



Más allá de RGB: Teledetección urbana de muy alta resolución con redes profundas multimodales

Nicolas Audebert a,b,* Bertrand Le Sauxa, Sébastien Lefèvre b

aONERA, El Laboratorio Aeroespacial
Francés F-91761 Palaiseau, Francia

bUniv. Bretagne-Sud, UMR 6074, IRISA, F-56000

Vannes, Francia

INFORMACIÓN DEL

Historial del artículo
Recibido el 28 de febrero de 2016
Revisado en forma revisada el 11 de noviembre de 2017
Disponible en línea el 23 de noviembre de 2017

Palabras clave:
Aprendizaje
profundo
Teledetección
Mapeo semántico
Fusión de
datos

ABSTRACTO

En este trabajo, investigamos varios métodos para tratar el etiquetado semántico de datos de teledetección multimodal de muy alta resolución. Especialmente, estudiamos cómo se pueden adaptar las redes profundas totalmente convolucionales para tratar con datos de teledetección multimodales y multiescala para el etiquetado semántico. Nuestras contribuciones son tres: (a) presentamos un enfoque multiescala eficiente para aprovechar tanto un gran contexto espacial como los datos de alta resolución, (b) investigamos la fusión temprana y tardía de datos Lidar y multiespectrales, (c) validamos nuestros métodos en dos conjuntos de datos públicos con resultados de última generación. Nuestros resultados indican que la fusión tardía permite recuperar errores a partir de datos ambiguos, mientras que la fusión temprana permite un mejor aprendizaje de características conjuntas, pero a costa de una mayor sensibilidad a los datos faltantes. © 2017 Sociedad Internacional de Fotogrametría y Teledetección, Inc. (ISPRS). Publicado por Elsevier B.V. Todos los derechos reservados.

1. Introducción

La teledetección se ha beneficiado mucho del aprendizaje profundo en los últimos años, principalmente gracias a los avances logrados en la comunidad de visión artificial sobre imágenes RGB naturales. De hecho, la mayoría de las arquitecturas de aprendizaje profundo diseñadas para la visión multimedia se pueden utilizar en imágenes ópticas de teledetección. Esto dio lugar a mejoras significativas en muchas tareas de teledetección, como la detección de vehículos (Chen et al., 2014), el etiquetado semántico (Audebert et al., 2016; Marmanis et al., 2016; Maggiori et al., 2017; Sherrah, 2016) y la clasificación de la cobertura/uso del suelo (Penatti et al., 2015; Nogueira et al., 2017). Sin embargo, estas mejoras se han limitado principalmente a las imágenes RGB tradicionales de 3 canales, ya sea en forma principal o por ejemplo, datos visión multiespectral o de otro (por ejemplo, una nube de puntos Lidar). Sin embargo, la adaptación de las redes profundas basadas en visión a estos datos más grandes no es trivial, ya que esto requiere trabajar con nuevas estructuras de datos que no comparten las mismas propiedades físicas y numéricas subyacentes. No obstante, todas estas fuentes proporcionan información complementaria que debe utilizarse conjuntamente para maximizar la precisión del etiquetado.

* Autor para correspondencia en: ONERA, Laboratorio Aeroespacial Francés, F-91761 Palaiseau, Francia.
Direcciones de correo electrónico: nicolas.audebert@onera.fr (N. Audebert), bertrand.le_saux@onera.fr (B. Le <https://doi.org/10.1016/j.isprsjprs.2017.01.011>).
© 2017, Sociedad Internacional de Fotogrametría y Teledetección, Inc. (ISPRS). Publicado por Elsevier B.V. Todos los derechos reservados.

En este trabajo presentamos cómo construir un modelo integral de aprendizaje profundo para aprovechar los datos de teledetección multimodal de alta resolución, con el ejemplo del etiquetado semántico de datos Lidar y multiespectrales en áreas urbanas. Nuestras aportaciones son las siguientes:

- Mostramos cómo implementar una red neuronal multiescala eficiente y totalmente convolucional profunda utilizando SegNet (Badrinarayanan et al., 2017) y ResNet (He et al., 2016).
- Investigamos la fusión temprana de datos de teledetección multimodal basados en el principio de FuseNet (Hazirbas et al., 2016). Demostramos que, si bien la fusión temprana mejora significativamente la semántica al permitir que la red aprenda conjuntamente características multimodales más fuertes, también induce una mayor sensibilidad a los datos faltantes o ruidosos.
- Investigamos la fusión tardía de datos de teledetección multimodal basados en la estrategia de corrección residual (Audebert et al., 2016). Demostramos que, aunque no funciona tan bien como la fusión temprana, la corrección residual mejora el etiquetado semántico y permite recuperar algunos errores críticos en píxeles duros.
- Validamos con éxito nuestros métodos en los conjuntos de datos del ISPRS Semantic Labeling Challenge de Cramer (2010), con resultados que sitúan nuestros métodos entre los mejores del estado del arte.

2. Trabajos conexos

El etiquetado semántico de los datos de teledetección se relaciona con la clasificación densa de las imágenes en forma de píxeles, que se denomina "semántica"



segmentación" o "comprensión de escenas" en la comunidad de visión artificial. El aprendizaje profundo ha demostrado ser eficaz y popular en esta tarea, especialmente desde la introducción de las redes totalmente convolucionales (FCN) (Long et al., 2015). Al reemplazar las capas estándar completamente conectadas de las redes neuronales convolucionales (CNN) tradicionales por capas convolucionales, fue posible densificar la salida de un solo vector de la CNN para lograr una clasificación densa a una resolución de 1:8. El primer modelo FCN ha sido rápidamente mejorado y declinado en varias variantes. Algunas mejoras se han basado en autocodificadores convolucionales con una arquitectura simétrica como SegNet (Badrinarayanan et al., 2017) y DeconvNet (Noh et al., 2015). Ambos utilizan una arquitectura de cuello de botella en la que los mapas de características se submuestrean para que coincidan con la resolución de entrada original, por lo que realizan predicciones píxeles a partir de píxeles a una resolución 1:1. Sin embargo, estos modelos han sido superados en imágenes multimedia por enfoques más sofisticados, como la eliminación de las capas de agrupación de la CNN estándar y el uso de convoluciones dilatadas (Yu y Koltun, 2015) para preservar la mayor parte de la información espacial de entrada, lo que dio lugar a modelos como el DeepLab multiescala (Chen et al., 2015) que realiza predicciones a varias resoluciones utilizando ramas separadas y produce predicciones 1:8. Por último, el auge de las redes residuales (He et al., 2016) pronto fue seguido por nuevas arquitecturas derivadas de ResNet (Pohlen et al., 2017; Zhao et al., 2017). Estas arquitecturas aprovechan la eficacia de última generación del aprendizaje residual para la clasificación de imágenes adaptándolas a la segmentación semántica, de nuevo por otro lado, también se ha investigado el aprendizaje profundo para el procesamiento de datos multimodal. Utilizando arquitecturas de doble flujo (Mnih et al., 2015), el procesamiento semántico puede beneficiarse de la retroalimentación de la estructura con detección de errores, como para el DCGAN (Engelbrecht et al., 2014). Además, el procesamiento de datos RGB-D (o 2.5D) tiene un interés significativo para las comunidades de visión por computadora y robótica, ya que muchos sensores integrados pueden detectar información óptica y de profundidad. Las arquitecturas relevantes incluyen dos redes paralelas CNN que se fusionan en las mismas capas totalmente conectadas (Eitel et al., 2015) (para la clasificación de datos RGB-D) y dos flujos CNN que se fusionan en el medio (Guo et al., 2016) (para la detección con la punta de los dedos). FuseNet (Hazirbas et al., 2016) extendió esta idea a redes totalmente convolucionales para la segmentación semántica de datos RGB-D mediante la integración de un esquema de fusión temprano en la arquitectura SegNet. Finalmente, el trabajo reciente de (Park et al., 2017) se basa en la arquitectura FuseNet para incorporar el aprendizaje residual y múltiples etapas de refinamiento para obtener datos RGB-D de predicciones multimodales de alta resolución. Estos modelos se pueden utilizar para el aprendizaje profundo de imágenes de varias fuentes de información, tales como las tareas de visión por computadora multimedia. La teledetección adoptó esas técnicas y las redes profundas se han utilizado a menudo para la observación de la Tierra. Desde el primer uso exitoso de la CNN basada en parches para la extracción de carreteras y edificios (Mnih y Hinton, 2010), muchos modelos se basaron en la canalización del aprendizaje profundo para procesar datos de teledetección. Por ejemplo, (Saito et al., 2016) realizaron predicciones de múltiples etiquetas (es decir, carreteras y edificios) en una sola CNN. (Vakalopoulou et al., 2015) extendió el enfoque a imágenes multiespectrales, incluidas las bandas visible e infrarroja. Aunque tiene éxito, el enfoque de clasificación basado en parches solo produce mapas generales, ya que un parche completo se asocia con una sola etiqueta. Los mapas densos se pueden obtener deslizando una ventana sobre toda la entrada, pero este es un proceso costoso y lento. Por lo tanto, para escenas urbanas con etiquetado denso en muy alta resolución, clasificación basada en superpíxeles

(Campos-Taberner et al., 2016) de imágenes de teledetección urbana fue un enfoque exitoso que clasificó regiones homogéneas para producir mapas densos, ya que combina el enfoque basado en parches con una presegmentación no supervisada. Gracias a la concatenación de características alimentadas al clasificador SVM, (Audebert et al., 2016; Lagrange et al., 2015) lograron extender este marco al procesamiento multiescala utilizando un enfoque piramidal basado en superpíxeles. Otros enfoques para la segmentación semántica incluyeron la predicción basada en parches con características profundas y expertas mixtas (Paisitkriangkrai et al., 2015), que utilizó conocimientos previos e ingeniería de características para mejorar las predicciones de la red profunda. Liu et al. (2016) han investigado las predicciones de CNN multiescala con una pirámide de imágenes utilizada como entrada para un conjunto de CNN para la clasificación del uso de la cobertura del suelo, mientras que (Chen et al., 2014) utilizaron bloques convolucionales variables para procesar múltiples escalas. Últimamente, el etiquetado semántico de las imágenes aéreas se ha trasladado a los modelos FCN (Sherrah, 2016; Maggiori et al., 2017; Volpi y Tuia, 2017). De hecho, las redes totalmente convolucionales como SegNet o DeconvNet son muy adecuadas para el mapeo semántico de datos de observación de la Tierra, ya que pueden capturar las dependencias espaciales entre clases sin la necesidad de un procesamiento previo, como una segmentación de superpíxeles, y producen predicciones de alta resolución. Estos enfoques se han extendido nuevamente para el procesamiento sofisticado de múltiples escalas en Marmanis et al. (2016) utilizando tanto el costoso enfoque piramidal con un FCN como la salida de múltiples resoluciones inspirada en Chen et al. (2015). Las escalas múltiples permiten que el modelo capture relaciones espaciales para objetos de diferentes tamaños, desde grandes arreglos de edificios hasta árboles individuales, lo que permite una mejor comprensión de la escena. Para imponer una mejor regularidad espacial, se han utilizado modelos gráficos probabilísticos como el postprocesamiento de campos aleatorios condicionales (CRF) para modelar las relaciones entre píxeles vecinos e integrar estos prioris en la predicción (Lin et al., 2015; Sherrah, 2016; Liu et al., 2017), aunque esto añade costosas composiciones que ralentizan significativamente la ejecución. De hecho, se ha utilizado sensores complementarios en la misma escena para medir varias propiedades que proporcionan diferentes perspectivas sobre la estructura espacial de las predicciones. Sin embargo, estos esquemas explícitos de regularización espacial son costosos. En este trabajo pretendemos demostrar que estos no son necesarios para obtener resultados de etiquetado semántico que sean competitivos con el estado del arte. Paisitkriangkrai et al. (2015) fusionaron datos ópticos y Lidar mediante la concatenación de características profundas y expertas como entradas para bosques aleatorios. De manera similar, Liu et al. (2017) integran características expertas de los datos auxiliares (Lidar y NDVI) en su CRF de orden superior para mejorar la red principal de clasificación óptica. El trabajo de Audebert et al. (2016) investigó la fusión tardía de datos Lidar y ópticos para la segmentación semántica utilizando la fusión de predicciones que no requirió ingeniería de características mediante la combinación de dos clasificadores con un enfoque de extremo a extremo de aprendizaje profundo. Esto también se investigó en Audebert et al. (2017) para fusionar el óptico y OpenStreetMap para el etiquetado semántico. Durante el Data Fusion Contest (DFC) 2015, Lagrange et al. (2015) propusieron un esquema de fusión temprana de datos ópticos y Lidar basado en una pila de características profundas para la clasificación basada en superpíxeles de datos urbanos de teledetección. En el DFC 2016, Mou y Zhu (2016) realizaron la clasificación de la cobertura del suelo y el análisis del tráfico mediante la fusión de datos multiespectrales y de video en una etapa tardía. Nuestro objetivo es estudiar a fondo los enfoques de aprendizaje profundo de extremo a extremo para la fusión de datos multimodal y comparar las estrategias de fusión temprana y tardía para esta tarea.

3.Descripción del método

3.1.Segmentación semántica de imágenes

aéreo etiquetado semántico de las imágenes aéreas requiere una clasificación densa de píxeles de las imágenes. Por lo tanto, podemos utilizar la arquitectura FCN para lograr esto, utilizando las mismas técnicas que son efectivas para las imágenes naturales. En este trabajo elegimos el modelo SegNet (Badrinarayanan et al., 2017) como red base. SegNet se basa en una arquitectura de codificador-decodificador que produce una salida con la misma resolución que la entrada, como se ilustra en la Fig. 1. Esta es una propiedad deseable ya que queremos etiquetar los datos a la resolución de la imagen original, por lo tanto, producir mapas a una resolución 1:1 en comparación con la entrada. SegNet permite realizar dicha tarea, ya que el decodificador es capaz de aumentar la muestreo de los mapas de características mediante la operación de desagrupación. También comparamos esta red base con una versión modificada de la red ResNet-34 (He et al., 2016) adaptada para la segmentación semántica.

El codificador de SegNet se basa en las capas convolucionales de VGG-16 (Simonyan y Zisserman, 2014). Tiene 5 bloques de convolución, cada uno de los cuales contiene 2 o 3 capas convolucionales de kernel 3x3 con un relleno de 1 seguido de una unidad lineal rectificada (ReLU) y una normalización por lotes (BN) (Ioffe y Szegedy, 2015). A cada bloque de convolución le sigue una capa de agrupación máxima de tamaño 2x2. Por lo tanto, al final del codificador, los mapas de características son cada uno $M2 \times 4/2$ donde la imagen original tiene una resolución $W \times H$. El decodificador realiza tanto el sobremuestreo como la clasificación. Aprende a restaurar la resolución espacial completa mientras transforma los mapas de características codificadas en las etiquetas finales. Su estructura es simétrica con respecto al codificador. Las capas de agrupación se sustituyen por capas de agrupación, como se describe en Zeiler y Fergus (2014). La agrupación reubica la activación de los mapas de entidades más pequeños en un mapa con muestreo ascendente rellenado con cero. Las actividades se reubican en los índices calculados en las fases de agrupación,

es decir, el argmax del max-pooling (véase la Fig. 2). Esta desagrupación permite reemplazar las características altamente abstractas del decodificador a los puntos de prominencia de los mapas de características geométricas de bajo nivel del codificador. Esto es especialmente efectivo en objetos pequeños que, de otro modo, se extraviarían o clasificarían incorrectamente. Después de la agrupación, los bloques de convolución densifican los mapas de entidades dispersas. Este proceso se repite hasta que los mapas de características alcanzan la resolución de entrada. De acuerdo con He et al. (2016), el aprendizaje residual ayuda a entrenar redes deep y logró un nuevo rendimiento de clasificación de última generación en ImageNet, así como resultados de segmentación semántica de última generación en el conjunto de datos COCO. En consecuencia, también comparamos nuestros métodos aplicados a la arquitectura ResNet-34. El modelo ResNet-34 utiliza cuatro bloques residuales. Cada bloque se compone de 2 o 3 convoluciones de núcleos de 3x3 y la entrada del bloque se suma a la salida mediante una conexión de salto. Al igual que en SegNet, las convoluciones van seguidas de las capas de normalización por lotes y activación de ReLU. La conexión de omisión puede ser la identidad si las formas del tensor coinciden, o una convolución de 1 x 1 que proyecta los mapas de características de entrada en el mismo espacio que los de salida si cambia el número de planos de convolución. En nuestro caso, para mantener la mayor parte de la resolución espacial, mantenemos la agrupación máxima inicial de 2 x 2, pero reducimos el paso de todas las convoluciones a 1. Por lo tanto, la salida del modelo ResNet-34 es un mapa de predicción 1:2. Para aumentar la muestra de este mapa a resolución completa, realizamos un unpooling seguido de un bloque convolucional estándar.

Por último, ambas redes utilizan una capa softmax para calcular la pérdida logística multinomial, promediada a lo largo de todo el parche:

$$\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(z_i)}{\sum_{j=1}^y \exp(z_j)} \quad (1)$$

donde N es el número de píxeles de la imagen de entrada, k el número de clases y, para un píxel especificado i, y^i denotan su etiqueta y (z_1, \dots, z_k) el vector de predicción. Esto significa que solo minimizamos el

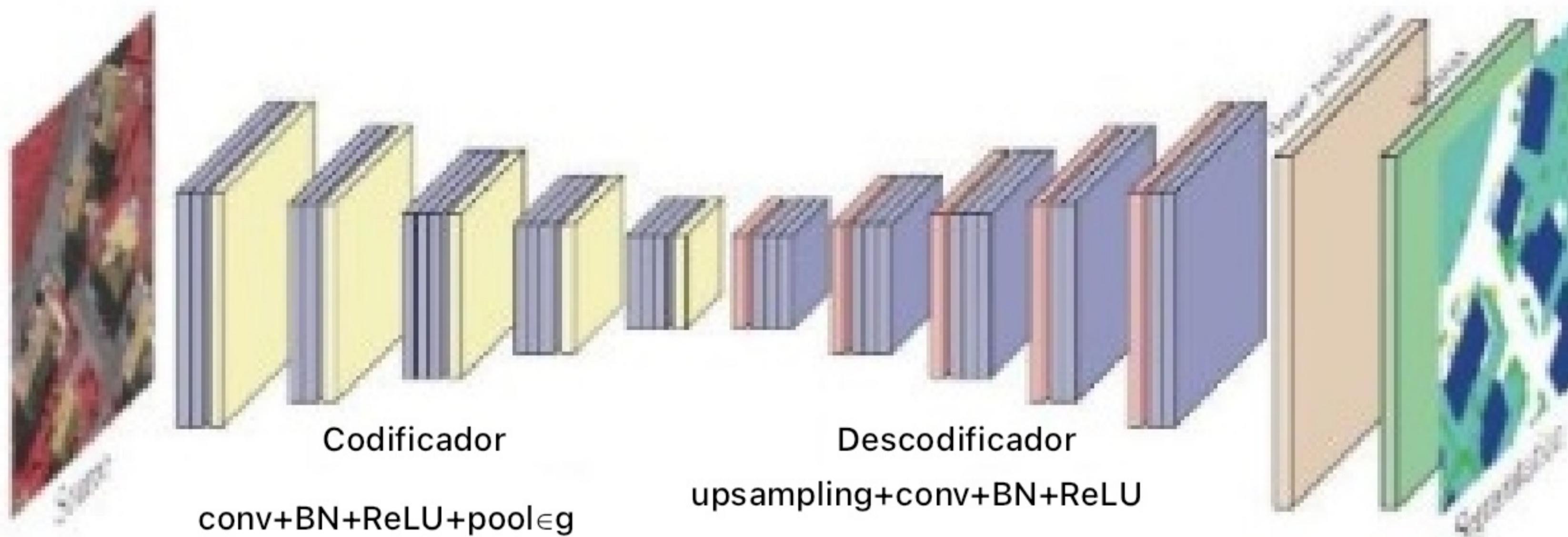


Figura 1. Arquitectura SegNet (Badrinarayanan et al., 2017) para el etiquetado semántico de datos de teledetección. Consulte el texto para obtener explicaciones más detalladas de cada capa.

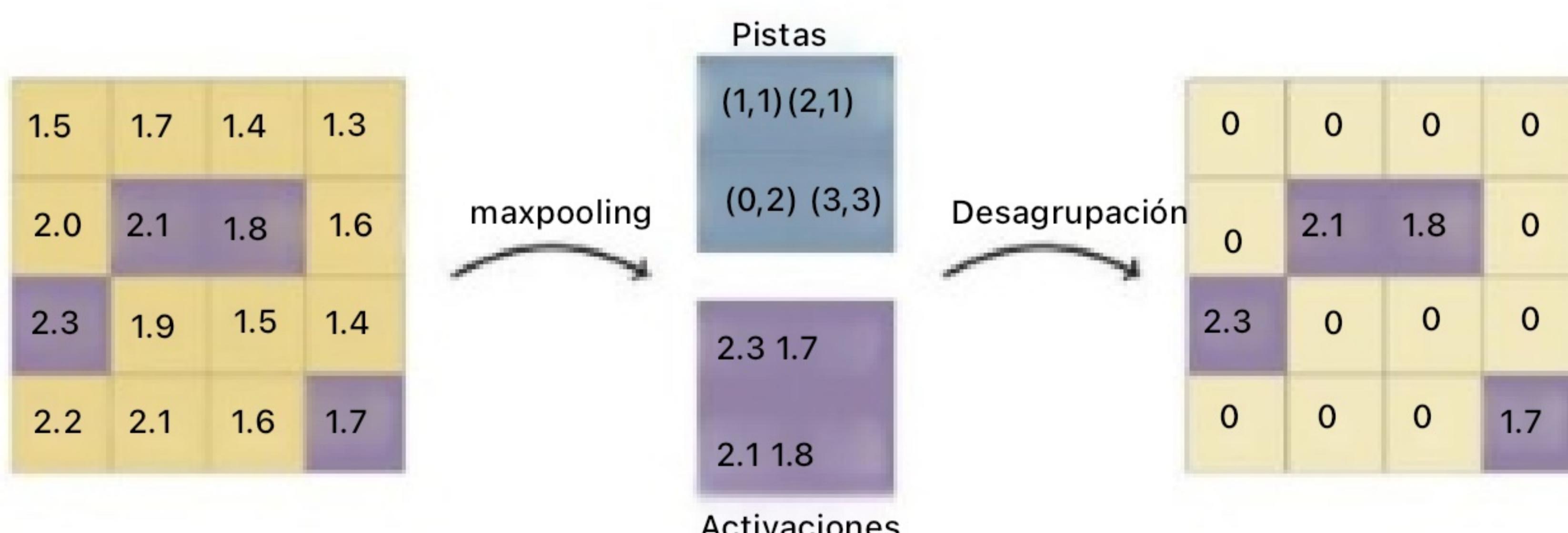


Figura 2. Ilustración de los efectos de las operaciones de agrupación máxima y desagrupación en un mapa de entidades de 4 x 4.

Pérdida de clasificación de píxeles $\frac{2018}{2018} \times \frac{20}{20} \times \frac{32}{32}$ ninguna regularización espacial, ya que será aprendida por la red durante el entrenamiento. No utilizamos ningún procesamiento posterior, por ejemplo, un CRF, ya que ralentizaría significativamente los cálculos con poca o ninguna ganancia.

3.2. Aspectos multiescala

A menudo, el procesamiento multiescala se aborda utilizando un enfoque piramidal: diferentes tamaños de contexto y diferentes resoluciones se alimentan como entradas paralelas a uno o varios clasificadores. Nuestra primera contribución es el estudio de un enfoque alternativo que consiste en ramificar nuestra red profunda para generar predicciones de salida a varias resoluciones. Cada salida tiene su propia pérdida que se retroalimenta a las capas anteriores de la red, de la misma manera que cuando se realiza una supervisión profunda (Lee et al., 2015). Este es el enfoque que se ha utilizado para la arquitectura DeepLab (Chen et al., 2015).

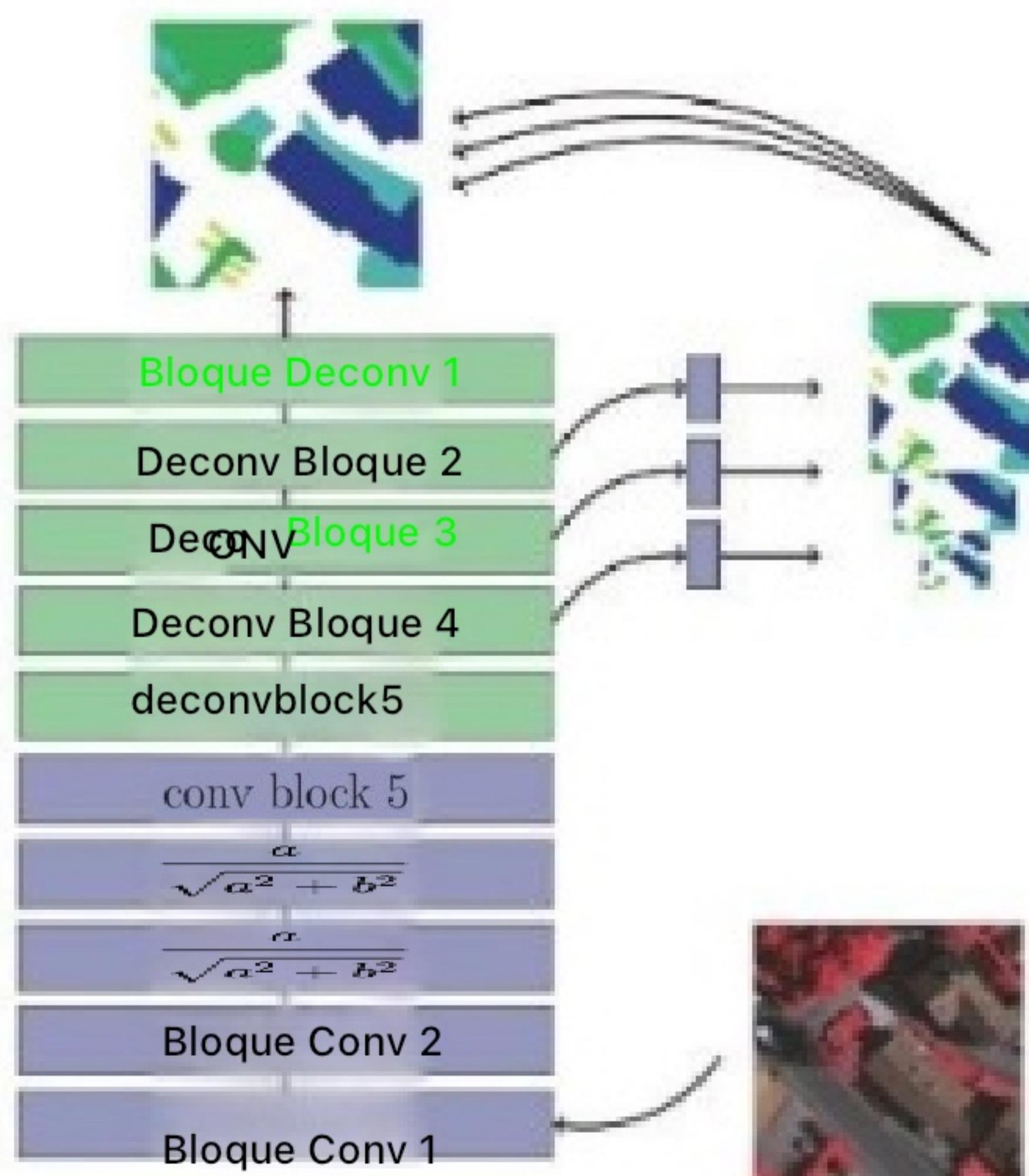
Por lo tanto, considerando nuestro modelo SegNet, no solo predecimos un mapa semántico a resolución completa, sino que también ramificamos el modelo anteriormente en el decodificador para predecir mapas de resoluciones más pequeñas. Después del bloque convolucional p-ésimo del decodificador, añadimos una capa de convolución que proyecta los mapas de características en el espacio de etiquetas, con una resolución de $20w \times 22$, como se ilustra en la Fig. 3. A continuación, esos mapas más pequeños se interpolan a resolución completa y se promedian para obtener el mapa semántico final de resolución completa.

Sea P_{full} la predicción de resolución completa, P_{down_d} las predicciones en el factor de escala descendente d y f_d la interpolación bilineal que sobremuestrea un mapa en un factor d . Por lo tanto, podemos agregar nuestras predicciones de múltiples resoluciones usando una suma simple (con, por ejemplo, si usamos cuatro escalas):

$$P_{full} = \sum_{d=0,2,4,8} f_d(P_{down_d}) = P_0 + f_2(P_2) + f_4(P_4) + f_8(P_8). \quad (2)$$

Durante la retropropagación, cada rama recibirá dos contribuciones:

- La contribución proveniente de la pérdida de la predicción promedio.
- La contribución proviene de su propia pérdida a escala reducida.



a) Predicción multiescala mediante SegNet.

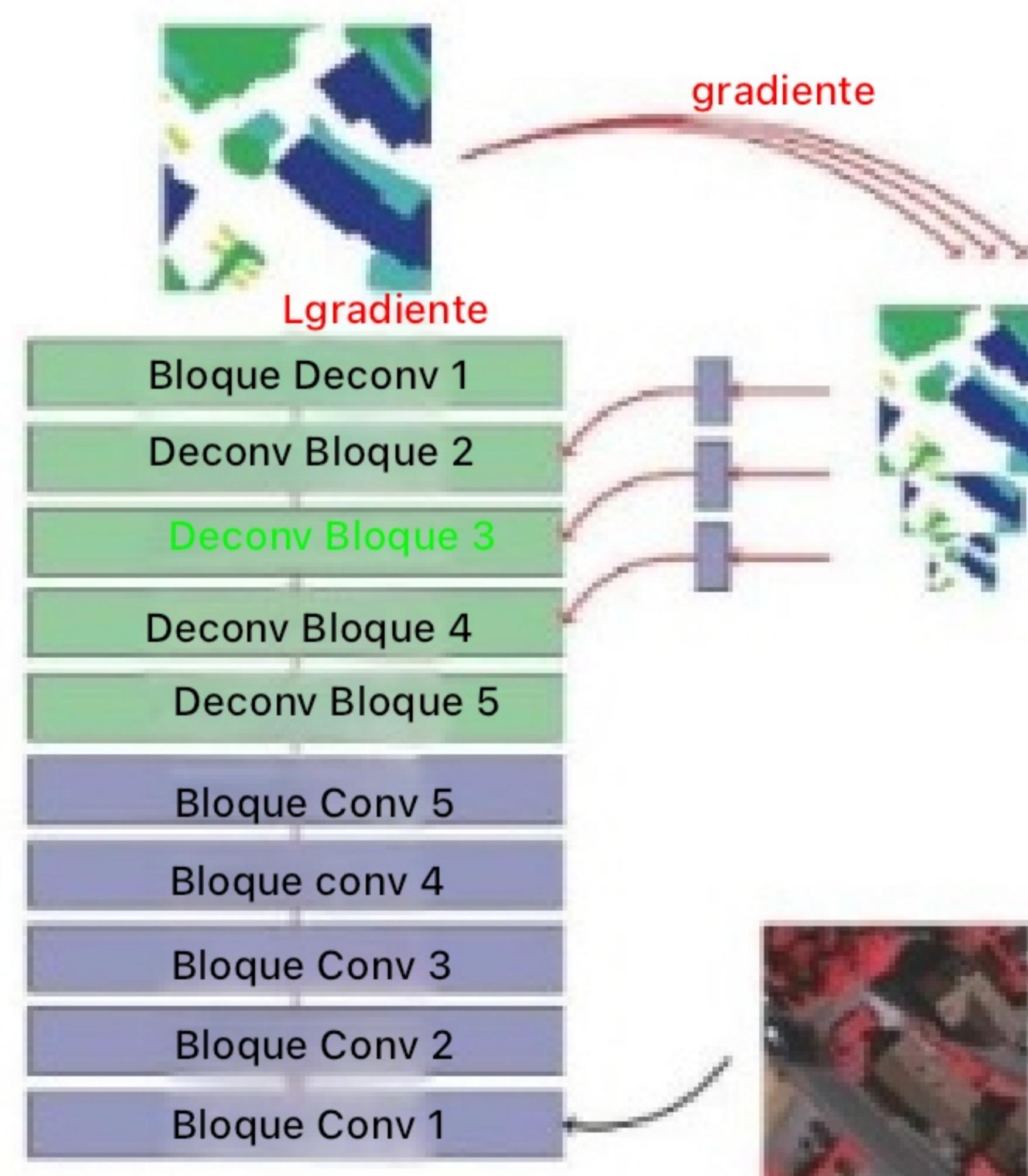
Esto garantiza que las capas anteriores sigan teniendo un gradiente significativo, incluso cuando la optimización global está convergiendo. Como se argumentó en Lin et al. (2017), las capas más profundas ahora solo tienen que aprender a refinar las predicciones más gruesas de las resoluciones más bajas, lo que ayuda al proceso de aprendizaje general.

3.3. Fusión temprana

En la comunidad de visión artificial, las imágenes RGB-D a menudo se denominan imágenes 2.5D. La integración de estos datos en modelos de aprendizaje profundo ha demostrado ser todo un reto, ya que el enfoque ingenuo del apilamiento no funciona bien en la práctica. Se han propuesto varios esquemas de fusión de datos para sortear este obstáculo. El enfoque de FuseNet (Hazirbas et al., 2016) utiliza la arquitectura de SegNet en un contexto multimodal. Como se ilustra en la Fig. 4a, codifica conjuntamente la información RGB y la información de profundidad utilizando dos codificadores cuyas contribuciones se suman después de cada bloque convolucional. A continuación, un único decodificador sobremuestrea la representación conjunta codificada de nuevo en el espacio de probabilidad de la etiqueta. Este enfoque de fusión de datos también se puede adaptar a otras redes neuronales profundas, como las redes residuales, como se ilustra en la Fig. 4b.

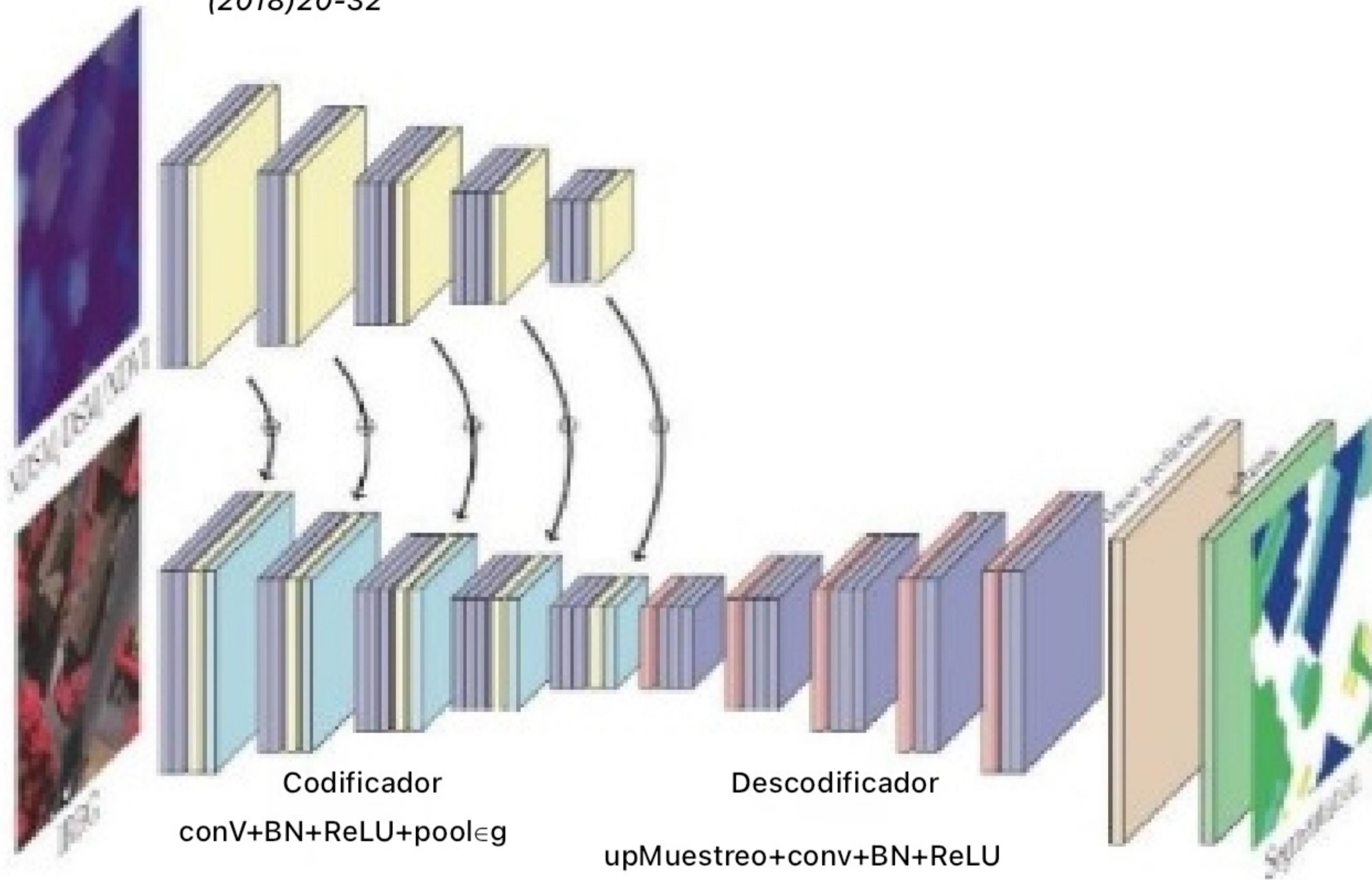
Sin embargo, en esta arquitectura los datos de profundidad se tratan como de segunda mano. De hecho, las dos ramas no son exactamente simétricas: la rama de profundidad trabaja solo con información relacionada con la profundidad, mientras que la rama óptica en realidad trata con una combinación de datos de profundidad y ópticos. Además, en el proceso de muestreo ascendente, solo se utilizarán los índices de la rama principal. Por lo tanto, es necesario elegir qué fuente de datos será la primaria y cuál será la auxiliar (véase la Fig. 5a). Existe un desequilibrio conceptual en la forma en que se tratan las fuentes two. Sugerimos una arquitectura alternativa con una tercera rama "virtual" que no tenga este desequilibrio, lo que podría mejorar el rendimiento.

En lugar de calcular la suma de los dos conjuntos de mapas de características, sugerimos un proceso de fusión alternativo para obtener las características de unión multimodales. Introducimos un tercer codificador que no corresponde a ninguna modalidad real, sino a una fuente de datos virtual fusionada. En la etapa n , el codificador virtual toma como entrada sus activaciones anteriores concatenadas con ambas activaciones de los otros codificadores. Estos mapas de características se pasan a través de un mapa convolucional

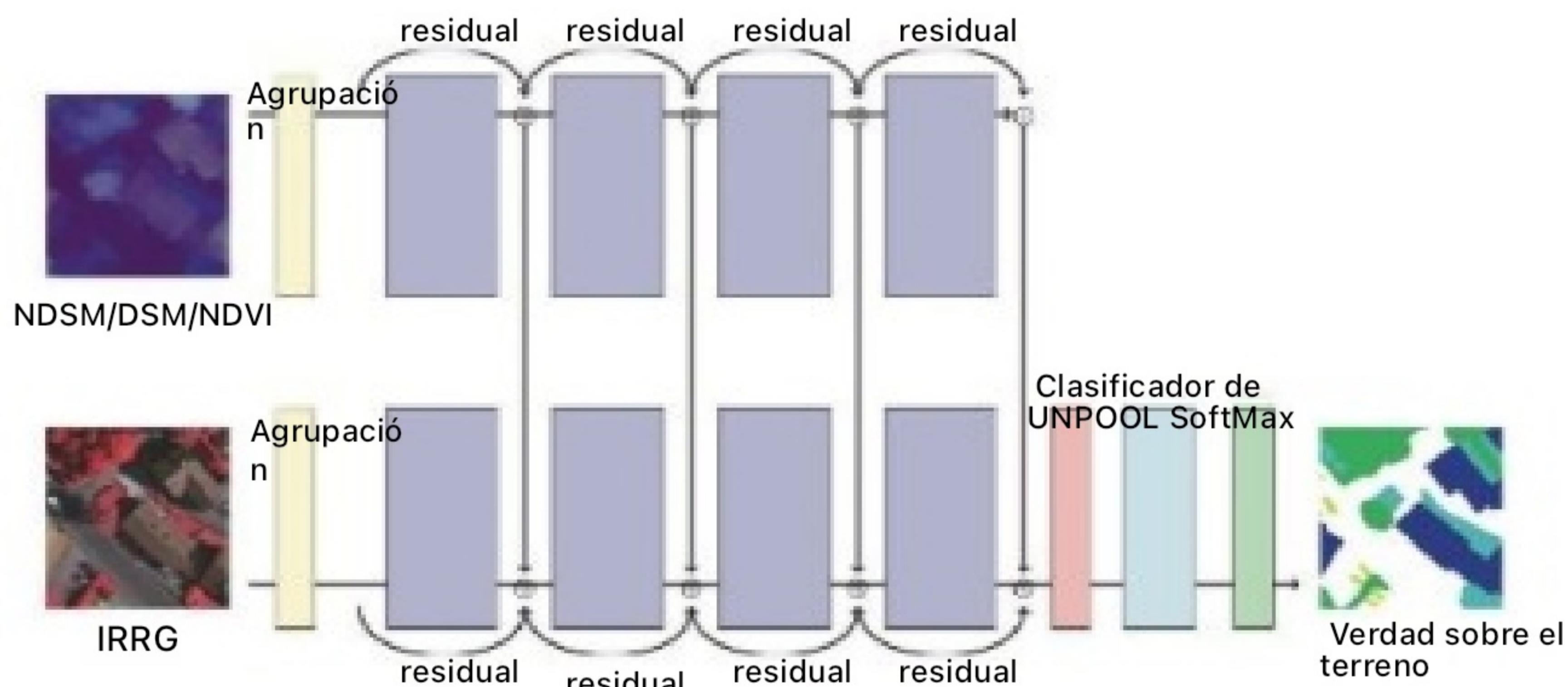


b) Retropropagación a múltiples escalas.

Fig.3. Supervisión profunda multiescala de SegNet con 3 ramas sobre datos de teledetección.

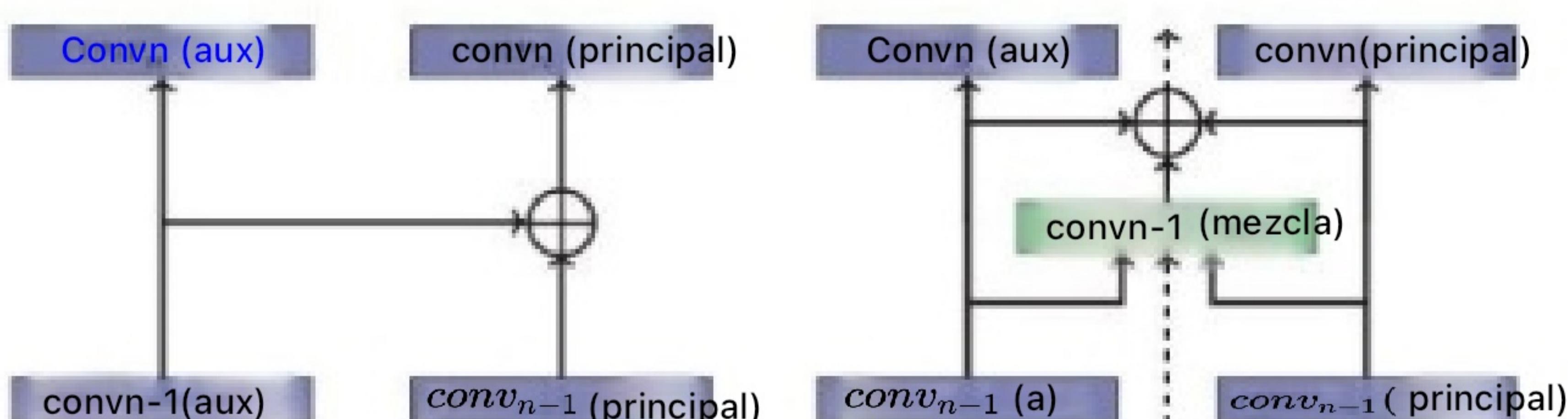


(a) Arquitectura FuseNet [10] para la fusión temprana de datos de teledetección.



(b) FusResNet: la arquitectura de FuseNet adaptada a una red residual.

Figura 4. Arquitecturas de líneas base FCN alteradas para adaptarse al marco de FuseNet.



(a) FuseNet original: los fusibles contribuciones mediante la suma de activaciones auxiliares en la rama principal.

(b) Nuestra FuseNet: fusiona las contribuciones con un bloque convolucional seguido de una sumatoria.

Figura 5. Estrategias de fusión para la arquitectura de FuseNet.

para aprender un residuo que se suma con los mapas de características promedio de los otros codificadores. Esto se ilustra en la Fig. 5b. Esta estrategia hace que FuseNet sea simétrico y, por lo tanto, nos libera de la elección de la fuente principal, que sería un hiperparámetro adicional a ajustar. Esta arquitectura se denominará V-FuseNet en el resto del documento para Virtual-FuseNet.

3.4. Fusión tardía

Una advertencia del enfoque de FuseNet es que se espera que ambos flujos sean topológicamente compatibles para fusionar los codificadores. Sin embargo, este puede no ser siempre el caso, especialmente cuando se trata de datos que no poseen la misma estructura (por ejemplo, imágenes 2D y una nube de puntos 3D). Por lo tanto, proponemos una técnica de fusión alternativa que se basa solo en los mapas de características tardías sin suposición sobre los modelos. Específicamente, en lugar de investigar la fusión a nivel de datos, trabajamos en torno a la heterogeneidad de los datos tratando de lograr la fusión de predicciones. Este proceso se investigó en Audebert et al. (2016), donde se introdujo un módulo de corrección residual. Este módulo consiste en una red neuronal convolucional residual que toma como entrada los últimos mapas de características de dos redes profundas. Esas redes pueden ser topológicamente idénticas o no. En nuestro caso, cada red profunda es una red totalmente convolucional que ha sido entrenada en lo óptico o en lo auxiliar

fuente de datos. Cada FCN genera una predicción. Primero, promediamos las dos predicciones para obtener un mapa de clasificación suave. A continuación, volvemos a entrenar el módulo de corrección de forma residual. Por lo tanto, la red de corrección residual aprende un pequeño desplazamiento para aplicarlo a cada probabilidad de píxel. Esto se ilustra en la Fig. 6a para la arquitectura SegNet y en la Fig. 6b para la arquitectura ResNet.

Sea R el número de salidas en las que se debe realizar la corrección residual, P_0 la verdad fundamental, P_i la predicción y ϵ_i el término de error de P_i con la verdad fundamental. Predecimos P' , la suma de las predicciones promediadas y el término de corrección c que se infiere de la red de fusión:

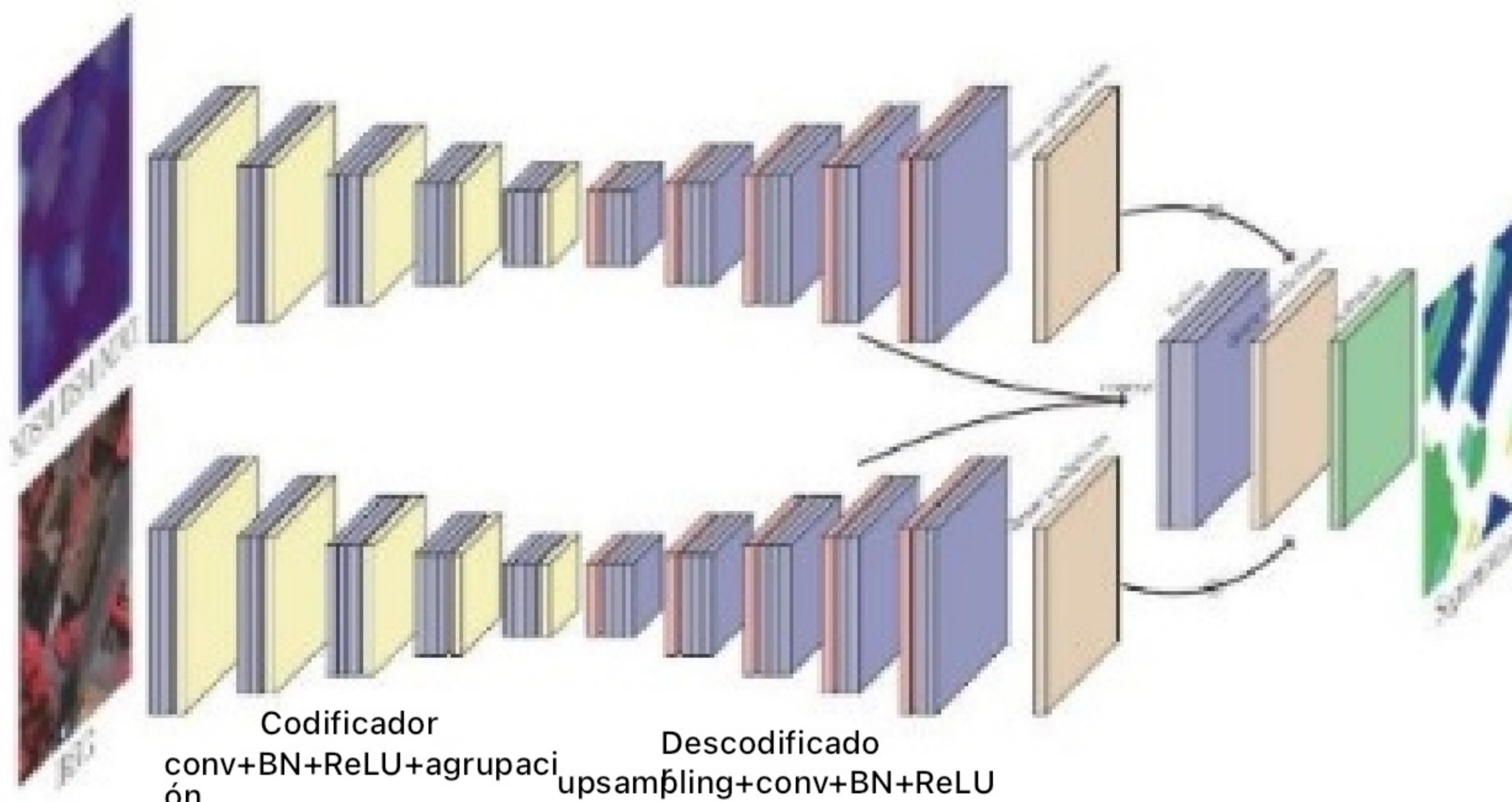
$$P' = P_{avg} + c = \frac{1}{R} \sum_{i=1}^R P_i + c = P_0 + \frac{1}{R} \sum_{i=1}^R \epsilon_i + c, \quad (3)$$

Como nuestro módulo de corrección residual está optimizado para minimizar la pérdida, aplicamos:

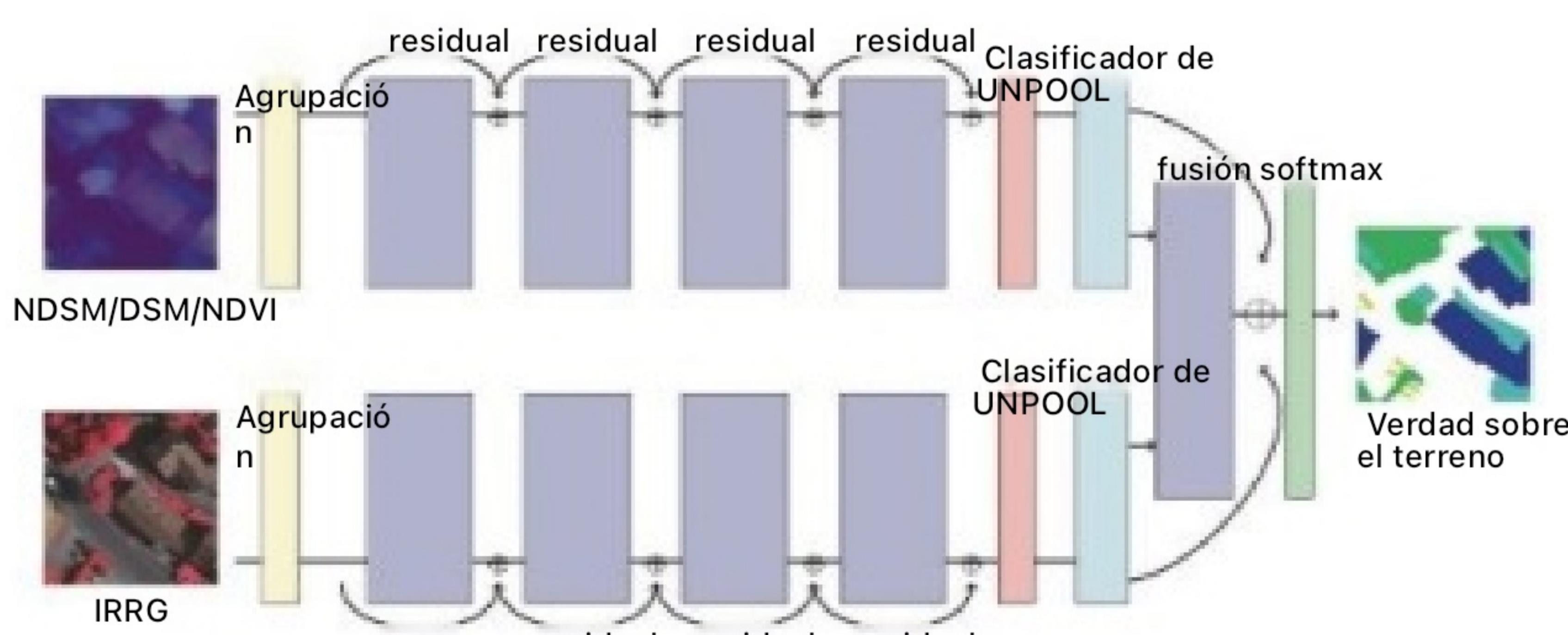
$$\|P' - P_0\| \rightarrow 0 \quad (4)$$

Lo que se traduce en una restricción en C y 6:

$$\left\| \frac{1}{R} \sum_{i=1}^R \epsilon_i - c \right\| \rightarrow 0. \quad (5)$$



(a) Corrección residual [2] para fusión tardía utilizando dos SegNets.



(b) Corrección residual [2] para fusión tardía utilizando dos ResNets.

Figura 6. Arquitecturas de líneas de base modificadas FCN para adaptarse al marco de corrección residual.

A medida que este desplazamiento se aprende de forma supervisada, la red puede inferir en qué entrada confiar en función de las clases predichas. Por ejemplo, si los datos auxiliares son mejores para la detección de vegetación, la corrección residual atribuirá más peso a la predicción que sale de la SegNet auxiliar. Este módulo se puede generalizar a n entradas, incluso con diferentes arquitecturas de red. Esta arquitectura se denominará SegNet-RC (por SegNet-Residual Correction) en el resto del documento.

4.3. Equilibrio de clases

Los conjuntos de datos de teledetección (descritos más adelante en la sección 4.1) que consideramos tienen clases semánticas desequilibradas. De hecho, las estructuras relevantes en las áreas urbanas no ocupan la misma superficie (es decir, el mismo número de píxeles) en las imágenes. Por lo tanto, al realizar una segmentación semántica, las frecuencias de clase pueden ser muy poco homogéneas. Para mejorar la precisión del promedio de la clase, equilibraremos la pérdida utilizando las frecuencias de la clase inversa. Sin embargo, como una de las clases consideradas es una clase de rechazo ("clutter") que también es muy rara, no usamos la frecuencia de clase inversa para esta. En cambio, aplicamos el mismo peso en esta clase que el peso más bajo en todas las demás clases. Esto tiene en cuenta que, en los conjuntos de datos, la clase de desorden es un problema mal planteado.

4.1. Conjuntos de datos

Validamos nuestro método en los dos conjuntos de imágenes del ISPRS 2D Semantic Labeling Challenge». Estos conjuntos de datos se componen de imágenes aéreas de muy alta resolución de dos ciudades de Alemania: Vaihingen y Potsdam. El objetivo es realizar el etiquetado semántico de las imágenes en seis clases: edificios, superficies impermeables (por ejemplo, carreteras), baja vegetación, árboles, coches y desorden. Hay dos tablas de clasificación en línea disponibles (una para cada ciudad) e informan las métricas de prueba obtenidas en las imágenes de [Artefactos de Vaihingen](#).

El conjunto de datos de Vaihingen tiene una resolución de 9 cmn/píxel con mosaicos de aproximadamente 2100 x 2100 píxeles. Hay 33 imágenes, de las cuales 16 tienen una verdad pública. Las teselas consisten en imágenes infrarrojas-rojas-verdes (IRRG) y datos DSM extraídos de la nube de puntos Lidar. También utilizamos el DSM normalizado (nDSM) de Gerke (2015).

4.1.2. ISPRS Potsdam

El conjunto de datos de Potsdam tiene una resolución de 5 cm/píxel con mosaicos de 6000x6000. Hay 38 imágenes, de las cuales 24 tienen una verdad pública. Las teselas consisten en imágenes multiespectrales infrarrojas-rojas-verdes-azules (IRRGB) y datos DSM extraídos de la nube de puntos Lidar. Los nDSM también se incluyen en el conjunto de datos con dos métodos diferentes.

4.2. Configuración

Para cada imagen óptica, calculamos el NDVI utilizando la siguiente fórmula:

$$NDVI = \frac{IR - R}{IR + R}. \quad (6)$$

A continuación, construimos una imagen compuesta compuesta por el DSM, el nDSM y el NDVI apilados.

Como las baldosas son de muy alta resolución, no podemos procesarlas directamente en nuestras redes profundas.

Utilizamos un enfoque de ventana deslizante para extraer parches de 128 x128. El paso de la ventana corredera también define el tamaño de las regiones superpuestas entre dos parches consecutivos. En el momento del entrenamiento, una zancada más pequeña nos permite extraer más muestras de entrenamiento y actúa como aumento de datos. En el momento de la prueba, un paso más pequeño nos permite promediar las predicciones en las regiones superpuestas, lo que reduce los efectos de borde y mejora la precisión general. Durante el entrenamiento, usamos una zancada de 64 px para Pots-dam y una zancada de 32 px para Vaihingen. Los modelos se implementan utilizando el marco Caffe. Entrenamos todos nuestros modelos utilizando el descenso de gradiente estocástico (SGD) con una tasa de aprendizaje base de 0,01, un momento de 0,9, una disminución del peso de 0,0005 y un tamaño de lote de 10. Para las arquitecturas basadas en SegNet, los pesos del codificador en SegNet se inicializan con los de VGG-16 entrenado en ImageNet, mientras que los pesos del decodificador se inician aleatoriamente utilizando la política de He et al. (2015). Dividimos la tasa de aprendizaje por 10 después de 5, 10 y 15 épocas. En el caso de los modelos basados en ResNet, los cuatro bloques convolucionales se inicializan mediante ponderaciones de ResNet-34 entrenadas en ImageNet, y las demás ponderaciones se inicializan mediante la misma política. Dividimos la tasa de aprendizaje por 10 después de 20 y 40 épocas. Los resultados se validan de forma cruzada en cada conjunto de datos mediante divisiones de 3 veces. Los modelos finales para las tres arquitecturas se combinan para obtener un resultado final en todo el conjunto de entrenamiento.

4.3. Resultados

En la Tabla 1 se detallan los resultados validados de nuestros métodos en el conjunto de datos de Vaihingen. Mostramos la precisión de píxeles y la puntuación F1 promedio en todas las clases de all. La puntuación F1 de una clase se define por:

$$F1_i = 2 \frac{precision_i \times recall_i}{precision_i + recall_i}, \quad (7)$$

$$recall_i = \frac{tp_i}{C_i}, precision_i = \frac{tp_i}{P_i}, \quad (8)$$

donde tp, el número de verdaderos positivos para la clase i, Ci el número de píxeles que pertenecen a la clase i y Pi el número de píxeles atribuidos a la clase i por el modelo. De acuerdo con las instrucciones de evaluación de los organizadores del desafío, estas métricas se calculan después de erosionar los bordes en un círculo de 3 px de radio y descartar esos píxeles.

En la Tabla 2 se detallan los resultados del enfoque multiescalas. "Sin rama" denota el modelo SegNet de escala única de referencia. La primera rama se añadió después del 4º bloque convolucional del decodificador (downscale =2), la segunda rama después de la 3ª (downscale =4) y la tercera rama después de la 2ª (downscale =8).

Tabla 1
Resultados de la validación
en Vaihingen.

Modelo	Precisión general	Promedio
SegNet (IRRG)	90.2±1.4	89.3±1.2
SegNet (compuesto)	88,3±0,9	81,6±0,8
SegNet-RC	90,6±1,4	89.2±1.2
FuseNet	90,8±1,4	90.1±1.2
V-FuseNet	91.1±1.5	90.3±1.2
ResNet-34 (IRRG)	90,3±1,0	89,1±0,7
ResNet-34	88,8±1,1	83,4±1,3
ResNet-34-RC	90,8±1,0	89.1±1.1
FusResNet	90.6±1.1	89.3±0,7

Los mejores resultados se resaltan en negrita.

Tabla 2

Resultados multiescala en Vaihingen.

Número de sucursales	Imp.surf.	Edificios	Verduras bajas.	Árboles	Coches	En general
Sin sucursal	92.2	95.5	82.6	88.1	88.2	90.2±1.4
1 rama	92.4	95.7	82.3	87.9	88.5	90.3±1.5
2 ramas	92.5	95.8	82.4	87.8	87.6	90.3±1.4
3 ramas	92.7	95.8	82.6	88.1	88.1	90.5±1.5

Los mejores resultados se resaltan en negrita.

Cuadro 3

Resultados finales del conjunto de datos de Vaihingen.

Método	Imp.surf.	Edificios	Verduras bajas.	Árboles	Coches	En general
FCN (Sherrah,2016)	90.5	93.7	83.4	89.2	72.6	89.1
FCN+ fusión + límites (Marmanis et al., 2016)	92.3	95.2	84.1	90.0	79.3	90.3
SegNet (IRRG)	91.5	94.3	82.7	89.3	85.7	89.4
SegNet-RC	91.0	94.5	84.4	89.9	77.8	89.8
FuseNet	91.3	94.3	84.8	89.9	85.9	90.1
V-FuseNet	91.0	94.4	84.5	89.9	86.3	90.0

Los mejores resultados se resaltan en negrita.

Cuadro 4

Resultados finales en el conjunto de datos de Potsdam.

Método	Imp.surf.	Edificios	Verduras bajas.	Árboles	Coches	En general
FCN + CRF + características expertas (Liu et al., 2017)	91.2	94.6	85.1	85.1	92.8	88.4
FCN (Sherrah,2016)	92.5	96.4	86.7	88.0	94.7	90.3
SegNet (IRRG)	92.4	95.8	86.7	87.4	95.1	90.0
SegNet-RC	91.3	95.9	86.2	85.6	94.8	89.0
V-FuseNet	92.7	96.3	87.3	88.5	95.4	90.6

Los mejores resultados se resaltan en negrita.

Las Tablas 3 y 4 muestran los resultados finales de nuestros métodos sobre los datos de prueba retenidos de los conjuntos de datos de Vaihingen y Potsdam, respectivamente.

5. Discusión

5.1. Líneas de base y experimentos

Como se ilustra en la Tabla 1, ResNet-34 tiene un rendimiento ligeramente mejor en precisión general y obtiene resultados más estables en comparación con Seg-Net. Esto se debe probablemente a una mejor capacidad de generalización de ResNet que hace que el modelo esté menos sujeto a sobreajustes. En general, ResNet y SegNet obtienen resultados similares, siendo ResNet más estable. Sin embargo, ResNet requiere significativamente más memoria en comparación con SegNet, especialmente cuando se utilizan los esquemas de fusión. Cabe destacar que no pudimos utilizar el esquema V-

FuseNet con ResNet-34 debido a la limitación de memoria (12 Gb) de nuestras GPU. No obstante, estos resultados muestran que las estrategias de fusión de datos investigadas se pueden aplicar a varios tipos de redes totalmente convolucionales y que nuestros hallazgos deberían generalizarse a otras redes base del estado del arte.

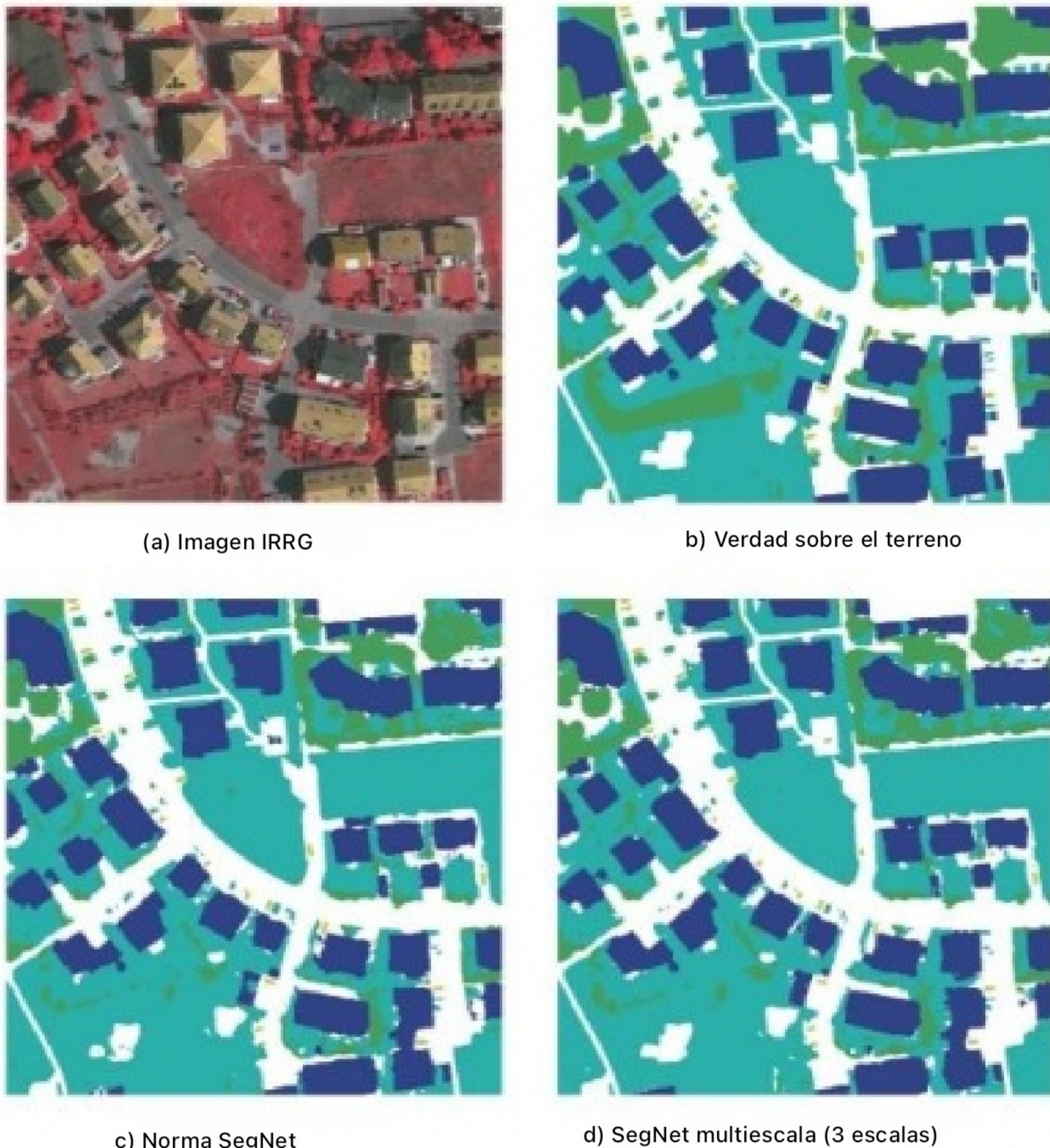
5.2. Efectos de la estrategia multiescala

La ganancia utilizando el enfoque multiescala es pequeña, aunque es prácticamente gratuita, ya que solo requiere unas pocas convoluciones adicionales

para extraer mapas a escala reducida de las capas inferiores. Como era de esperar, las grandes estructuras, como las carreteras y los edificios, se benefician de las predicciones a menor escala, mientras que los coches están ligeramente menos dimensionados en resoluciones más bajas. Suponemos que la vegetación no está estructurada y, por lo tanto, el enfoque multiescala no ayuda aquí, sino que aumenta la confusión entre vegetación baja y arbórea. El aumento del número de ramas mejora la clasificación general, pero por un margen menor cada vez, lo que es de esperar ya que las predicciones reducidas se vuelven muy toscas con una resolución de 1:16 o 1:32. Finalmente, aunque las mejoras cuantitativas son bajas, una evaluación visual de los mapas inferidos muestra que la mejora cualitativa no es despreciable. Como se ilustra en la Fig. 7, la predicción multiescala regulariza y reduce el ruido en las predicciones. Esto facilita la posterior interpretación humana o el uso efectivo secundario de esta investigación. Nuestras pruebas mostraron que los resultados reducidos siguen siendo bastante precisos. Por ejemplo, la predicción reducida en un factor 8 tenía una precisión media de sólo un 0,5% por debajo de la predicción de resolución completa, y la diferencia residía principalmente en la clase de "coche". Esto no es sorprendente, ya que los autos suelen tener ~ 30 px de largo en el mosaico de resolución completa y, por lo tanto, cubren solo 3-4 píxeles en la predicción a escala reducida, lo que los hace más difíciles de ver. Sin embargo, la buena precisión media de las salidas a escala reducida parece indicar que el decodificador de SegNet podría reducirse a su primer bloque convolucional sin perder demasiada precisión. Esta técnica podría usarse para reducir el tiempo de inferencia cuando los objetos pequeños son irrelevantes mientras se mantiene una buena precisión en las otras clases.

fusión

Como se esperaba, ambos métodos de fusión mejoran la precisión de la clasificación en los dos conjuntos de datos, como se ilustra en la Fig. 8. Mostramos algunos ejemplos de parches mal clasificados que se corrigen usando



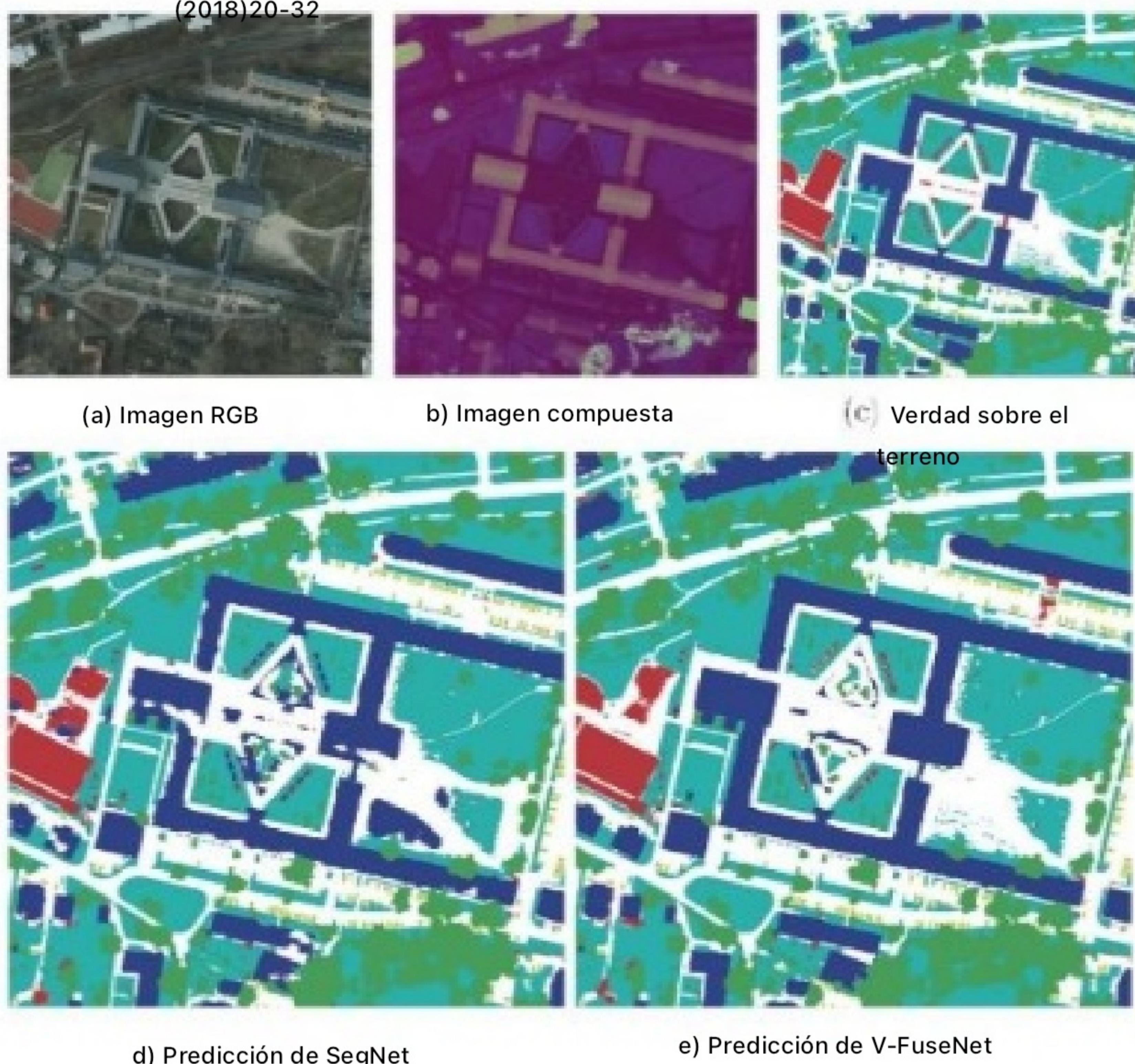


Fig.8. Efecto de la estrategia de fusión en un extracto del conjunto de datos ISPRS Potsdam. La confusión entre superficies impermeables y edificios se reduce significativamente gracias a la contribución del nDSM en la estrategia V-FuseNet. (blanco: carreteras, azul: edificios, cian: vegetación baja, verde: árboles, amarillo: coches). (Para la interpretación de las referencias al color en esta leyenda de la figura, se remite al lector a la versión web de este artículo).



Figura 9. Inconsistencias discutibles entre nuestras predicciones y la verdad sobre el terreno. (blanco: carreteras, azul: edificios, cian: vegetación baja, verde: árboles, amarillo: coches). (Para la interpretación de las referencias al color en esta leyenda de la figura, se remite al lector a la versión web de este artículo).

En conclusión, las dos estrategias de fusión se pueden utilizar para diferentes casos de uso. La fusión tardía por corrección residual es más adecuada para combinar varios clasificadores fuertes que confían en sus predicciones, mientras que el esquema de fusión temprana de FuseNet es más adecuado para integrar datos auxiliares más débiles en el proceso de aprendizaje principal. En el conjunto de pruebas retenido, la estrategia V-FuseNet no funciona tan bien como se esperaba. Su precisión global está marginalmente por debajo del modelo original de FuseNet, aunque las puntuaciones de F1 en clases más pequeñas y difíciles han mejorado, especialmente el "desorden", que mejora del 49,3% al 51,0%. Como la clase "clutter" se ignora en las métricas del conjunto de datos, esto no se refleja en la precisión final.

5.4. Solidez ante las incertidumbres y los datos faltantes

Al igual que para todos los conjuntos de datos, las etiquetas semánticas ISPRS en la verdad fundamental sufren algunas limitaciones. Esto puede causar errores de etiquetado erróneos injustos causados por objetos faltantes en la realidad del terreno o transiciones bruscas que no reflejan la imagen verdadera (cf. Fig.9a). Sin embargo, incluso los datos brutos (ópticos y DSM) pueden ser engañosos. De hecho, los artefactos geométricos del proceso de costura también tienen un impacto negativo en la segmentación, ya que nuestro modelo se sobreajusta en esos píxeles deformes (cf. Fig.9b).

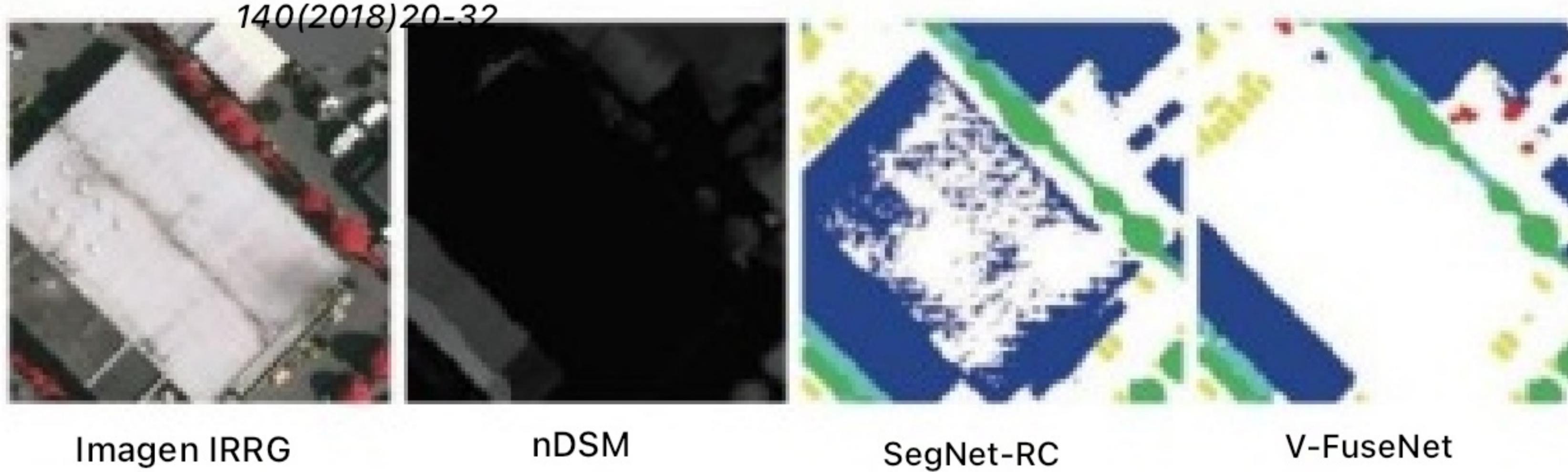
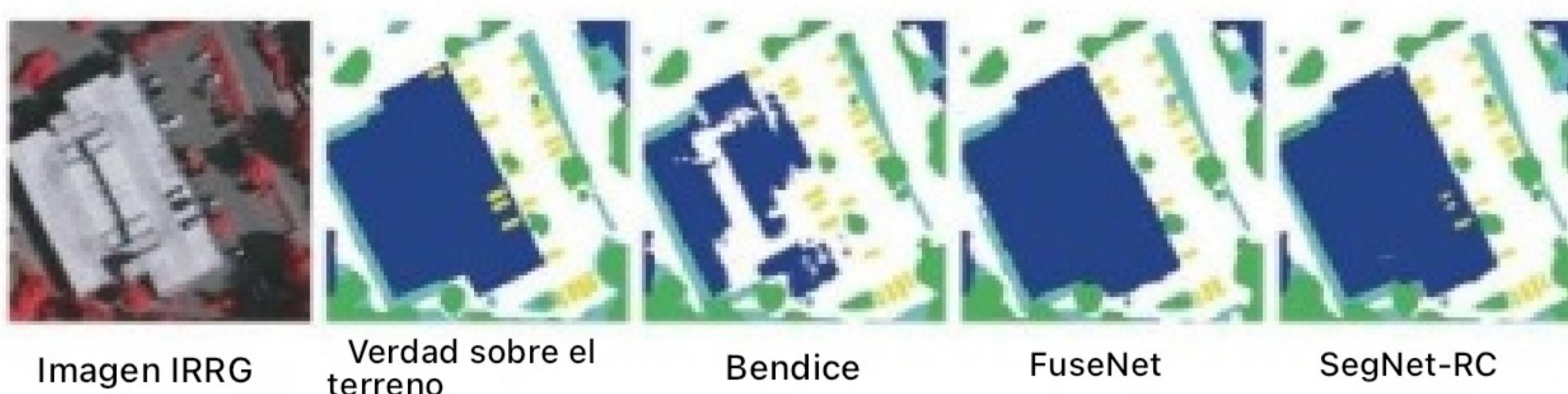
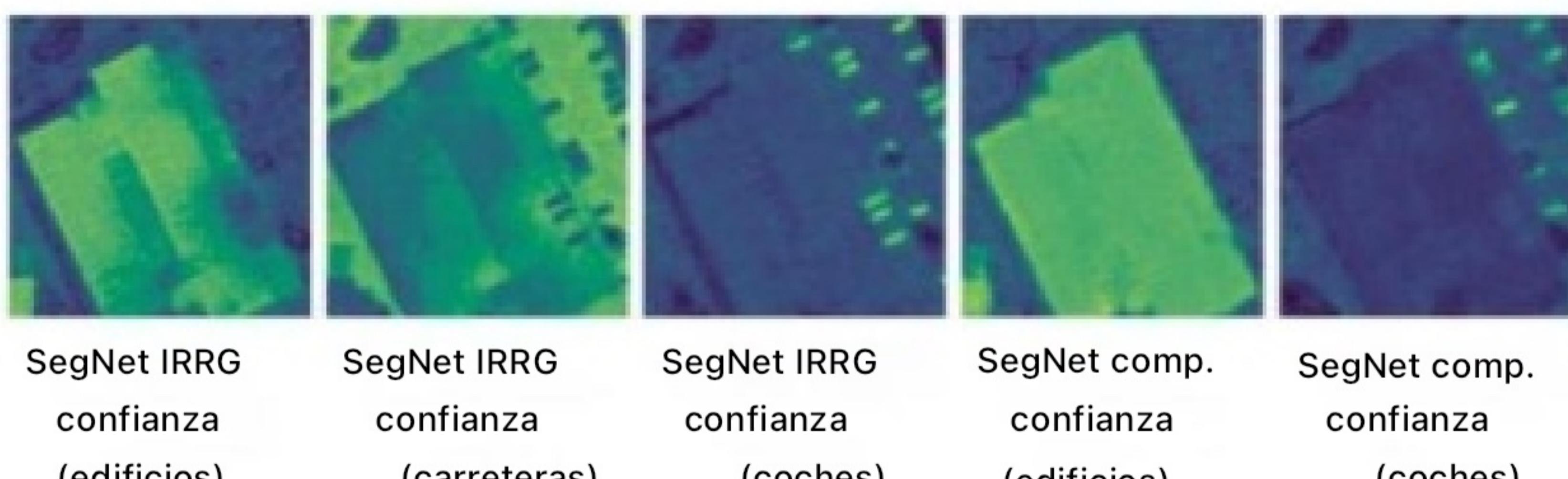


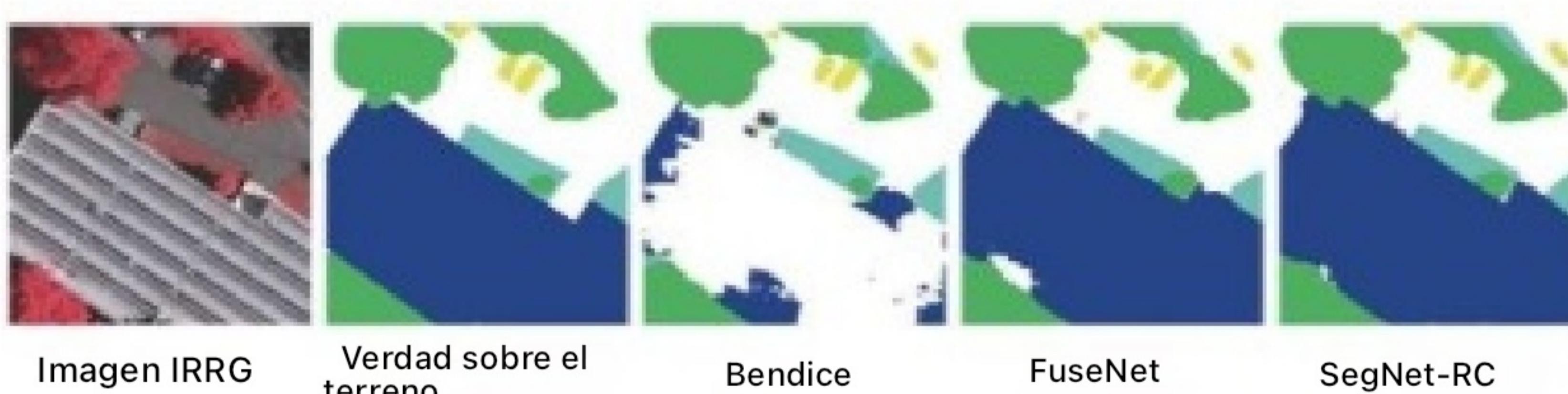
Figura 10. Los errores en el nDSM de Vaihingen son mal manejados por ambos métodos de fusión. Aquí, un edificio entero desaparece.



(a) Predicciones de varios modelos en un parche del conjunto de datos de Vaihingen.



(b) Mapas de calor de confianza SegNet para varias clases utilizando varias entradas.



(c) Predicciones de varios modelos en un parche del conjunto de datos de Vaihingen.

Fig.11. Predicciones exitosas utilizando las estrategias de fusión.

Finalmente, debido a las limitaciones y el ruido en la nube de puntos Lidar, como puntos faltantes o aberrantes, el DSM y, posteriormente, el nDSM presentan algunos artefactos. Como se informó en Marmanis et al. (2016), algunos edificios desaparecen en el nDSM y a los píxeles relevantes se les atribuye falsamente una altura de 0. Esto causa una incorrecta clasificación significativa en la imagen compuesta que es mal manejada por ambos métodos de fusión, como se ilustra en la Fig. 10. Marmanis et al. (2016) solucionaron este problema corrigiendo manualmente el nDSM, aunque este método no se escala a conjuntos de datos más grandes. Por lo tanto

podría ser útil mejorar el método para que sea resistente a datos y artefactos impuros, por ejemplo, mediante el uso de redes de alucinación (Hoffman et al., 2016) para inferir la modalidad faltante, como se propone en Kampffmeyer et al. (2016). Creemos que la arquitectura de V-FuseNet podría adaptarse para tal propósito utilizando la rama virtual para codificar los datos que faltan. Además, el trabajo reciente sobre modelos generativos podría ayudar a aliviar el sobreajuste y mejorar la robustez mediante el entrenamiento con datos sintéticos, como se propone en Xie et al. (2017).

6. Conclusión

En este trabajo, investigamos redes neuronales profundas para el etiquetado semántico de datos de teledetección urbana multimodales de muy alta resolución. Especialmente, demostramos que las redes totalmente convolucionales son adecuadas para la tarea y obtienen excelentes resultados. Presentamos un sencillo truco de supervisión profunda que extrae mapas semánticos a múltiples resoluciones, lo que ayuda a entrenar la red y mejora la clasificación general. A continuación, ampliamos nuestro trabajo a los datos no ópticos mediante la integración de un modelo digital de superficie extraído de nubes de puntos Lidar. Estudiamos dos métodos para el procesamiento de datos de teledetección multimodal con redes profundas: la fusión temprana con FuseNet y la fusión tardía mediante corrección residual. Demostramos que ambos métodos pueden aprovechar eficientemente la complementariedad de los datos heterogéneos, aunque en diferentes casos de uso. Mientras que la fusión temprana permite que la red aprenda características más fuertes, la fusión tardía permite procesar imágenes de nubes de puntos Lidar de forma más eficiente. Por tanto, nuestros resultados demuestran que tanto la fusión temprana como la fusión tardía tienen sus ventajas y desventajas y que deben ser utilizadas juntas para obtener los mejores resultados.

Nicolas Audebert agradece a los miembros del equipo de investigación de la Universidad de Vaihingen y Potsdam por su apoyo y agradecimientos.

Referencias

Audebert, N., Le Saux, B., Lefèvre, S., 2016. Segmentación semántica de datos de observación de la Tierra utilizando redes profundas multimodales y multiescala. En: Conferencia Asiática sobre Visión por Computador (ACCV'16), Taiwán, Taiwán, 2016. Audebert, N., Le Saux, B., Lefèvre, S., 2016. ¿Qué aporta la clasificación regional de imágenes de teledetección en un marco de aprendizaje profundo? En: Simposio Internacional de Aprendizaje y Teledetección IEEE, 2016 (IGARSS), Beijing, China, pp. 5091-5094.

Audebert, N., Le Saux, B., Lefèvre, S., 2016. Aprendizaje conjunto de la observación de la Tierra y los datos de OpenStreetMap para obtener mapas semánticos mejores y más rápidos. En: Actas de los talleres de la Conferencia IEEE sobre Visión por Computador y Reconocimiento de Patrones (CVPR), Honolulu, EE.UU., 2017. Audebert, N., Le Saux, B., Lefèvre, S., 2016. Aprendizaje profundo para la segmentación de imágenes. Patrón. IEEE Trans. Pattern Anal. Mach. Intell. 38, 1619–1631.

Camps-Taberner, M., Romero-Soriano, A., Gatta, C., Camps-Valls, G., Lagrange, A., Le Saux, B., Beaupère, A., Boulch, A., Chan-Hon-Tong, A., Herbin, S., Randrianarivo, H., Ferecatu, M., Shimoni, M., Moser, G., Tuia, D., 2016. Procesamiento de datos LiDAR y RGB de extremadamente alta resolución: resultado del concurso de fusión. Xie, X., IEEE GRSS 2014, Parte A, 2014. Deteción de vehículos en imágenes satelitales mediante redes neuronales convolucionales profundas híbridas. IEEE Geosci. Rem. Sens. Letters 12, 1011–1015.

Chen, L., Papandreou, G., Murphy, K., Yuille, A.L., 2015. Segmentación semántica de imágenes con detección de bordes específicos de la tarea mediante CNN y una transformación de dominio entrenada discriminativamente. En: Accesos M., 2016. Encuentro sobre Representaciones del Aprendizaje Automático, San Diego, EE.UU.

Chen, L., Papandreou, G., Murphy, K., Yuille, A.L., 2015. Descripción general y diseño de pruebas. Fotogrametría-Fernerkundung-Geoinformación 2, 73–82.

Fitzek, A., Springenberg, J., Riedmiller, M., Burgard, W., 2015. Aprendizaje automático modal para el reconocimiento robusto de objetos RGB-D. En: Actas de la Conferencia Internacional sobre Robots y Sistemas Inteligentes. IEEE, Hamburg, Alemania, pp. 681–687.

Gerke, M., 2015. Uso de la biblioteca Star Vision dentro del ISPRS 2d Semantic Labeling Benchmark (Vaihingen).

Guo, H., Wang, G., Chen, X., 2016. Red neuronal convolucional de dos flujos para la detección precisa de la punta de los dedos RGB-D utilizando información de profundidad y borde. En: 2016 IEEE International Conference on Image Processing (ICIP). IEEE, Phoenix, EE.UU., pp. 2608–2612.

- Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2016. FuseNet: incorporación de profundidad en la segmentación semántica a través de la arquitectura CNN basada en fusión. En: Actas de la Conferencia IEEE sobre Visión por Computador y Reconocimiento de Patrones (CVPR), 2015. Computación profunda para los Técnicos: superando el rendimiento a nivel humano en la clasificación de ImageNet. En: Actas de la Conferencia Internacional IEEE sobre Visión por Computador, 2016. Aprendizaje residual profundo para el reconocimiento de imágenes. En: Actas de la Conferencia IEEE sobre Visión por Computador y Reconocimiento de Patrones, 2016. Evolución de la modalidad alucinación. En: Actas de la Conferencia IEEE sobre Visión por Computador y Reconocimiento de Patrones, 2015. Nuevas vías de estudio; pp. 1826–1834.
- Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2016. Segmentación semántica de objetos pequeños y modelado de la incertidumbre en imágenes de teledetección urbana utilizando redes neuronales convolucionales profundas. En: Actas de la Conferencia IEEE sobre Talleres de Visión por Computador y Reconocimiento de Patrones, Las Vegas, EE.UU., pp. 1–9.
- Hong, A., Herbin, S., Randrianarivo, H., Ferecatu, M., 2015. Clasificación comparativa de los datos de observación de la Tierra: desde el aprendizaje de características explícitas hasta las técnicas de aprendizaje de imágenes. En: Conferencia internacional de aprendizaje de la IEEE (ICML), 2015. Intel, pp. 1417–1426.
- Liu, F.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, L.P., 2014. Microsoft COCO: objetos comunes en contexto. En: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision-ECCV 2014, Lecture Notes in Computer Science, vol. 8693, Springer, Berlin, pp. 201–215.
- Maggioni, E., Tarabalka, Y., Charpiat, G., Alirez, P., 2017. Redes neuronales convolucionales para la clasificación de imágenes de teledetección a gran escala. IEEE Trans. Geosci. Rem. Sens. 55, 645–657.
- Marmanis, D., Wegner, J.D., Galliani, S., Schindler, K., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Maragos, P., Paragios, N. (Eds.), Computer Vision NEUCLV, 2010, Xext, 2010. Notas de la conferencia sobre ciencias visuales 2010.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Clasificación con un borde: mejora de la segmentación semántica de imágenes con detección de bordes. Disponible en: arXiv: <1612.01337>.
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2016. Segmentación semántica de imágenes aéreas con un conjunto de CNNs. ISPRS Ann. Photogram. Rem. Sens. Espacial Informar. Ciencia 3, 473–480.
- Marmanis, D., Schindler, K., Weg

Saito, S., Yamashita, T., Aoki, Y., 2016. Extracción de objetos múltiples a partir de imágenes aéreas con redes neuronales. *ISPRS Journal of Photogrammetry and Remote Sensing* 110, 20–32.

Shen, J., 2016. Redes convolucionales para el etiquetado semántico denso de imágenes aéreas de alta resolución. Disponible en: arXiv:1606.02585.

Simonyan, K., Zisserman, A., 2014. Redes convolucionales muy profundas para el reconocimiento de imágenes a gran escala. Disponible en: arXiv:1409.1556.

Vakalopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N., 2015. Edificio Detección en datos multiespectrales de muy alta resolución con funciones de Deep Learning. En: Simposio de Geociencias y Teledetección (IGARSS), 2015 IEEE, Volpi, M., Tufa, D., 2017. Etiquetado semántico denso de imágenes de resolución subdecimétrica con redes neuronales convolucionales. *IEEE Trans. Geosci. Rem. Sens.* 55(2), 881–893.

Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A., 2017. Ejemplos antagónicos para la segmentación semántica y la detección de objetos. En: Conferencia Internacional IEEE sobre Visión por Computador (ICCV).

Yifan, Koltun, V., 2015. Agregación de contexto multiescalas por convoluciones dilatadas. En: Actas de la Conferencia Internacional sobre Representaciones del Aprendizaje, San Diego, EE.UU.

Zeller, M.D., Fergus, R., 2014. Visualización y comprensión de redes convolucionales. En: Computer Vision-ECCV 2014. Springer, Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Red de análisis de escenas piramidales. En: Actas de los talleres de la Conferencia IEEE sobre Visión por Computador y Reconocimiento de Patrones (CVPR), Honolulu, EE. UU.