

## Article

# RFA2Net: A Receptive Field and Global Attention Enhanced Model for Semantic Segmentation of High-Resolution Remote-Sensing Images

Xingyi Zhong <sup>1</sup>, Junhao Liu <sup>2</sup>, Yiqiu Mao <sup>2</sup>, Yubin Zhong <sup>2,\*</sup> and Guanquan Zhu <sup>3,\*</sup>

<sup>1</sup> School of Architectural Design and Art, Guangdong Construction Polytechnic, Guangzhou 510440, China; zhongxingyi@gdcvi.edu.cn

<sup>2</sup> School of Mathematics and Information Science, Guangzhou University, Guangzhou 510006, China; 32315300029@e.gzhu.edu.cn (J.L.); 205366@e.gzhu.edu.cn (Y.M.)

<sup>3</sup> College of Future Information Technology, Fudan University, Shanghai 200433, China

\* Correspondence: zhong\_yb@gzhu.edu.cn (Y.Z.); guanquanzhu@163.com (G.Z.)

## Abstract

Semantic segmentation of high-resolution remote-sensing images is critical for urban planning, land-cover mapping, and ecological monitoring. However, existing methods face limitations in handling complex land-cover types, multi-scale objects, and modeling long-range dependencies. To address these challenges, we propose RFA2Net, an enhanced semantic segmentation model based on the DeepLabv3+ framework. The key innovations include the integration of the RFCSA-Conv module into the ResNet101 backbone to enhance feature representation and long-range dependency modeling, the design of the RFA-DASPP structure built upon the Dense ASPP framework with the novel RFCA-DConv dilated convolution module to reduce information loss during multi-scale feature fusion and enhance the model's ability to perceive long-range directional structures, and the introduction of a Dual-Branch Fusion Network to improve segmentation accuracy for small-scale objects. Experimental results on the ISPRS Potsdam and LoveDA datasets demonstrate that RFA2Net outperforms several CNN and Transformer-based models, achieving 78.94% and 59.46% mean intersection over union (mIoU) on the ISPRS Potsdam and LoveDA datasets, respectively, with improvements of 3.19% and 3.08% over the original DeepLabv3+. Ablation studies and comparative experiments further confirm the model's effectiveness, robustness, and practical applicability in high-resolution remote-sensing image segmentation, with particular relevance to environmental monitoring and sustainable energy applications.

**Keywords:** computer vision; semantic segmentation; RFA2Net; Artificial Intelligence



Academic Editor: Arslan Munir

Received: 22 February 2026

Revised: 23 April 2026

Accepted: 27 April 2026

Published: 29 April 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

High-resolution remote-sensing images refers to Earth observation data acquired by satellite, airborne, or unmanned aerial vehicle (UAV) platforms, with spatial resolutions ranging from sub-meter to several meters [1]. Compared with medium- and low-resolution images, such data are able to capture finer geometric structures, spatial patterns, and texture details, enabling a more accurate representation of complex land-cover features, including buildings, transportation networks, and surface materials [2]. With continuous advances in sensor capabilities and Earth observation technologies, high-resolution remote-sensing images has been increasingly applied in a wide range of fields, such as urban planning,

land-use and land-cover change analysis, and environmental and ecological assessment, as well as disaster monitoring and management [3–6].

Semantic segmentation is a fundamental image analysis task that provides pixel-level interpretation by assigning a semantic label to each pixel in an image [7]. Compared with image classification and object detection, semantic segmentation enables more detailed and comprehensive scene understanding. It is important for the interpretation of high-resolution remote-sensing images [8]. By means of semantic segmentation techniques, major land-cover types such as buildings, roads, vegetation, and water bodies can be accurately delineated and identified. It can provide essential data support for land-cover mapping [9,10], urban structure analysis [11,12], and environmental monitoring [13,14]. However, high-resolution remote-sensing images are characterized by wide variations in object scale, complex boundary shapes, and diverse spatial arrangements, which pose significant challenges to achieving high segmentation accuracy.

Early semantic segmentation methods relied on hand-crafted features and domain-specific knowledge, which limited their robustness and generalization ability when applied to complex high-resolution remote-sensing images [15,16]. With the rapid development of deep learning, CNN-based approaches have gradually become dominant. Chen et al. [17] proposed Deeplabv3+, which employs an encoder–decoder framework for semantic segmentation. Compared with Transformer-based models, DeepLabv3+ achieves competitive segmentation performance with lower computational complexity, making it suitable for medium-scale remote-sensing datasets and practical engineering applications [18]. MGNet introduces a multi-scale global modeling mechanism within an encoder–decoder CNN framework to enhance the representation of objects with varying scales in remote-sensing images [19]. Band-combination-based CNN models leverage spectral characteristics of remote-sensing data to improve feature discrimination and segmentation accuracy [20]. To strengthen multi-scale modeling capability, several studies have incorporated multi-scale supervision and contextual fusion structures into remote-sensing segmentation networks. BEMS-UNetFormer incorporates boundary-aware modules and multi-scale feature modeling, significantly improving segmentation performance in high-resolution imagery [21]. MST-DeepLabv3+ enhances segmentation consistency across objects of varying sizes by integrating multi-scale feature extraction mechanisms [22]. Attention mechanisms have also proven effective in handling the complex spatial structures of remote-sensing images. For instance, feature information entropy-based channel attention improves feature weighting by leveraging information content, leading to better segmentation accuracy [23]. Some studies also emphasize the joint modeling of global and local features. CM-UNet combines CNN-based local feature extraction with global sequence modeling to achieve effective feature fusion and improved segmentation results [24]. Despite these advances, CNN-based methods still face notable challenges in high-resolution scenarios. It mainly includes weak capability of extracting complex features, limited ability to model long-range dependencies, poor performance in multi-scale object segmentation, and insufficient utilization of shallow features [25,26].

High-resolution remote-sensing image semantic segmentation is a dense prediction task in which effectively modeling global context is essential for maintaining category consistency and accurately understanding complex scenes. Due to the inherently local receptive fields of convolutional operations, conventional CNNs often fail to capture long-range dependencies across distant regions. To overcome this limitation, researchers have developed a variety of methods aimed at enhancing global semantic representation. SENet [27] strengthens global feature representation by performing channel-wise recalibration through global average pooling. CBAM [28] applies channel and spatial attention to jointly determine “which” features and “where” in the spatial domain should

be emphasized. DANet [29] features parallel position and channel attention branches that explicitly capture long-range dependencies in both spatial and channel dimensions. Coordinate attention (CA) [30] embeds positional information into channel attention to enhance directional awareness while maintaining lightweight global context modeling. Hybrid architectures combining convolution with self-attention have also been explored to balance local detail extraction with global dependency modeling. Li et al. incorporates convolutional operations into attention mechanisms, enabling models to simultaneously capture local and global information [31]. DETR [32] utilizes CNNs for feature extraction and a Transformer encoder–decoder structure to capture global relationships across the entire image. Liang et al. proposed a global–local deep fusion strategy that integrates Transformer-based global context modeling with fine-scale local details [33].

In semantic segmentation of high-resolution remote-sensing images, the complex spatial layouts and rich texture information of surface objects impose high demands on feature extraction. A variety of enhanced convolutional modules have been proposed to enhance feature representation capability. Depth-wise separable convolution decomposes standard convolution into depth-wise and point-wise operations, significantly reducing computational cost and parameter complexity while maintaining strong feature learning capacity [17]. Deformable convolution introduces learnable offset parameters to dynamically adjust sampling locations, enabling more flexible modeling of objects with irregular shapes and varying orientations [34,35]. Res2Net, proposed by Shang-Hua et al., splits convolutional features into multiple scale-wise subsets and connects them through hierarchical residual-like connections, strengthening local multi-scale feature extraction [36]. RFAConv, proposed by Zhang et al. [37], dynamically generates non-overlapping receptive field spatial features that match the convolution kernel size and learns independent attention weights for each receptive field region. It enables differentiated enhancement of features across spatial locations and scales. Although these methods improve feature extraction capability, they still struggle to achieve adaptive feature modeling and selection based on global dependency information when dealing with complex object shapes, multi-scale targets, and noise interference. This limitation constrains their further performance improvement in high-resolution remote-sensing semantic segmentation tasks [38].

To address the challenges of semantic segmentation in high-resolution remote-sensing images, this study proposes an improved model RFA2Net based on DeepLabv3+. We design a novel convolutional module, RFCSA-Conv, which enhances feature representation in complex high-resolution remote-sensing scenes and strengthens the modeling of long-range dependencies. Furthermore, by integrating the Dense ASPP framework with a newly dilated convolution RFCA-DConv, we construct the RFA-DASPP structure. The incorporation of this module mitigates information loss during multi-scale feature fusion and improves the model's sensitivity to positional cues and long-range contextual relationships. Finally, a Dual-Branch Fusion Network (DFN) is proposed to fuse features from the first three stages of the backbone, and the aggregated features are fed into the decoder. This strategy effectively enhances the recognition and segmentation performance of small-scale objects.

The main contributions of this paper are summarized as follows:

- (1) This paper proposes a RFCSA-Conv module and integrates it into ResNet101, forming RFA-ResNet. It enhances the model's ability to extract complex structural features and improves its capability to model long-range dependencies.
- (2) This paper introduces a RFA-DASPP structure, which builds upon the Dense ASPP framework and incorporates the novel dilated convolution module RFCA-DConv. The proposed design effectively alleviates information loss during multi-scale feature fusion, expands the effective receptive field and enhances the model's ability to perceive long-range directional structures.

- (3) This paper develops a Dual-Branch Fusion Network (DFN). Through a bidirectional modulation mechanism between high-level semantic features and low-level detailed features, the proposed network effectively enhances the model's capability for small-target segmentation.
- (4) An improved semantic segmentation model, RFA2Net, is proposed to address problems including inadequate complex feature extraction and limited capability to model long-range dependencies. The proposed model achieves enhanced performance in high-resolution remote-sensing image semantic segmentation.

The remainder of this paper is organized as follows. Section 2 introduces the materials and methodologies of the proposed RFA2Net. Section 3 presents comprehensive experimental results on two benchmark remote-sensing datasets. Section 4 discusses the effectiveness and limitations of the proposed method and provides directions for future research. Finally, Section 5 concludes this paper.

## 2. Materials and Methods

### 2.1. Architecture of the Proposed RFA2Net

The proposed RFA2Net adopts an encoder–decoder architecture, as illustrated in Figure 1. The encoder is responsible for extracting hierarchical feature representations from the input image, while the decoder progressively restores spatial resolution to produce pixel-wise segmentation outputs.

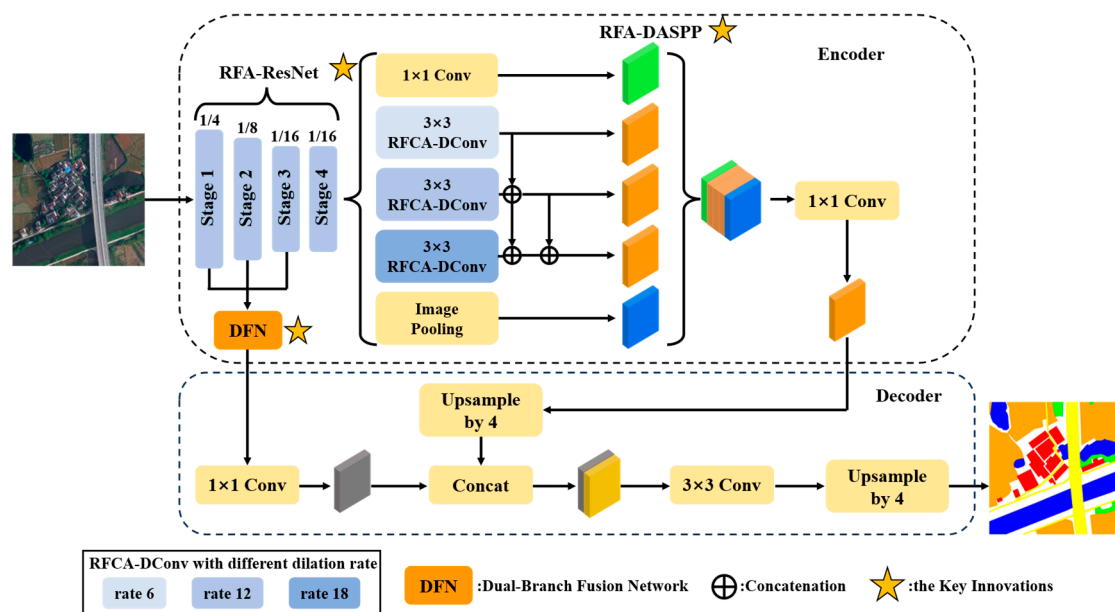


**Figure 1.** Architecture of the proposed RFA2Net.

Specifically, RFA-ResNet is adopted as the backbone of the encoder to enhance feature extraction capability and facilitate the modeling of long-range dependencies. It processes the input image to generate hierarchical feature representations for subsequent modules. The resulting high-level features are then fed into the RFA-DASPP module, which is designed to improve multi-scale feature fusion. By aggregating contextual information at multiple scales, it produces enriched feature representations for the decoder. Meanwhile, shallow features from the early three stages of the backbone are processed by the Dual-Branch Fusion Network (DFN), which enhances the utilization of low-level features. The fused features are subsequently provided to the decoder to preserve high-resolution spatial details and improve the segmentation of small-scale objects.

Based on the overall architecture, the three key components of the proposed model, namely RFA-ResNet, RFA-DASPP, and DFN, are introduced in detail in the following subsections.

### 2.1.1. RFA-ResNet

In high-resolution remote-sensing images, objects with similar local visual characteristics may belong to entirely different semantic categories due to variations in their spatial locations or surrounding contexts. The ResNet101 backbone adopted in DeepLabv3+ primarily relies on shared parameter  $3 \times 3$  convolutional layers for feature extraction. While this design is effective in capturing fine-grained local details, the use of spatially shared kernel weights limits the network's ability to adaptively extract discriminative features from complex regions. Recently, Zhang et al. proposed RFAConv, a convolutional module that incorporates receptive field attention [37]. By applying the attention mechanism at the receptive field level and generating distinct attention weights for each element within the receptive field, RFAConv can dynamically adjust the convolution kernel based on local features from different receptive fields, thereby enhancing its ability to extract local features. This module has demonstrated notable performance improvements in tasks such as image classification and object detection. However, RFAConv mainly focuses on local receptive field information when generating attention maps and does not explicitly model long-range spatial dependencies. This limitation restricts its effectiveness in semantic segmentation.

To address this issue, this study proposes the novel convolutional module RFCSA-Conv, as illustrated in Figure 2. Channel attention and spatial attention mechanisms are integrated into RFAConv, so that the attention weights of each receptive field incorporate not only local information within the receptive field but also long-range dependency information. This design preserves the strong local feature representation capability of RFAConv and enhances the model's awareness of global structures and semantic consistency, making it more suitable for semantic segmentation of high-resolution remote-sensing images.

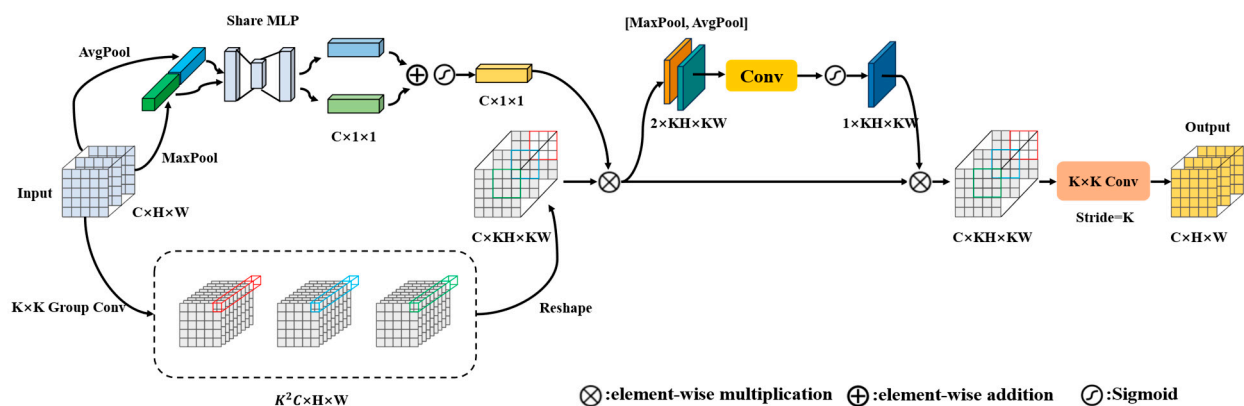


**Figure 2.** Framework of the proposed RFCSA-Conv.

First, the input feature  $X \in \mathbb{R}^{C \times H \times W}$  is processed using a  $K \times K$  group convolution with channel grouping. It should be noted that the stride and padding of the grouped convolution are equal to those of the entire RFCSA-Conv module. The input channels are divided into  $C$  groups along the channel dimension, and each group is convolved with one corresponding  $K \times K$  convolution kernel, respectively. Different from standard group convolution, we stack the extracted features within a same convolution slider in the same group along the channel dimension to form a  $K^2 \times 1 \times 1$  feature. In this way, one group can obtain  $H \times W$  such stacked features. These stacked features from the same group are then assembled according to the spatial positions of their corresponding convolution slider to form a  $K^2 \times H \times W$  feature map. Finally, the resulting features from all  $C$  groups are

concatenated along the channel dimension to form the feature  $X' \in \mathbb{R}^{K^2 C \times H \times W}$ . After batch normalization and ReLU activation, a reshape operation is applied to obtain the receptive field feature representation  $X_{rf} \in \mathbb{R}^{C \times KH \times KW}$ . The process can be formulated as follows:

$$X_{rf} = \text{Reshape} \left( \text{ReLU} \left( \text{BN} \left( \text{Group Conv}^{K \times K}(X) \right) \right) \right) \quad (1)$$

The reshape operation first splits the feature  $X'$  obtained after batch normalization and ReLU activation into  $C$  groups along the channel dimension, which can be expressed as:

$$X' \rightarrow \left\{ X'_g \right\}_{g=1}^C, \quad X'_g \in \mathbb{R}^{K^2 \times H \times W} \quad (2)$$

For each group  $g$  and each spatial position  $(i, j)$ , the  $K^2$ -dimensional feature vector at that position is unfolded into a  $1 \times K \times K$  feature block:

$$\text{unfold} \left( X'_g[:, i, j] \right) \in \mathbb{R}^{1 \times K \times K} \quad (3)$$

Subsequently, all of these unfolded  $1 \times K \times K$  blocks within the same group are concatenated in spatial order to construct a feature map,  $F_g \in \mathbb{R}^{1 \times KH \times KW}$ :

$$F_g = \text{concat}_{(i,j)} \left( \text{unfold} \left( X'_g[:, i, j] \right) \right) \quad (4)$$

Finally, the refined features from all  $C$  groups are concatenated along the channel dimension, yielding the final output feature  $X_{rf} \in \mathbb{R}^{C \times KH \times KW}$ :

$$X_{rf} = \text{concat}_{g=1}^C (F_g) \quad (5)$$

Meanwhile, the channel attention weights  $M_c \in \mathbb{R}^{C \times 1 \times 1}$  are computed from the input feature  $X$ . The channel attention weights are then multiplied by the receptive field features  $X_{rf}$  to obtain the channel-refined receptive field features  $X_{rfc} \in \mathbb{R}^{C \times KH \times KW}$ . This process can be formulated as follows:

$$M_c(X) = \sigma \left( \text{MLP}(\text{AvgPool}(X)) + \text{MLP}(\text{MaxPool}(X)) \right) \quad (6)$$

$$X_{rfc} = M_c(X) \otimes X_{rf} \quad (7)$$

where  $\sigma(\cdot)$  denotes the sigmoid activation function,  $\otimes$  denotes element-wise multiplication and MLP represents a shared fully connected layer operation.

Subsequently, spatial attention weights  $M_s \in \mathbb{R}^{C \times 1 \times 1}$  are computed for  $X_{rfc}$ . The spatial attention weights are then multiplied by  $X_{rfc}$  to obtain the spatially refined receptive field features  $X_{rfcs} \in \mathbb{R}^{C \times KH \times KW}$ . This process can be expressed as follows:

$$M_s(X_{rfc}) = \sigma \left( f^{7 \times 7} \left( \left[ \text{AvgPool}(X_{rfc}); \text{MaxPool}(X_{rfc}) \right] \right) \right) \quad (8)$$

$$X_{rfcs} = M_s(X_{rfc}) \otimes X_{rfc} \quad (9)$$

where  $\sigma(\cdot)$  denotes the sigmoid activation function,  $f^{7 \times 7}$  represents a  $7 \times 7$  convolutional operator and  $\otimes$  denotes element-wise multiplication.

Finally, a  $K \times K$  convolution operation with a stride of  $K$  is applied to  $X_{rfcs}$  to generate the output feature map  $O \in \mathbb{R}^{C \times H \times W}$ . Through the above operations, each element within different receptive fields is assigned distinct global attention weights. This design preserves the original local feature representation capability while endowing the original RFACConv with explicit global feature modeling ability.

ResNet101 is composed of a series of Bottleneck blocks, as illustrated in Figure 3a. In this study, the proposed RFCSA-Conv is used to replace the standard  $3 \times 3$  convolution within the Bottleneck structure, as shown in Figure 3b, with its padding and stride kept consistent with the original  $3 \times 3$  convolution to preserve spatial feature size. This enables feature extraction to be guided jointly by local receptive field attention and long-range information, allowing the model to capture complex structures more precisely. It also improves the overall consistency of the segmentation results, reducing misclassification caused by local ambiguity.

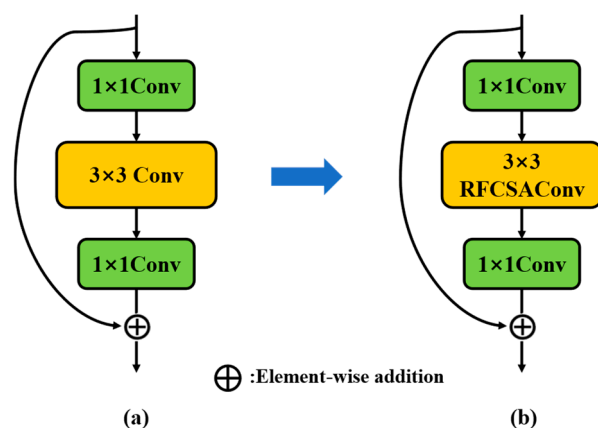


**Figure 3.** Framework of Bottleneck. (a) Bottleneck in ResNet101. (b) Bottleneck in RFA-ResNet.

### 2.1.2. RFA-DASPP

In high-resolution remote-sensing images, objects often exhibit complex shapes with strong directionality and spatial continuity, such as roads, rivers, and building boundaries. Moreover, a single remote-sensing image may contain objects at multiple scales. In the original DeepLabv3+, there is a lack of explicit modeling of spatial positional information and long-range dependencies. Meanwhile, the original ASPP structure extracts multi-scale contextual information through parallel dilated convolutions with different dilation rates, as shown in Figure 4a. However, since each branch operates independently, the fusion of multi-scale features lacks continuity, which may result in partial feature loss. To address these limitations, this paper proposes an improved RFA-DASPP structure, as illustrated in Figure 4b. Based on the original ASPP module, the proposed structure incorporates the Dense ASPP framework [39] together with the proposed RFCA-DConv module.

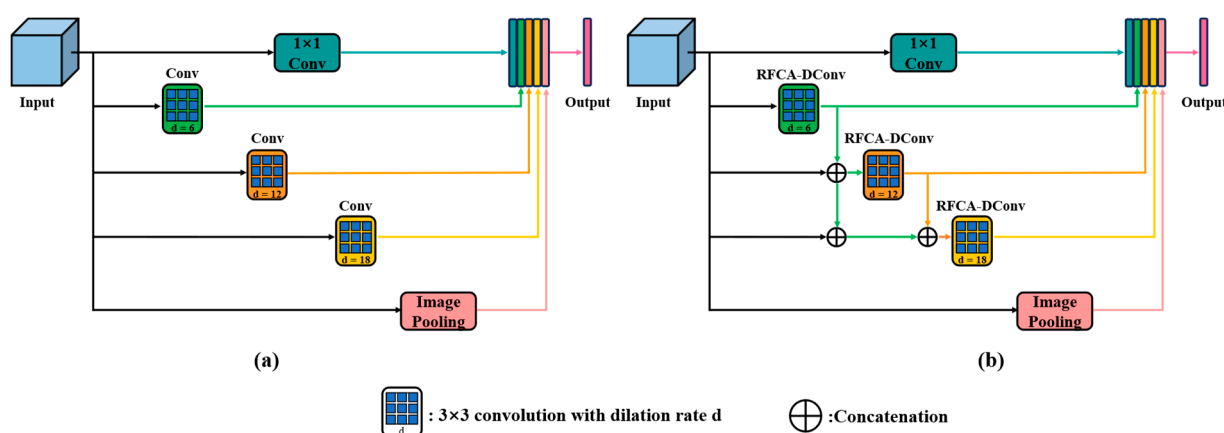


**Figure 4.** ASPP and RFA-DASPP structure. (a) ASPP. (b) RFA-DASPP.

The Dense ASPP architecture progressively fuses features with different dilation rates. Specifically, for the atrous convolution branch with a dilation rate of 12, the input feature is

first concatenated with the output feature of the atrous convolution branch with a dilation rate of 6 along the channel dimension, and the concatenated feature is then processed by the atrous convolution. For the atrous convolution branch with a dilation rate of 18, the input feature is first concatenated with the output feature of the atrous convolution branches with dilation rates of 6 and 12 along the channel dimension, and the fused feature is subsequently fed into the atrous convolution operation. This dense connection enables information to flow continuously across multiple scales, effectively alleviating feature discontinuity and reducing information loss during multi-scale fusion. Then, features yielded by the  $1 \times 1$  convolution, pooling layer and three dilated branches with distinct dilation ratios are concatenated in the channel dimension. Afterwards, a  $1 \times 1$  convolution is utilized to reduce channel dimensions, and the refined feature map serves as the high-level input of the decoder. Dense ASPP architecture enlarges the receptive field, allowing the network to capture the global structure and spatial relationships of large objects. This helps to reduce local misclassifications and boundary fragmentation, improving the completeness and accuracy of segmentation results in complex scenes.

This paper modifies the RFCANet proposed by Zhang et al. [37] and introduces its dilated convolution variant RFCA-DConv. The structure of this module is illustrated in Figure 5. By incorporating the coordinate attention mechanism [30], the feature extraction process is guided by both the importance of features within the receptive field and channel attention that integrates long-range positional information. This design enhances the model's ability to represent complex boundaries and long-range directional structures in remote-sensing images.

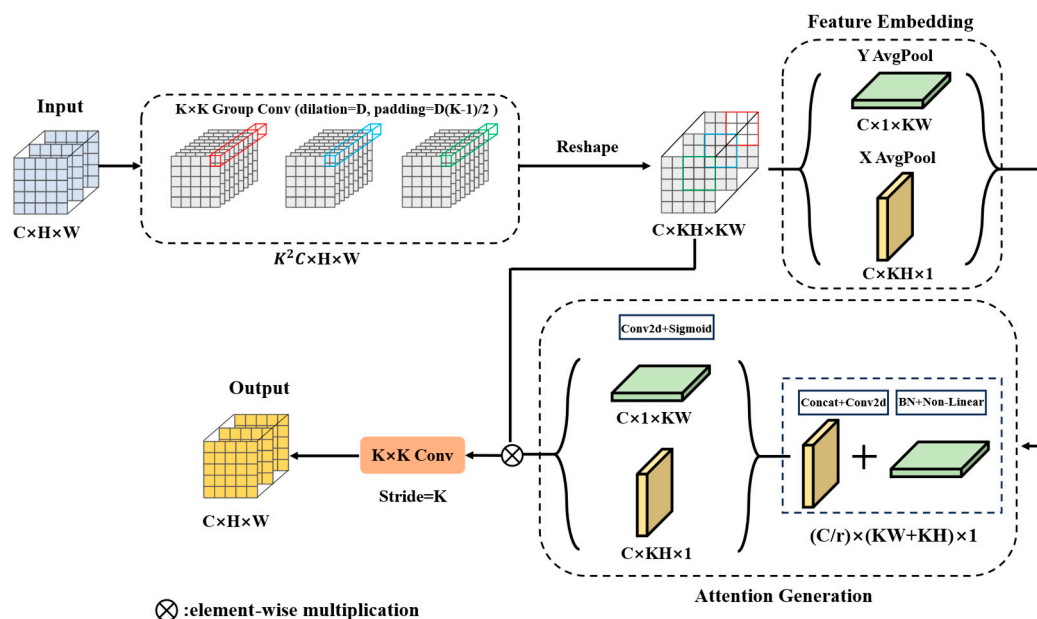


**Figure 5.** Framework of the proposed RFCA-DConv.

The input feature  $X$  is processed by a  $K \times K$  grouped convolution (with dilation rate  $D$  and padding  $D(K-1)/2$ ), generating a feature map of size  $K^2 C \times H \times W$ . After batch normalization and ReLU activation, it is reshaped into receptive field features  $X_{rf}$  with size  $C \times KH \times KW$ . Subsequently,  $X_{rf}$  is fed into the coordinate attention module. As shown in Figure 5, to avoid compressing all spatial information into the channel dimension through global pooling, causing the loss of long-range positional information, the global average

pooling operation is decomposed into horizontal and vertical directions. The calculation formulas are as follows:

$$z_c^h = \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i), z_c^w = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w) \quad (10)$$

where  $W$  and  $H$  denote the width and height of the input feature.  $z_c^h$  represents the mean value of all elements in the  $h$ -th row of the  $c$ -th channel, and  $z_c^w$  represents the mean value of all elements in the  $w$ -th column of the  $c$ -th channel.  $x_c$  denotes the  $c$ -th channel of the input feature.

$$f = \delta \left( \text{Conv} \left( \left[ z^h, z^w \right] \right) \right) \quad (11)$$

where  $[\cdot, \cdot]$  denotes concatenation along the spatial dimension,  $\delta$  represents a nonlinear activation function, and  $f$  is the intermediate feature map obtained by combining the vertically and horizontally pooled features.  $\text{Conv}$  denotes the convolution operation.

$$g^h = \sigma \left( \text{Conv} \left( f^h \right) \right), g^w = \sigma \left( \text{Conv} \left( f^w \right) \right) \quad (12)$$

where  $g^h$  and  $g^w$  are feature tensors with the same number of channels as the input, which are obtained by transforming  $f^h$  and  $f^w$ , respectively. The sigmoid function is denoted by  $\sigma$ , and  $f^h$  and  $f^w$  are two tensors derived by decomposing the intermediate feature map  $f$  along the spatial dimensions.

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (13)$$

where the output of the  $c$ -th channel is denoted as  $y_c$ .

Finally, the CA-weighted feature map is fed into a convolutional layer with a kernel size of  $K \times K$  and a stride of  $K$ , as shown in the equation:

$$O = \text{Conv} \left( M_{CA} \otimes X_{rf} \right) \quad (14)$$

where  $O$  denotes the output feature of RFCA-DConv,  $M_{CA}$  represents the attention weights computed by the coordinate attention module and  $\otimes$  denotes element-wise multiplication.

### 2.1.3. Dual-Branch Fusion Network

In the original DeepLabv3+ network, only the output features from the first stage of the backbone are used as shallow inputs to the decoder. Although these shallow features retain relatively high spatial resolution, their limited semantic information results in insufficient segmentation accuracy for small objects and fine-grained land-cover categories. To address this limitation, this study proposes a Dual-Branch Fusion Network (DFN). The mechanism fuses the outputs from the first three stages of the backbone and uses the fused features as shallow input features for the decoder. The fused features retain high-resolution spatial information while incorporating intermediate-level semantic guidance, enabling the decoder to more accurately capture the boundaries and details of small-scale objects and significantly improving segmentation performance for small targets in high-resolution remote-sensing images.

The architecture of the Dual-Branch Fusion Network is illustrated in Figure 6. We define the input image size to the model as  $3 \times H \times W$ . The output of the first stage of the backbone is used as the input to the shallow feature branch, denoted as  $F_l$ , with 256 channels. The output of the third stage of the backbone is upsampled and passed through a  $1 \times 1$  convolution to adjust the channel dimension to 512. It is then added element-wise to the output of the second stage of the backbone. The fused feature is subsequently

upsampled again and processed by another  $1 \times 1$  convolution to reduce the channel dimension to 256, which serves as the input to the high-level semantic branch, denoted as  $F_h$ . High-level semantic information is leveraged to guide the low-level feature branch to focus on small-target features that are prone to being overlooked. Low-level feature information is employed to reversely modulate the high-level semantic branch, preventing the semantic representations of small-scale targets from being suppressed or diluted. This interaction enhances the attention of high-level semantics toward small targets. The overall process can be formulated as follows:

$$F_{h-l} = \sigma(F_h) \otimes F_l \quad (15)$$

$$F_l^* = F_{l-h} \oplus F_l \quad (16)$$

$$F_{l-h} = \sigma(F_l) \otimes F_h \quad (17)$$

$$F_h^* = F_{h-l} \oplus F_h \quad (18)$$

where  $F_{l-h}$  denotes the guidance information transmitted from the low-level feature branch to the high-level semantic branch, and  $F_h^*$  represents the refined feature of the high-level semantic branch after interaction.  $F_{h-l}$  denotes the guidance information from the high-level semantic branch to the low-level feature branch, and  $F_l^*$  represents the refined feature of the low-level feature branch after interaction.  $\sigma(\cdot)$  denotes the sigmoid activation function. The operator  $\otimes$  indicates element-wise multiplication, while  $\oplus$  denotes element-wise addition.

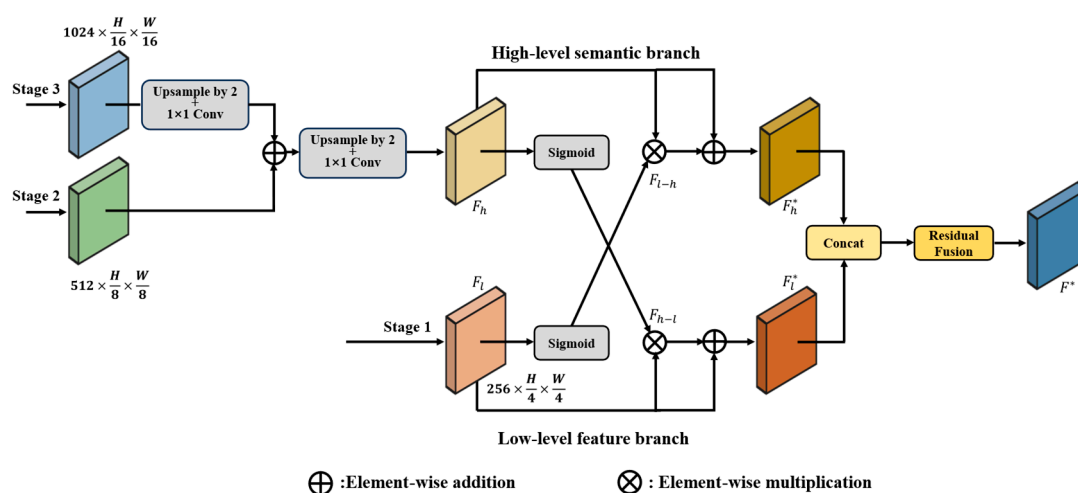


**Figure 6.** Framework of the proposed Dual-Branch Fusion Network.

Subsequently, the two branches are integrated through a residual fusion module, as illustrated in Figure 7. The high-level semantic branch feature  $F_h^*$  and the low-level feature branch feature  $F_l^*$  are first concatenated in the channel dimension. The main fusion pathway consists of two  $3 \times 3$  convolutional layers with padding = 1 and stride = 1 to maintain spatial dimensions, each followed by batch normalization. The first  $3 \times 3$  convolution reduces the channel dimension to 256. This pathway is designed to learn complex nonlinear feature transformations from input to output. The shortcut pathway employs a  $1 \times 1$  convolutional layer to project the input features to 256 channels. The outputs of the main pathway and the shortcut pathway are combined via element-wise addition. The fused features are activated by a ReLU function. The adoption of the residual fusion module effectively alleviates the gradient vanishing problem in deep networks, leading to more stable training and more

comprehensive feature integration, while simultaneously enhancing the representation of texture-related features. The formulation of this process is given as follows:

$$BN(X) = \gamma \frac{X - \mu}{\sqrt{\sigma_B^2 + \epsilon}} + \beta \quad (19)$$

$$X_{main} = BN\left(f^{3 \times 3}\left(ReLU\left(BN\left(f^{3 \times 3}(X)\right)\right)\right)\right) \quad (20)$$

$$O = ReLU\left(X_{main} \oplus X\right) \quad (21)$$

where  $BN$  denotes the batch normalization layer.  $X$  represents the input feature, and  $\mu$  and  $\sigma_B^2$  denote the mean and variance of the batch.  $\epsilon$  is a small constant added for numerical stability. The parameters  $\gamma$  and  $\beta$  are learnable scaling and shifting coefficients.  $X_{main}$  denotes the output feature of the main fusion pathway, and  $O$  represents the output feature.  $f^{3 \times 3}(\cdot)$  denotes a  $3 \times 3$  convolution operation and  $\oplus$  denotes element-wise addition.

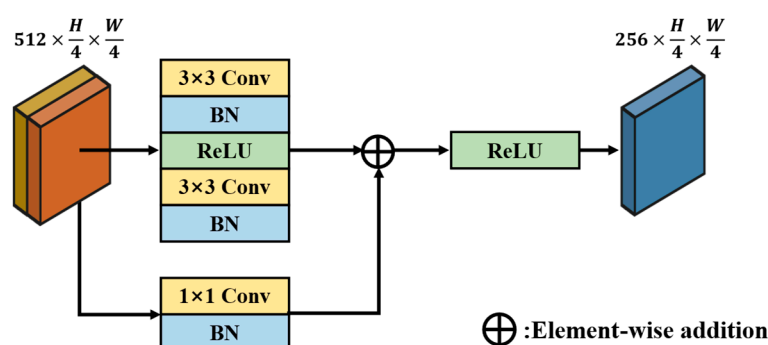


**Figure 7.** Structure of the residual fusion module.

## 2.2. Evaluation Indicators

In this study, the proposed RFA2Net model was comprehensively evaluated in terms of both segmentation accuracy and computational efficiency. To assess segmentation performance, overall accuracy (OA), mean F1-score (mF1) and mean intersection over union (mIoU) were selected as the primary evaluation metrics. OA reflects the proportion of correctly classified pixels across the entire image, providing an overall measure of pixel-level classification performance. mIoU is commonly used in semantic segmentation and is calculated as the average intersection over union between the predicted segmentation and the ground truth. mF1 (mean F1-score) averages the class-wise F1 scores in segmentation. It balances precision and recall for each category, fairly evaluating overall performance. The predicted results are compared with the dataset annotations to obtain four basic quantities: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Based on these values, OA and mIoU can be computed. The formulas used to calculate these metrics are as follows:

$$mIoU = \frac{1}{n} \sum_{i=1}^n \frac{TP}{FN + FP + TP} \quad (22)$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (23)$$

$$mF1 = \frac{1}{n} \sum_{i=1}^n \frac{2 \times TP}{(2 \times TP + FN + FP)} \quad (24)$$

In addition to segmentation accuracy, the computational efficiency and parameter scale of the model are critical for evaluating its practical applicability. In this study, floating

point operations (FLOPs) and the total number of model parameters (Params) are used to quantify these aspects. FLOPs represent the total number of floating point operations required for a single forward pass. It can be used to estimate computational cost and inference complexity. The number of parameters indicates the total learnable weights in the network, providing insight into storage requirements and overall model size. Generally, models with lower FLOPs and fewer parameters exhibit higher computational efficiency and are more amenable to deployment.

### 3. Results

#### 3.1. Dataset

The performance of the proposed RFA2Net model was assessed using two publicly accessible and challenging benchmark datasets: the ISPRS Potsdam dataset and the LoveDA dataset. Both datasets focus on semantic segmentation tasks for high-resolution remote-sensing images, and several representative samples are presented in Figure 8. A detailed overview of the content and specific characteristics of these datasets is provided below.

**ISPRS Potsdam dataset:** The ISPRS Potsdam dataset [40] consists of 38 high-resolution aerial images of the city of Potsdam, Germany, each with a spatial resolution of 0.05 m and dimensions of  $6000 \times 6000$  pixels. In addition to standard RGB images, the dataset also provides Digital Surface Model (DSM) and True Orthophoto (TOP) information, available in three spectral combinations: IR-R-G, R-G-B, and R-G-B-IR. For the purposes of this study, only the RGB channels are utilized, while DSM and other spectral bands are excluded. Due to the large size of the original images, training directly at full resolution would require considerable computational resources. To address this, each image was divided into smaller patches of  $512 \times 512$  pixels, yielding a total of 4598 patches, which were employed for both model training and evaluation.

**LoveDA dataset:** The LoveDA dataset [41] contains high-resolution satellite images from several major cities in China, including Nanjing, Changzhou, and Wuhan. It includes a total of 4191 images, each with a spatial resolution of 0.3 m and dimensions of  $1024 \times 1024$  pixels. The dataset is divided into urban and rural regions. Urban areas are characterized by densely packed manmade structures with diverse colors and generally smaller object sizes, whereas rural areas mainly feature natural landscapes with more uniform tones and larger-scale objects. This variation in object size and appearance introduces significant challenges for model training, making LoveDA an important benchmark for evaluating the generalization, adaptability, and robustness of segmentation methods across different environmental contexts.

For both datasets, images are split into training, validation, and test sets at a ratio of 6:2:2, with the specific quantities and details provided in Table 1. To ensure fairness, the LoveDA dataset was cropped into  $512 \times 512$  patches using only a sliding window approach, and no additional data augmentation methods were applied during model training.

**Table 1.** Detailed information about the ISPRS Potsdam and LoveDA datasets.

| Dataset       | GSD    | DSM   | Classes | Images | Pixels             | Train | Val | Test |
|---------------|--------|-------|---------|--------|--------------------|-------|-----|------|
| ISPRS Potsdam | 0.05 m | 71.48 | 6       | 4598   | $6000 \times 6000$ | 2759  | 920 | 919  |
| LoveDA        | 0.3 m  | 78.23 | 7       | 4191   | $1024 \times 1024$ | 2515  | 838 | 838  |



**Figure 8.** Some samples from the ISPRS Potsdam and LoveDA datasets.

### 3.2. Experiment Details

The details of the experimental platform's hardware and software parameters used for the experiment can be seen in Table 2.

**Table 2.** Experimental platform's environment settings.

| Parameter               | Configuration                        |
|-------------------------|--------------------------------------|
| CPU                     | Intel Xeon Platinum 8352V @ 2.10 GHz |
| GPU                     | NVIDIA GeForce RTX 4090              |
| CUDA version            | CUDA 12.1                            |
| Operating system        | Ubuntu22.04                          |
| Programming language    | Python 3.12.3                        |
| Deep learning framework | Pytorch 2.3.0                        |

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA) and an Intel Xeon Platinum 8352V CPU (Intel Corporation, Santa Clara, CA, USA) operating at 2.10 GHz. All models were implemented based on PyTorch 2.3.0 and trained for 300 epochs using the Adam optimizer with a batch size of 8. An early stopping strategy with a patience of 30 was applied. The initial learning rate was set to  $1 \times 10^{-2}$  for CNN-based models and  $5 \times 10^{-4}$  for Transformer-based models. A cosine annealing schedule was applied to adjust the learning rate throughout training. Method-specific hyperparameters for baseline models were selected according to the recommendations in the original publications or official code repositories.

### 3.3. Baseline Model Description

- **FCN:** FCN [42] replaces fully connected layers with convolutional layers, enabling it to handle input images of varying sizes and providing a fundamental baseline for research in semantic segmentation.
- **PSPNet:** PSPNet [43] integrates a pyramid pooling module to capture global contextual information through multi-scale pooling operations, improving the representation of objects at different scales.
- **HRNet:** HRNet [44] is designed to preserve high-resolution feature representations by fusing multi-resolution features within a parallel multi-branch architecture, maintaining fine spatial details.
- **PIDNet:** PIDNet [45] is a lightweight network that employs parallel high-resolution and low-resolution context branches along with a dual-path fusion module. It enables efficient modeling of global context while preserving fine spatial details.

- **SegFormer:** SegFormer [46] is a Transformer-based semantic segmentation model that employs a hierarchical hybrid Transformer encoder to extract multi-scale features and capture long-range dependencies.
- **BiFormer:** BiFormer [47] incorporates a bi-level attention mechanism that combines local window attention with global sparse attention. This architecture captures long-range dependencies while reducing computational requirements.

### 3.4. Ablation Analysis

To evaluate the individual contributions of the RFA-ResNet backbone, RFA-DASPP structure and DFN structure to the overall performance of the proposed model, a series of ablation experiments were carried out on the ISPRS Potsdam dataset, followed by detailed analysis and validation.

#### 3.4.1. Quantitative Analysis

Table 3 summarizes a series of ablation study results on the ISPRS Potsdam dataset, where all metrics are reported as the averages over five random seeds. The baseline model (Method 1) achieves an mIoU of 75.75%, an mF1 of 85.82% and an OA of 88.39%, with 58.75 M parameters and 79.29 G FLOPs. In Method (2), the ResNet101 backbone is replaced with RFA-ResNet, increasing the mIoU, mF1 and OA to 76.96% (+1.21%), 86.60% (+0.78%) and 89.07% (+0.68%). And it only slightly increases the parameters and FLOPs by 1.07 M and 1.53 G. This indicates that RFA-ResNet enhances the model's ability to extract complex structural features with only a minor computational overhead. Method (3) replaces the original ASPP with the RFA-DASPP structure, improving the mIoU, mF1 and OA to 77.42% (+1.67%), 86.89% (+1.07%) and 89.39% (+1.00%). This demonstrates that the RFA-DASPP structure can reduce information loss and enhance the model's sensitivity to directional structures. In Method (4), a Dual-Branch Fusion Network (DFN) is employed to process the output features from the first three stages of the backbone, resulting in a final mIoU of 76.94% (+1.19%), an mF1 of 86.59% (+0.77%) and an OA of 89.02% (+0.63%). This suggests that the approach effectively leverages shallow features to improve segmentation accuracy. The complete RFA2Net model (Method 8) achieves a 3.19% improvement in mIoU, a 1.89% improvement in mF1 and a 1.58% improvement in OA compared to the original DeepLabv3+. These quantitative results demonstrate that the proposed improvements are effective not only when applied individually but also exhibit complementary benefits when integrated, jointly enhancing the model's capability for high-resolution remote-sensing image segmentation.

**Table 3.** Ablation experiment of the proposed models on the ISPRS Potsdam dataset (mean  $\pm$  std over five seeds).

| Method | RFA-ResNet | RFA-DASPP | DFN | mIoU (%)         | mF1 (%)          | OA (%)           | Params (M) | FLOPs (G) |
|--------|------------|-----------|-----|------------------|------------------|------------------|------------|-----------|
| (1)    |            |           |     | 75.75 $\pm$ 0.27 | 85.82 $\pm$ 0.32 | 88.39 $\pm$ 0.31 | 58.75      | 79.29     |
| (2)    | ✓          |           |     | 76.96 $\pm$ 0.88 | 86.60 $\pm$ 0.92 | 89.07 $\pm$ 1.04 | 59.82      | 80.82     |
| (3)    |            | ✓         |     | 77.42 $\pm$ 0.31 | 86.89 $\pm$ 0.71 | 89.39 $\pm$ 0.35 | 61.43      | 81.94     |
| (4)    |            |           | ✓   | 76.94 $\pm$ 0.18 | 86.59 $\pm$ 0.45 | 89.02 $\pm$ 0.21 | 59.32      | 80.26     |
| (5)    | ✓          | ✓         |     | 78.13 $\pm$ 0.79 | 87.37 $\pm$ 0.70 | 89.73 $\pm$ 0.82 | 62.51      | 84.93     |
| (6)    | ✓          |           | ✓   | 78.09 $\pm$ 0.49 | 87.29 $\pm$ 0.37 | 89.67 $\pm$ 0.69 | 60.39      | 82.86     |
| (7)    |            | ✓         | ✓   | 78.26 $\pm$ 0.26 | 87.54 $\pm$ 0.35 | 89.79 $\pm$ 0.32 | 62.01      | 84.50     |
| (8)    | ✓          | ✓         | ✓   | 78.94 $\pm$ 0.37 | 87.71 $\pm$ 0.23 | 89.97 $\pm$ 0.39 | 63.07      | 85.56     |

Note: The symbol “✓” indicates that the corresponding module is added.

#### 3.4.2. Qualitative Analysis

In Section 3.4.1, the quantitative impact of each optimization strategy on segmentation performance was evaluated using standard metrics. To further investigate the individual

effects of these strategies, this subsection presents a visual comparison of segmentation results produced by the baseline model and models incorporating each optimization independently on the ISPRS Potsdam dataset. For rigor and fairness, all test images used in this evaluation were excluded from the training set.

Figure 9 illustrates the segmentation results of the baseline model and the DeepLabv3+ model with the backbone replaced by RFA-ResNet. In the first row, the RFA-ResNet backbone achieves more accurate segmentation of low vegetation boundaries around vehicles, effectively avoiding misclassification as clutter. As shown in the second row, the improved model is able to better distinguish low vegetation from the background while more precisely delineating the complex boundaries of clutter regions. In the third row, misclassifications within building areas are substantially reduced. The fourth row demonstrates a clear improvement in boundary delineation within complex regions. These results indicate that incorporating the RFA-ResNet backbone enables feature extraction to be guided jointly by local receptive field attention and global semantic context, significantly improving segmentation consistency and facilitating accurate delineation of complex boundaries in high-resolution remote-sensing images.

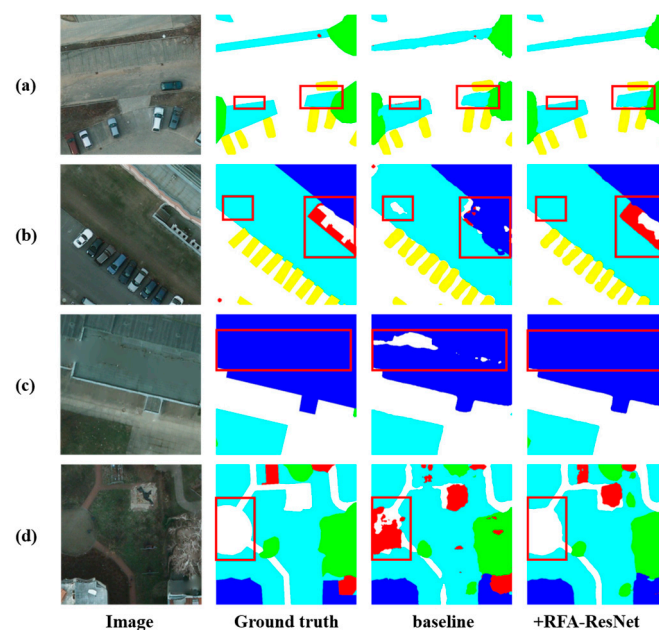


**Figure 9.** Comparison of the baseline before and after using RFA-ResNet. (a–d) denote different examples, with no specific classification criteria in their selection.

Figure 10 presents the segmentation results obtained by integrating only the RFA-DASPP structure into the model. In the first row, the introduction of RFA-DASPP reduces misclassification at the boundaries between different land-cover types and prevents low vegetation with distinct color characteristics from being incorrectly labeled as buildings. In the second row, the baseline model shows weak performance in segmenting large buildings and fails to successfully identify small cluttered objects. After incorporating the RFA-DASPP structure, these issues are effectively alleviated. These visual results demonstrate that RFA-DASPP promotes continuous multi-scale feature propagation through dense connections, minimizing feature loss and enhancing the model's ability to segment objects at different scales. In addition, the RFA-DASPP module significantly enlarges the receptive field, which helps to reduce local misclassification and improves the continuity of boundary segmentation for large targets. Meanwhile, the RFCA-DConv module within RFA-DASPP enhances the network's capability to segment oriented structures and boundary regions. In the third row, the improved model achieves more accurate segmenta-

tion of slender, directionally structured cluttered objects. In the fourth row, the enhanced model better delineates the long, oriented boundaries of large buildings while avoiding local misclassification.

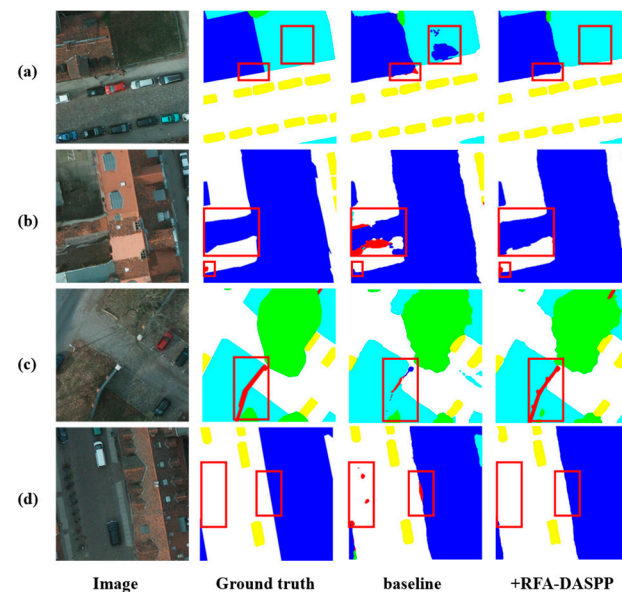


**Figure 10.** Comparison of the baseline before and after using the RFA-DASPP. (a–d) denote different examples, with no specific classification criteria in their selection.

The DFN structure improves the decoder’s capability to capture fine details and boundaries of small objects by integrating high-level semantic information with the high-resolution, low-level features used as decoder inputs. This fusion notably enhances the detection and segmentation of small-scale targets. Figure 11 shows the segmentation results when only the DFN is added to the model. In the first row, the enhanced model successfully segments very small clutter objects. The second row illustrates that small-targets, such as vehicles, are segmented more accurately, with clearer boundaries between adjacent objects. In the third row, the improved model correctly identifies a car partially occluded by trees, while the baseline model incorrectly classifies it as part of the tree category.

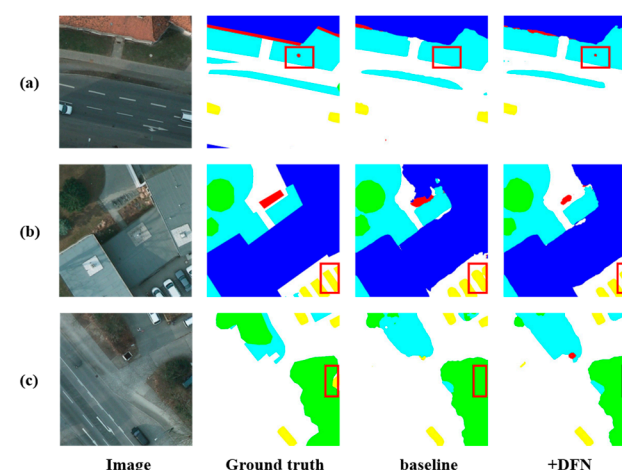


**Figure 11.** Comparison of the baseline before and after using the DFN. (a–c) denote different examples, with no specific classification criteria in their selection.

To further investigate the underlying mechanisms of the three key modules in RFA2Net, we conduct a qualitative analysis of the attention distribution through visualization. Specifically, we employ Grad-CAM [48] to compare the baseline model with its

variants obtained by individually incorporating the three proposed modules. Grad-CAM utilizes gradient information to compute the importance weights of feature maps and projects them back onto the input image to generate heatmaps, where colors from blue to red indicate increasing levels of attention.

As shown in the third row of Figure 12, after replacing the backbone of the baseline model with RFA-ResNet, the model exhibits significantly enhanced responses over large-area objects such as buildings and forests (columns **a** and **e**), indicating stronger attention compared to the baseline. This suggests that RFA-ResNet effectively improves the model's ability to capture global contextual information. Meanwhile, for regions with complex distributions such as low vegetation (column **b**), the model produces more precise activations, demonstrating improved representation capability for complex features. From the fourth row (columns **c** and **d**), it can be observed that after introducing the RFA-DASPP module, the model not only produces more concentrated responses over large-scale directional impervious surfaces, but also shows increased attention to small objects such as cars. This indicates that RFA-DASPP effectively alleviates feature loss during multi-scale feature fusion, while enhancing the model's sensitivity to targets with long-range structural characteristics. As shown in the fifth row (column **d**), the incorporation of the DFN module further strengthens the model's response to small-scale objects such as cars, suggesting that this module plays a positive role in improving small-object segmentation performance.

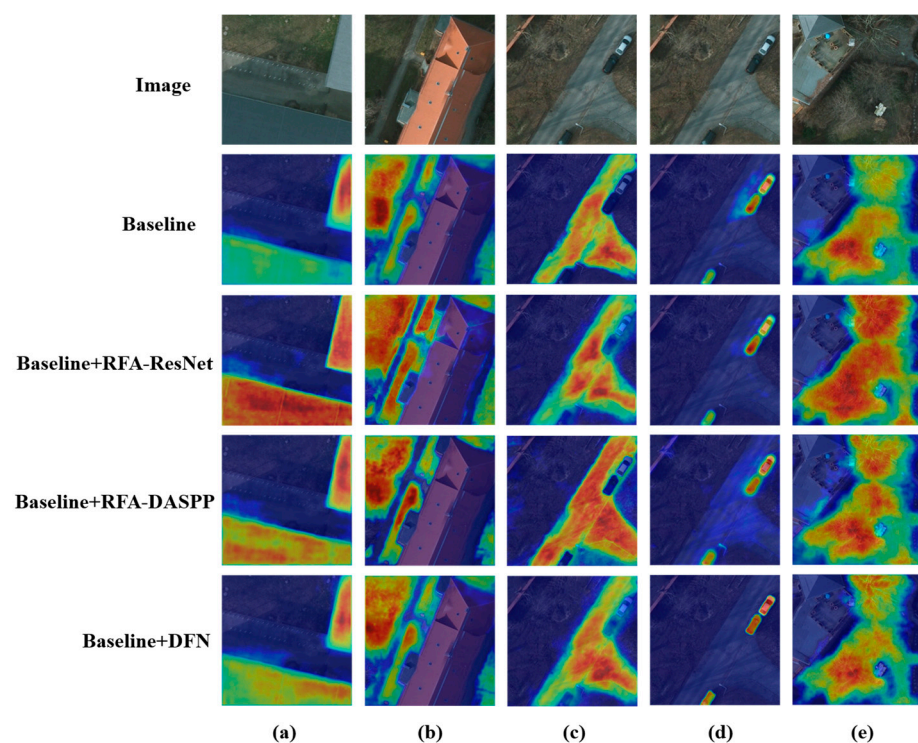


**Figure 12.** Visualization of heatmap for RFA-ResNet, RFA-DASPP, and DFN integrated into the baseline model. (a) Building. (b) Low vegetation. (c) Impervious surface. (d) Car. (e) Tree.

### 3.5. Comparative Analysis

To validate the superiority of the proposed RFA2Net, we compared it with several existing semantic segmentation methods. The experimental results on the Potsdam and LoveDA datasets are presented in Tables 4 and 5, respectively. The proposed RFA2Net achieves the highest mIoU and overall accuracy on both datasets, demonstrating its superior segmentation performance.

**Table 4.** Comparison of segmentation results on the ISPRS Potsdam dataset. mIoU, mF1, OA and class-wise IoU (%) are listed (mean  $\pm$  std over five seeds).

| Model       | mIoU                    | mF1                     | OA                      | Imp. Surf.              | Building                | Low Veg.                | Tree                    | Car                     | Back./Clu.              |
|-------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| FCN         | 72.70 $\pm$ 0.09        | 84.22 $\pm$ 0.51        | 87.33 $\pm$ 0.13        | 80.10 $\pm$ 0.44        | 89.63 $\pm$ 0.61        | 72.07 $\pm$ 0.88        | 71.82 $\pm$ 1.50        | 69.91 $\pm$ 0.34        | 52.94 $\pm$ 0.82        |
| PSPNet      | 74.63 $\pm$ 0.08        | 85.2 $\pm$ 0.11         | 88.11 $\pm$ 0.10        | 81.64 $\pm$ 0.81        | 90.15 $\pm$ 0.29        | 72.61 $\pm$ 0.11        | 72.94 $\pm$ 1.02        | 74.08 $\pm$ 0.46        | 56.35 $\pm$ 0.96        |
| HRNet       | 75.35 $\pm$ 0.11        | 85.59 $\pm$ 0.42        | 88.02 $\pm$ 0.15        | 81.32 $\pm$ 0.12        | 90.14 $\pm$ 0.65        | 72.63 $\pm$ 0.06        | 72.50 $\pm$ 2.95        | 80.24 $\pm$ 0.15        | 55.25 $\pm$ 1.54        |
| PIDNet      | 76.04 $\pm$ 0.10        | 86.04 $\pm$ 0.39        | 88.62 $\pm$ 0.14        | 82.29 $\pm$ 0.18        | 89.95 $\pm$ 0.44        | 74.64 $\pm$ 0.06        | 73.20 $\pm$ 1.24        | 78.59 $\pm$ 0.05        | 57.61 $\pm$ 1.47        |
| SegFormer   | 75.77 $\pm$ 0.57        | 85.63 $\pm$ 2.47        | 88.67 $\pm$ 0.77        | 82.53 $\pm$ 0.45        | 90.03 $\pm$ 2.69        | <u>74.90</u> $\pm$ 3.78 | <u>73.75</u> $\pm$ 1.93 | 79.07 $\pm$ 1.12        | 54.37 $\pm$ 2.74        |
| BiFormer    | 76.12 $\pm$ 0.13        | 86.53 $\pm$ 0.42        | 88.68 $\pm$ 1.07        | 82.72 $\pm$ 0.24        | 90.92 $\pm$ 1.08        | 73.99 $\pm$ 0.08        | 72.82 $\pm$ 2.10        | 79.18 $\pm$ 0.41        | 57.11 $\pm$ 1.18        |
| Mask2Former | 75.39 $\pm$ 0.75        | 85.61 $\pm$ 0.97        | 88.15 $\pm$ 0.89        | 81.61 $\pm$ 1.69        | 89.83 $\pm$ 1.31        | 73.28 $\pm$ 3.09        | 72.18 $\pm$ 4.21        | 77.36 $\pm$ 2.62        | 58.05 $\pm$ 1.66        |
| DeepLabv3+  | 75.75 $\pm$ 0.27        | 85.82 $\pm$ 0.32        | 88.39 $\pm$ 0.31        | 82.10 $\pm$ 0.75        | 90.08 $\pm$ 0.87        | 73.86 $\pm$ 1.28        | 72.52 $\pm$ 1.32        | 78.32 $\pm$ 0.13        | 57.61 $\pm$ 1.82        |
| RFA2Net     | <b>78.94</b> $\pm$ 0.37 | <b>87.71</b> $\pm$ 0.23 | <b>89.97</b> $\pm$ 0.39 | <b>84.37</b> $\pm$ 0.96 | <b>92.32</b> $\pm$ 0.43 | <b>76.31</b> $\pm$ 2.33 | <b>74.41</b> $\pm$ 2.76 | <b>83.81</b> $\pm$ 0.54 | <b>62.41</b> $\pm$ 1.47 |

Note: The best results are highlighted in bold, and the second-best results are underlined.

**Table 5.** Comparison of segmentation results on the LoveDA dataset. mIoU, mF1, OA and class-wise IoU are listed (mean  $\pm$  std over five seeds).

| Model       | mIoU                    | mF1                     | OA                      | Back.                   | Build.                  | Road                    | Water                   | Barren                  | Forest                  | Agri.                   |
|-------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| FCN         | 52.52 $\pm$ 0.59        | 65.17 $\pm$ 1.22        | 70.01 $\pm$ 0.89        | 46.00 $\pm$ 1.67        | 55.69 $\pm$ 1.11        | 54.95 $\pm$ 0.85        | 70.79 $\pm$ 1.18        | 35.51 $\pm$ 1.15        | 45.27 $\pm$ 1.68        | 59.46 $\pm$ 1.30        |
| PSPNet      | 55.71 $\pm$ 0.75        | 67.11 $\pm$ 1.58        | 72.58 $\pm$ 1.33        | 46.19 $\pm$ 2.29        | 60.28 $\pm$ 0.22        | 57.82 $\pm$ 1.01        | 72.82 $\pm$ 0.56        | 39.10 $\pm$ 1.04        | 50.35 $\pm$ 0.82        | 63.40 $\pm$ 0.04        |
| HRNet       | 56.09 $\pm$ 3.44        | 67.73 $\pm$ 3.48        | 71.96 $\pm$ 3.29        | 47.11 $\pm$ 4.63        | 60.71 $\pm$ 1.74        | 58.83 $\pm$ 2.02        | 72.13 $\pm$ 1.88        | 41.78 $\pm$ 1.30        | 52.58 $\pm$ 2.53        | 59.52 $\pm$ 0.02        |
| PIDNet      | 55.57 $\pm$ 1.25        | 66.87 $\pm$ 1.05        | 70.96 $\pm$ 0.89        | 48.21 $\pm$ 0.92        | 62.90 $\pm$ 0.53        | 58.04 $\pm$ 0.18        | 71.45 $\pm$ 0.02        | 43.95 $\pm$ 1.65        | <u>52.60</u> $\pm$ 0.85 | 51.85 $\pm$ 3.03        |
| SegFormer   | 56.65 $\pm$ 3.41        | 68.33 $\pm$ 2.93        | 73.06 $\pm$ 3.32        | 48.91 $\pm$ 6.74        | 63.19 $\pm$ 3.05        | 58.26 $\pm$ 4.66        | 70.89 $\pm$ 2.94        | 44.34 $\pm$ 4.99        | 51.91 $\pm$ 6.83        | 59.03 $\pm$ 5.93        |
| BiFormer    | 57.86 $\pm$ 3.84        | 69.78 $\pm$ 1.60        | 73.62 $\pm$ 3.95        | 50.00 $\pm$ 6.32        | 63.80 $\pm$ 4.73        | 60.39 $\pm$ 4.09        | 71.56 $\pm$ 6.31        | <b>46.37</b> $\pm$ 4.68 | 51.41 $\pm$ 5.44        | 61.52 $\pm$ 3.91        |
| Mask2Former | 56.57 $\pm$ 3.47        | 68.03 $\pm$ 3.31        | 73.21 $\pm$ 2.86        | 47.42 $\pm$ 3.94        | 60.98 $\pm$ 4.89        | 60.70 $\pm$ 3.73        | 74.07 $\pm$ 1.50        | 38.45 $\pm$ 6.77        | 50.83 $\pm$ 4.36        | <b>63.53</b> $\pm$ 2.43 |
| DeepLabv3+  | 56.38 $\pm$ 1.12        | 67.57 $\pm$ 2.03        | 72.11 $\pm$ 1.29        | 48.00 $\pm$ 1.17        | 62.90 $\pm$ 1.04        | 59.12 $\pm$ 1.73        | 70.19 $\pm$ 2.63        | 45.45 $\pm$ 3.05        | 50.84 $\pm$ 2.39        | 58.18 $\pm$ 0.75        |
| RFA2Net     | <b>59.46</b> $\pm$ 1.80 | <b>70.17</b> $\pm$ 2.25 | <b>74.63</b> $\pm$ 1.59 | <b>50.48</b> $\pm$ 4.02 | <b>65.04</b> $\pm$ 2.47 | <b>62.56</b> $\pm$ 2.87 | <b>75.51</b> $\pm$ 2.14 | <u>45.92</u> $\pm$ 2.17 | <b>53.78</b> $\pm$ 3.21 | 62.97 $\pm$ 0.39        |

Note: The best results are highlighted in bold, and the second-best results are underlined.

### 3.5.1. Comparative Results on the ISPRS Potsdam Dataset

Table 4 reports the performance of all baseline models and the proposed RFA2Net on the Potsdam dataset, where all metrics are reported as averages over five random seeds. By leveraging the ASPP module to enlarge the receptive field and capture multi-scale contextual information, DeepLabv3+ outperforms CNN-based models such as FCN, PSPNet, and HRNet, as well as the Transformer-based Mask2Former, achieving an mIoU of 75.75%. Among the CNN series, PIDNet achieves the best performance. The BiFormer model, which incorporates a bi-level routing attention mechanism, attains an mIoU of 76.12%, an mF1 of 86.53% and an OA of 88.68%, making it the best baseline among the baseline methods. The proposed RFA2Net surpasses all baseline models, achieving an mIoU of 78.94%, an mF1 of 87.71% and an OA of 89.97%. Compared with the original DeepLabv3+, this corresponds to improvements of 3.19% in mIoU, 1.89% in mF1 and 1.58% in OA, while gains of 2.82%, 1.18% and 1.29% are observed over the second best BiFormer model. Notably, RFA2Net achieves the highest IoU across all object categories, with particularly significant improvements in the car class and the background or clutter class, increasing by 5.49% and 4.80%, respectively. It can be seen from Figure 13a that for categories with a small proportion in all classes such as car and clutter, although RFA2Net does not achieve IoU scores comparable to those of high-proportion categories like building, it does not suffer from very severe class imbalance. Meanwhile, RFA2Net achieves small standard deviations across all metrics, demonstrating excellent robustness. These results quantitatively demonstrate that the proposed RFA2Net achieves substantial performance gains in segmenting small objects and multi-scale targets.

A further qualitative analysis of the segmentation results is conducted, as shown in Figure 14. In the fourth group of Figure 14, other baseline methods tend to misclassify pixels within impervious surfaces as background or clutter. But RFA2Net accurately segments these regions, demonstrating strong intra-class consistency. RFA2Net also provides precise boundary delineation for large objects. In the first and fourth groups, Transformer-based models benefiting from global context modeling are able to produce relatively complete segmentations of long boundaries and complex edges of large buildings. In the fourth group, RFA2Net achieves a comparable performance in capturing complex boundaries,

while in the first group, it outperforms these Transformer-based models in segmenting long building boundaries, highlighting its capability to capture long-range dependencies. In the second group, Transformer-based models do not exhibit clear advantages over CNN-based models in small-object segmentation and none of these methods are able to precisely delineate small targets. In contrast, RFA2Net successfully segments small clutter objects within low-vegetation areas, reflecting its strong ability to extract complex features. As illustrated in the third group of Figure 14, RFA2Net can accurately distinguish visually similar regions of low vegetation and trees. Figure 15 presents the segmentation results of RFA2Net on the ISPRS Potsdam dataset using the whole-image size setting ( $6000 \times 6000$  pixels). In such remote-sensing images, categories such as clutter and vehicles are distributed over extensive regions, posing significant challenges in accurate segmentation. Despite the small spatial scale of these objects, RFA2Net achieves relatively complete and precise segmentation. It also demonstrates superior performance in segmenting large target regions. These results validate the excellent scalability of RFA2Net for very large remote-sensing image.

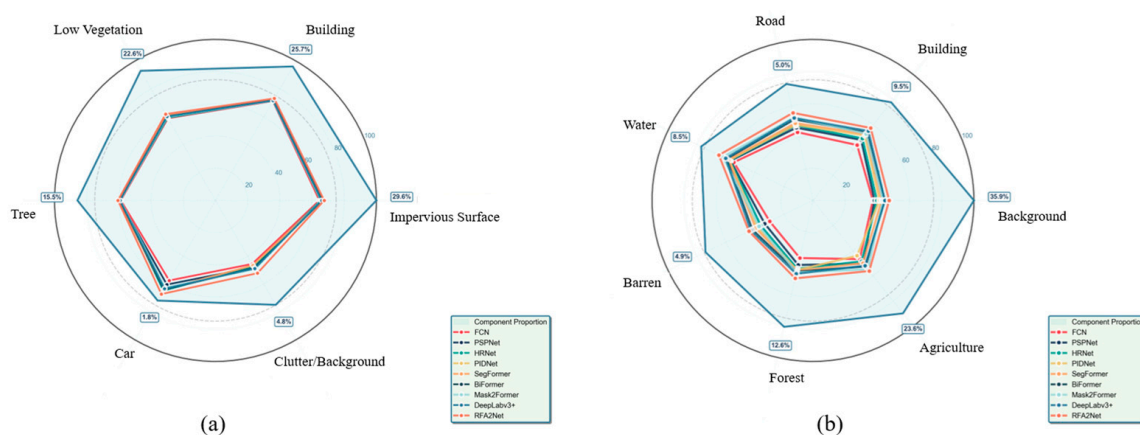


**Figure 13.** Comparison of class-wise IoU between the comparative models. (a) ISPRS Potsdam dataset. (b) LoveDA dataset.

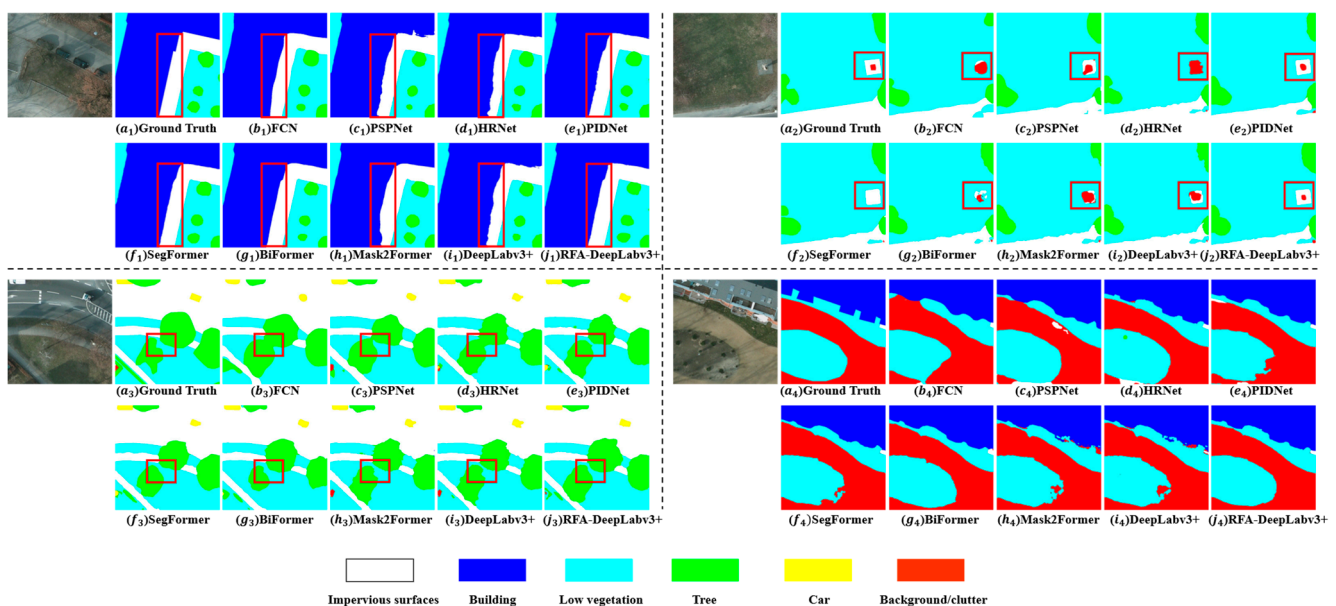


**Figure 14.** Visualization of segmentation on the ISPRS Potsdam dataset.

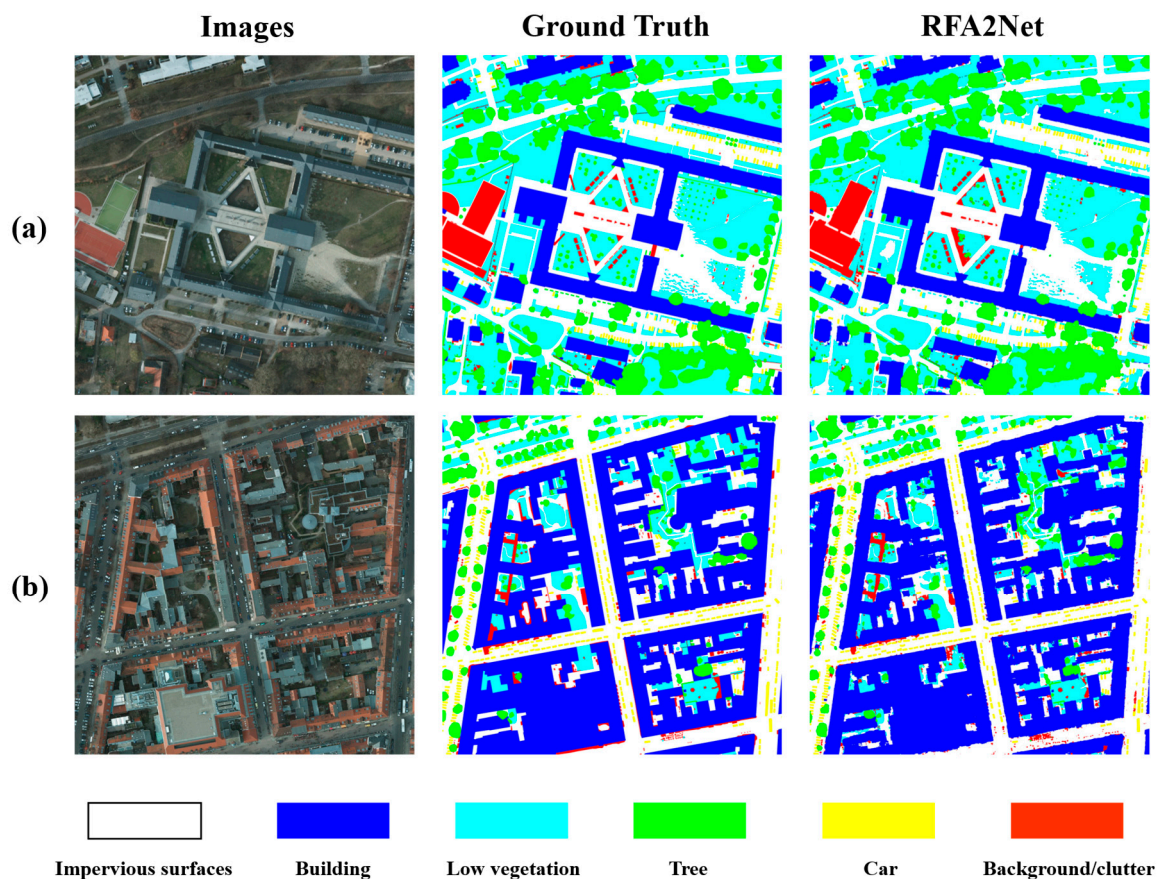


**Figure 15.** Visualization of segmentation on the ISPRS Potsdam dataset with the whole-image size setting. (a,b) denote different examples, with no specific classification criteria in their selection.

### 3.5.2. Comparative Results on the LoveDA Dataset

Table 5 presents the performance comparison of all baseline models and the proposed RFA2Net on the LoveDA dataset. Compared with the Potsdam dataset, LoveDA has a larger ground sampling distance (GSD) and more diverse and complex scene types, which increases the difficulty of the semantic segmentation task. RFA2Net achieves the best segmentation performance on LoveDA despite these challenges, with a mIoU of 59.46%, an mF1 of 71.17% and an OA of 74.63%. In this complex dataset, Transformer-based models generally outperform CNN-based models, especially in large scale object categories such as agriculture and water bodies. Among the compared methods, BiFormer achieves the second best performance, with mIoU, mF1 and OA of 57.86%, 69.78% and 73.62%, respectively. Compared with the original DeepLabv3+, the proposed RFA2Net improves mIoU, mF1 and OA by 3.08%, 2.60% and 2.52%, respectively. Notably, the IoU improvements are most significant in the water and agriculture categories, with gains of 5.32% and 4.79%, respectively. As shown in Figure 13b, RFA2Net achieves highly competitive IoU scores on minority classes such as water and road. Although it delivers relatively inferior segmentation performance on sparse and infrequent classes like barren, no significant class imbalance is observed. These results indicate that the proposed improvements effectively enhance the model's global perception of large-scale object structures, thereby improving overall segmentation performance in complex high-resolution remote-sensing scenes.

Due to the severe domain shift between urban and rural scenarios and the highly heterogeneous visual features in the LoveDA dataset, experimental fluctuations are intensified and the standard deviations of all comparative models increase to a certain extent. We observe that Transformer-based models exhibit larger standard deviations, which may be attributed to their weaker inductive bias and higher sensitivity to data distribution shifts.

Although the proposed RFA2Net yields a slightly larger standard deviation compared with DeepLabv3+, it remains within an acceptable range, demonstrating the strong robustness of RFA2Net.

Figure 16 presents the semantic segmentation results of all models on the LoveDA dataset. In the first group of examples, due to the relatively large ground sampling distance (GSD), the background regions between water occupy only a small number of pixels. Except for BiFormer and RFA2Net, the other models fail to correctly segment these narrow regions. In the second group, RFA2Net achieves more accurate segmentation in densely built-up areas, precisely delineating building boundaries and their spatial layouts. This demonstrates the model's strong capability in local feature extraction and multi-scale feature representation. In the third and fourth groups, models with strong global context modeling ability, such as BiFormer and SegFormer, perform better in segmenting large and spatially continuous farmland regions. Benefiting from the enhanced backbone and the RFA-DASPP structure, RFA2Net produces more accurate segmentation results in these extensive farmland areas than several Transformer-based models, highlighting its effective global feature modeling capability.

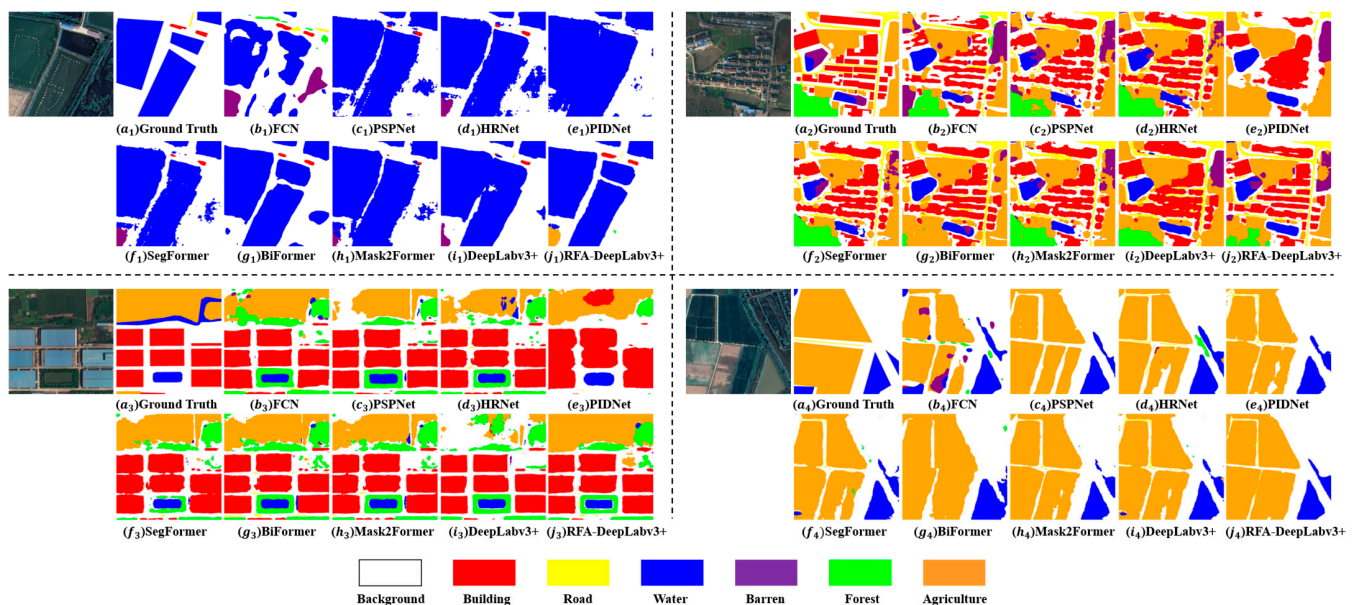


**Figure 16.** Visualization of segmentation on the LoveDA dataset.

### 3.6. Efficiency Analysis

In this subsection, we evaluate the model size and computational cost using two metrics, namely the number of parameters (Params) and floating-point operations (FLOPs). For all models, FLOPs are computed based on an input size of  $512 \times 512 \times 3$ . As shown in Table 6 compared with the original DeepLabv3+, RFA2Net increases the Params and FLOPs by only 4.32 M and 6.36 G, respectively, indicating a relatively modest computational overhead. Benefiting from its functionally decoupled multi-branch design, PIDNet records the smallest number of parameters and FLOPs among all evaluated models, indicating high computational efficiency. In contrast, Transformer-based models rely on extensive global token interactions, which leads to substantially higher computational cost in terms of FLOPs. Overall, these results indicate that RFA2Net achieves a favorable balance between segmentation accuracy on high-resolution remote-sensing images and computational efficiency, maintaining a moderate model size while delivering significant performance improvements.

**Table 6.** Comparison with other networks on FLOPs and Params.

| Model       | Params (M)   | FLOPs (G)    |
|-------------|--------------|--------------|
| FCN         | 68.48        | 153.07       |
| PSPNet      | 46.71        | 118.43       |
| HRNet       | <b>29.54</b> | 90.99        |
| PIDNet      | 37.31        | <b>34.46</b> |
| SegFormer   | 85.80        | 183.30       |
| BiFormer    | 55.22        | 231.27       |
| Mask2Former | 63.00        | 85.61        |
| DeepLabv3+  | 58.75        | 79.20        |
| RFA2Net     | 63.07        | 85.56        |

Note: The best results are highlighted in bold.

#### 4. Discussion

High-resolution remote-sensing image semantic segmentation is a key technology for extracting detailed surface information. It plays a crucial role in applications such as urban planning, land-cover mapping, ecological monitoring, and disaster assessment. However, it also presents significant challenges, including large variations in object scales, complex boundaries, and highly heterogeneous scene semantics. To address these issues, this study proposes an improved model RFA2Net, which introduces systematic enhancements at four aspects: backbone feature extraction, long-range dependency modeling, multi-scale feature fusion, and shallow feature utilization. The proposed model achieves significant performance improvements on two representative high-resolution datasets, ISPRS Potsdam and LoveDA.

RFAConv, proposed by Zhang et al. [37], achieves more refined feature extraction by introducing receptive field attention. This attention mechanism is generated only through interactions within the receptive field and between corresponding positions of adjacent channels within the same receptive field. As a result, its ability to model long-range dependencies remains limited. To address this issue, this paper proposes RFCSA-Conv. Building upon the original receptive field attention mechanism, channel attention and spatial attention are incorporated, enabling the convolution kernel weights to be dynamically reweighted under the joint guidance of receptive field attention and global dependency information. By integrating this module into the backbone, the model's capability to extract complex features is enhanced, while effectively alleviating the inherent limitation of CNNs in modeling global dependencies.

Furthermore, the proposed RFA-DASPP structure further validates the importance of continuous multi-scale information propagation. Compared with the parallel ASPP structure, the Dense ASPP structure of RFA-DASPP establishes a continuous information flow among features at different scales, effectively alleviating the problem of information loss during multi-scale feature fusion and further enlarging the receptive field. Combined with the improved RFCA-DConv module, which integrates receptive field attention and coordinate attention during feature extraction, the model becomes more sensitive to positional information and directional structures. This significantly enhances the segmentation accuracy of directional structures, such as rivers and building boundaries.

Finally, a DFN is introduced to fuse shallow features with high-level semantic information and serve as the shallow input to decoder. This ensures that the input features contain both high-resolution local detail information and high-level semantic information. Guided by semantic information, the decoder can more accurately recognize local details, thereby improving its ability to segment small objects. The ablation study results demonstrate that each improved module independently contributes to performance gains and exhibits complementary effects when used jointly. The increases in the number of param-

ters and FLOPs are relatively limited, indicating that the performance improvements do not stem from model scale expansion, but from the effectiveness of the structural design.

In comparative experiments against a range of CNN-based and Transformer-based semantic segmentation methods, RFA2Net consistently achieves the best performance on both the LoveDA and ISPRS Potsdam datasets. Compared with Transformer-based models, the proposed method demonstrates a clear advantage in extracting complex features while retaining strong global context modeling capability, with lower computational cost and inference overhead. In contrast to CNN-based approaches, it exhibits superior multi-scale feature representation and stronger adaptability to complex scene structures. These results indicate that through the effective integration of attention mechanisms, multi-scale contextual modeling and improved convolutional operations, CNN-based architectures still hold substantial potential for practical applications in high-resolution remote-sensing semantic segmentation.

Although RFA2Net demonstrates strong segmentation performance, several limitations remain. First, the model employs a relatively deep backbone, which may restrict inference speed and efficiency in real time scenarios. Moreover, the current experiments are mainly conducted on RGB remote-sensing images, without fully exploiting auxiliary information such as multispectral channels or digital surface models (DSMs). As a result, there is still room for further improvement in the model's ability to represent complex terrain characteristics and three-dimensional structures.

Future research can proceed along several directions. First, more lightweight designs of RFCSA-Conv can be explored to improve the model's applicability in practical engineering deployments. Second, the collaborative modeling between RFCSA-Conv and Transformer modules warrants further investigation to enhance the representation of long-range dependencies. Third, the proposed approach can be extended to multi-modal remote-sensing semantic segmentation by integrating multispectral and elevation information, aiming to improve generalization and robustness in complex scenarios.

## 5. Conclusions

To address the challenges of large-scale variations, complex spatial structures and the difficulty of capturing long-range dependencies in high-resolution remote-sensing semantic segmentation, this study proposes an enhanced model RFA2Net based on DeepLabv3+. This study proposes a novel RFCSA-Conv module and integrates it into the backbone of the DeepLabv3+ framework, forming RFA-ResNet. This design effectively combines local adaptive feature extraction with global contextual attention. And the RFA-DASPP structure is designed to strengthen multi-scale feature fusion and improve the segmentation performance of directional structures and complex boundaries. In addition, a Dual-Branch Fusion Network is proposed to enhance the model's capability of segmenting small objects.

Extensive experiments conducted on the ISPRS Potsdam and LoveDA datasets demonstrate that RFA2Net outperforms a variety of CNN-based and Transformer-based methods in terms of mean intersection over union (mIoU), mean F1-score (mF1) and overall accuracy (OA). The proposed model shows clear advantages in segmenting small objects, complex boundaries, and large continuous regions. Ablation studies further verify the effectiveness and complementarity of each optimization strategy, while evaluations on specific scene subsets highlight the robustness and generalization capability of the model.

RFA2Net achieves a strong trade-off between segmentation accuracy and computational efficiency, providing a practical, high-performance solution for high-resolution remote-sensing semantic segmentation and offering insights for the design of efficient models tailored to complex remote-sensing environments.

**Author Contributions:** Conceptualization, X.Z. and J.L.; Methodology, X.Z., J.L. and G.Z.; Validation, X.Z.; Investigation, X.Z., J.L., Y.M., Y.Z. and G.Z.; Resources, X.Z., J.L., Y.M., Y.Z. and G.Z.; Data curation, X.Z., J.L., Y.M., Y.Z. and G.Z.; Writing—original draft preparation, X.Z. and J.L.; Writing—review and editing, Y.Z. and G.Z.; Visualization, X.Z.; Supervision, Y.Z. and G.Z.; Project administration, Y.Z. and G.Z.; Funding acquisition, Y.Z. and G.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the National Science Foundation for Young Scientists of China (Grant No. 12301131).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data and the code of this study are available from the corresponding author upon request.

**Conflicts of Interest:** There are no conflicts of interest for this article.

## References

- Huang, M.; Dai, W.; Yan, W.; Wang, J. High-Resolution Remote Sensing Image Segmentation Algorithm Based on Improved Feature Extraction and Hybrid Attention Mechanism. *Electronics* **2023**, *12*, 3660. [\[CrossRef\]](#)
- Singh, N.J.; Nongmeikapam, K. Semantic Segmentation of Satellite Images Using Deep-Unet. *Arab. J. Sci. Eng.* **2023**, *48*, 1193–1205. [\[CrossRef\]](#)
- Tokarczyk, P.; Wegner, J.D.; Walk, S.; Schindler, K. Features, Color Spaces, and Boosting: New Insights on Semantic Classification of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 280–295. [\[CrossRef\]](#)
- Tang, Y.; Zhang, L. Urban Change Analysis with Multi-Sensor Multispectral Imagery. *Remote Sens.* **2017**, *9*, 252. [\[CrossRef\]](#)
- Wu, L.; Lu, M.; Fang, L. Deep Covariance Alignment for Domain Adaptive Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [\[CrossRef\]](#)
- Liu, J.; Wu, J.; Xie, H.; Xiao, D.; Ran, M. Semantic Segmentation of Urban Remote Sensing Images Based on Deep Learning. *Appl. Sci.* **2024**, *14*, 7499. [\[CrossRef\]](#)
- Wang, Z.; Wang, J.; Yang, K.; Wang, L.; Su, F.; Chen, X. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+. *Comput. Geosci.* **2022**, *158*, 104969. [\[CrossRef\]](#)
- Zhang, H.; Jiang, Z.; Zheng, G.; Yao, X. Semantic Segmentation of High-Resolution Remote Sensing Images with Improved U-Net Based on Transfer Learning. *Int. J. Comput. Intell. Syst.* **2023**, *16*, 181. [\[CrossRef\]](#)
- Costa, H.; Foody, G.M.; Boyd, D.S. Supervised methods of image segmentation accuracy assessment in land cover mapping. *Remote Sens. Environ.* **2018**, *205*, 338–351. [\[CrossRef\]](#)
- van der Sande, C.J.; de Jong, S.M.; de Roo, A.P.J. A segmentation and classification approach of IKONOS-2 imagery for land cover mapping to assist flood risk and flood damage assessment. *Int. J. Appl. Earth Obs. Geoinf.* **2003**, *4*, 217–229. [\[CrossRef\]](#)
- Tang, Z.; Tang, H.; He, S.; Mao, T. Object-based change detection model using correlation analysis and classification for VHR image. In *Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015*; IEEE: Piscataway, NJ, USA, 2015; pp. 4840–4843.
- Liu, M.; Wang, P.; Hu, K.; Gu, C.; Jin, S.; Chen, L. A Method for Extracting High-Resolution Building Height Information in Rural Areas Using GF-7 Data. *Sensors* **2024**, *24*, 6076. [\[CrossRef\]](#)
- Li, P.; Guo, J.; Song, B.; Xiao, X. A Multilevel Hierarchical Image Segmentation Method for Urban Impervious Surface Mapping Using Very High Resolution Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *4*, 103–116. [\[CrossRef\]](#)
- Miao, Z.; Fu, K.; Sun, H.; Sun, X.; Yan, M. Automatic Water-Body Segmentation from High-Resolution Satellite Images via Deep Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 602–606. [\[CrossRef\]](#)
- Liu, M.Y.; Tuzel, O.; Ramalingam, S.; Chellappa, R. Entropy rate superpixel segmentation. In *Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011*; IEEE: Piscataway, NJ, USA, 2011; pp. 2097–2104.
- Radman, A.; Zainal, N.; Suandi, S.A. Automated segmentation of iris images acquired in an unconstrained environment using HOG-SVM and GrowCut. *Digit. Signal Process.* **2017**, *64*, 60–70. [\[CrossRef\]](#)
- Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the Computer Vision—ECCV 2018, Cham, Switzerland, 8–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 833–851.
- Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 645–657. [\[CrossRef\]](#)

19. Gao, Y.; Luo, X.; Gao, X.; Yan, W.; Pan, X.; Fu, X. Semantic segmentation of remote sensing images based on multiscale features and global information modeling. *Expert Syst. Appl.* **2024**, *249*, 123616. [\[CrossRef\]](#)
20. Jia, P.; Chen, C.; Zhang, D.; Sang, Y.; Zhang, L. Semantic segmentation of deep learning remote sensing images based on band combination principle: Application in urban planning and land use. *Comput. Commun.* **2024**, *217*, 97–106. [\[CrossRef\]](#)
21. Wang, J.; Chen, T.; Zheng, L.; Tie, J.; Zhang, Y.; Chen, P.; Luo, Z.; Song, Q. A multi-scale remote sensing semantic segmentation model with boundary enhancement based on UNetFormer. *Sci. Rep.* **2025**, *15*, 14737. [\[CrossRef\]](#)
22. Wang, Y.; Yang, L.; Liu, X.; Yan, P. An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3. *Sci. Rep.* **2024**, *14*, 9716. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Duan, S.; Zhao, J.; Huang, X.; Zhao, S. Semantic Segmentation of Remote Sensing Data Based on Channel Attention and Feature Information Entropy. *Sensors* **2024**, *24*, 1324. [\[CrossRef\]](#)
24. Liu, M.; Dan, J.; Lu, Z.; Yu, Y.; Li, Y.; Li, X. CM-UNet: Hybrid CNN-Mamba UNet for Remote Sensing Image Semantic Segmentation. *arXiv* **2024**, arXiv:2405.10530. [\[CrossRef\]](#)
25. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [\[CrossRef\]](#)
26. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [\[CrossRef\]](#)
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141.
28. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the Computer Vision—ECCV 2018, Cham, Switzerland, 8–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–19.
29. 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 (CVPR), Long Beach, CA, USA, 15–20 June 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 3141–3149.
30. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 13708–13717.
31. Li, M.; Zhang, X.; Huang, Y.; Oymak, S. On the power of convolution-augmented transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2025; pp. 18393–18402.
32. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 213–229.
33. Liang, C.; Huang, K.; Mao, J. Global-Local Deep Fusion: Semantic Integration with Enhanced Transformer in Dual-Branch Networks for Ultra-High Resolution Image Segmentation. *Appl. Sci.* **2024**, *14*, 5443. [\[CrossRef\]](#)
34. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 764–773.
35. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets v2: More Deformable, Better Results. *arXiv* **2018**, arXiv:1811.11168. [\[CrossRef\]](#)
36. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [\[CrossRef\]](#)
37. Zhang, X.; Liu, C.; Yang, D.; Song, T.; Ye, Y.; Li, K.; Song, Y. RFACnv: Innovating Spatial Attention and Standard Convolutional Operation. *arXiv* **2023**, arXiv:2304.03198. [\[CrossRef\]](#)
38. Chen, L.; Li, C.; Gao, Y.; Chang, Y.; Jin, S.; Wang, Z.; Ma, X.; Jia, L. Research on Remote Sensing Image Object Segmentation Using a Hybrid Multi-Attention Mechanism. *Appl. Sci.* **2026**, *16*, 695. [\[CrossRef\]](#)
39. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. DenseASPP for Semantic Segmentation in Street Scenes. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 3684–3692.
40. ISPRS Test Project on Urban Classification, 3D Building Reconstruction and Semantic Labeling. Available online: <https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/> (accessed on 20 December 2025).
41. 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. [\[CrossRef\]](#)
42. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. *arXiv* **2016**, arXiv:1612.01105. [\[CrossRef\]](#)
44. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [\[CrossRef\]](#) [\[PubMed\]](#)

45. Xu, J.; Xiong, Z.; Bhattacharyya, S. PIDNet: A Real-time Semantic Segmentation Network Inspired from PID Controller. *arXiv* **2022**, arXiv:2206.02066.
46. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203. [[CrossRef](#)]
47. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R. BiFormer: Vision Transformer with Bi-Level Routing Attention. *arXiv* **2023**, arXiv:2303.08810. [[CrossRef](#)]
48. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.