remote sensing

MDPI

Article

Semantic Labeling of High-Resolution Images Combining a Self-Cascaded Multimodal Fully Convolution Neural Network with Fully Conditional Random Field

Qiongqiong Hu [1], Feiting Wang [2], Jiangtao Fang [2] and Ying Li [1,*]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China; qionghu@mail.nwpu.edu.cn
[2] Department of Computer Technology and Application, Qinghai University, Xi'ning 810016, China; ys220854040283@qhu.edu.cn (F.W.); ys230854100348@qhu.edu.cn (J.F.)
* Correspondence: lybyp@nwpu.edu.cn

Abstract: Semantic labeling of very high-resolution remote sensing images (VHRRSI) has emerged as a crucial research area in remote sensing image interpretation. However, challenges arise due to significant variations in target orientation and scale, particularly for small targets that are more prone to obscuration and misidentification. The high interclass similarity and low intraclass similarity further exacerbate difficulties in distinguishing objects with similar color and geographic location. To address this concern, we introduce a self-cascading multiscale network (ScasMNet) based on a fully convolutional network, aimed at enhancing the segmentation precision for each category in remote sensing images (RSIs). In ScasMNet, cropped Digital Surface Model (DSM) data and corresponding RGB data are fed into the network via two distinct paths. In the encoder stage, one branch utilizes convolution to extract height information from DSM images layer by layer, enabling better differentiation of trees and low vegetation with similar color and geographic location. A parallel branch extracts spatial, color, and texture information from the RGB data. By cascading the features of different layers, the heterogeneous data are fused to generate complementary discriminative characteristics. Lastly, to refine segmented edges, fully conditional random fields (DenseCRFs) are employed for postprocessing presegmented images. Experimental findings showcase that ScasMNet achieves an overall accuracy (OA) of 92.74% on two challenging benchmarks, demonstrating its outstanding performance, particularly for small-scale objects. This demonstrates that ScasMNet ranks among the state-of-the-art methods in addressing challenges related to semantic segmentation in RSIs.

Keywords: semantic labeling; RSIs; fully convolutional neural network; dilated convolution; DenseCRF

Citation: Hu, Q.; Wang, F.; Fang, J.; Li, Y. Semantic Labeling of High-Resolution Images Combining a Self-Cascaded Multimodal Fully Convolution Neural Network with Fully Conditional Random Field. *Remote Sens.* **2024**, *16*, 3300. https://doi.org/10.3390/rs16173300

Academic Editors: Mohammad Awrangjeb and Salah Bourennane

Received: 11 April 2024
Revised: 19 August 2024
Accepted: 3 September 2024
Published: 5 September 2024

1. Introduction

Lately, the interpretation of remote sensing images (RSIs) has become a focal point in the realm of Earth observation, propelled by the ongoing and swift progress in remote sensing technology. Furthermore, the obtention of very high-resolution RSIs (VHRRSIs) has become progressively convenient and cost-effective, leading to unprecedented growth in analytical techniques for RSIs across various research directions, including land use classification [1,2], target detection [3,4], autonomous driving, and three-dimensional reconstruction [5]. The semantic labeling of VHRRSIs involves semantically annotating all pixels in an image simultaneously. This process is a crucial step in image interpretation. The accuracy of this segmentation directly impacts subsequent processing tasks.

VHRRSIs are typically highly accurate, containing hundreds or thousands of pixels, with varying scales and orientations of objects. These images often display rich details, complex textures, strong spatial correlations, and a broad range of categories. The complex

imaging principle underlying RSI renders them highly ambiguous and uncertain, significantly increasing the difficulty of semantic labeling. As a result, semantic labeling has become among the most pivotal yet arduous tasks within the realm of computer vision. Lastly, the items of focus on an RSI are often visually small (e.g., cars) and densely distributed. These small objects are more prone to occlusion and misclassification, resulting in a substantial reduction in average segmentation accuracy.

To tackle these challenges, our proposal introduces ScasMNet for the semantic labeling of VHR images, with the aim of improving segmentation precision for small objects in RSI while preserving accuracy for other categories. Our approach comprises the following key components:

(1) We utilize a dual-input fully convolutional network (FCN) [6] equipped with an encoder–decoder architecture, given that FCN-based networks are particularly adept at handling semantic labeling tasks within a fully supervised learning framework. In the encoding phase, we integrate heterogeneous data by fusing features derived from both spectral channel inputs and Digital Surface Model (DSM) [7] data, thereby augmenting the complementarity of the data features.

(2) Instead of employing traditional up-sampling following the downsampling operation of the maximum pooling layer, we adopt dilated convolution [8] with a range of dilation rates. This approach extends the receptive field without a concomitant increase in the number of parameters or computational overhead. By implementing dilated convolutions with varying dilation rates, we facilitate the resampling of contextual information across multiple specialized layers, which allows for the extraction of a more comprehensive spectrum of distinctive features.

(3) Following the extraction of feature maps from the network, we apply dense conditional random fields (DenseCRFs) [9] to refine object boundaries, thereby enhancing the segmentation accuracy.

The structure of the subsequent sections of this manuscript is as follows: Section 2 provides a comprehensive review of pertinent literature on semantic labeling techniques for VHRRSI. In Section 3, we delineate our proposed method in depth, which encompasses the dual-path fully convolutional network architecture designed for the integration of heterogeneous data from DSM and optical imagery, the introduction of dilated convolution for feature extraction at various scales, and the utilization of DenseCRF to refine class boundaries and improve segmentation outcomes. Section 4 presents the experimental findings and a comparative analysis with other state-of-the-art deep learning methodologies. The manuscript concludes with a summary of the study and final remarks in Section 5.

2. Related Work

Recently, deep learning has achieved remarkable success within the field of computer vision, leading to a multitude of milestone accomplishments. Semantic labeling, also referred to as semantic segmentation in the computer vision literature, represents a core challenge in image analysis and plays a crucial role in the broader domain of computer vision. As a form of pixel-level classification, semantic labeling aims to assign semantic annotations to each pixel in an image, thereby distinguishing various categories through the delineation of segmented regions, each represented by a unique color. Semantic labeling is a core undertaking in the realm of VHRRSIs processing, and it serves as a critical technology for remote sensing application systems.

A. Single-Modal Semantic Labeling

Convolutional neural networks (CNNs) [10] constitute one of the most of prevalent architectural frameworks for deep learning within the domain of semantic segmentation for remote sensing imagery. Scholars have successfully enhanced the accuracy of semantic segmentation in RSIs by refining the conventional CNN architecture, including strategies such as deepening the network architecture and incorporating residual connections. FCN has revolutionized the approach by substituting the traditional CNN's fully connected layers with convolutional layers. Thereby facilitating end-to-end pixel-level classification.

Koltun et al. [11] enhanced FCN network performance by incorporating pyramid downsampling and deconvolution layers, albeit with less than optimal label accuracy. Inspired by FCNs, researchers have proposed a variety of improved FCN structures, such as U-Net [12], DeepLab series [13,14], etc., to further enhance the segmentation performance. A large number of encoder–decoder-based network structures are used for RSIs semantic labeling task, this type of network extracts image features by encoder and then realizes the reconstruction of feature maps by decoder so as to realize high-precision semantic segmentation, e.g., SegNet [15], and so on. The SegNet model constructed an encoder–decoder symmetric structure based on the FCN architecture to accomplish end-to-end pixel-level image segmentation, uniquely utilizing the decoder to upscale its lower-resolution input feature map. Chen et al. [16] expanded filter support and minimized input feature map downsampling for dense labeling. These methods have proven effective in pixel-level classification of RSIs, showcasing superiority over traditional pixel-level classification approaches that depend on manual feature descriptors. To address the limitations of some semantic segmentation methods that suffer from reduced image resolution due to convolution or pooling layers, the proposed spatial pyramid pooling module, called as PSPNet can introduce more contextual and multiscale information to minimize mis-segmentation probabilities. RefineNet [17] is a multipath reinforcement network that leverages all down-sampling process information and achieves high-resolution prediction through remote residual connections. Drawing inspiration from deep networks with stochastic depth, a Dropout-like approach has been proposed to enhance ResNet in DenseNet [18], significantly boosting its generalization capability [19]. The integration of Atrous Spatial Pyramid Pooling (ASPP) from the DeepLab series and dense connections from DenseNet in DenseASPP [20] yields a larger capture field and more dense sampling points, achieving state-of-the-art labeling on CityScape. FastFCN [21], proposed in 2019, improves semantic segmentation by incorporating the JPU (Joint Pyramid Upsampling) module into the semantic segmentation model. The UNet3+ [22] model is specifically designed for segmenting and labeling buildings in RSI.

More recently, Vision Transformer (ViT) [23] was proposed, which is an innovative approach to apply the Transformer architecture to computer vision, which achieves pixel-level prediction by categorizing each image block, such as SETR (Segment Transformer) [24,25] and TransUNet. Combining the advantages of ViT with those of CNNs improves the segmentation accuracy, which has also inspired many following works [25–28]. However, ViT is computationally and memory intensive, and is not friendly to mobile terminal deployment of algorithms, especially for high-resolution semantic labeling. And a multi-stage attention resu-net [29] is proposed for semantic segmentation of high-resolution RSIs. Swin-Unet [30] is proposed for medical image segmentation by Unet-like pure transformer. And swin transformer embedding UNet is used for RSIs semantic labeling.

B. Multimodal Semantic Labeling Mulitimodal remote sensing technology can fuse data from multiple sensors, such as optical, LiDAR, thermal infrared, etc., to provide richer and complementary feature information, thus improving the recognition accuracy of feature targets. In recent years, with the development of deep learning, multimodal RSIs semantic segmentation has made significant progress. Optical images can provide features, such as texture, color, etc., while DSM data from LiDAR represent height information of ground objects. For roads and buildings with more regular shapes, as well as trees and low vegetation with similar colors and geographic locations, the involvement of elevation information from DSM data makes it possible to better distinguish the two from each other in terms of differences in height. With the acquisition of DSM data no longer difficult, the study of optical image fusion of DSM data has gained more attention. Both FuseNet [31] and ResUNet-a [32] designed a dual-input deep learning network structure that fuses RGB data and DSM data, allowing the network to extract complementary features to improve segmentation accuracy. vFuseNet [33] is a similar structure. GSCNN [34] employs two parallel CNN structures for regular extraction and boundary-related information extraction, utilizing a traditional semantic segmentation model-like Regular stream and a Shape stream

dedicated to boundary information acquisition. Finally, these two streams are fused to generate segmentation results. CMGFNet [35] proposed a gated fusion module to combine two modalities for building extraction. CIMFNet [36] designed the cross-layer gate fusion mechanism. ABHNet [37] explored feature fusion based on attention mechanisms and residual connections. DKDFN [38] is domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification. And other similar multimodal networks [39–41] for classification of RSIs have obtained better results.

However, the fact that these methods ignore long-range spatial dependencies makes them perform poorly in extracting global semantic information. The transformer architecture is capable of capturing global contextual information in images, which is essential for distinguishing targets (different categories of vegetation) having similar appearances but different categories. TransFuser [42] incorporates the attention mechanism of transformer in the feature extraction layers of different modalities to fuse global contextual information of 3D scenes and integrate it into end-to-end autonomous driving tasks. TransUNet is a network that combines the transformer and U-Net architectures. It utilizes transformer to extract global contextual information while utilizing the encoder–decoder architecture of U-Net to maintain spatial information for accurate segmentation. And in 2024, TransUNet [43] was used for medical image segmentation through the lens of transformer by rethinking the U-Net architecture design. STransFuse [44] fused a swin transformer and convolutional neural network for RSIs semantic labeling. Similar networks for semantic labeling of RSIs are CMFNet [45], EDFT [46], MFTransNet [47], and FTransUNet [48], and the last one is proposed to provide a robust and effective multimodal fusion backbone for semantic segmentation by integrating both CNN and Vit into one unified fusion framework.

C. Contional Random Fields for Postprocessing

Energy-based random fields, such as Markov random fields (MRFs) [49] and conditional random fields (CRFs) [50], have proven invaluable for extracting background information from natural and RSI. In 2001, Lafferty [51] proposed a CRF model for 1-D sequence data processing based on MRF theory, effectively overcoming the aforementioned limitations of MRF. Currently, researchers continue to explore the combination of deep learning models and CRF. In recent studies, CNNs or FCNs integrated with CRF have been employed to enhance RSI segmentation accuracy [52], improve road detection [53], building detectionli2018building, and water body detection [54] efficiency.

D. Semisupervised/Unsupervised Learning Methods

A high-resolution image has hundreds of thousands of pixels, and sometimes even more. In order to alleviate the dependence on labeled data in training, semisupervised and unsupervised learning methods all reduce the need for labeled data in training. Semisupervised learning methods [55,56] are able to achieve better performance using consistency regularization and average updates of pseudolabels. Unsupervised learning methods [57,58] can accomplish recognition and segmentation tasks without a large amount of labeled data, using techniques such as data preprocessing, feature extraction, clustering, optimization iteration, and postprocessing. The application of unsupervised learning methods in the semantic labeling of RSIs is feasible, especially in the case of scarce labeled data. However, the disadvantage of unsupervised learning is its low performance compared with supervised learning methods.

3. Proposed Method

In this section, we aim to offer a comprehensive introduction to ScasMNet. Firstly, we propose a novel dual-path data fusion network that seamlessly integrates optical images and DSM data within an end-to-end fully convolutional network. This approach enhances the effectiveness of semantic labeling for VHRRSIs by fusing heterogeneous features. Furthermore, we discussed the principle of dilated convolution and its benefits for semantic labeling. Our method employs multiscale dilated convolutions within a fully convolutional deep network to facilitate semantic information fusion. Lastly, building upon the multiscale semantic information fusion achieved via dilated convolutions, we utilize

a dense conditional random field (DenseCRF) [59] model to establish point-to-potential energy relationships between all pixel pairs within the image, which results in improved refinement and segmentation outcomes.

3.1. Dual-Path Fully Convolution Network

DSM is a ground elevation model encompassing the heights of structures, such as buildings, bridges, and trees on the ground. DSM genuinely depicts the undulations of the ground and finds applications in a broad range of industries. As remote sensing technology advances, obtaining DSM has become more convenient, rapid, and precise. Leveraging the land surface undulation information provided by DSM can significantly enhance the effectiveness of RSI analysis.

Nevertheless, combining three-channel optical data and DSM into a four-input dimension structure and feeding it into the network is not an optimal approach. This is due to the distinct information contained in these two data sources. The optical data from IRRG images, obtained from the same sensor, typically encompasses appearance information such as object color and texture. In contrast, DSM data, acquired using a different type of sensor, represents the height information of surface undulations. Consequently, our focus is on exploring how to process and fuse these heterogeneous source data to enhance model performance.

Drawing inspiration from the concept of multimodal fusion, we present a technique to enhance the semantic labeling accuracy of VHRRSIs. Our approach employs the FCN framework overall while constructing a multimodal network that incorporates dual-path data input to enhance feature diversity and expand the network's capabilities. The primary objective of our method is to enhance the accuracy of semantic labeling for RSIs. Figure 1 illustrates the complete structure of the proposed network.

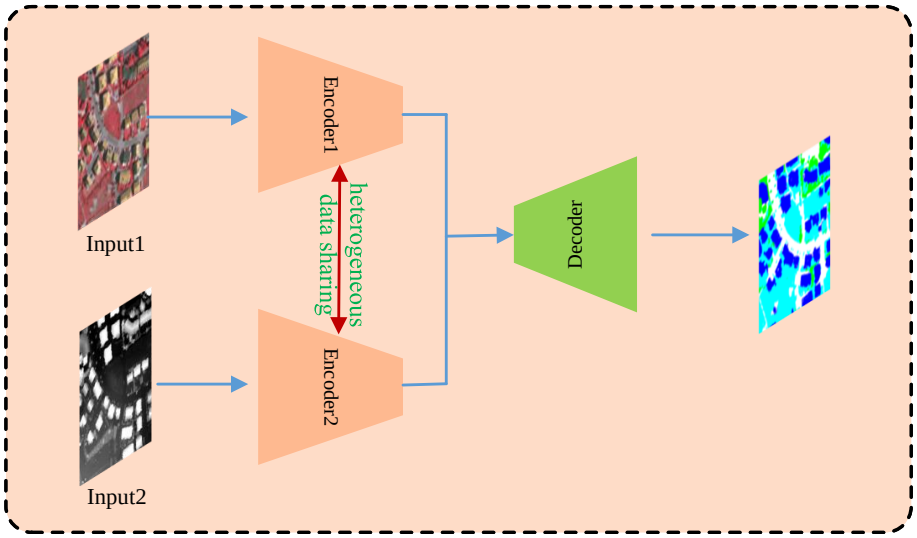

Figure 1. the overall architecture of the dual-path fully convolutional network.

3.2. Multiscale Feature Fusion

Dilated convolution, also referred to as atrous convolution, was designed to address image segmentation challenges. Traditional image segmentation algorithms often employ pooling layers and convolutional layers to broaden the receptive field, thereby reducing the size of the feature map. This is followed by upsampling to restore the original image size. However, the process of shrinking and expanding the feature map may result in accuracy losses. In contrast, dilated convolutional operations can increase the receptive field without altering the size of the feature map. This approach is utilized in this study to encompass a wider spectrum of information, thereby expanding the receptive field. Dilated convolution incorporates a hyperparameter known as the "dilation rate", which

determines the distance between values when the convolution kernel processes the data. In our research, we explored various dilation rates (rate = 6, rate = 12, rate = 18, rate = 24) to systematically aggregate multiscale contexts without sacrificing resolution. This approach allowed us to achieve high-precision dense predictions.

As depicted in Figure 2, dilated convolution utilizing the same feature map achieves a larger receptive field compared with basic convolution, resulting in denser data acquisition. A broader receptive field enhances the overall performance of small object recognition and segmentation tasks. Importantly, employing dilated convolution instead of downsampling or upsampling effectively preserves the spatial characteristics of the image without compromising information loss. When network layers demand a larger perceptual field but increasing the number or size of convolution kernels is impractical due to limited computational resources, the use of dilated convolution proves to be advantageous.

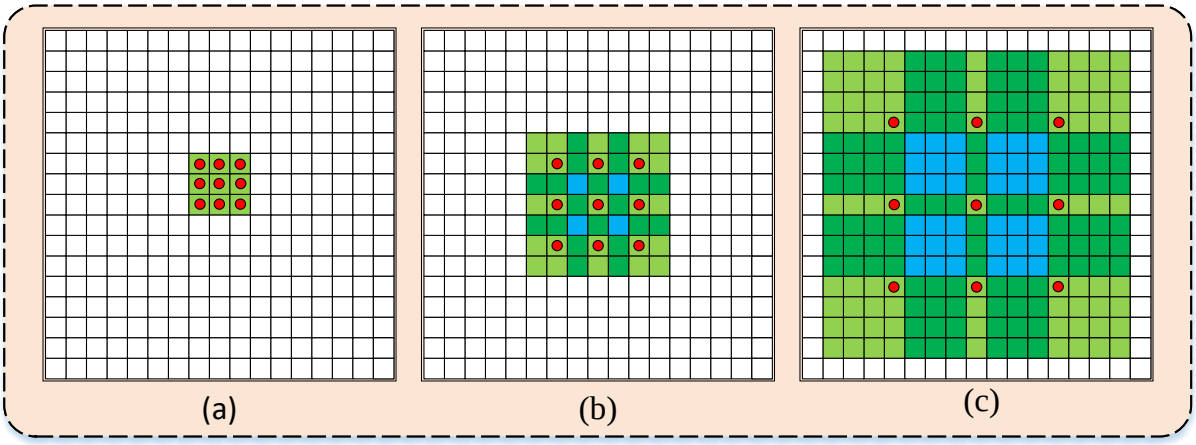

Figure 2. Illustration of the dilated convolution process: (**a**) is the general receptive field of 3 × 3 convolution. (**b**) is based on the convolution of (**a**), and the dilation rate is set to 2. On the basis of the original 3 × 3 convolution, the receptive field is expanded to 7 × 7 with the dilated convolution. (**c**) is based on the convolution of (**b**), and the dilated parameter is set to 4. On the basis of the original receptive field of 7 × 7, the receptive field is expanded to 15 × 15 by the whole convolution.

In this paper, we propose a parallel dilated convolution module with varying dilation rates to expand the receptive field and enhance the network's ability to extract features. This module, referred to as the multiscale convolutional block, is presented in Figure 3. The primary operation involves performing feature fusion in the final stage of the encoder, which serves as the input feature for the decoder stage. For the input feature map, we conduct four parallel dilated convolution operations with distinct dilation rates. Subsequently, we fuse the corresponding feature maps of the same size from the encoder and perform convolution operations with a 1 × 1 × 6 kernel size. Ultimately, the four branches of the parallel dilated convolution generate feature maps of identical scales. After fusion, the remaining operations in the decoder stage are executed. A comprehensive overview is provided in Figure 3.

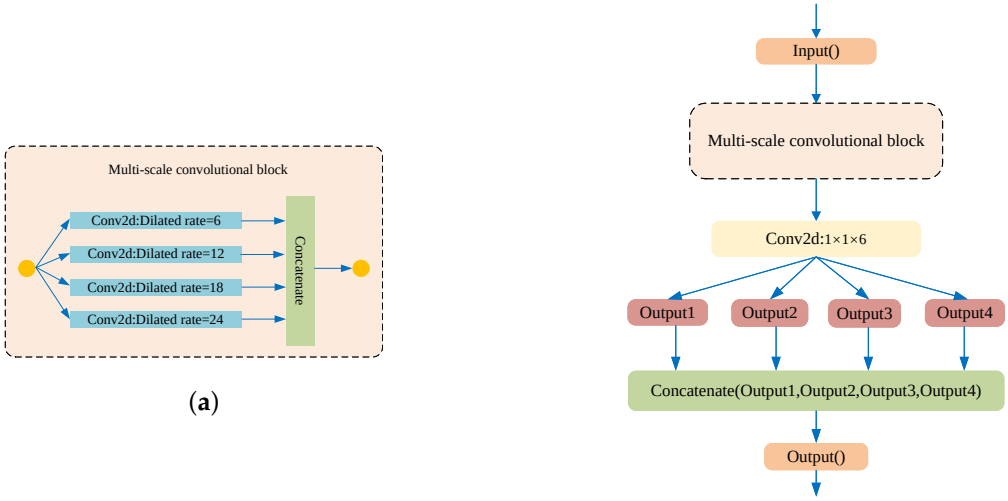

Figure 3. Illustration of the multiscale convolution: (**a**) the multiscale convolutional block and (**b**) overview of the ScasMNet.

3.3. *DenseCRF Model*

Our model employs several upsampling operations utilizing deconvolution, which not only resizes the feature map back to its original image dimensions but also leads to feature loss. This, in turn, generates blurred classification target boundaries. To achieve more precise final classification results, we incorporate the DenseCRF model following the presegmentation outcomes. DenseCRF is an enhanced version of CRF that optimizes the deep learning-based classification results by considering the relationships among all pixels in the original image. This approach corrects missegmented regions and provides more detailed segmentation boundaries.

In the DenseCRF model, the Gaussian kernel function of the pixel pair is expressed by Equation (1),

$$k(f_i, f_j) = \omega^{(1)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) + \omega^{(2)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right) C^{(1)} \quad (1)$$

where $-\frac{|p_i - p_j|^2}{2\theta_\alpha^2}$ takes into account the shape, texture, and $\frac{|I_i - I_j|^2}{2\theta_\beta^2}$color information of the pixel pairs in the image and considers the position information of the pixel pairs. As can be seen, the function considers both color and position information by incorporating $-\frac{|p_i - p_j|^2}{2\theta_\alpha^2}$ and $\frac{|I_i - I_j|^2}{2\theta_\beta^2}$. It encourages pixels with similar colors and close positions to be assigned the same label, while pixels with greater differences receive different labels. Consequently, the DenseCRF model can segment images along boundaries as accurately as possible, providing a comprehensive description of the relationships between pixels regarding color and position.

In our network architecture depicted in Figure 4, a dual-path fully convolutional network model is employed to fuse IRRG data and DSM data and appropriately cascade them to achieve feature complementarity. The encoder section comprises two input data paths, one for optical channel data and the other for DSM data. Given that the topologies of the two encoder branches are similar, cropping the input images of both branches to the same size and normalizing the DSM data to nDSM enables them to share the same range of values as the optical path images. In the final part of the encoder, a dilated convolution operation employing diverse dilation rates is utilized to acquire feature maps at various

scales, with the maps being cascaded at the same scale. The input of the dual-stream data fusion module consists of optical IRRG data and DSM data, referred to as a four-channel image. The feature map thus originates from the two aforementioned branches. The feature map is represented as $[a^T, bT]T$, where a and b denote the features learned from the IRRG path and DSM path, respectively. The output of the fusion module can be expressed as Equation (2):

$$x_f = \omega_1 a^T + \omega_2 b^T \tag{2}$$

where ω_1 and ω_2 denote the fusion weights from the respective streams. Feature fusion is executed throughout the training process, allowing the learned features and fusion weights to be adjusted and optimized together. Consequently, the fusion weight that can be learned ensures a more suitable fusion strategy by controlling the contribution of the two data sources to the segmentation target based on the difference in extracted characteristics from the two diverse branches of the data stream. In this paper, we utilize the network structure diagram presented in Table 1 as the foundation for relevant network fusion. Notably, the dual-stream network fusion is achieved by expanding and convolving the second half of the network based on the dual inputs' realization, thereby enabling parallelism at this stage. The two-stream network incorporates shallow-layer feature information, which is richer than the previous single-layer result. Redundant information is discarded through pooling operations, followed by expansion convolutions after fusion. The employment of dilated convolutions enlarges the receptive field and relatively enhances high-level semantic segmentation outcomes. Furthermore, dense CRF is incorporated into smooth segmentation edges and enhances segmentation accuracy.

Table 1. detailed architectures of our proposed dual-path model.

Name	Structure		Stride	Output
	Branch 1	Branch 2		
Input	256 × 256 × 3	256 × 256 × 3		
Conv_layerl	Conv2D	Conv2D		256 × 256 × 64
Conv_layer2	Conv2D	Conv2D		256 × 256 × 64
	Maxpooling2D	Maxpooling2D	2	128 × 128 × 64
	Conv2D	Conv2D		128 × 128 × 128
Conv_layer3	Conv2D	Conv2D		128 × 128 × 128
	Maxpooling2D	Maxpooling2D	2	64 × 64 × 128
	Conv2D	Conv2D		64 × 64 × 256
	Conv2D	Conv2D		64 × 64 × 256
	Conv2D	Conv2D		64 × 64 × 256
	Maxpooling2D	Maxpooling2D	2	32 × 32 × 256
Conv_layer4	Conv2D	Conv2D		32 × 32 × 512
	Conv2D	Conv2D		32 × 32 × 512
	Conv2D	Conv2D		32 × 32 × 512
	Maxpooling2D	Maxpooling2D	2	32 × 32 × 512
	AtrousConv2D	AtrousConv2D		32 × 32 × 512
	Maxpooling2D(model 1)	Maxpooling2D(model 1)		32 × 32 × 512
Concatenate	Concaenate(model 1.mode2)			32 × 32 × 1024
AtrousConv2D(b1)	AtrousConv2D			32 × 32 × 1024
	Conv2D			32 × 32 × 1024
	Conv2D			32 × 32 × 6

Table 1. *Cont.*

Name	Structure		Stride	Output
	Branch 1	Branch 2		
AtrousConv2D(b2)	AtrousConv2D Conv2D Conv2D			32 × 32 × 1024 32 × 32 × 1024 32 × 32 × 6
AtrousConv2D(b3)	AtrousConv2D Conv2D Conv2D			32 × 32 × 1024 32 × 32 × 1024 32 × 32 × 6
AtrousConv2D(b4)	AtrousConv2D Conv2D Conv2D			32 × 32 × 1024 32 × 32 × 1024 32 × 32 × 6
Concatenate	Concatenate(b1.b2.b3.b4)			32 × 32 × 24
UpSmapling2D	UpSmapling2D Conv2D		8	256 × 256 × 6 256 × 256 × 6
Output	Softmax layer			256 × 256 × 6

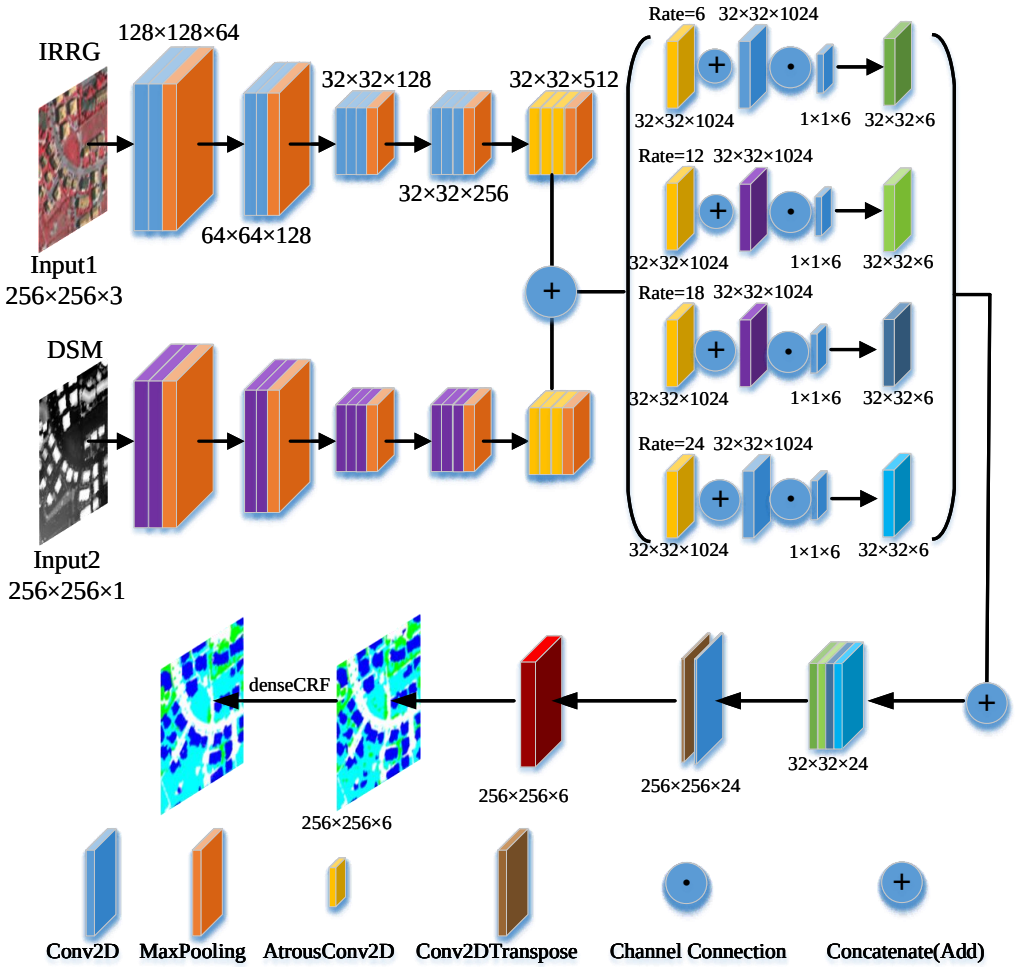

Figure 4. Architecture of the proposed ScasMNet. Different colors indicate different layer types. The ScasMNet contains two branches in the encoding part to extract complementary information from the RGB data and DSM data, respectively.

4. Experimental Results

In this segment of the document, we implemented the proposed method and assessed its performance on the two datasets provided by the ISPRS competition, aiming to verify its feasibility and effectiveness. We further compared it with several established deep learning models, including FastFCN, PSPNet, DeepLabv3+, MFTransUNet, and CMFNet. The section is structured as follows: It commences with a succinct overview of the utilized datasets and assessment metrics and subsequently proceeds to the comprehensive presentation of the overall outcomes. And finally, we show the ablation studies and conclusion.

4.1. Data Description

We conducted experimental assessments using datasets obtained from Vaihingen and Potsdam with a resolution finer than a decimeter. These datasets are state-of-the-art airborne image datasets provided by the ISPRS 2-D semantic labeling challenge, covering true orthophoto tiles of extremely high resolution and their corresponding DSMs generated through dense image-matching techniques. The Vaihingen dataset comprises 33 RSIs of varying dimensions, with an average size of 2494 $\times$ 2064. Each image is composed of three bands: Near Infrared (NIR), Red (R), and Green (G), offering a spatial resolution of 9 cm. Sixteen of these images have complete annotations for six primary land cover/land use classes (impervious surfaces, building, low vegetation, tree, car, and clutter/background). Additionally, the associated DSMs generated through dense image matching techniques are included, and they have undergone normalization to nDSMs, as discussed in [60]. The Potsdam dataset comprises 38 patches of size (6000 $\times$ 6000), with a ground sampling distance of 5 cm for both TOP and DSMs. The manually labeled categories and numbering of the Potsdam dataset are consistent with those of the Vaihingen dataset.

In this dataset, "building" and "impervious surface" are relatively easy to identify and segment accurately due to their regular shapes. However, distinguishing between "trees" and "low vegetation" can be challenging, as they share similar colors and are often geographically connected. "Car" segments exhibit the lowest accuracy due to their smaller target size and vulnerability to occlusion. These issues are indeed prevalent in other RSIs as well.

4.2. Evaluation Metrics

To facilitate the assessment of the model's impact on image segmentation accuracy, it is crucial to establish a unified standard for evaluating the model's accuracy. Within image semantic segmentation, frequently used performance assessment metrics encompass accuracy, recall, precision, F1, and MeanIoU (mIoU) . The calculation formulas for these metrics are provided below.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

$$\text{MeanIoU} = \frac{TP}{TP + FP + FN} \tag{7}$$

where TP (True Positive) signifies both the actual and predicted labels being positive, indicating a correct (true) prediction. FN (False Negative) represents a false prediction despite the actual label being true. FP (False Positive) indicates a positive prediction despite the actual label being negative, while TN (True Negative) denotes both the actual and predicted labels being negative. In our experiments, we employ three specific quantitative evaluation metrics: overall accuracy (OA), F1-score, and per-class average pixel-level accuracy, in compliance with dataset guidelines. OA functions as a comprehensive measure of segmentation accuracy, providing an overview of the proportion of accurately classified pixels. Nevertheless, a drawback of OA is its tendency to prioritize classes with a significant number of samples, potentially overshadowing the contributions of smaller classes with larger ones. Conversely, the F1-score is specific to each class and remains unaffected by class size. It represents the balanced average of precision and recall.

4.3. Implementation Details

VHRRSIs are typically comprising thousands of pixels or more. Transmitting such massive images to a deep learning network in a single go is challenging. Furthermore, the labeled data within a benchmark is often limited, and not all of the datasets in ISPRS are annotated. To address these issues, We partitioned the original image data into a sequence of uniform-sized overlapping patches employing a sliding window method. This technique not only enlarged the training set but also enabled the deep learning network to undergo batch training, thus reducing computational demands. Following the aforementioned processing steps, we obtained numerous fixed-size training datasets and their corresponding labeled data. Prior to training, we subjected them to random transformations to augment the dataset and increase the randomness of the data. These transformations included rotating the training images by 90°, 180°, and 270°, randomly scaling them, adding noise, and horizontally/vertically flipping them. In our experimental set-up, the initial training images, which include IRRG and nDSM data, are standardized to have a mean of zero and a variance of one. Both the raw images and the labeled ones used for training are divided into a series of patches measuring 256 × 256, with corresponding numbers marked. For the Vinhingen dataset, 10 original images are employed for training, resulting in 40,000 patches from the training set. Similarly, for the Potsdam dataset, 10 original images are used for training, yielding 10,000 patches from the training set. The respective numbers of images for the training and test data used in this study are presented in Table 2, with a ratio of 3 training patches to 1 validation patch.

Table 2. Datasets descriptions and training details.

Property	Vaihingen Dataset	Pot Sdam Dataset
Training set (ID)	1, 2, 3, 4, 5, 6, 7, 11, 23, 32	2_10, 2_11, 2_12, 3_11, 3_12, 4_11, 4_12, 6_9, 6_10, 7_10
Prediction set (ID)	13, 15, 17, 26, 27, 28, 30, 34, 37, 38	5_11, 5_12, 6_7, 6_11, 6_12, 7_7, 7_8, 7_9, 7_11, 7_12
Average size	6000 × 6000	2560 × 2046
Training image size	256 × 256	256 × 256
Batch size	4	2

This experimental environment utilizes Python 3.5, with the network application built on the TensorFlow and Keras framework. The hardware platform is an NVIDIA-SMI 440.44 GPU, employing Cuda 10.2 for accelerated calculations, and a GPU memory of 11.91 GiB. Throughout the network's application process, numerous experiments revealed that a learning rate of lr = 1×10^{-4} yielded the best results for the fusion network. To prevent premature convergence caused by the network's deep structure, the Batch Normalization (BN) layer and the Rectified Linear Unit (ReLu) layer were added to the

convolutional layer. Additionally, issues, such as the convergence phenomenon and slowed network training speed due to the extensive workload were addressed. Simultaneously, a basic semantic labeling network was compared and tested on the corresponding dataset, including FCN-16s, SegNet, U-Net, ICNet, DeepLabV3+, PspNet, and other network applications, all based on the TensorFlow and Keras framework. All optimizers used in this study are Adam optimizers, aiming to achieve the control variable method comparison experiment's objectives.

4.4. Experimental Results and Analysis

In this section, we showcase our experimental outcomes obtained using our proposed method on two datasets sourced from the ISPRS competition. These include numerical and visual results, along with a comparison of our approach with other classical models from previous literature. For the alignment of data in all tables, impervious surfaces are abbreviated with Imp. surf, and low vegetation is represented by Low veg.

4.4.1. Experimental Results on the Vaihingen Dataset

Table 3 presents the results obtained from the Vaihingen datasets, with bold numbers indicating superior performance. The second to sixth rows display the test results of five comparison methods utilizing classical semantic segmentation models, while the last row showcases the outcomes achieved by our proposed method. By neglecting the "clutter" category, our approach achieved the best performance across various evaluation metrics, demonstrating excellent performance in five distinct categories. Notably, the accuracy in classifying the "building" category achieved 94.82%, marking the highest among all classes. This indicates that a majority of the pixels belonging to the "building" class were accurately classified. This is attributed to the relatively simple texture information of this category, making it easier for deep learning models to extract relevant feature details and produce accurate segmentation. In comparison, the "car" class achieved the lowest segmentation accuracy, indicating that objects with dense distributions and small targets are less likely to be precisely segmented. This highlights one of the research challenges in VHRRSIs semantic labeling task.

Table 3. Classification accuracy in the Vaihingen dataset.

Methods \ Ouantitative Metrics (%)	IoU					F1	OA	MIoU
	Imp.Surf	Building	Low Veg.	Tree	Car			
DeepLabv3+ [14]	92.55	93.34	82.93	89.29	73.35	85.34	89.63	86.29
PspNet [17]	87.33	90.17	83.69	80.50	71.22	85.12	86.61	82.58
FastFCN [22]	87.14	89.06	81.37	82.54	70.82	83.95	84.40	82.19
CIMFNet [47]	90.21	91.52	85.32	84.23	74.68	87.22	88.53	85.19
TransUNet [49]	93.03	92.76	86.03	86.57	81.65	88.67	89.22	88.01
ScasMNet	93.36	94.82	86.85	90.29	86.73	**88.67**	**90.38**	**90.41**

As shown in Table 3, it can be seen that the segmentation performance of the multimodal fusion methods is better than those of single-modal segmentation methods. However, our proposed ScasMNet achieved optimal results. MFTransUNet is able to recover local information due to its powerful coding capability, making its segmentation performance suboptimal. The computational cost required for each method is discussed in Section 4.7. The segmentation accuracy is around 90% for both regular-shaped buildings and impervious surfaces. Especially for trees and low vegetation with similar colors and locations, the segmentation performance of the multimodal approach is substantially improved. And for densely distributed small-scale targets, such as cars, the segmentation accuracy of our method reached 86.73%, which is five percentage points higher than MFTransUNet.

Since DSM data are a single-channel grayscale image, it is not conducive to visual observation. Therefore, the original DSM images are not shown in the subsequent visualization results.

Experimental results for three scenarios are presented in Figure 5. Most of the pixels in the first row belong to either the "building" class (labeled in blue) or the "impervious surface" class (labeled in black). It can be seen that our method can segment all pixels correctly, and fewer pixels are wrongly segmented. Of course, these two classes are also the easiest to segment. However, there are still obvious missegmentations used PSPNet and CMFNet. The circular marking results are the most obvious, but not only limited to there.

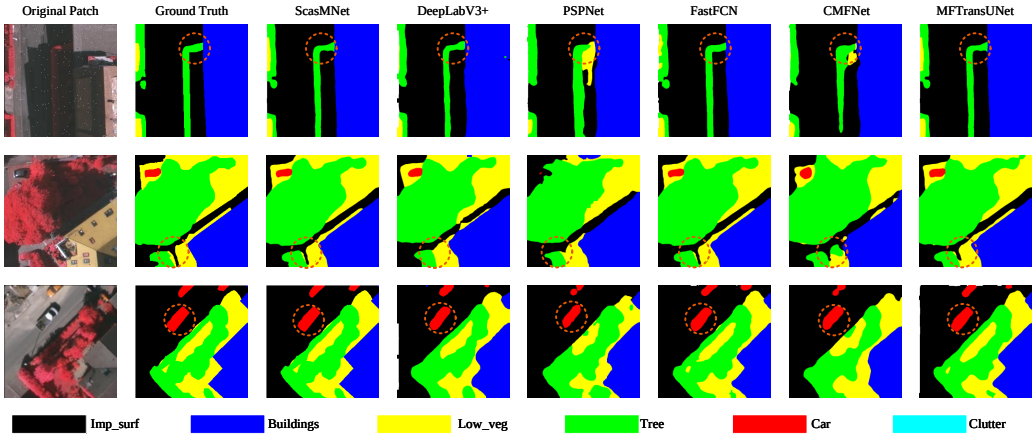

Figure 5. Example classification patches of Vaihingen validation datasets with the comparison architectures. Annotation: White: impervious surfaces. Blue: building. Cyan: low vegetation. Green: trees. Yellow: cars. Red: clutter, background.

Most pixels in the second row belong to either the "tree" class (green labeling) or the "low vegetation" class (yellow labeling), which are also difficult to segment due to their similar colors and locations. However, our method performed optimally to segment all boundaries. The DSM data are input to be able to extract complementary features to further distinguish these two objects from each other in terms of height.

The last row has a small-scale target object (Car), for which segmentation is most difficult. For this reason, we use multiscale dilated convolution (rate = 1, 6, 12, 18) in our method, with the aim of being able to extract features of large-scale targets (buildings and impervious surfaces) while not ignoring the presence of the small-scale targets.

In conclusion, we can see that the qualitative and quantitative results are consistent.

4.4.2. Experimental Results on the Potsdam Dataset

For the Potsdam dataset, we replicated the comparison experiments by training the same network model using identical training data.

Table 4 displays the performance of six different models, with the last row showcasing the results of our proposed ScasMNet model. Our model evidently showcases superior performance compared with the previously mentioned models. The segmentation accuracy of the "Low veg." class has been improved by 3.1% at least, and the "Tree" class has seen an improvement of 1.3% at least. This demonstrated that DSM possesses superior recognition ability for objects with similar spectral information but varying height profiles. Most notably, the segmentation accuracy of the "car" class has been enhanced by nearly 10% compared with signal-modal-based FastFCN. We can thus infer that employing multiple atrous convolutions with distinct atrous rates contributes to extracting richer features of small target objects. Simultaneously, the complementary DSM has played an indispensable role. Overall, our proposed ScasMNet model yields better performance for semantic segmentation of RSI compared with other models inferred in our experiments.

Table 4. Classification accuracy in the Potsdam dataset.

Methods \ Ouantitative Metrics (%)	IoU					F1	OA	MIoU
	Imp.Surf	Building	Low Veg.	Tree	Car			
DeepLabv3+ [14]	92.46	93.27	87.28	86.12	79.54	85.92	87.55	87.73
PspNet [17]	84.33	92.35	81.83	79.88	73.28	84.67	87.68	82.33
FastFCN [22]	86.94	90.28	84.59	84.61	78.22	85.64	86.77	84.93
CIMFNet [47]	89.30	91.54	87.42	86.59	75.23	88.70	89.58	86.02
TransUNet [49]	93.22	92.56	85.26	88.17	84.22	89.31	90.55	88.69
ScasMNet	93.36	94.82	95.30	95.76	88.60	**89.67**	**92.46**	**91.74**

The qualitative results for three different scenarios are presented in Figure 6. Most of the pixels in the first row of patches belong to either impervious surfaces (labeled in black) or buildings (labeled in blue). It can be seen that our proposed method correctly segmented the edges and contours of these two classes with almost no misclassification. As for the other five methods, all of them have some misclassification. Most of the pixels in the second row of the patch belong to the "tree" class (labeled in green) or the "low vegetation" class (labeled in yellow). Our proposed ScasMNet's performance is the best, almost the same as the ground truth. For example, FastFCN and MFTransUNet do not segment the trees in the upper right corner correctly, while the misclassification of CMFNet is also obvious. The third row of paths covers cars, which are densely distributed and small. Our method also correctly segmented the targets, checking the places marked by circles in each patch. In summary, the qualitative visualization and quantitative results remain completely consistent.

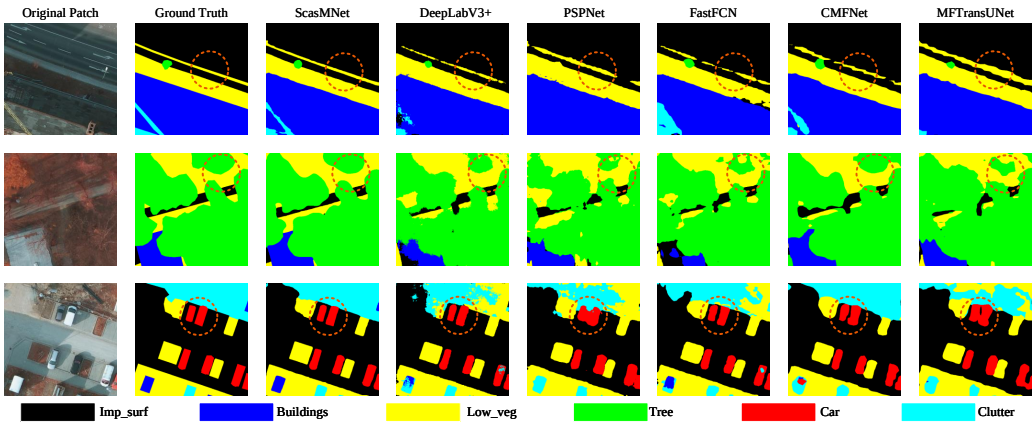

Figure 6. Example classification patches of Potsdam validation datasets with the comparison architectures. Annotation: White: impervious surfaces. Blue: building. Cyan: low vegetation. Green: trees. Yellow: cars. Red: clutter, background.

4.5. Effect of the Size of a Patch

To achieve optimal experimental results, we randomly and repetitively divide the training samples into patches of 128 × 128, 256 × 256, and 512 × 512 and send them to the training network for comparative experimental analysis. As evident from the data in Figures 7 and 8, it can be seen that when the training samples are cropped to 256 × 256, the highest values for OA, F1, and MIoU are obtained. In contrast, when cropping to 128 × 128 or 512 × 512, the evaluation metrics did not improve but decreased. Consequently, in all experiments conducted throughout this paper, we randomly and repetitively crop the original image to a size of 256 × 256.

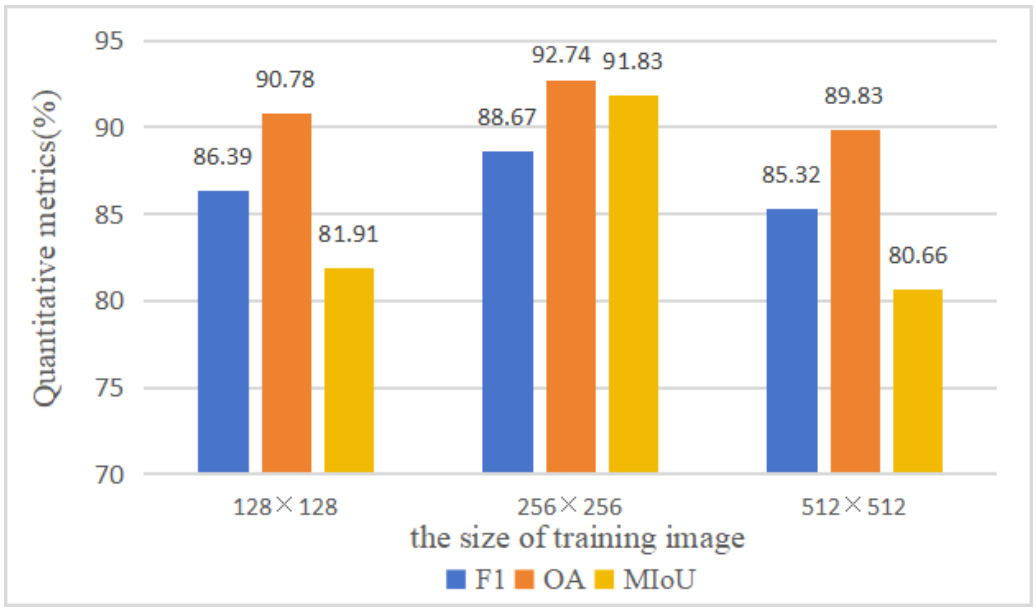

Figure 7. Segmentation accuracy (%) of the proposed algorithm for the Vinhingen dataset with training image sizes of 128 × 128, 256 × 256, and 512 × 512.

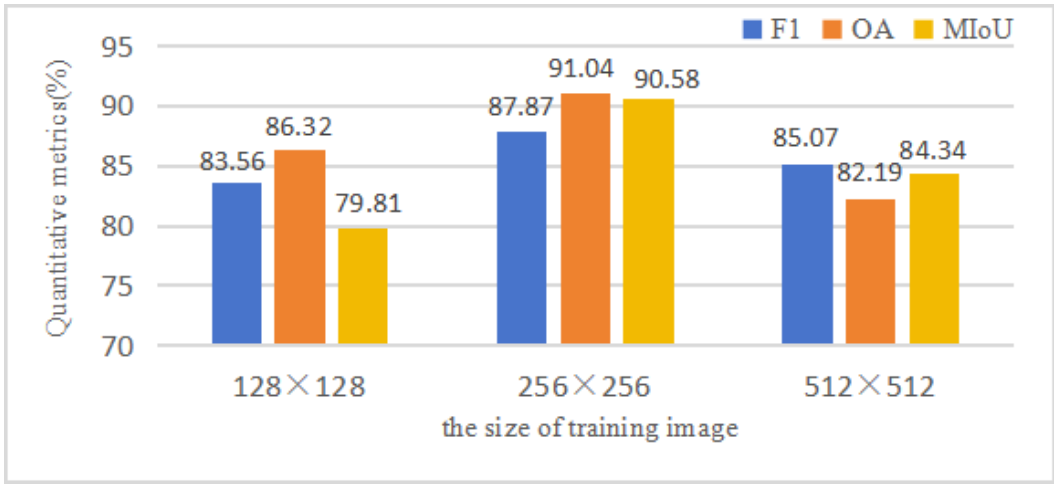

Figure 8. Segmentation accuracy (%) of the proposed algorithm for Potsdam dataset with training image sizes of 128 × 128, 256 × 256, and 512 × 512.

4.6. Ablation Study for ScasMNet

To demonstrate the substantiality of ScasMNet, we conduct ablation experiments on the aforementioned two datasets, with the results presented in Tables 5 and 6. These findings demonstrate that when training samples are cropped to 256 × 256 and RGB and nDSM are employed as inputs to ScasMNet utilizing a dual-path data strategy, the segmentation accuracy achieves optimal performance.

As depicted in Table 5, when employing the identical network architecture with RGB single-path data input and training sample sizes of 128 × 128 or 256 × 256, the values of OA, F1, and MIoU witness minor fluctuations, yet these changes are not particularly significant. For the Potsdam dataset, Table 6 presents analysis results that are virtually identical to those in Table 5. By utilizing a dual-path input and 256 × 256 training samples, the segmentation accuracy achieves the highest level throughout the entire experiment, essentially registering a five percentage point increase.

Table 5. Ablation study of Vinhingen dataset.

Quantitative Metrics		F1 (%)	OA(%)	MIoU (%)
ScasMNet (only IRRG)	128 × 128	84.94	82.67	83.68
	256 × 256	85.29	83.34	83.74
ScasMNet (IRRG + DSM)	128 × 128	83.22	85.92	83.77
	256 × 256	**88.67**	**90.38**	**90.41**

Table 6. Ablation study of Vinhingen dataset.

Quantitative Metrics		F1 (%)	OA (%)	MIoU (%)
ScasMNet (only IRRG)	128 × 128	82.20	83.06	82.95
	256 × 256	84.82	86.31	84.85
ScasMNet (IRRG + DSM)	128 × 128	82.91	85.17	84.30
	256 × 256	**89.67**	**92.74**	**91.74**

To validate the effectiveness of the DenseCRF module, that is the postprocessing operation, we also implemented ablation experiments. In our proposed ScasMNet model, the ablation experiments were performed with patch sizes of 256 × 256 as input, which were removed from and added to the DenseCRF, respectively. The results on the Vaihingen dataset and the Potsdam dataset are shown in Table 7 and Table 8, respectively. The symbol "+" represents the addition of a DenseCRF module, and the symbol "-" represents the removal of the corresponding module.

Table 7. Ablation study of Vaihingen dataset.

Methods \ Ouantitative Metrics (%)	IoU					F1	OA	MIoU
	Imp.Surf	Building	Low Veg.	Tree	Car			
ScasMNet-DenseCRF	92.56	93.88	85.40	88.97	85.41	87.44	88.56	89.24
ScasMNet+DenseCRF	93.36	94.82	86.85	90.29	86.73	88.67	90.38	90.41

Table 8. Ablation study of Potsdam dataset.

Methods \ Ouantitative Metrics (%)	IoU					F1	OA	MIoU
	Imp.Surf	Building	Low Veg.	Tree	Car			
ScasMNet-DenseCRF	93.87	94.93	87.88	89.50	86.44	88.64	91.82	90.52
ScasMNet+DenseCRF	95.30	95.76	88.60	90.71	88.32	89.67	92.74	91.74

From the results shown in Tables 7 and 8, we can summarize that adding the post-processing operation in the DenseCRF module can effectively improve the segmentation accuracy by at least 1%, as can be seen from the index values of F1, OA, and MIoU. In order to avoid repetition, the textural content will not be repeated.

4.7. Model Complexity Analysis

We evaluate the computational complexity of the proposed ScasMNet using the floating point operation count (FLOPs) and the number of model parameters. FLOPs are used to evaluate the model complexity whereas the number of model parameters. Ideally, an efficient model should have a smaller value in the FLOPs and the number of model parameters.

Table 9 showed the complexity analysis results of all comparing methods considered in this paper. Table 9 indicates that the proposed ScasMNet exhibited lower FLOPs, fewer

parameters, and smaller memory occupancy than conventional CMFNet and TransUNet. It is observed that the proposed ScasMNet demonstrated better performance than other methods. Single-modal methods have lower FLOPs and fewer parameters than those of multimodal methods because the former just have one modal input and less computational complexity.

Table 9. Comparison of different methods.

Method	DeepLabV3	PSPNet	FastFCN	CMF Net	TransUNet	ScasMNet
Multimodal	N	N	N	Y	Y	Y
FLOPs (G)	46.54	43.57	48.54	8026	75.06	59.23
Parameter (M)	36.81	50.93	70.33	130.02	96.18	78.11
Memory (MB)	2765	3089	3167	3757	3355	3209
mIoU (%)	86.29	82.58	82.19	85.19	88.01	90.41

5. Conclusions

In this paper, a novel FCN-based Self-Cascaded Multi-Modal and Multi-Scale Fully Convolutional Neural Network was proposed for the semantic segmentation of VHRRSI. Our framework boasts three significant advantages. First, the dual-channel input framework is employed to facilitate information complementarity between the two-channel data, resulting in richer extracted features. Second, our approach enhances the complementarity of features at different scales between layers through a multiscale feature fusion mechanism, allowing the network to accurately and efficiently extract rich and useful features. Lastly, DenseCRF is applied to presegmentation results, taking into full consideration the spatial consistency relationship between pixels and thereby improving segmentation accuracy.

Experimental findings indicate that the ScasMNet model design enhances the segmentation accuracy of trees and low vegetation with similar color and geographic location. This is attributed to the fact that elevation information from DSM complementarily augments spatial information from RGB, as evident from ablation experiment results. Furthermore, the incorporation of both the multiscale module and DenseCRF leads to improved segmentation accuracy for other categories, with optimal performance observed for small-sized cars.

However, there are several extensions of this study that can be further explored. In particular, distinguishing trees from low vegetation remains challenging. Therefore, it is of interest to develop new strategies for ground targets with similar colors and irregular boundaries, without degrading the segmentation accuracy of small-scale target objects. In addition, due to the high resolution of remote sensing images, it is of great relevance to explore image-based elevation estimation for downstream remote sensing tasks, such as crop identification and planting decision implementation and plant disease identification and growth monitoring. Finally, research on incorporating large-scale models, such as the segment anything model (SAM), into the semantic segmentation framework in remote sensing is needed.

Author Contributions: Conceptualization, Q.H. and Y.L.; methodology, Q.H.; software, Q.H.; validation, Q.H., J.F. and F.W.; formal analysis, F.W.; investigation, Q.H.; resources, Q.H.; data curation, J.F.; writing—original draft preparation, J.F.; writing—review and editing, J.F.; visualization, J.F.; supervision, J.F.; project administration, Q.H.; funding acquisition, F.W. All authors have read and agreed to the published version of the manuscript.

Funding: The National Natural Science Foundation of China (62271400), and the Shaanxi Provincial Key R&D Program, China (2023-GHZD-02).

Data Availability Statement: It can be provided upon request from corresponding author.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Deep learning for hyperspectral image classification: An overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709. [CrossRef]
2. Yang, X.; Ye, Y.; Li, X.; Lau, R.Y.; Zhang, X.; Huang, X. Hyperspectral image classification with deep learning models. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5408–5423. [CrossRef]
3. Yao, H.; Yu, Q.; Xing, X.; He, F.; Ma, J. Deep-learning-based moving target detection for unmanned air vehicles. In Proceedings of the 2017 36th Chinese Control Conference (CCC), Dalian, China, 26–28 July 2017; pp. 11459–11463.
4. Khan, M.J.; Yousaf, A.; Javed, N.; Nadeem, S.; Khurshid, K. Automatic target detection in satellite images using deep learning. *J. Space Technol.* **2017**, *7*, 44–49.
5. Chen, Y.; Cheng, L.; Li, M.; Wang, J.; Tong, L.; Yang, K. Multiscale grid method for detection and reconstruction of building roofs from airborne LiDAR data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4081–4094. [CrossRef]
6. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
7. Järnstedt, J.; Pekkarinen, A.; Tuominen, S.; Ginzler, C.; Holopainen, M.; Viitala, R. Forest variable estimation using a high-resolution digital surface model. *ISPRS J. Photogramm. Remote Sens.* **2012**, *74*, 78–84. [CrossRef]
8. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
9. Desmaison, A.; Bunel, R.; Kohli, P.; Torr, P.H.; Kumar, M.P. Efficient continuous relaxations for dense CRF. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 818–833.
10. Guo, T.; Dong, J.; Li, H.; Gao, Y. Simple convolutional neural network on image classification. In Proceedings of the 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, 10–12 March 2017; pp. 721–724.
11. Alam, M.; Wang, J.F.; Guangpei, C.; Yunrong, L.; Chen, Y. Convolutional neural network for the semantic segmentation of remote sensing images. *Mob. Netw. Appl.* **2021**, *26*, 200–215. [CrossRef]
12. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th international conference, Munich, Germany, 5–9 October 2015; proceedings, part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
13. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
14. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
15. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
16. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
17. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
18. Guo, X.; Chen, Z.; Wang, C. Fully convolutional DenseNet with adversarial training for semantic segmentation of high-resolution remote sensing images. *J. Appl. Remote Sens.* **2021**, *15*, 016520. [CrossRef]
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
20. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.
21. Wu, H.; Zhang, J.; Huang, K.; Liang, K.; Yu, Y. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv* **2019**, arXiv:1903.11816.
22. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and sIgnal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
24. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
25. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
26. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
27. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G.; Zhang, D. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–15. [CrossRef]

28. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.R. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. *Proc. Aaai Conf. Artif. Intell.* **2022**, *36*, 2441–2449. [CrossRef]
29. Li, R.; Zheng, S.; Duan, C.; Su, J.; Zhang, C. Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
30. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
31. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In Proceedings of the Computer Vision—ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Revised Selected Papers, Part I 13; Springer: Berlin/Heidelberg, Germany, 2017; pp. 213–228.
32. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
33. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [CrossRef]
34. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5229–5238.
35. Hosseinpour, H.; Samadzadegan, F.; Javan, F.D. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 96–115. [CrossRef]
36. Zhou, W.; Jin, J.; Lei, J.; Yu, L. CIMFNet: Cross-layer interaction and multiscale fusion network for semantic segmentation of high-resolution remote sensing images. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 666–676. [CrossRef]
37. Ma, J.; Zhou, W.; Lei, J.; Yu, L. Adjacent bi-hierarchical network for scene parsing of remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
38. Li, Y.; Zhou, Y.; Zhang, Y.; Zhong, L.; Wang, J.; Chen, J. DKDFN: Domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification. *ISPRS J. Photogramm. Remote Sens.* **2022**, *186*, 170–189. [CrossRef]
39. Roy, S.K.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; Chanussot, J. Multimodal fusion transformer for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–20. [CrossRef]
40. Hong, D.; Gao, L.; Yokoya, N.; Yao, J.; Chanussot, J.; Du, Q.; Zhang, B. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4340–4354. [CrossRef]
41. He, Q.; Sun, X.; Diao, W.; Yan, Z.; Yao, F.; Fu, K. Multimodal remote sensing image segmentation with intuition-inspired hypergraph modeling. *IEEE Trans. Image Process.* **2023**, *32*, 1474–1487. [CrossRef]
42. Prakash, A.; Chitta, K.; Geiger, A. Multi-modal fusion transformer for end-to-end autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 7077–7087.
43. Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **2024**, *97*, 103280. [CrossRef]
44. Gao, L.; Liu, H.; Yang, M.; Chen, L.; Wan, Y.; Xiao, Z.; Qian, Y. STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10990–11003. [CrossRef]
45. Ma, X.; Zhang, X.; Pun, M.O. A crossmodal multiscale fusion network for semantic segmentation of remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 3463–3474. [CrossRef]
46. Yan, L.; Huang, J.; Xie, H.; Wei, P.; Gao, Z. Efficient depth fusion transformer for aerial image semantic segmentation. *Remote Sens.* **2022**, *14*, 1294. [CrossRef]
47. He, S.; Yang, H.; Zhang, X.; Li, X. MFTransNet: A multi-modal fusion with CNN-transformer network for semantic segmentation of HSR remote sensing images. *Mathematics* **2023**, *11*, 722. [CrossRef]
48. Ma, X.; Zhang, X.; Pun, M.O.; Liu, M. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5403215. [CrossRef]
49. Li, S.Z. Markov random field models in computer vision. In Proceedings of the Computer Vision—ECCV'94: Third European Conference on Computer Vision Stockholm, Sweden, 2–6 May 1994; Proceedings, Volume II 3; Springer: Berlin/Heidelberg, Germany, 1994; pp. 361–370.
50. Artieres, T. Neural conditional random fields. In JMLR Workshop and Conference Proceedings, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; pp. 177–184.
51. Lafferty, J.; McCallum, A.; Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Icml, Williamstown, MA, USA, 28 June–1 July 2001; Volume 1, p. 3.
52. Lu, X.; Yuan, Y.; Zheng, X. Joint dictionary learning for multispectral change detection. *IEEE Trans. Cybern.* **2016**, *47*, 884–897. [CrossRef]
53. Rao, Y.; Liu, W.; Pu, J.; Deng, J.; Wang, Q. Roads detection of aerial image with FCN-CRF model. In Proceedings of the 2018 IEEE Visual Communications and Image Processing (VCIP), Taichung, Taiwan, 9–12 December 2018; pp. 1–4.
54. Li, Z.; Wang, R.; Zhang, W.; Hu, F.; Meng, L. Multiscale features supported DeepLabV3+ optimization scheme for accurate water semantic segmentation. *IEEE Access* **2019**, *7*, 155787–155804. [CrossRef]

55. Wang, J.; HQ Ding, C.; Chen, S.; He, C.; Luo, B. Semi-supervised remote sensing image semantic segmentation via consistency regularization and average update of pseudo-label. *Remote Sens.* **2020**, *12*, 3603. [CrossRef]
56. Li, L.; Zhang, W.; Zhang, X.; Emam, M.; Jing, W. Semi-supervised remote sensing image semantic segmentation method based on deep learning. *Electronics* **2023**, *12*, 348. [CrossRef]
57. Zhu, J.; Guo, Y.; Sun, G.; Yang, L.; Deng, M.; Chen, J. Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5603518. [CrossRef]
58. Ma, X.; Zhang, X.; Wang, Z.; Pun, M.O. Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5400515. [CrossRef]
59. Reddy, N.D.; Singhal, P.; Krishna, K.M. Semantic motion segmentation using dense CRF formulation. In Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing, Bangalore, India, 14–18 December 2014; pp. 1–8.
60. Markus Gerke, I.T.C. Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen). 2014. Available online: https://www.researchgate.net/publication/270104226_Use_of_the_Stair_Vision_Library_within_the_ISPRS_2D_Semantic_Labeling_Benchmark_Vaihingen?channel=doi&linkId=54ae59c50cf2828b29fcdf4b&showFulltext=true (accessed on 2 September 2024).