones	0.9549	0.9280	0.9278	0.9636	**0.9765**	0.9478	0.9186	0.9685
Vegetación Baja	0.7008	0.8065	**0.8172**	0.7159	0.7556	0.8132	0.7989	0.7729
Arboles	0.8475	0.8710	0.8638	0.8814	0.8502	0.8240	0.8719	**0.8886**
Carros	0.9002	0.8788	0.8821	0.8678	0.9368	**0.9423**	0.8207	0.9389
Otros	0.2574	0.4046	0.4404	0.3561	0.3754	0.3098	0.5339	**0.5586**

Tabla 5. Exactitud por clase.

	U-net RGB	U-net RGB-IR	U-net RGB-NDVI	U-net RGB-nDSM	U-net RGB-IR-nDSM-NDVI	U-net IR-RG	U-net VGG16 IR-RG	U-net VGG16 Late Fusion
Superficie imp.	0.8369	0.8445	0.8495	0.8499	0.8923	0.8737	0.8456	**0.9008**
Edificaciones	0.92.4	0.9199	0.9242	0.9152	0.9467	0.9411	0.9024	**0.9584**
Vegetación Baja	0.7605	0.8341	0.8352	0.7840	0.8008	0.8192	**0.8446**	0.8290
Arboles	0.7915	0.8236	0.8210	0.8084	0.8249	0.8258	0.8074	**0.8631**
Carros	0.8934	0.8759	0.8760	0.8855	0.9236	0.9269	0.8869	**0.9532**
Otros	0.3585	0.5527	0.5727	0.4937	0.4808	0.4206	**0.6356**	0.5144

Tabla 6. Métrica F1 por clase.

	U-net RGB	U-net RGB-IR	U-net RGB-NDVI	U-net RGB-nDSM	U-net RGB-IR-nDSM-NDVI	U-net IR-RG	U-net VGG16 IR-RG	U-net VGG16 Late Fusion
Superficie imp.	0.7196	0.7305	0.7384	0.7389	0.8055	0.7757	0.7325	**0.8195**
Edificaciones	0.8526	0.8516	0.8592	0.8437	0.8988	0.8888	0.8021	**0.9202**
Vegetación Baja	0.6136	0.7154	0.7172	0.6447	0.6678	0.6938	**0.7310**	0.7079
Arboles	0.6549	0.7001	0.6964	0.6784	0.7020	0.7032	0.6770	**0.7591**
Carros	0.8073	0.7792	0.7793	0.7945	0.8580	0.8638	0.7967	**0.9106**
Otros	0.2184	0.3819	0.4012	0.3278	0.3165	0.2663	**0.4659**	0.3462

Tabla 7. Métrica IoU por clase.

5.5 Resultados Cualitativos

En la figura 14 se observa resultados obtenidos por los diferentes modelos probados. Para el caso de la clase edificaciones y superficies impermeables se observan que los mejores resultados se dan para los dos enfoques de fusión. Tambien se puede observar los errores al segmentar la clase **otros** (clutter) debido a la dificultad que representa este complemento de elementos heterogéneos.

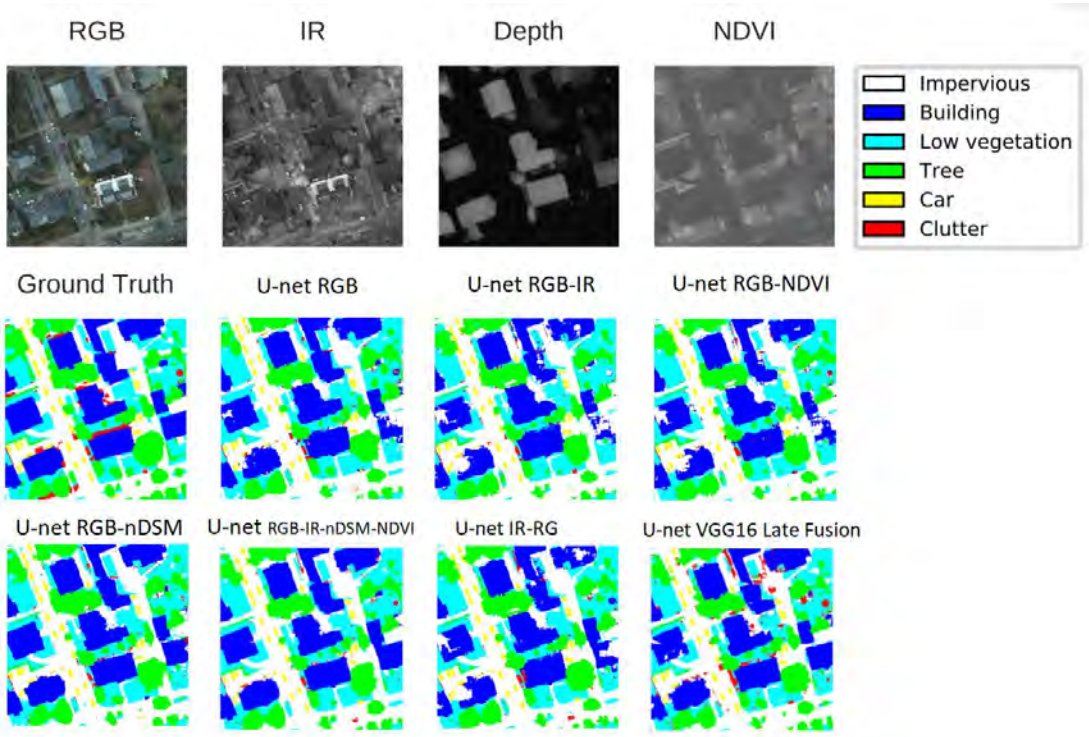

Figura 14. Resultados obtenidos al ejecutar los diferentes modelos.

Capítulo 6

Conclusiones

En el presente trabajo se resume la investigación sobre diferentes enfoques de fusión de datos utilizando aprendizaje profundo aplicando en el área de la teledetección urbana. Dos enfoques han sido investigados: apilamiento de canales de entrada y fusión tardía con corrección residual. Ambos enfoques se implementaron tomando como base la arquitectura U-net. El conjunto de datos ISPRS Potsdam contiene imágenes rgb, infrarrojo cercano y DSM que permite experimentar sobre el problema planteado.

La evaluación de las redes implementadas se realizó en forma cuantitativa y cualitativa. Se realizó entrenamientos de extremo a extremo y basados en transferencia de aprendizaje. Se obtuvo resultados aceptables para todas las clases comparado con el estado del arte. Estos resultados reflejan la necesidad de entrenar con un tamaño de batch mayor para poder mejorar los resultados lo cual involucra la utilización de hardware de alta prestaciones. Otra opción de mejora es la utilización de pesos de redes pre-entranadas de mayor complejidad como Resnet-50 o Inception.

Se pudo observar que la fuente de datos utilizadas fue etiquetada burdamente respecto a algunas clases como la de árboles y que la superficie DSM contiene varias incongruencias. Esta mala calidad en los datos y la escasez de nuevos datos necesarios para enfrentar estos tipos de problema limita la investigación por lo cual la generación de nuevos conjuntos de datos es una tarea pendiente.

Referencias

[1] A. Boulch, B. L. Saux y N. Audebert, «Unstructured point cloud semantic labeling using deep segmentation networks,» de *Eurographics Workshop on 3D Object Retrieval*, 2017.

[2] Anonymous, «United Nations Department of Economic and Social Affairs,» 2014.

[3] A. M. Ramiya, R. R. Nidamanuri y R. Krishnan, «Object-oriented semantic labelling of spectral–spatial LiDAR point cloud for urban land cover classification and buildings detection,» *Geocarto International,* vol. 31, nº 2, pp. 121-139, 2016.

[4] Y. Liu, S. Piramanayagam, S. Monteiro y S. E, «Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order crfs,» de *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Honolulu, USA, 2017.

[5] L. Wald, «Some terms of reference in data fusion,» *IEEE Transactions on geoscience and remote sensing,* vol. 37, nº 3, pp. 1190 - 1193, 1999.

[6] N. Audebert, B. Le Saux y S. Lefèvre, «Semantic segmentation of earth observation data using multimodal and multi-scale deep networks,» de *In Asian Conference on Computer Vision*, Cham, Springer, 2016, pp. pp. 180-196.

[7] N. Audebert, B. Le Saux y S. Lefèvre, «Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks,» *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 140, pp. 20-32, 2018.

[8] K. Liu y G. Mattyus, «Fast Multiclass Vehicle Detection on Aerial Images,» *IEEE Geosci. Remote Sensing Lett.,* vol. 12, nº 9, pp. 1938-1942, 2015.

[9] N. Audebert, B. Le Saux y S. Lefèvre, «Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images,» *Remote Sensing,* vol. 9, nº 4, p. 368, 2017.

[10] M. Dalla Mura, S. Prasad, F. Pacifici, P. Gamba, J. Chanussot y J. Benediktsson, «Challenges and opportunities of multimodality and data fusion in remote sensing,» *Proceedings of the IEEE,* vol. 103, nº 9, pp. 1585-1601, 2015.

[11] N. Chang y K. Bai, Multisensor Data Fusion and Machine Learning for Environmental Remote Sensing, Taylor & Francis, 2018.

[12] «ISPRS,» [En línea]. Available: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html. [Último acceso: 14 10 2018].

[13] G. Pajares, «Overview and Current Status of Remote Sensing Applications Based on Unmanned Aerial Vehicles (UAVs),» *Photogrammetric Engineering & Remote Sensing,* vol. 81, nº 4, pp. 281-330, 2015.

[14] DGAC, «Norma Técnica Complementaria 001,» 2015.

[15] DJI, «dji,» 2018. [En línea]. Available: https://www.dji.com/phantom-4. [Último acceso: Octubre 2018].

[16] D. L. Hall y L. J, «An introduction to multisensor data fusion,» *Proceedings of the IEEE,* vol. 85, nº 1, pp. 6-23, 1997.

[17] M. Schmitt y X. X. Zhu, «Data Fusion and Remote Sensing: An ever-growing relationship,» *IEEE Geoscience and Remote Sensing Magazine,* vol. 4, nº 4, pp. 6-23, 2016.

[18] R. Zhang, S. A. V. K. Candra y A. Zakhor, «Sensor Fusion for Semantic Segmentation of Urban Scenes,» de *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[19] L. Shapiro y G. Stockman, Computer Vision, New Jersey: Prentice-Hall, 2001.

[20] L. Mou y X. Zhu, «RiFCN: Recurrent Network in Fully Convolutional Network for Semantic Segmentation of High Resolution Remote Sensing Images,» *arXiv preprint,* 2018.

[21] D. Poole y A. Mackworth, Artificial Intelligence: foundations of computational agents, Cambridge University Press, 2010.

[22] A. Samuel, «Some Studies in Machine Learning Using the Game of Checkers,» *IBM Journal of Research and Development,* vol. 3, nº 3, p. 210–229, 1959.

[23] Y. Lecun, B. Y y H. G, «Deep learning,» *Nature,* vol. 521, nº 7553, p. 436, 2015.

[24] L. Capelo, Beginning Application Development with TensorFlow and Keras, Packt, 2018.

[25] C. Sundelius, «Deep Fusion of Imaging Modalities for Semantic Segmentation of Satellite Imagery,» Computer Vision Laboratory, Department of Electrical Engineering, Linköping University, Linköping, Sweeden, 2017.

[26] I. Goodfellow, Y. Bengio y C. Aaron, Deep Learning, MIT, 2016.

[27] A. Karpathy, «CS231n: Convolutional Neural Networks for Visual Recognition.,» [En línea]. Available: http://cs231n.github.io/convolutional-networks/. [Último acceso: 15 Octubre 2018].

[28] M. Sewak, M. R. Karim y P. Pujari, Practical Convolutional Neural Networks, Packt, 2018.

[29] I. Zafar, G. Tzanidou, R. Burton, N. Patel y A. Leonardo., Hands-On Convolutional Neural Networks with TensorFlow, Packt, 2018.

[30] M. Thoma, «A Survey of Semantic Segmentation,» *arxiv.org,* pp. 1-16, 2016.

[31] P. Student y J. Nahas, «A Survey of Artificial Neural Networks and Semantic Segmentation,» vol. 8, nº 5, p. 2590–2596, 2017.

[32] M. Kass, A. Witkin y T. D., «Snakes: Active contour models,» *International journal of computer vision,* vol. 1, nº 4, p. 321–331, 1988.

[33] R. J y M. A., «The watershed transform: Definitions, algorithms and parallelization

strategies,» *Fundam. Inform.,* vol. 41, nº 1-2, pp. 187-228, 2000.

[34] M. Pal, «Random forest classifier for remote sensing classification,» *International Journal of Remote Sensing,* vol. 26, nº 1, 2005.

[35] B. Lowe y A. Kulkarni, «Multispectral image analysis using random forest.,» 2015.

[36] P. Mitra, B. U. Shankar y S. K. Pal, «Segmentation of multispectral remote sensing images using active support vector machines,» *Pattern recognition letters,* vol. 25, nº 9, pp. 1067-1074, 2004.

[37] T. Sziranyi y M. Shadaydeh, «Segmentation of remote sensing images using similarity-measure-based fusion-MRF model,» *IEEE geoscience and remote sensing letters,* vol. 11, nº 9, pp. 1544-1548, 2014.

[38] M. Volpi, Ferrari y M., «Semantic segmentation of urban scenes by learning local class interactions,» de *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015.

[39] J. Long, E. Shelhamer y D. T, «Fully convolutional networks for semantic segmentation,» de *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[40] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy y A. Yuille, «Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs,» de *Proceedings of the International Conference on Learning Representations*, 2015.

[41] F. Yu y V. Koltun, «Multi-Scale Context Aggregation by Dilated Convolutions,» de *Proceedings of the International Conference on Learning Representations*, 2015.

[42] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang y P. Torr, «Conditional Random Fields as Recurrent Neural Networks,» de *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[43] T. P, «Higher Order Conditional Random Fields in Deep Neural Networks,» *arXiv,* 2015.

[44] K. He, X. Zhang, S. Ren y J. Sun, «Deep Residual Learning for Image Recognition,» de *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[45] Z. Wu, C. Shen y V. D. H. A, «High-performance Semantic Segmentation Using Very Deep Fully Convolutional Networks,» *arXiv,* 2016.

[46] Z. Yan, H. Zhang, Y. Jia, T. Breuel y Y. Yu, «Combining the Best of Convolutional Layers and Recurrent Layers: A Hybrid Network for Semantic Segmentation,» *arXiv,* 2016.

[47] J. Zhao, M. Mathieu, R. Goroshin y Y. LeCun, «Stacked What-Where Autoencoders,» de *Proceedings of the International Conference on Learning Representations*, 2015.

[48] H. Noh, S. Hong y H. B, «Learning Deconvolution Network for Semantic Segmentation,» de *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[49] V. Badrinarayanan, A. Kendall y R. Cipolla, «SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,» *arXiv,* 2015.

[50] V. Mnih y G. Hinton, «Learning to Detect Roads in High-Resolution Aerial Images,» de *Computer Vision ECCV 2010. Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2010.

[51] O. Penatti, K. Nogueira y J. Dos Santos, «Do deep features generalize from everyday

objects to remote sensing and aerial scenes domains?,» de *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015.

[52] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. Le Saux, A. B. A. C.-H.-T. H. S. Beaupere, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser y D. Tuia, «Processing of Extremely High-Resolution LiDAR and RGB Data,» de *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2016.

[53] A. Lagrange, B. Le Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, C. Herbin, H. Randrianarivo y M. Ferecatu, «Benchmarking classification of earth observation data: From learning explicit features to convolutional networks,» de *IEEE International Geosciences and Remote Sensing Symposium (IGARSS)*, 2015.

[54] K. Nogueira, O. Penatti y J. Dos Santos, «Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classification,» *arXiv,* 2016.

[55] W. D. S. Zhao, «Learning multiscale and deep representations for classifying remotely sensed imagery,» de *ISPRS Journal of Photogrammetry and Remote Sensing*, 2016.

[56] D. Marmanis, J. G. S. Wegner, K. Schindler, M. Datcu y U. Stilla, «Semantic Segmentation of Aerial Images with an Ensemble of CNNs,» de *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016.

[57] J. Sherrah, «Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery,» de *arXiv*, 2016.

[58] J. Marcello y A. E. F. Medina, «Evaluation of spatial and spectral effectiveness of pixel-level fusion techniques,» *IEEE Geosci. Remote Sens. Lett.,* 2013.

[59] W. Li, S. Prasad y J. Fowler, «Decision fusion in kernelinduced spaces for hyperspectral image classification,» *IEEE Geosci. Remote Sens. Lett.,* 2014.

[60] C. Pohl y J. Van Genderen, «Review article multisensory image fusion in remote sensing: concepts, methods and applications,» *Int. J. Remote Sens.,* 1998.

[61] J. Benediktsson, P. Swain y O. Ersoy, «Neural network approaches versus statistical methods in classification of multisource remote sensing data,» *IEEE Trans. Geosci. Remote Sens.,* 1990.

[62] N. Audebert, B. L. Saux y S. Lefèvre, «Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks,» 2016.

[63] A. Valada, G. L. Oliveira, T. Brox y W. Burgard, «Deep multispectral semantic scene understanding of forested environments using multimodal fusion,» de *International Symposium on Experimental Robotics*, 2016.

[64] O. Ronneberger, P. Fischer y T. Brox, « U-net: Convolutional networks for biomedical image segmentation,» de *International Conference on Medical image computing and computer-assisted intervention*, 2015.