remote sensing

MDPI

Article

A Spectral–Spatial Context-Boosted Network for Semantic Segmentation of Remote Sensing Images

Xin Li [1,†], Xi Yong [2,†], Tao Li [3,4,†], Yao Tong [5,6,*], Hongmin Gao [1,7], Xinyuan Wang [1], Zhennan Xu [1], Yiwei Fang [1], Qian You [1] and Xin Lyu [1,7]

1 College of Computer Science and Software Engineering, Hohai University, Nanjing 211100, China; li-xin@hhu.edu.cn (X.L.); gaohongmin@hhu.edu.cn (H.G.); wxyhhu@hhu.edu.cn (X.W.); zhennanxu@hhu.edu.cn (Z.X.); fangyiwei@hhu.edu.cn (Y.F.); 221307040027@hhu.edu.cn (Q.Y.); lvxin@hhu.edu.cn (X.L.)
2 Information Center, Ministry of Water Resources, Beijing 100053, China; yongxi@mwr.gov.cn
3 Engineering Technology Center of Henan Province Smart Water Conservancy, Yellow River Institute of Hydraulic Research, Yellow River Conservancy Commission, Zhengzhou 450003, China; litao@hky.yrcc.gov.cn
4 Information Engineering Center, Yellow River Institute of Hydraulic Research, Yellow River Conservancy Commission, Zhengzhou 450003, China
5 School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing 210023, China
6 Jiangsu Province Engineering Research Center of TCM Intelligence Health Service, Nanjing University of Chinese Medicine, Nanjing 210023, China
7 Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing 211100, China
* Correspondence: yaotong@njucm.edu.cn; Tel.: +86-190-0251-8010
† These authors contributed equally to this work.

Abstract: Semantic segmentation of remote sensing images (RSIs) is pivotal for numerous applications in urban planning, agricultural monitoring, and environmental conservation. However, traditional approaches have primarily emphasized learning within the spatial domain, which frequently leads to less than optimal discrimination of features. Considering the inherent spectral qualities of RSIs, it is essential to bolster these representations by incorporating the spectral context in conjunction with spatial information to improve discriminative capacity. In this paper, we introduce the spectral–spatial context-boosted network (SSCBNet), an innovative network designed to enhance the accuracy semantic segmentation in RSIs. SSCBNet integrates synergetic attention (SYA) layers and cross-fusion modules (CFMs) to harness both spectral and spatial information, addressing the intrinsic complexities of urban and natural landscapes within RSIs. Extensive experiments on the ISPRS Potsdam and LoveDA datasets reveal that SSCBNet surpasses existing state-of-the-art models, achieving remarkable results in F_1-scores, overall accuracy (OA), and mean intersection over union (mIoU). Ablation studies confirm the significant contribution of SYA layers and CFMs to the model's performance, emphasizing the effectiveness of these components in capturing detailed contextual cues.

Keywords: semantic segmentation; remote sensing images; spectral–spatial context; synergetic attention; cross-fusion module

Citation: Li, X. ; Yong, X.; Li, T.; Tong, Y.; Gao, H.; Wang, X.; Xu, Z.; Fang, Y.; You, Q.; Lyu, X. A Spectral–Spatial Context-Boosted Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2024**, *16*, 1214. https://doi.org/10.3390/rs16071214

Academic Editors: Le Sun, Yuhui Zheng, Guoqing Zhang and Byeungwoo Jeon

Received: 9 January 2024
Revised: 18 March 2024
Accepted: 24 March 2024
Published: 29 March 2024

1. Introduction

Semantic segmentation of remote sensing images (RSIs) stands as a critical and challenging task within remote sensing applications [1]. This process involves not just delineating object boundaries but also assigning a specific category to each pixel in an image. The complexity escalates in urban scenarios due to the resemblance among varied objects and the diversity within similar ones [2]. Consequently, semantic segmentation technologies hold considerable importance across various societal sectors, including water resources management [3,4], agricultural data analysis [5,6], and military applications [7,8].

Traditional approaches to semantic segmentation have relied on low-level features derived from color, shape, and texture [9–11]. Subsequently, these features undergo clustering and classification in a higher-dimensional space for complete image segmentation. The limitations of this methodology are evident: first, model efficacy is largely contingent on manually designed features, leading to constrained model expressiveness; second, the low-level features lack sufficient expressivity for segmentation tasks. Consequently, there is a need to extract more abstract, and potentially hierarchical, features from images for vision-related tasks.

Over the past decade, convolutional neural network (CNN) has achieved significant advancements in remote sensing, as evidenced in several studies [12–15]. These advancements have led to markedly improved performance in various areas, including scene classification, object detection, change detection, and image fusion. Concurrently, in computer vision, semantic segmentation technology has rapidly evolved. While the results from fully convolutional networks (FCNs) [16] were promising, their overly simplistic decoder design resulted in coarse outcomes. This led to the development of more sophisticated encoder–decoder architectures, enhancing accuracy [17–20]. Nonetheless, in the context of RSIs, with their intricate backgrounds and large sizes, the reliance on FCN-based methods limits long-range dependency, which is crucial for unconstrained scene image segmentation [21].

RSIs, characterized by high intraclass variability and low interclass variability, present significant labeling challenges compared with standard dataset images. Context utilization has emerged as a viable approach to enhance the discriminative power of learned representations. One method involves integrating multiscale feature encoding through varied-scale dilated convolutional layers or pooling functions [22]. For instance, MLCRNet [23], designed for RSI semantic segmentation, demonstrated impressive performance on the ISPRS Potsdam [24] and Vaihingen [25] benchmarks through multilevel context aggregation. Additionally, Zhang et al. [26] designed a multiscale feature fusion network for land cover and land use classification. Du et al. [27] utilized multiscale aggregation to accurately map urban functional zones from very high resolution images.

Another effective strategy involves the integration of attention modules to capture long-range dependencies [28,29]. The attention mechanism, fundamental in human cognition and survival, allows for selective concentration on specific information aspects while disregarding others, i.e., CBAM [30], DANet [31], and OCRNet [32]. This mechanism has substantially improved segmentation accuracy in networks by focusing on information-rich areas [33,34]. Various attention-based methods have emerged in RSI segmentation. For example, Li et al. [35] introduced dual attention and deep fusion strategies for large-scale satellite RSI segmentation. Other notable developments include HCANet [36], which combines cross-level contextual and attentive representations. HMANet [37] and LANet [38] are also notable for their effective capture of spatial, channel, and category correlations, and for bridging the gap between high-level and low-level features with patch attention modules. Overall, the attention mechanism has proven highly effective in the RSI domain, aiding models in recognizing diverse intraclass variance and subtle interclass variance.

Despite the advancements in processing RSIs and feature learning within the spatial domain, previous methods have often overlooked the spectral characteristics of these images. Image processing fundamentally associates inner body features with low-frequency components and edges with high-frequency components. This concept is visually depicted in Figure 1, where an RSI is transformed to the frequency domain by discrete wavelet transformation (DWT). Specifically, LL displays the low-frequency component, and HH illustrates the high-frequency component. This fact reveals that different components carry different information. A robust model should learn both low-frequency and high-frequency details concerning all spectral bands. Additionally, while self-attention mechanisms are designed to improve object internal consistency through similarity measurements, they apply the same learnable parameters across all frequency components. This approach limits their effectiveness in enhancing both internal consistency and edge delineation between

objects. Consequently, the emphasis on effectively leveraging frequency domain features, especially for learning the spectral context in RSIs, is crucial.

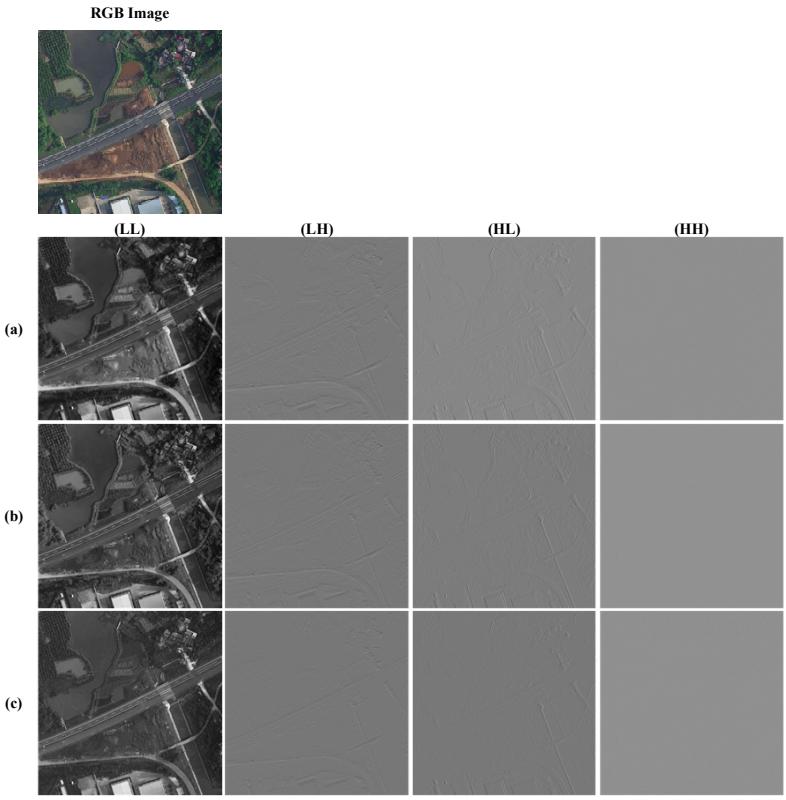

Figure 1. Visualizations of different frequency components: (**a**) components of R (red) band, (**b**) components of G (green) band, and (**c**) components of B (blue) band. RGB image comes from LoveDA dataset [39] ; *LL*, *LH*, *HL*, and *HH* represent low-frequency, horizontal, vertical, and high-frequency components (projected by discrete wavelet transformation), respectively.

Conclusively, this paper introduces the spectral–spatial context-boosted network (SSCBNet) for the semantic segmentation of RSIs. Building upon the encoder–decoder architecture, SSCBNet incorporates cross-fusion modules (CFMs) to seamlessly integrate encoded and decoded features. A pivotal element is the synergetic attention layer, which synergistically models a spectral–spatial context. This layer sequentially applies frequential spectral attention (FSA) and position-wise self-attention (PSA) in lieu of traditional convolutional layers. Notably, FSA emphasizes significant spectral bands by analyzing vertical and horizontal channel contexts. The paper's main contributions can be summarized as follows:

1. We propose a frequential spectral attention (FSA) mechanism, which focuses on the spectral context for feature enhancement. Specifically, FSA initially transforms feature maps to the frequency domain using DWT. Subsequently, the intracomponent spectral context is analyzed through a dual directional spectral attention (DDSA) mechanism, which concurrently learns columnwise and rowwise channel attentions, integrated using a fusion operator. DDSA is then applied to condensed full components to address the intercomponent spectral context, allowing FSA to effectively utilize the spectral property in remote sensing images.
2. Our design includes a synergetic attention (SYA) module, which consecutively applies frequential spectral attention (FSA) and position-wise self-attention (PSA) to refine representations. FSA enhances spectral contexts, while PSA further interprets position-wise spatial relationships. This approach enables the input feature maps

to be refined, ensuring the preservation of high-frequency details and promoting internal consistency.

3. Employing an encoder–decoder structure, we developed a spectral–spatial context-boosted network (SSCBNet) for the semantic segmentation of remote sensing images. The encoder, enhanced with the SYA layer in place of the conventional convolution layer, demonstrates superior capability in learning discriminative features. The novel cross-fusion module (CFM) efficiently connects the encoder and decoder, facilitating feature recovery with minimal loss.
4. SSCBNet underwent a comprehensive evaluation on two notable benchmarks, ISPRS Potsdam and LoveDA [39], encompassing both aerial and satellite imagery. The tests highlight SSCBNet's superior performance compared with various leading models. Additionally, the contributions of SYA and CFM are validated.

The paper is structured as follows: Section 2 provides an overview of related works in the semantic segmentation of RSI and the advanced methods based on frequency analysis. Section 3 introduces SSCBNet within its submodules. Section 4 presents the comparisons and ablation studies. Section 5 draws the conclusion of our work and points out future directions.

2. Related Works

2.1. Semantic Segmentation of Remote Sensing Images

The semantic segmentation of RSIs entails the pixel-level classification of images from satellites or aerial platforms into land cover or use categories [40]. The expansive coverage and limited spatial resolution of RSIs render many neural networks, tailored for specific terrestrial targets, inapplicable. The variability in the spectral, spatial, and temporal characteristics of RSIs introduces uncertainties in semantic labeling. Thus, there is a need for versatile models tailored to the unique aspects and varied conditions of RSIs [41].

State-of-the-art methods in remote sensing semantic segmentation often employ the encoder–decoder architecture, prized for its robust feature extraction and transformation abilities. For instance, Mboga et al. [42] explored the efficacy of deep fully convolutional networks in urban aerial image classification in Goma, Democratic Republic of Congo, comparing their performance with a semiautomatic geographic object-based image analysis (GEOBIA) system. Wang et al. [43] developed an encoder–decoder network integrating inner convolution with directional conditional random fields for postprocessing, demonstrating impressive segmentation results on the Massachusetts road dataset. Liu et al. [44] adapted U-Net for the pixel-level identification and grading of maize drought from aerial imagery. Sun et al. [45] examined the structural limitations and learning deficiencies of encoder–decoder networks in RSI segmentation, proposing ensemble training and inference techniques to mitigate these issues. Tan et al. [46] introduced an end-to-end road segmentation approach, leveraging convolutional layers at different levels to precisely delineate road edges and shapes. A scale-sensitive module in this method employs a weight tensor to assess the significance of fused features. U-Net, a widely explored model for RSIs, is augmented in various adaptations. TreeUNet [47] enhances pixel-level classification accuracy by incorporating a Tree-CNN block with a confusion matrix to integrate multiscale features. ResUNet-a [48], an extension of U-Net, incorporates residual connections, atrous convolutions, pyramid scene parsing, pooling, and multitasking inference, boasting substantial parameters and exceptional accuracy.

Recent advancements have seen the introduction of attention-based deep neural networks for RSI semantic segmentation, leveraging their capacity to model extensive contextual information [49,50]. For example, Li et al. [35] developed a model combining deep layer channel attention and shallow layer spatial attention for large-scale satellite imagery segmentation. Moreover, the SE design is extended to a spatial dimension to focus on patch-wise semantics, bridging the semantic gap between high-level and low-level features [51]. Sun et al. [52] combined the attention mechanism with deep multiscale feature extraction to enhance the representation of high- and low-level semantic features. AFNet [53] features

a multipath encoder structure, attention-fused modules, and a refinement block, tailored for very high-resolution RSIs. LANet [38] employs a patchwise attention module and a cross-level attention fusion mechanism. Other notable attention-focused models include A2FPN [54], SCAttNet [55], and RAANet [56]. More recently, Li et al. [57] identified that existing attention models, which separately address spatial and channel feature affinities before fusing them, often underperform. In response, they introduced the synergistic attention perception neural network (SAPNet) for RSI segmentation. Additionally, GPINet, guided by geometric priors, utilizes features from CNN and transformer branches to infer pixel semantics with increased accuracy [58].

In essence, effective segmentation hinges on extracting and utilizing informative cues. Traditional methods, primarily confined to the spatial domain, often neglect spectral characteristics, leading to distorted contextual representation, particularly in the spectral domain. Our proposed solution entails a combined modeling of spectral and spatial contexts in an attentive framework. This strategy refines representations to maintain both spectral and spatial integrity, offering valuable cues to enhance discriminative capabilities.

2.2. Learning in the Frequency Domain

Recent years have seen a growing focus on frequency domain learning in image processing, especially in the semantic segmentation of RSIs. This approach excels at differentiating frequency components within images, leading to a finer feature extraction. This precision is especially beneficial in RSIs, characterized by varied textures and patterns. In RSIs, frequency domain features maintain spectral properties with minimal distortion, providing an effective way to model the spectral context.

Xu et al. [59] conducted a theoretical evaluation of neural networks' spectral bias using Fourier analysis, highlighting their propensity for low-frequency functions and challenges in capturing high-frequency details. This insight spurred subsequent research to emphasize high-frequency elements. Azad et al. [60] enhanced the local feature capacity of transformer models through a redesigned self-attention map, decomposing features into the frequency domain to reveal finer details for enriched contextual cues. Ref. [61] utilized frequency domain features for image classification, while [62] focused on converting spatial-domain CNN models into the frequency domain. Another approach involves bypassing complex conversion processes by selecting frequency channels with SE-Blocks, and a frequency channel attention network was introduced to explicitly process spectral information less apparent in the spatial domain [63]. In addition, [64] introduces a frequency channel attention network. As can be observed, spectral analysis in the frequency domain is beneficial for distinguishing subtle spectral signatures of different land cover classes, enabling the model to differentiate between features with similar spatial but distinct spectral properties.

In processing RSIs, Su et al. [65] developed a complete frequency channel attention network (CFCANet) capable of handling noisy RS images without filtering. CFCANet selectively integrates low-frequency information with feature maps, assigning a discrete cosine transformation frequency component to each original feature map and extracting the most significant eigenvalue per channel. CFCANet shows superior noise resistance by leveraging low-frequency image information. For semantic segmentation, Li et al. [66] introduced the spectrum-space collaborative network (SSCNet), capturing both spectral and spatial dependencies to improve the RSI segmentation quality. SSCNet features a joint spectral–spatial attention module (JSSA) employing both spectral (SpeA) and spatial (SpaA) attention.

In summary, leveraging frequency domain representations for spectral analysis in RSIs is a potent method to harness inherent spectral attributes. This approach not only improves differentiation between land cover classes but also enhances model generalization and facilitates the integration of the spectral context in semantic segmentation. The result is more accurate and robust outcomes in remote sensing applications.

3. Method

3.1. Overview of SSCBNet

As depicted in Figure 2, the spectral–spatial context-boosted network (SSCBNet) adopts the encoder–decoder architecture, utilizing cross-fusion modules (CFMs) to effectively connect encoded and decoded features. Specifically, the network comprises four feature encoder blocks, each containing a convolution block (as shown in Figure 3), a synergetic attention (SYA) layer, and a pooling layer. The final encoder block consists of a convolution block and an SYA layer. During the decoding phase, each of the five blocks is equipped with a 1×1 convolution layer, an SYA layer, and a convolution block. Additionally, four CFMs are strategically placed to integrate features from the preceding decoder block and the corresponding encoder block, thereby minimizing information loss during feature map recovery. The application of a Softmax layer yields the final predictions. This innovative combination of SYA and CFM within SSCBNet redefines the encoder–decoder framework for semantic segmentation in RSIs, ensuring that both spectral and spatial contexts are effectively considered and utilized. The subsequent subsections will provide an in-depth description of these submodules.

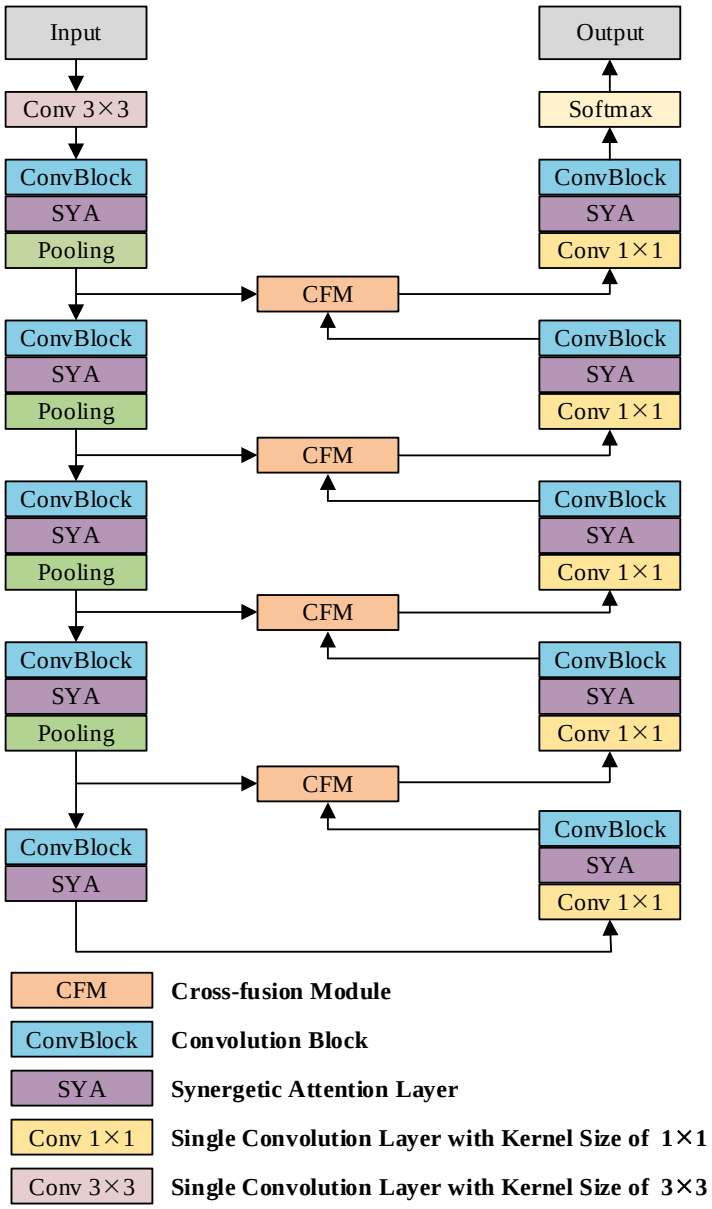

Figure 2. The topological framework of SSCBNet.

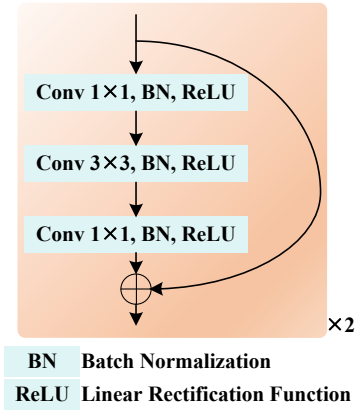

Figure 3. Pipeline of convolution block in SSCBNet.

3.2. Pipeline of SYA

Figure 4 illustrates how the SYA module in SSCBNet sequentially applies FSA and PSA, using element-wise summation as a connector. In FSA, the spectral context is represented and injected in the frequency domain, optimizing diverse frequency components for detailed spectral information using DDSA. Subsequently, a standard PSA is employed to exploit position-wise correlations, further refining the features. This combination in SYA effectively harnesses a comprehensive spectral–spatial context, significantly enhancing the discriminability of the learned representations.

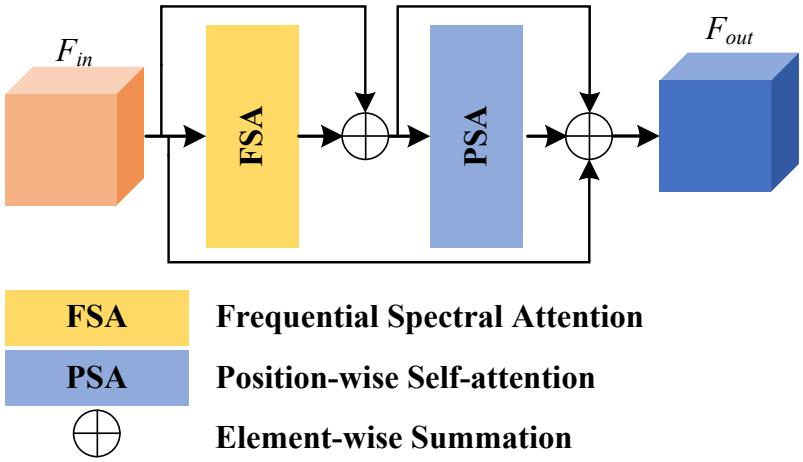

Figure 4. Pipeline of SYA.

The mechanism of FSA is detailed in Figure 5. For input features, $F_{in}(FSA) \in \mathbb{R}^{C\times H\times W}$, DWT initially projects them onto the frequency domain, resulting in four components: the low-frequency LL, the horizontal LH, the vertical HL, and the high-frequency HH, each with dimensions, $\mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$. These components are individually processed by DDSA, which captures the spectral context and accentuates informative bands within each component.

DDSA's functionality is further explained in Figure 6. Given the diverse properties of frequency domain features along horizontal and vertical axes, DDSA uses two parallel branches to capture frequential bandwise correlations. One branch analyzes features along the horizontal axis and the other along the vertical axis, resulting in two split features, $F_W \in \mathbb{R}^{W\times(H\times C)}$ and $F_H \in \mathbb{R}^{H\times(W\times C)}$. Matrix multiplication then generates initial attention maps, which are processed with an average operator and a Softmax function to produce refined feature maps. These are combined via element-wise summation to produce the DDSA output, where each spectral band is refined by a bandwise context:

$$AM_W = F_W^T \times F_W, \tag{1}$$

where $F_W^T \in \mathbb{R}^{W\times(C\times H)}$, $F_W \in \mathbb{R}^{W\times(H\times C)}$, and $AM_W \in \mathbb{R}^{W\times(C\times C)}$. Notably, F_W is the feature reshaped along the horizontal axe. In other words, F_W can be formed as a combination of W matrices of the size of $H \times C$. Then, the matrix $H \times C$ is transposed. This formula represents W times matrix multiplications between $C \times H$ and $H \times C$. Then, an average operator is applied on AM_W with regard to each position, generating an attention map by the following:

$$AM_v = \text{Softmax}(\text{Average}(AM_W)), \tag{2}$$

where $AM_v \in \mathbb{R}^{C\times C}$ is the attention map over the vertical axis. Likewise, we split and calculate the attention map AM_h as follows:

$$AM_h = \text{Softmax}\left(\text{Average}\left(F_H^T \times F_H\right)\right), \tag{3}$$

where $AM_v \in \mathbb{R}^{C\times C}$ is the attention map over the horizontal axis. Injecting the attention weights into convolved features, we have two refined feature maps. At last of DDSA, an element-wise summation is deployed:

$$F_{out}(DDSA) = Conv_{1\times1}(F_{in}(DDSA)) + F_{va} + F_{ha}, \tag{4}$$

where $F_{out}(DDSA) \in \mathbb{R}^{C\times H\times W}$ is the output of DDSA, in which each spectral band is refined by the bandwise context.

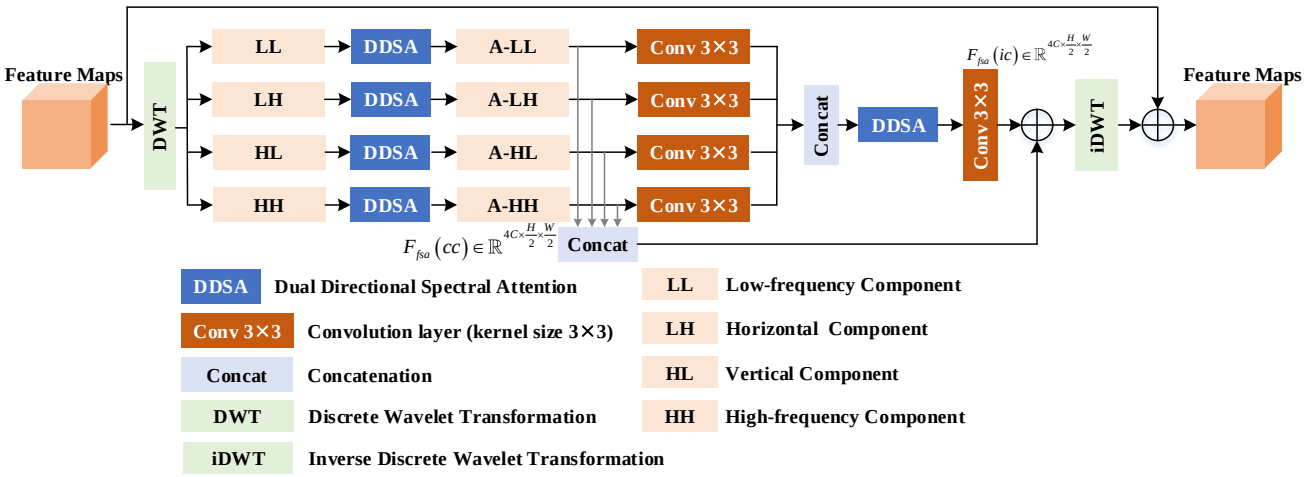

Figure 5. Pipeline of FSA.

During the initial phase of the FSA, four components are refined using DDSA modules. To be specific, DDSA-refined features, F_{A-LL}, F_{A-LH}, F_{A-HL}, and F_{A-HH}, are all with a size of $C \times \frac{H}{2} \times \frac{W}{2}$. To reduce computational costs while preserving feature details, we compress the four components by a convolution layer (kernel size = 3×3, stride = 1, padding = 1, $out_{channels}$ = 1). Then the concatenated feature is with a size of $1 \times \frac{H}{2} \times \frac{W}{2}$. With the refinement of the intercomponent context by DDSA, another convolution layer (kernel size = 3×3, stride = 1, padding = 1, $out_{channels}$ = 4C) is applied to expand the channels to 4C, generating $F_{fsa}(ic) \in \mathbb{R}^{4C\times\frac{H}{2}\times\frac{W}{2}}$.

On the other hand, these DDSA-refined components are fused by a concatenation layer, and individual convolution layers with a kernel size of 3×3 compress each component into a single dimension. This process is crucial for modeling the context between components. The formula for this concatenation is represented as follows:

$$F_{fsa}(cc) = \text{concat}(A-LL, A-LH, A-HL, A-HH), \tag{5}$$

where $F_{fsa}(cc)$ is defined as a tensor in $\mathbb{R}^{4C\times\frac{H}{2}\times\frac{W}{2}}$. Each component, $A-LL$, $A-LH$, $A-HL$, and $A-HH$, undergoes convolution, resulting in a dimensionally reduced form of $1\times\frac{H}{2}\times\frac{W}{2}$. Consequently, a feature map of dimensions $4\times\frac{H}{2}\times\frac{W}{2}$ is obtained through concatenation, where each channel signifies the abstracted features of the corresponding frequency component. Following this, DDSA is applied to model and utilize the intercomponent context to produce $F_{fsa}(ic)$. Prior to iDWT, we sum $F_{fsa}(ic)$ with $F_{fsa}(cc)$. Finally, these reprojected feature maps are combined with the initial FSA input, generating spectral attention-refined features, denoted as $F_{FSA}^{output}\in\mathbb{R}^{C\times H\times W}$.

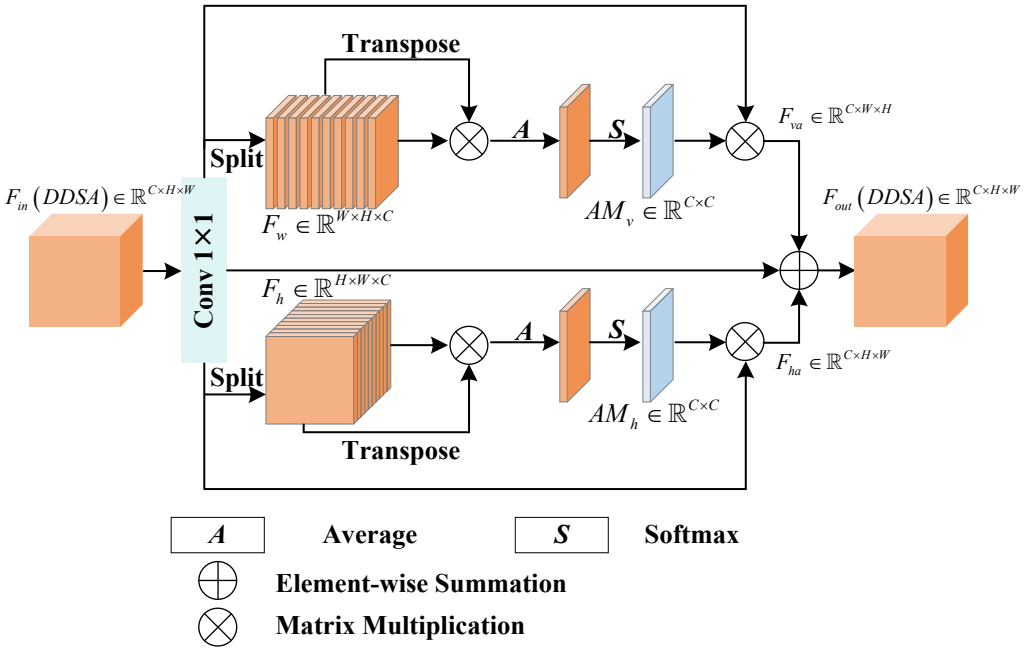

Figure 6. Pipeline of DDSA.

It is noted that PSA in SYA follows the design in DANet [30] (see details in Figure 7). More concretely, given the input features from the previous layer, with a size of $F_{PSA}^{input}\in\mathbb{R}^{C\times H\times W}$, PSA refines them with the position-wise context. Initially, input features are projected by three convolution layers. At the top branch, we have $F_1\in\mathbb{R}^{C\times N}$ with $N=H\times W$. Similarly, $F_2\in\mathbb{R}^{N\times C}$ is generated. By applying matrix multiplication, followed by a Softmax function, the position-wise attention maps are produced with $A_p\in\mathbb{R}^{N\times N}$. Then, a matrix multiplication and reshape operation is implemented to produce $F_4\in\mathbb{R}^{C\times H\times W}$. At last, an element-wise summation with F_{PSA}^{input} helps us to have the PSA-refined feature maps $F_{PSA}^{output}\in\mathbb{R}^{C\times H\times W}$.

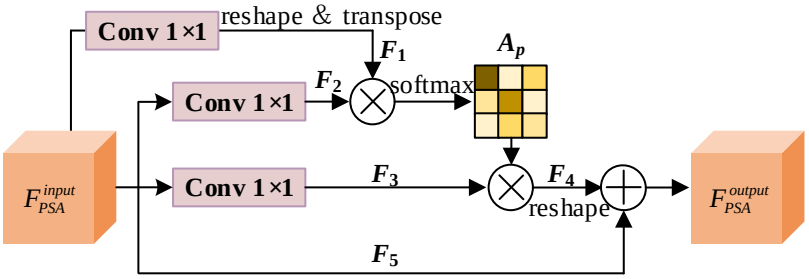

Figure 7. Pipeline of PSA.

Overall, SYA is formulated as follows:

$$F_{out}=F_{in}+\left(F_{FSA}^{output}+F_{in}\right)+F_{PSA}^{output}. \tag{6}$$

This design first emphasizes intracomponent spectral bands in FSA, followed by modeling the component-wise context. The integration of vanilla PSA and residual connections produces a feature map with a fully realized spectral–spatial context.

3.3. Pipeline of CFM

Enhancing contextual cues in feature extraction and transformation is further augmented by fusing feature maps from corresponding encoder and decoder blocks. This approach has been demonstrated to effectively minimize distortions and maintain the integrity of original expressions. Consequently, this fusion ensures that the recovery path retains contextual information to the greatest extent possible.

Figure 8 illustrates the integration of a straightforward cross-attention mechanism within the cross-fusion module (CFM). In this setup, the decoded feature serves as the query component, while the encoded feature is employed as both the key and value components. The process can be formally expressed as follows:

$$F_{dec}^{in}(i) = \text{Softmax}(F_{enc}(i) \times F_{dec}(i+1)) \times F_{enc}(i) + F_{enc}(i), \tag{7}$$

where i represents the index of the encoder and decoder block, $F_{enc}(i)$ denotes the feature map from the corresponding encoder block, and $F_{dec}(i+1)$ is sourced from the preceding decoder block. Notably, $F_{enc}(i)$ and $F_{dec}(i+1)$ are with similar sizes. Let $F_{enc}(i)$ and $F_{dec}(i+1)$ be $C \times H \times W$; with matrix multiplication and Softmax function, the contextual correlations between $F_{enc}(i)$ and $F_{dec}(i+1)$ are modeled. AttnMap is with a size of $(H \times W) \times (H \times W)$. Then, a matrix multiplication is applied to inject such contextual information stored in AttnMap to form $F_{cf} \in \mathbb{R}^{C \times H \times W}$. Finally, a element summation is deployed to generate $F_{dec}^{in}(i)$.

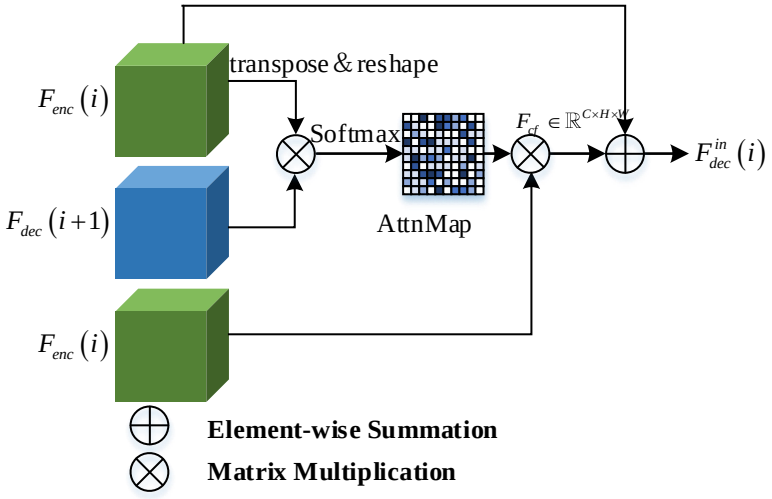

Figure 8. Pipeline of CFM.

4. Experiments

4.1. Datasets

4.1.1. ISPRS Potsdam Dataset

The Potsdam dataset comprises 38 tiles of TOP images with very fine spatial resolution (ground sample distance (GSD) of 5 cm), each measuring 6000 × 6000 pixels. It shares the same category information as the Vaihingen dataset. Included in this dataset are four multispectral bands (red, green, blue, and near-infrared) along with digital surface model (DSM) and normalized digital surface model (NDSM) data. For our study, IDs 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, and 6_15, along with 7_13, were designated for testing; ID 2_10 for validation; and the remaining 22 images (excluding image 7_10 due to erroneous annotations) for training purposes. We utilized

only the red, green, and blue bands and segmented the original image tiles into patches of 256 × 256 pixels for experimental analysis. A random sample is shown in Figure 9.

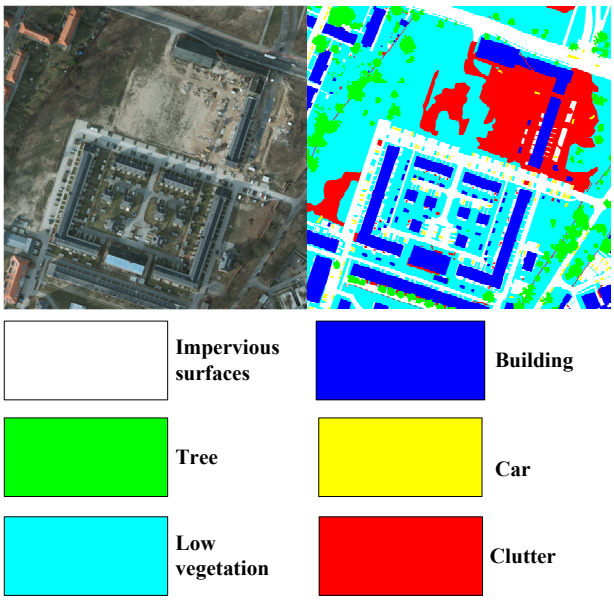

Figure 9. Example of ISPRS Potsdam dataset.

4.1.2. LoveDA Dataset

The LoveDA dataset, described in [39], encompasses 5987 fine-resolution optical remote sensing images (GSD of 0.3 m), each with dimensions of 1024 × 1024 pixels. This dataset includes seven land cover categories: building, road, water, barren, forest, agriculture, and background. It allocates 2522 images for training, 1669 for validation, and 1796 officially provided images for testing. The LoveDA dataset presents images from two distinct scenes—urban and rural, collected from three cities in China: Nanjing, Changzhou, and Wuhan. The dataset poses significant challenges due to multiscale objects, complex backgrounds, and inconsistent class distributions in these images. A random sample is shown in Figure 10.

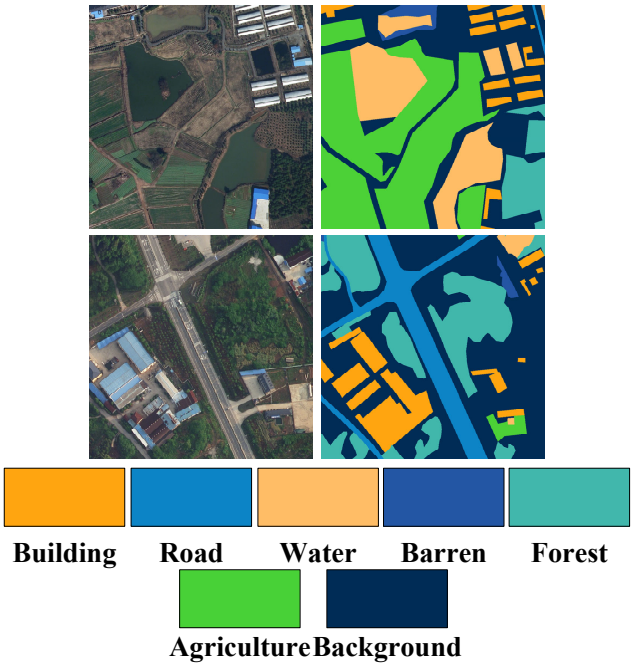

Figure 10. Example of LoveDA dataset.

4.2. Implement Details

The experimental framework, as summarized in Table 1, details the settings employed in this study. Our proposed SSCBNet and other semantic segmentation models were implemented using the PyTorch framework on a Linux operating system, with an NVIDIA A40 GPU (Santa Clara, CA, USA) for computational support. To enhance the model's robustness, data augmentation techniques such as random flipping and cropping were applied across all datasets and networks. The training protocol involved a fixed initial learning rate of 0.02 and a maximum of 500 epochs. We utilized the Adam optimization algorithm, with a learning strategy defined by polynomial decay and a momentum setting of 0.9. The model parameters yielding the lowest validation loss were preserved for further evaluation.

For comparative analysis, eleven advanced approaches were selected: U-Net [20], DeepLabV3+ [22], DANet [30], ResUNet-a [48], DASSN [35], HCANet [36], RAANet [56], SCAttNet [55], A2-FPN [54], LANet [38], and SAPNet [57]. Notably, U-Net [20], DeepLabV3+ [22], and DANet [30] were initially developed for natural image segmentation, while ResUNet-a [48], DASSN [35], HCANet [36], RAANet [56], SCAttNet [55], A2-FPN [54], LANet [38], MSAFNet [67], CLCFormer [68], and SAPNet [57] represent recent state-of-the-art methodologies specifically designed for RSI segmentation.

Table 1. Experimental settings.

Items	Settings
Learning strategy	Poly decay
Initial learning rate	0.002
Loss function	Cross-entropy
Max epoch	500
GPU memory	48 GB
Optimizer	Adam
Subpatch size	256×256
Batch size	32

4.3. Evaluation Metrics

In this research, we evaluated the performance of our predictions on the test set using standard evaluation metrics. These include the classwise F_1 score (F_1), the average F_1 score across all classes (AF), overall accuracy (OA), and mean intersection over union (mIoU). The F_1 score, representing the harmonic mean of precision and recall, assesses the balance between false positives and false negatives. OA quantifies the proportion of correctly classified pixels out of the total number of pixels. mIoU serves as a comprehensive metric for accuracy assessment. The formulations for F_1, OA, and IoU are specified in Equations (8), (9), and (10), respectively:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}, \tag{8}$$

$$\mathrm{OA} = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \tag{9}$$

$$\mathrm{IoU} = \frac{TP}{(TP + FP + FN)}, \tag{10}$$

where *precision* and *recall* are computed as follows:

$$precision = \frac{TP}{TP + FP}, \tag{11}$$

$$recall = \frac{TP}{TP + FN}, \tag{12}$$

where TP, TN, FP, and FN denote the counts of true positives, true negatives, false positives, and false negatives, respectively. mIoU is calculated as the average IoU across all classes.

4.4. Compared with State-of-the-Art Models

4.4.1. Numerical Evaluation of ISPRS Potsdam

The results presented in Table 2 for the ISPRS Potsdam dataset provide a comprehensive overview of the performance of various semantic segmentation methods, including the proposed SSCBNet. A detailed analysis of these results yields several insightful observations.

First, SSCBNet exhibits superior performance across most metrics when compared with the other evaluated methods. Notably, it achieves the highest scores in terms of impervious surfaces (93.52), building (97.23), and car (96.42) categories, as well as in the OA and mIoU metrics. This indicates that SSCBNet effectively captures and integrates detailed features from the dataset, leading to more accurate segmentation, particularly in distinguishing between complex urban features like buildings and cars. The performance in the low vegetation and tree categories, while not the highest, is still commendably high, suggesting a well-rounded capability in handling various land cover types.

Comparatively, other advanced methods such as HCANet, LANet, and SAPNet also demonstrate robust performance, with HCANet showing particular strength in categories like impervious surfaces and building. However, SSCBNet's balanced approach across all classes, as evidenced by its high AF score, signifies its comprehensive efficacy in semantic segmentation tasks. The lower performance of models like U-Net and DeepLabV3+ in certain categories, despite their overall good results, underscores the advancements in more recent methods that focus on specific challenges in remote sensing image analysis.

Specifically, when compared with MSAFNet and CLCFormer, two recent and robust models represented in blue, our approach demonstrates superior performance. SSCBNet achieves a remarkable classwise F_1-score for 'impervious surfaces' at 93.52%, outpacing MSAFNet's 91.61% and CLCFormer's 92.66%. Moreover, the OA and mIoU of SSCBNet are 91.31% and 83.09%, respectively, which show significant improvements over MSAFNet's OA of 88.90% and mIoU of 80.68%, and even CLCFormer's OA of 89.97% and mIoU of 81.68%.

It is also noteworthy to observe the performance trends in different land cover categories. While most methods perform well in detecting impervious surfaces and buildings, categories such as low vegetation and car tend to have more variation in performance. This could be attributed to the intrinsic challenges associated with these categories, such as the complex textures in vegetation and the smaller size and diverse shapes of cars. The high performance of SSCBNet in these challenging categories highlights the effectiveness of its design in dealing with intricate and varied features in RSIs.

In conclusion, the experimental results demonstrate the effectiveness of SSCBNet in semantic segmentation tasks on the ISPRS Potsdam dataset. Its ability to outperform other state-of-the-art methods across multiple metrics underscores the benefits of its spectral–spatial contextual approach. This analysis not only validates the superiority of SSCBNet but also provides insights into the evolving landscape of semantic segmentation methodologies in the field of remote sensing.

Table 2. Results on the ISPRS Potsdam dataset. Classwise F_1-score, AF, OA, and mIoU are listed, where the bold text indicates the best results.

Methods	Impervious Surfaces	Building	Low Vegetation	Tree	Car	AF	OA	mIoU
U-Net	87.51	89.32	73.98	87.03	48.19	77.21	75.83	70.02
DeepLabV3+	84.49	86.13	77.37	77.37	85.47	82.17	80.82	73.59
DANet	86.96	92.04	80.14	79.95	89.36	85.69	83.99	77.13
ResUNet-a	89.16	92.76	86.80	83.02	79.25	86.20	85.90	77.99

Table 2. *Cont.*

Methods	Impervious Surfaces	Building	Low Vegetation	Tree	Car	AF	OA	mIoU
DASSN	87.54	92.39	81.62	82.49	83.56	85.52	85.21	77.36
HCANet	92.88	96.90	87.25	88.15	93.88	91.81	90.67	81.92
RAANet	90.40	95.70	87.20	81.80	77.60	86.54	85.21	77.95
SCAttNet	91.87	96.90	85.24	87.05	92.78	90.77	89.06	80.83
A2FPN	90.67	95.69	85.54	85.44	91.55	89.78	88.25	80.67
LANet	92.41	96.38	86.82	87.77	93.44	91.36	90.26	81.19
MSAFNet	91.61	96.63	85.00	86.80	92.51	90.51	88.90	80.68
CLCFormer	92.66	96.64	87.05	88.00	93.69	91.61	89.97	81.68
SAPNet	93.00	96.25	86.49	88.27	94.57	91.72	90.44	81.71
SSCBNet (ours)	**93.52**	**97.23**	**87.90**	**88.61**	**96.42**	**92.73**	**91.31**	**83.09**

4.4.2. Visual Inspections of ISPRS Potsdam

The visual representations provided in Figure 11 complement the quantitative results from the ISPRS Potsdam dataset discussed earlier. These visualizations enable a qualitative assessment of each method's performance by comparing the predicted segmentation against the ground truth.

In the first row of images, the segmentation results are juxtaposed with the original infrared–red–green (IRRG) images and the ground truth. It is immediately noticeable that models like SSCBNet (our method), SAPNet, and HCANet demonstrate closer alignment with the ground truth, particularly in complex areas where multiple classes intersect or are in close proximity. The fidelity of SSCBNet's predictions to the ground truth, especially in delineating the 'car' and 'building' classes, corresponds with the high F_1 scores and mIoU values reported in the table.

In contrast, models like U-Net and DeepLabV3+ show a tendency to misclassify certain regions, as evidenced by the apparent discrepancies between their predicted segments and the ground truth. For example, areas classified as 'low vegetation' or 'tree' are sometimes incorrectly labeled as 'building' or 'impervious surfaces', which can be attributed to the challenges of differentiating between similar spectral signatures in these classes.

The middle and bottom rows of images further illustrate the segmentation performance across different urban landscapes. It is evident that the more advanced models, including ours, tend to produce smoother and more coherent segment shapes that are more representative of the actual land cover, whereas simpler models display fragmented or mis-shapen segments.

Overall, the visualizations provide compelling evidence of the advancements in semantic segmentation for RSIs. The superior performance of SSCBNet in maintaining structural integrity and class distinction, as visualized, reaffirms the quantitative results. These visual insights into model performance highlight the importance of considering both spectral and spatial contexts, as effectively performed by SSCBNet, in achieving high accuracy in complex urban environments.

4.4.3. Numerical Evaluation of LoveDA

The results from the LoveDA dataset as presented in Table 3 showcase the performance of various semantic segmentation methods, with a particular emphasis on SSCBNet. SSCBNet outperforms other methods in several key metrics, indicating its robustness and efficacy in RSI segmentation.

SSCBNet demonstrates the highest F_1-scores across most land cover categories, notably in 'building' and 'road', which are critical classes for urban planning and monitoring applications. The increase in F_1-score for these categories suggests that SSCBNet is particularly adept at handling urban features, which are often characterized by complex structures and varying spectral signatures. In 'water', 'forest', and 'agriculture' classes, SSCBNet also performs impressively, showing its versatility across diverse natural landscapes. The

superior performance in 'water' and 'forest' is notable, given that these classes can be challenging due to their spectral similarity to other classes. The 'agriculture' class often presents difficulties due to the heterogeneous nature of agricultural fields, yet SSCBNet manages to achieve high accuracy, as reflected in the F_1-score and mIoU metrics.

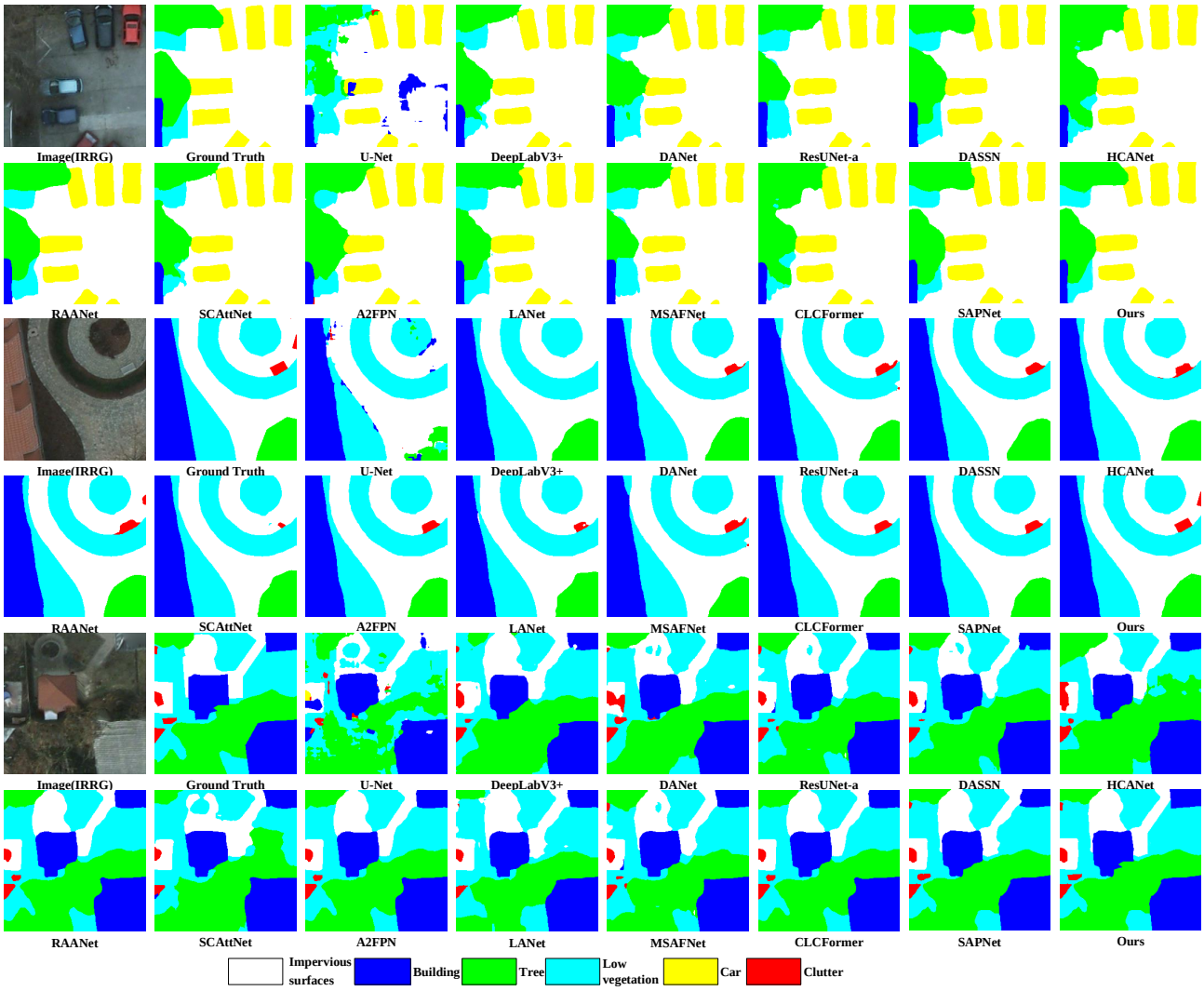

Figure 11. Visual inspections of random samples from ISPRS Potsdam test set.

OA and mIoU are critical global metrics for evaluating segmentation performance. SSCBNet's leading scores in both OA and mIoU indicate that it not only correctly classifies a higher proportion of pixels overall but also maintains high precision and recall balance across classes. This balance is essential in applications where misclassification can lead to significant consequences, such as land use change detection and environmental monitoring.

Comparing SSCBNet with other advanced methods like SAPNet, LANet, and HCANet reveals that while these methods also achieve commendable results, SSCBNet's innovative approach to integrating spectral–spatial contexts gives it an edge. The results suggest that SSCBNet's architecture effectively captures and integrates detailed spatial and spectral features, leading to improved performance, especially in distinguishing between closely related categories. Notably, SSCBNet's classwise F_1-score for 'building' is 80.18%, which is a significant improvement over MSAFNet's 73.71% and CLCFormer's 74.34%. Furthermore, the OA and mIoU obtained by SSCBNet are 72.44% and 65.92%, respectively, surpassing the 67.17% OA and 60.76% mIoU of MSAFNet and the 70.45% OA and 62.55% mIoU of CLCFormer.

In summary, SSCBNet establishes new benchmarks on the LoveDA dataset, underscoring the potential of leveraging both spectral and spatial information in semantic segmentation tasks. These results validate the effectiveness of SSCBNet in handling the complex nature of remote sensing data and highlight the importance of continued research in this area to further enhance segmentation accuracy and reliability.

Table 3. Results on the LoveDA dataset. Classwise F_1-score, AF, OA, and mIoU are listed, where the bold text indicates the best results.

Methods	Background	Building	Road	Water	Barren	Forest	Agriculture	AF	OA	mIoU
U-Net	50.21	54.74	56.38	77.12	18.09	48.93	66.05	53.07	51.81	47.84
DeepLabV3+	52.29	54.99	57.16	77.96	16.11	48.18	67.79	53.50	52.30	47.62
DANet	54.47	61.02	63.37	79.17	26.63	52.28	70.02	58.14	54.64	50.18
ResUNet-a	59.16	64.08	66.73	81.01	32.23	55.81	75.79	62.12	59.65	54.16
DASSN	57.95	66.90	68.63	76.64	44.35	54.96	70.49	62.85	60.35	55.42
HCANet	66.39	70.76	75.11	88.29	51.14	63.92	81.07	70.95	69.47	62.77
RAANet	55.02	62.19	65.58	81.03	29.25	54.11	74.07	60.18	58.95	53.93
SCAttNet	65.95	71.88	77.04	86.61	50.79	61.19	82.00	70.78	67.31	61.09
A2FPN	65.17	73.32	75.19	88.01	48.82	59.96	79.71	70.03	66.89	61.14
LANet	67.04	74.19	77.54	87.54	52.23	64.78	80.80	72.02	69.11	62.16
MSAFNet	65.51	73.71	75.59	88.47	49.08	60.28	80.13	70.40	67.17	60.76
CLCFormer	67.17	74.34	77.69	87.71	52.34	64.91	80.96	72.16	70.45	62.55
SAPNet	68.27	76.81	78.77	**92.20**	51.15	62.82	83.51	73.36	71.83	64.90
SSCBNet (ours)	**69.45**	**80.18**	**82.25**	91.86	**53.16**	**65.87**	**84.48**	**75.32**	**72.44**	**65.92**

4.4.4. Visual Inspections of LoveDA

The visualizations provided in Figure 12 correspond with the quantitative data from Table 3, offering a qualitative perspective on the performance of various semantic segmentation models on the LoveDA dataset.

At a glance, the visual comparisons between the predicted results and the ground truth for each method reveal significant differences in segmentation accuracy. SSCBNet, identified as 'Ours' in the figure, shows remarkable congruence with the ground truth, particularly in delineating complex and mixed land cover classes. The model appears to accurately segment the intricate patterns of agricultural fields and urban structures, which is consistent with the high F_1-scores and mIoU reported in the table. This visual alignment with the ground truth underscores the effectiveness of SSCBNet's approach to integrating spectral and spatial details, leading to superior performance in semantic segmentation tasks.

Other methods such as HCANet, LANet, and SAPNet also demonstrate strong performance, with a notably clear segmentation of larger areas like water bodies and roads. However, they show varying degrees of fragmentation and misclassification in more challenging areas, such as the transition zones between urban and natural environments. For example, in the areas where buildings are adjacent to vegetation, there is a noticeable difference in the precision of the boundaries as delineated by SSCBNet compared with other models.

The lower rows in the figure emphasize the models' performance in highly heterogeneous scenes. The ability of SSCBNet to maintain consistency in segmentation across different scenes, evident in its visual output, correlates with the robustness indicated by the overall accuracy metric. Models like U-Net and DeepLabV3+ show less precise segmentation in these areas, which may be attributed to the difficulty of segmenting small or narrow features such as roads and waterways amidst varied background classes.

In summary, the visualization of segmentation results provides a clear illustration of the strengths and weaknesses of each model, with SSCBNet displaying a high level of accuracy and detail retention across diverse land cover categories. The visual data support

the quantitative findings from the LoveDA dataset, reaffirming SSCBNet's proficiency in semantic segmentation and highlighting the importance of advanced modeling techniques that can handle the complexity and variability inherent in remote sensing imagery.

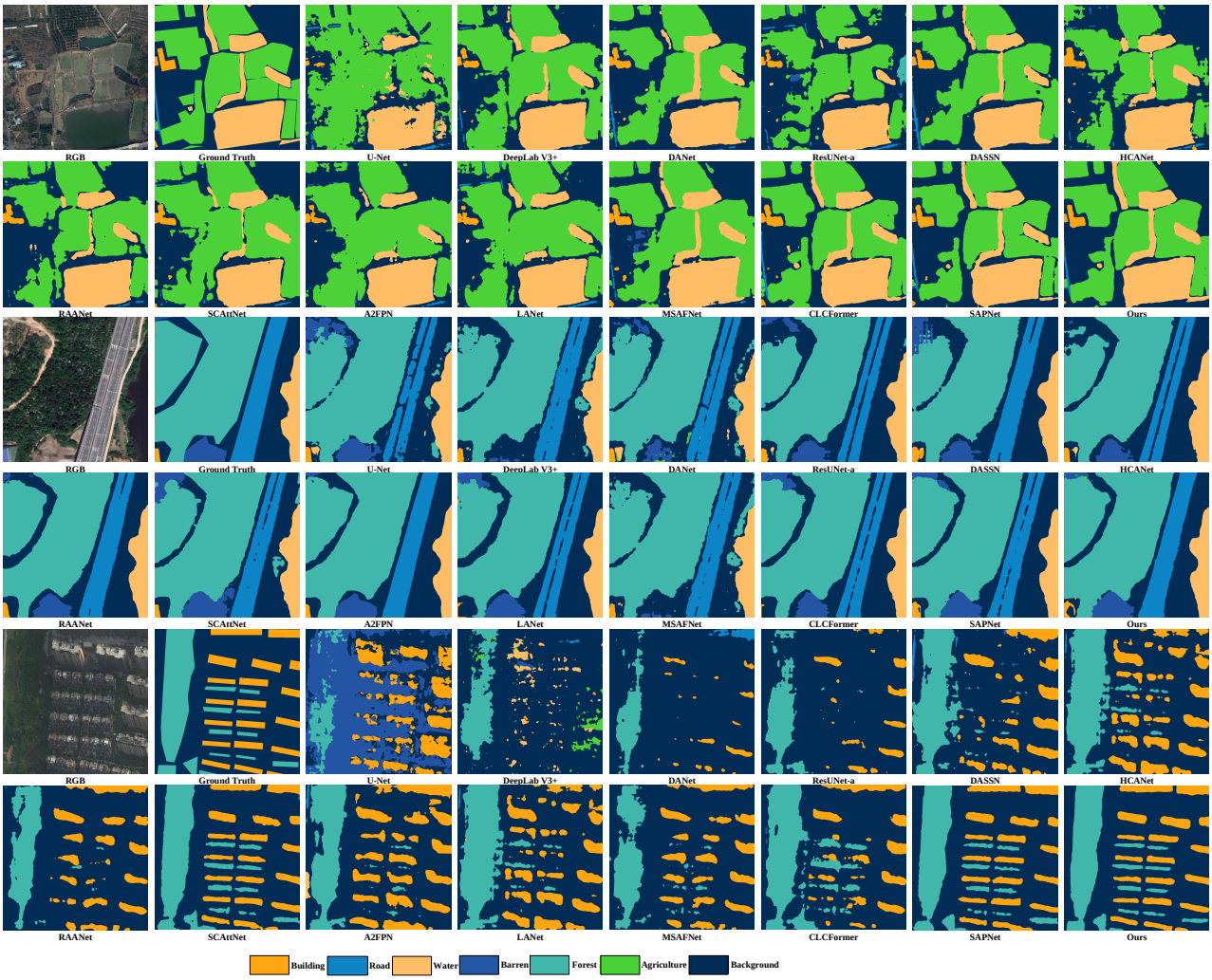

Figure 12. Visual inspections of random samples from LoveDA test set.

4.5. Impacts of SYA

By replacing the SYA layer with a convolution block (see Figure 3), we study the effects of SYA. The ablation study presented in Table 4 provides insightful findings on the impact of the SYA layer within SSCBNet. The study compares the performance metrics of the original SSCBNet against a variant where the SYA layer is replaced with a standard convolution block. The results clearly indicate the significance of the SYA layer in enhancing the network's performance on both the ISPRS Potsdam and LoveDA datasets.

For the ISPRS Potsdam dataset, the presence of the SYA layer in SSCBNet leads to a notable improvement in AF by approximately 5.93%, OA by 5.84%, and mIoU by 5.31%. Similarly, on the LoveDA dataset, the improvements are 5.91 in AF, 5.69 in OA, and 5.18 in mIoU. These increments underscore the efficacy of the SYA layer in boosting the discriminative capability of the network, leading to more precise segmentation results.

The decrease in performance when the SYA layer is omitted suggests that the ability of SSCBNet to accurately capture and utilize contextual information is significantly diminished. The SYA layer, designed to leverage both spectral and spatial contexts by modeling long-range dependencies and interclass relationships, evidently plays a crucial role in the network's ability to discern and classify complex features within the images. This is particularly reflected in the mIoU metric, which measures the accuracy of the segmentation

by considering both true-positive predictions and the overlap between the predicted and ground truth segments.

Furthermore, the reduction in AF across both datasets when the SYA layer is absent indicates that the model without SYA struggles to maintain a balance between precision and recall, leading to a higher rate of false positives and negatives. This could be particularly detrimental in applications where accurate segmentation is critical, such as land cover mapping for environmental monitoring or urban planning. Additionally, as shown in Figures 13 and 14, we visualize the predictions on two test sets. We zoom in specific areas in images to make a clear comparison to validate the effects of SYA.

In conclusion, the SYA layer is a vital component of SSCBNet, as evidenced by the ablation study. Its inclusion facilitates a significant enhancement in the network's segmentation accuracy and the ability to handle the diverse challenges presented by different landscapes. These results not only validate the design choices behind SSCBNet but also highlight the importance of attention mechanisms in advancing the field of semantic segmentation for RSIs. The study reinforces the notion that careful architectural considerations can lead to substantial improvements in model performance.

Table 4. Ablation study of SYA. Results are in form of AF/OA/mIoU.

Models	ISPRS Potsdam	LoveDA
SSCBNet	92.73/91.31/83.09	75.32/72.44/65.92
SSCBNet w/o SYA	86.80/85.47/77.78	69.41/66.75/60.74

4.6. Impacts of CFM

By removing CFM in SSCBNet, we examine the effects of CFM in our network. The ablation study detailed in Table 5 assesses the impact of CFM on the performance of SSCBNet. CFM is integral to SSCBNet's architecture, designed to fuse features from encoder and decoder blocks to enhance the network's representational power. By comparing the standard SSCBNet against a version without CFM, we can discern the contribution of CFM to the model's overall effectiveness.

The results indicate that CFM plays a significant role in the network's performance. In the ISPRS Potsdam dataset, the removal of CFM leads to a decrease in AF by 2.40%, OA by 2.37%, and mIoU by 2.16%. A similar trend is observed in the LoveDA dataset, with reductions of 1.95% in AF, 1.88% in OA, and 1.71% in mIoU. These results suggest that CFM is pivotal in enhancing the segmentation accuracy and the model's ability to generalize across different classes.

Without CFM, SSCBNet experiences a notable drop in its ability to accurately classify pixels, especially in complex urban environments where different classes share similar spectral characteristics. CFM's role in effectively merging low-level and high-level features is crucial for maintaining the spatial hierarchy of features, which is particularly important for distinguishing fine-grained details in high-resolution imagery. The decrease in mIoU further emphasizes that without CFM, the model's precision in identifying the overlap between predicted segmentation and the ground truth is compromised.

The impact of CFM removal is also evident in the network's reduced capacity to handle class imbalances and varied object scales, challenges that are inherently present in remote sensing data. CFM assists in bridging the gap between the detailed, local information captured by the encoder and the broader, contextual information reconstructed by the decoder. This symbiosis is essential for achieving high-quality segmentation that accurately reflects the diverse landscapes within RSIs. Additionally, Figures 13 and 14 visualize the results to examine the effects of CFM.

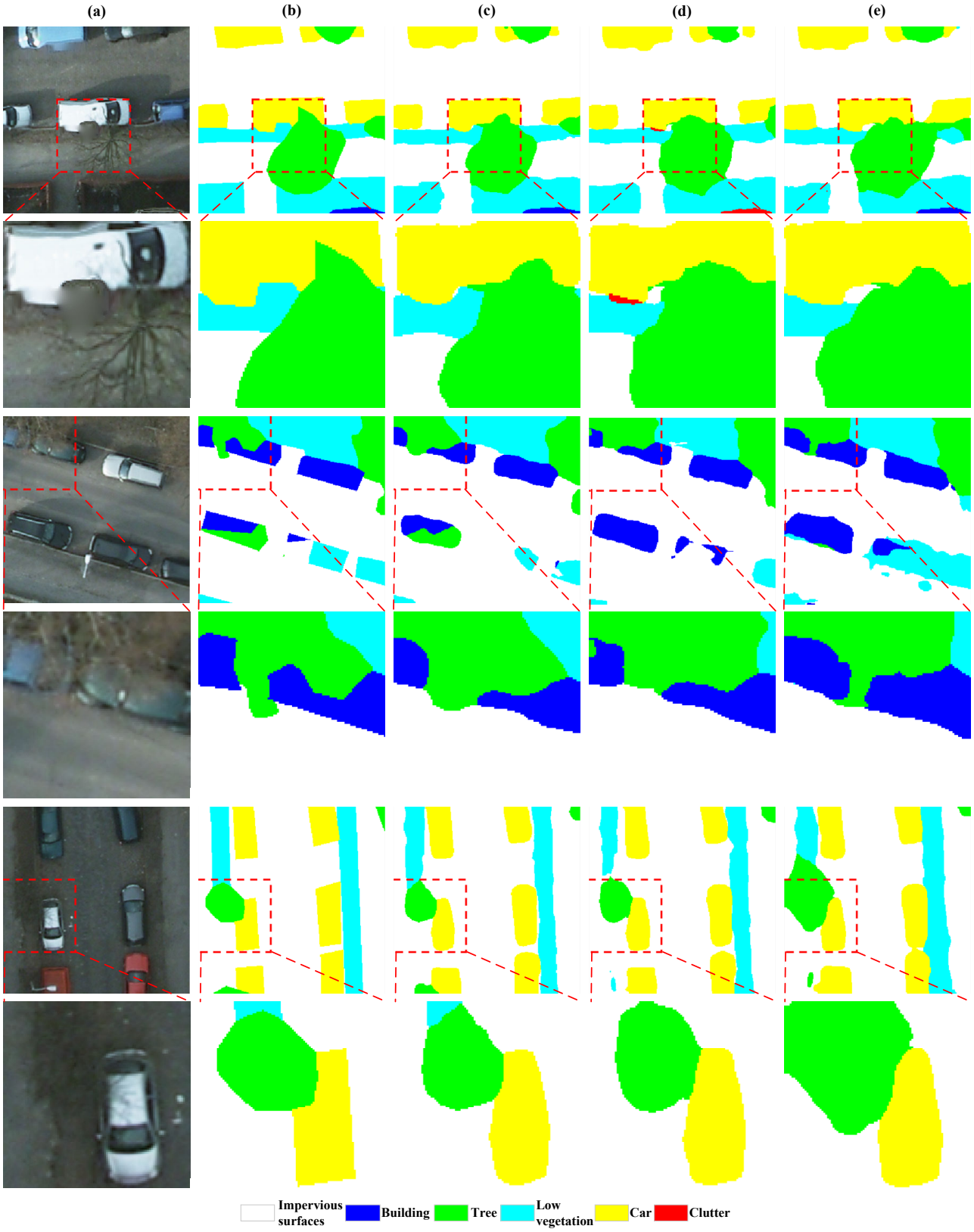

Figure 13. Visual inspections on validating SYA and CFM (random samples from ISPRS Potsdam test set): (**a**) input image, (**b**) ground truth, (**c**) predicted by SSCBNet, (**d**) predicted by SSCBNet without SYA, (**e**) predicted by SSCBNet without CFM.

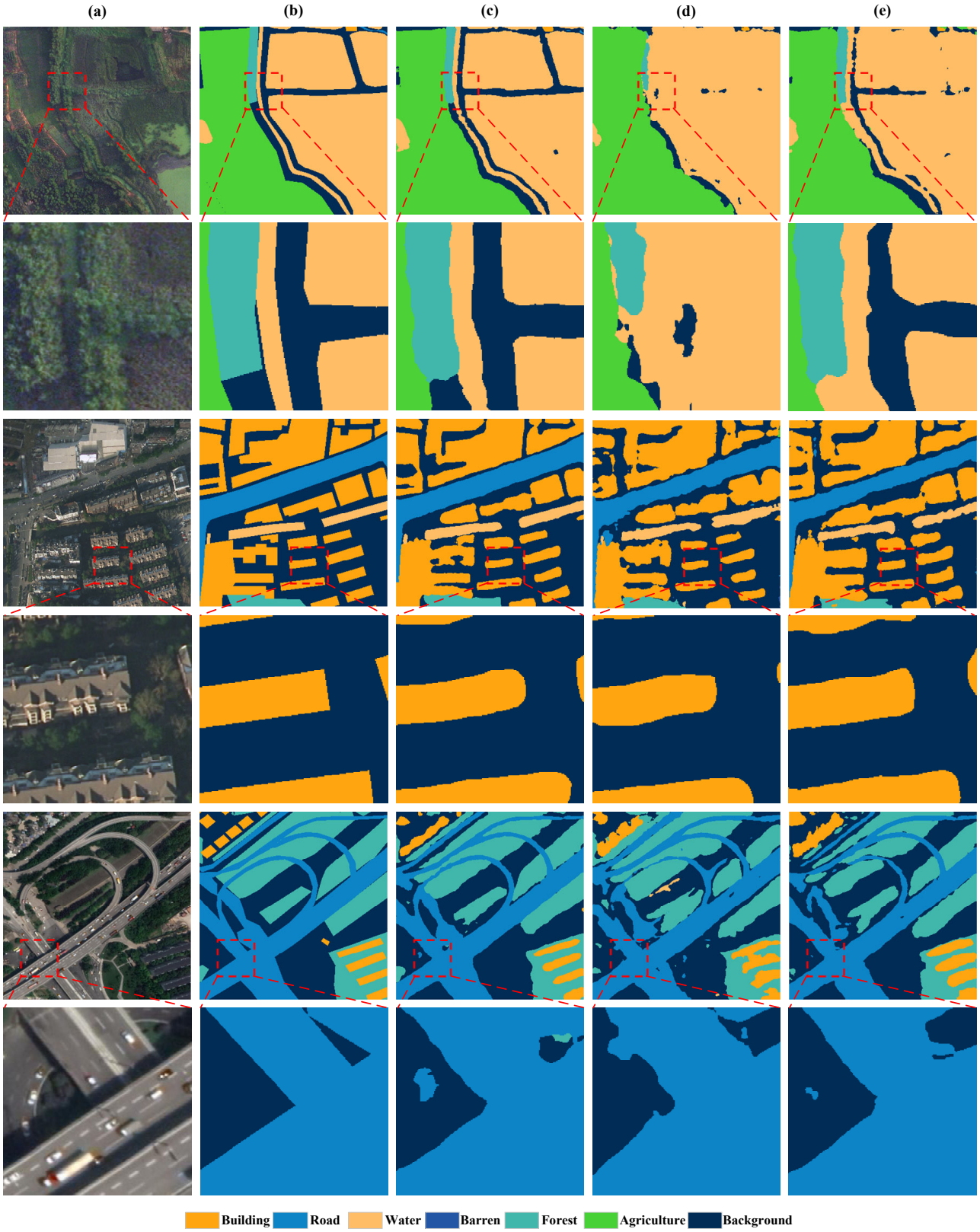

Figure 14. Visual inspections on validating SYA and CFM (random samples from LoveDA test set): (**a**) input image, (**b**) ground truth, (**c**) predicted by SSCBNet, (**d**) predicted by SSCBNet without SYA, (**e**) predicted by SSCBNet without CFM.

In conclusion, CFM is a critical component of SSCBNet, enhancing both the quality and the accuracy of the semantic segmentation. The ablation study confirms that CFM

substantially contributes to the network's state-of-the-art performance on both the ISPRS Potsdam and LoveDA datasets. This discussion underscores the importance of feature fusion in deep learning architectures, particularly for complex tasks such as semantic segmentation in remote sensing. The study not only validates the architectural design of SSCBNet but also serves as a testament to the effectiveness of integrating sophisticated modules like CFM for improving model performance.

Table 5. Ablation study of CFM. Results are in form of AF/OA/mIoU.

Models	ISPRS Potsdam	LoveDA
SSCBNet	92.73/91.31/83.09	75.32/72.44/65.92
SSCBNet w/o CFM	90.33/88.94/80.93	73.37/70.56/64.21

5. Conclusions

This study introduced SSCBNet, a novel approach for the semantic segmentation of RSIs that effectively integrates both spectral and spatial contexts. Our comprehensive experiments, conducted on the ISPRS Potsdam and LoveDA datasets, demonstrate that SSCBNet outperforms existing state-of-the-art methods across various metrics. The model's superior performance is particularly noteworthy in urban and agricultural landscapes, where the intricate interplay of spectral signatures poses significant segmentation challenges.

The proposed SSCBNet showcases an advanced capability to maintain the structural integrity of segmented objects and offers high precision in delineating complex urban features. The integration of the SYA layer and CFMs within the SSCBNet architecture has been pivotal in achieving these results. Ablation studies have further confirmed the importance of these components, indicating that their removal leads to a marked decrease in accuracy and model performance.

Looking ahead, the success of SSCBNet opens new avenues for research in the remote sensing domain. There is potential for adapting the network to other challenging scenarios within RSIs, such as those with higher levels of class imbalance or varying object scales. Additionally, future work could explore the integration of SSCBNet with other deep learning architectures or its application to other forms of environmental monitoring and land management tasks. Moreover, this study has centered on the methodological advancements and segmentation accuracy of SSCBNet, setting the stage for future investigations into the model's computational efficiency, which we plan to rigorously evaluate to balance performance with practical applicability.

As remote sensing technology continues to advance, the need for highly accurate and reliable semantic segmentation models becomes ever more critical. SSCBNet represents a significant step forward in meeting this need, offering a robust framework that leverages deep learning to understand and interpret the complex data our world presents.

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

Funding: This work was supported in part by the National Key Research and Development Program of China (Grant Nos. 2023YFC3209201 and 2023YFC3209305), the Special Funds for Basic Research Operating Expenses of Central-Level Public Welfare Research Institutes (Grant No. HKY-JBYW-2023-20), the Fundamental Research Funds for the Central Universities (Grant Nos. B230201007, B230204009, and B220206006), the National Natural Science Foundation of China (Grant Nos. 42101343, 42104033,

and 82004498), the Project of Water Science and Technology of Jiangsu Province (Grant No. 2021080), and the Excellent Postdoctoral Program of Jiangsu Province (Grant No. 2022ZB166).

Data Availability Statement: Public datasets were used in this paper. The download links are [ftp://wg-3-4-benchmark:LVK4jvv7mk@ftp.ipi.uni-hannover.de], accessed on 2 December 2022, and [https://github.com/Junjue-Wang/LoveDA], accessed on 2 December 2022. The source code and trained models will be released at [https://github.com/JacHothanLee/SSCBNet] (accessed on 1 January 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
2. Hong, D.; Zhang, B.; Li, H.; Li, Y.; Yao, J.; Li, C.; Werner, M.; Chanussot, J.; Zipf, A.; Zhu, X.X. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sens. Environ.* **2023**, *299*, 113856. [CrossRef]
3. Wieland, M.; Martinis, S.; Kiefl, R.; Gstaiger, V. Semantic segmentation of water bodies in very high-resolution satellite and aerial images. *Remote Sens. Environ.* **2023**, *287*, 113452. [CrossRef]
4. Xiang, D.; Zhang, X.; Wu, W.; Liu, H. Denseppmunet-a: A robust deep learning network for segmenting water bodies from aerial images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4202611. [CrossRef]
5. Luo, Z.; Yang, W.; Yuan, Y.; Gou, R.; Li, X. Semantic segmentation of agricultural images: A survey. *Inf. Process. Agric.* **2023**, *in press*. [CrossRef]
6. Xu, H.; Song, J.; Zhu, Y. Evaluation and Comparison of Semantic Segmentation Networks for Rice Identification Based on Sentinel-2 Imagery. *Remote Sens.* **2023**, *15*, 1499. [CrossRef]
7. Bandarupally, H.; Talusani, H.R.; Sridevi, T. Detection of Military Targets from Satellite Images using Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 30–31 October 2020; pp. 531–535.
8. Fourati, F.; Alouini, M.S. Artificial intelligence for satellite communication: A review. *Intell. Converg. Netw.* **2021**, *2*, 213–243. [CrossRef]
9. Lizarazo, I. SVM-based segmentation and classification of remotely sensed data. *Int. J. Remote Sens.* **2008**, *29*, 7277–7283. [CrossRef]
10. Zhou, C.; Liang, D.; Yang, X.; Yang, H.; Yue, J.; Yang, G. Wheat ears counting in field conditions based on multi-feature optimization and TWSVM. *Front. Plant Sci.* **2018**, *9*, 1024. [CrossRef]
11. Wang, D.; Liu, S.; Zhang, C.; Xu, M.; Yang, J.; Yasir, M.; Wan, J. An improved semantic segmentation model based on SVM for marine oil spill detection using SAR image. *Mar. Pollut. Bull.* **2023**, *192*, 114981. [CrossRef]
12. Bai, T.; Wang, L.; Yin, D.; Sun, K.; Chen, Y.; Li, W.; Li, D. Deep learning for change detection in remote sensing: A review. *Geo-Spat. Inf. Sci.* **2023**, *26*, 262–288. [CrossRef]
13. Gao, L.; Liu, B.; Fu, P.; Xu, M. Adaptive spatial tokenization transformer for salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602915. [CrossRef]
14. Thapa, A.; Horanont, T.; Neupane, B.; Aryal, J. Deep Learning for Remote Sensing Image Scene Classification: A Review and Meta-Analysis. *Remote Sens.* **2023**, *15*, 4804. [CrossRef]
15. Ciotola, M.; Scarpa, G. Fast Full-Resolution Target-Adaptive CNN-Based Pansharpening Framework. *Remote Sens.* **2023**, *15*, 319. [CrossRef]
16. 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.
17. Chen, G.; Tan, X.; Guo, B.; Zhu, K.; Liao, P.; Wang, T.; Wang, Q.; Zhang, X. SDFCNv2: An improved FCN framework for remote sensing images semantic segmentation. *Remote Sens.* **2021**, *13*, 4902. [CrossRef]
18. Chen, L.; Dou, X.; Peng, J.; Li, W.; Sun, B.; Li, H. EFCNet: Ensemble Full Convolutional Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8011705. [CrossRef]
19. 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]
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
21. Zhao, Q.; Liu, J.; Li, Y.; Zhang, H. Semantic segmentation with attention mechanism for remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5403913. [CrossRef]
22. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]

23. Huang, Z.; Zhang, Q.; Zhang, G. MLCRNet: Multi-level context refinement for semantic segmentation in aerial images. *Remote Sens.* **2022**, *14*, 1498. [CrossRef]
24. ISPRS 2D Semantic Labeling—Potsdam. Available online: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx (accessed on 20 October 2021).
25. ISPRS 2D Semantic Labeling—Vaihingen. Available online: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx (accessed on 20 October 2021).
26. Zhang, C.; Chen, Y.; Yang, X.; Gao, S.; Li, F.; Kong, A.; Zu, D.; Sun, L. Improved remote sensing image classification based on multi-scale feature fusion. *Remote Sens.* **2020**, *12*, 213. [CrossRef]
27. Du, S.; Du, S.; Liu, B.; Zhang, X. Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network and object based approach. *Remote Sens. Environ.* **2021**, *261*, 112480. [CrossRef]
28. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
29. Brauwers, G.; Frasincar, F. A General Survey on Attention Mechanisms in Deep Learning. *IEEE Trans. Knowl. Data. Eng.* **2023**, *35*, 3279–3298. [CrossRef]
30. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149. [CrossRef]
31. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
32. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 173–190.
33. Zhao, D.; Wang, C.; Gao, Y.; Shi, Z.; Xie, F. Semantic segmentation of remote sensing image based on regional self-attention mechanism. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8010305. [CrossRef]
34. Nan, Z.; Peng, J.; Jiang, J.; Chen, H.; Yang, B.; Xin, J.; Zheng, N. A joint object detection and semantic segmentation model with cross-attention and inner-attention mechanisms. *Neurocomputing* **2021**, *463*, 212–225. [CrossRef]
35. Li, X.; Xu, F.; Lyu, X.; Gao, H.; Tong, Y.; Cai, S.; Li, S.; Liu, D. Dual attention deep fusion semantic segmentation networks of large-scale satellite remote-sensing images. *Int. J. Remote Sens.* **2021**, *42*, 3583–3610. [CrossRef]
36. Li, X.; Xu, F.; Xia, R.; Lyu, X.; Gao, H.; Tong, Y. Hybridizing Cross-Level Contextual and Attentive Representations for Remote Sensing Imagery Semantic Segmentation. *Remote Sens.* **2021**, *13*, 2986. [CrossRef]
37. Niu, R.; Sun, X.; Tian, Y.; Diao, W.; Chen, K.; Fu, K. Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 3065112. [CrossRef]
38. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *59*, 426–435. [CrossRef]
39. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.
40. Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 60–77. [CrossRef]
41. Kotaridis, I.; Lazaridou, M. Remote sensing image segmentation advances: A meta-analysis. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 309–322. [CrossRef]
42. Mboga, N.; Georganos, S.; Grippa, T.; Lennert, M.; Vanhuysse, S.; Wolff, E. Fully Convolutional Networks and Geographic Object-Based Image Analysis for the Classification of VHR Imagery. *Remote Sens.* **2019**, *11*, 597. [CrossRef]
43. He, H.; Yang, D.; Wang, S.; Wang, S.; Li, Y. Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder-Decoder Network and Structural Similarity Loss. *Remote Sens.* **2019**, *11*, 1015. [CrossRef]
44. Liu, C.; Li, H.; Su, A.; Chen, S.; Li, W. Identification and Grading of Maize Drought on RGB Images of UAV Based on Improved U-Net. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 198–202. [CrossRef]
45. Sun, Y.; Tian, Y.; Xu, Y. Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning. *Neurocomputing* **2019**, *330*, 297–304. [CrossRef]
46. Tan, X.; Xiao, Z.; Wan, Q.; Shao, W. Scale Sensitive Neural Network for Road Segmentation in High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 533–537. [CrossRef]
47. Yue, K.; Yang, L.; Li, R.; Hu, W.; Zhang, F.; Li, W. TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation. *ISPRS J. Photogramm. Remote Sens.* **2019**, *156*, 1–13. [CrossRef]
48. 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]
49. Cheng, S.; Li, B.; Sun, L.; Chen, Y. HRRNet: Hierarchical Refinement Residual Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1244. [CrossRef]
50. Sun, L.; Fang, Y.; Chen, Y.; Huang, W.; Wu, Z.; Jeon, B. Multi-Structure KELM With Attention Fusion Strategy for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5539217. [CrossRef]
51. Huang, J.; Zhang, X.; Sun, Y.; Xin, Q. Attention-Guided Label Refinement Network for Semantic Segmentation of Very High Resolution Aerial Orthoimages. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4490–4503. [CrossRef]

52. Sun, L.; Cheng, S.; Zheng, Y.; Wu, Z.; Zhang, J. SPANet: Successive Pooling Attention Network for Semantic Segmentation of Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4045–4057. [CrossRef]
53. Yang, X.; Li, S.; Chen, Z.; Chanussot, J.; Jia, X.; Zhang, B.; Li, B.; Chen, P. An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 238–262. [CrossRef]
54. Li, R.; Wang, L.; Zhang, C.; Duan, C.; Zheng, S. A2-FPN for semantic segmentation of fine-resolution remotely sensed images. *Int. J. Remote Sens.* **2022**, *43*, 1131–1155. [CrossRef]
55. Li, H.; Qiu, K.; Chen, L.; Mei, X.; Hong, L.; Tao, C. SCAttNet: Semantic Segmentation Network With Spatial and Channel Attention Mechanism for High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 905–909. [CrossRef]
56. Liu, R.; Tao, F.; Liu, X.; Na, J.; Leng, H.; Wu, J.; Zhou, T. RAANet: A Residual ASPP with Attention Framework for Semantic Segmentation of High-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3109. [CrossRef]
57. Li, X.; Xu, F.; Liu, F.; Lyu, X.; Tong, Y.; Xu, Z.; Zhou, J. A Synergistical Attention Model for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5400916. [CrossRef]
58. Li, X.; Xu, F.; Liu, F.; Tong, Y.; Lyu, X.; Zhou, J. Semantic Segmentation of Remote Sensing Images by Interactive Representation Refinement and Geometric Prior-Guided Inference. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5400318. [CrossRef]
59. Xu, Z.Q.J.; Zhang, Y.; Luo, T. Overview frequency principle/spectral bias in deep learning. *arXiv* **2022**, arXiv:2201.07395.
60. Azad, R.; Kazerouni, A.; Sulaiman, A.; Bozorgpour, A.; Aghdam, E.K.; Jose, A.; Merhof, D. Unlocking Fine-Grained Details with Wavelet-Based High-Frequency Enhancement in Transformers. In Proceedings of the International Workshop on Machine Learning in Medical Imaging, Vancouver, BC, Canada, 8 October 2023; Springer: Cham, Switzerland, 2023; pp. 207–216.
61. Gueguen, L.; Sergeev, A.; Kadlec, B.; Liu, R.; Yosinski, J. Faster neural networks straight from jpeg. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
62. Ehrlich, M.; Davis, L.S. Deep residual learning in the jpeg transform domain. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3484–3493.
63. Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.K.; Ren, F. Learning in the frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1740–1749.
64. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF international conference on computer vision, Virtual Conference, 11–17 October 2021; pp. 783–792.
65. Su, B.; Liu, J.; Su, X.; Luo, B.; Wang, Q. Cfcanet: A complete frequency channel attention network for sar image scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11750–11763. [CrossRef]
66. Li, X.; Xu, F.; Yong, X.; Chen, D.; Xia, R.; Ye, B.; Gao, H.; Chen, Z.; Lyu, X. SSCNet: A Spectrum-Space Collaborative Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 5610. [CrossRef]
67. Lyu, X.; Jiang, W.; Li, X.; Fang, Y.; Xu, Z.; Wang, X. MSAFNet: Multiscale Successive Attention Fusion Network for Water Body Extraction of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 3121. [CrossRef]
68. Long, J.; Li, M.; Wang, X. Integrating Spatial Details With Long-Range Contexts for Semantic Segmentation of Very High-Resolution Remote-Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2501605. [CrossRef]