remote sensing

MDPI

Technical Note

Semantic Segmentation of High-Resolution Remote Sensing Images Based on Sparse Self-Attention and Feature Alignment

Li Sun, Huanxin Zou *, Juan Wei, Xu Cao, Shitian He, Meilin Li and Shuo Liu

College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China

* Correspondence: zouhuanxin@nudt.edu.cn; Tel.: +86-731-8700-3288

Abstract: Semantic segmentation of high-resolution remote sensing images (HRSI) is significant, yet challenging. Recently, several research works have utilized the self-attention operation to capture global dependencies. HRSI have complex scenes and rich details, and the implementation of self-attention on a whole image will introduce redundant information and interfere with semantic segmentation. The detail recovery of HRSI is another challenging aspect of semantic segmentation. Several networks use up-sampling, skip-connections, parallel structure, and enhanced edge features to obtain more precise results. However, the above methods ignore the misalignment of features with different resolutions, which affects the accuracy of the segmentation results. To resolve these problems, this paper proposes a semantic segmentation network based on sparse self-attention and feature alignment (SAANet). Specifically, the sparse position self-attention module (SPAM) divides, rearranges, and resorts the feature maps in the position dimension and performs position attention operations (PAM) in rearranged and restored sub-regions, respectively. Meanwhile, the proposed sparse channel self-attention module (SCAM) groups, rearranges, and resorts the feature maps in the channel dimension and performs channel attention operations (CAM) in the rearranged and restored sub-channels, respectively. SPAM and SCAM effectively model long-range context information and interdependencies between channels, while reducing the introduction of redundant information. Finally, the feature alignment module (FAM) utilizes convolutions to obtain a learnable offset map and aligns feature maps with different resolutions, helping to recover details and refine feature representations. Extensive experiments conducted on the ISPRS Vaihingen, Potsdam, and LoveDA datasets demonstrate that the proposed method precedes general semantic segmentation- and self-attention-based networks.

Keywords: semantic segmentation; high-resolution remote sensing; self-attention; context modeling; feature alignment

Citation: Sun, L. ; Zou, H.; Wei, J.; Cao, X.; He, T.; Li, M.; Liu, S. Semantic Segmentation of High-Resolution Remote Sensing Images Based on Sparse Self-Attention and Feature Alignment. *Remote Sens.* **2023**, *15*, 1598. https://doi.org/10.3390/rs15061598

Academic Editors: Qian Du, Jiaojiao Li, Wei Li, Jocelyn Chanussot, Rui Song, Yunsong Li and Bobo Xi

Received: 19 January 2023
Revised: 8 March 2023
Accepted: 13 March 2023
Published: 15 March 2023

1. Introduction

Semantic segmentation predicts the semantic labels for each pixel in an image. Semantic segmentation of high-resolution remote sensing images (HRSI) is the cornerstone of remote sensing interpretation. It is of great importance in many fields, such as mapping, navigation, land resource management, etc. [1–3]. Specifically, land cover maps depict local and overall landscape conditions, from which environmental change trends can be obtained. Semantic segmentation can be used to assess urban development and estimate the impact of natural disasters. Since remote sensing technology has advanced, HRSI with more complex pixel representation have become more readily available. Semantic segmentation is more crucial and challenging for HRSI. Traditional semantic segmentation methods [4–6] rely on expert experience and complex human-designs. Moreover, the segmentation performance relies on the accuracy and suitability of manually designed features. With robust feature modeling capabilities, deep learning technology has become an effective method used for semantic segmentation of HRSI, and researchers have applied deep learning technology

to this operation. Specifically, a convolutional neural network (CNN) [7] has been widely used in semantic segmentation and achieved satisfactory results. To further enhance the accuracy of semantic segmentation, researchers focus on both contextual information fusion and the refinement of segmentation results.

To achieve contextual information fusion, several network variants are proposed to enhance contextual aggregation. PSPNet [8] developed spatial pyramid pooling to acquire a rich, multi-scale context. The Deeplab series [9–11] utilized the atrous spatial pyramid pooling (ASPP) to gather contextual clues, which consisted of parallel atrous convolutions with different dilated rates. GCN [12] removed the pooling in the network and developed a large decoupling convolution kernel to extract features. The large convolution kernel can obtain a large receptive field and is beneficial to the capture of long-range contextual information. However, the above methods fail to model the global contextual dependencies across an entire image. Recently, self-attention mechanisms commonly used in natural language processing (NLP) have been widely used for visual tasks with exciting results. Wang et al. [13] first proposed self-attention to capture global dependencies. Fu et al. [14] developed DANet to model non-local dependencies in position and channel dimensions. Instead of calculating self-attention at each point, EAMNet [15] utilized the expectation-maximization iteration manner to learn a more compact basis set, and then carried out self-attention. To model spatial long-range dependencies, CCNet [16] proposed recurrent a criss-cross attention module. Yuan et al. [17] developed OCNet with interlaced sparse self-attention. The above methods show that the self-attention operation is an effective way to capture global dependencies. Thus, several studies have used the self-attention mechanism for semantic segmentation of HRSI. Shi et al. [18] combined self-attention and atrous convolution with different atrous rates to capture spatially adaptive global context information. Li et al. [19] proposed kernel attention with linear complexity and combined it with the standard dot product attention. However, the above methods ignore a key problem: due to the complex background and rich details of HRSI, standard self-attention will introduce redundant information and interfere with semantic segmentation. To solve this problem, this paper proposes the sparse position self-attention module (SPAM) and sparse channel self-attention module (SCAM), which not only captures the global information, but also reduces the interference of redundant information.

For the refinement of segmentation results, the current semantic segmentation network uses several strategies. One is to obtain the high-level semantic information gradually via down-sampling and then integrate the features of various levels through the decoder to recover the details. For example, Long et al. [20] proposed fully convolutional networks (FCNs) that restored the original image size by incorporating the low-level features and high-level features. SegNet [21] retained the index of the maximum position when pooling, and the index was reused when upsampling. U-Net [22] adopted skip-connections to connect shallow layers and deep layers. RefineNet [23] utilized a Laplacian image pyramid to explicitly model the available information during downsampling and predictions from coarse to fine. Another potential strategy is to learn semantic information while maintaining high resolution feature maps. For example, HRNet [24] proposed a parallel structure backbone network, which maintained high resolution characteristics during the entire process. Additionally, several networks refine the segmentation edges to obtain more precise semantic segmentation results. Gated-SCNN [25] deconstructed the edge information from the regular features and used a shape branch to focus on semantic boundary information. SegFix [26] proposed a post-processing method to refine the boundaries of semantic segmentation results. ERN [27] developed the edge enhancement structure and the loss function used to supervise the edge to enhance the segmentation accuracy. Zheng et al. [28] developed a Dice-based edge-aware loss function to refine edge information directly from semantic segmentation prediction. Li et al. [29] highlighted the edge distribution of the feature map in a self-attention fashion. The above methods recover the details and improve the edge segmentation performance to some extent. However, the issue of feature maps with different resolutions being misaligned

is ignored. To solve this problem, this paper proposes the feature alignment module (FAM), which generates a learnable offset map to align feature maps with different resolutions.

HRSI generally have complex background information and abundant details, which makes semantic segmentation more challenging. The standard self-attention and excessive fusion of long-range context information may introduce redundant information and cause interference to object segmentation. This paper proposes SPAM and SCAM to effectively model the position global context and channel-wise dependencies. Additionally, feature maps with different resolutions are not aligned. Features from shadow layers and deep layers are directly fused and, thus, fail to obtain higher-quality segmentation results. This paper proposes FAM, which combines low-level and high-level features with different resolutions. FAM is beneficial, as it refines segmentation results and improves the segmentation performance of an object edge. The contributions of this work are threefold:

1. The paper proposes SPAM and SCAM to efficiently model the position non-local information and channel-wise dependency, which reduces redundant information, contributing to the intraclass consistency of large objects and the segmentation accuracy of small objects.
2. The paper introduces FAM, which can align feature maps with different resolutions and further improve segmentation results.
3. Extensive experimental results demonstrate that SAANet achieves leading performance on ISPRS Vaihingen, Potsdam, and LoveDA datasets.

2. Materials and Methods

The particulars of the proposed semantic segmentation network based on sparse self-attention and feature alignment (SAANet) for semantic segmentation will be introduced. We first present the overall framework of our SAANet and then illustrate the details of the SPAM, SCAM, and FAM.

2.1. Overview

As shown in Figure 1, the proposed SAANet consists of a backbone, SPAM, SCAM, and FAM. Many studies have proved the good performance of a pretrained ResNet backbone in semantic segmentation tasks. First, the dilated ResNet-101 [30] is set as the backbone to extract features. The outputs of the dilated ResNet-101 in each stage are denoted as $\{S_1, S_2, S_3, S_4\}$. Due to the removal of down-sampling operations and adoption of dilated convolutions in the last two blocks, feature maps have strides of $\{4, 8, 8, 8\}$ pixels, with respect to the input image. Then, SPAM and SCAM take the feature map S_4 as input to model non-local dependencies in the position and channel dimensions. In addition, in order to achieve better feature representations, a feature pyramid network (FPN) [31] is used to combine low-level and high-level features and the outputs are denoted as $\{F_1, F_2, F_3, F_4\}$. Finally, feature maps F_2,F_3, and F_4 are up-sampled to the same size as feature map F_1 utilizing FAM. The four feature maps are concatenated to gain final pixel-level feature representations.

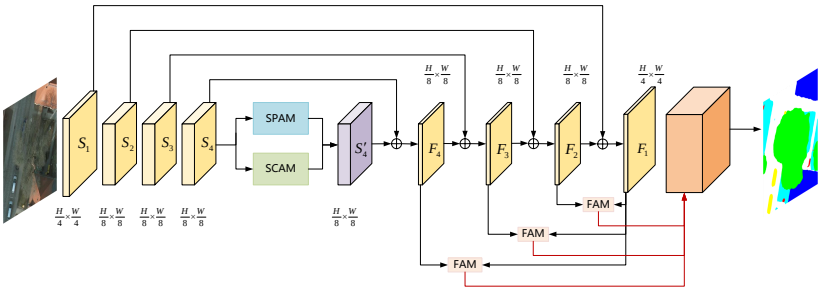

Figure 1. An overview of our proposed semantic segmentation network based on sparse self-attention and feature alignment (SAANet). H and W represent the height and width of the input image, respectively.

2.2. Sparse Position Self-Attention Module

Due to the complex scenes and rich details of HRSI, the implementation of a position self-attention module (PAM) on the whole image introduces redundant information and interferes with semantic segmentation. To capture long-range dependencies more efficiently and reduce redundant information, this paper proposes the SPAM, which is based on PAM.

2.2.1. Position Self-Attention Module

PAM is first introduced and shown in Figure 2. Given the feature map M, the features query (Q), key (K), and value (V) are first generated by convolutions, respectively, where $Q, K, V \in R^{C\times H\times W}$. C, H, and W denote the number of channels of the feature, image height, and image width, respectively. Then, they are reshaped to $Q_p, K_p, V_p \in R^{C\times N}$, where $N = H \times W$ is the number of pixels. Next, Q_p is multiplied by the transpose of K_p, and the softmax layer is applied to calculate the position attention map $A_p \in R^{N\times N}$:

$$A_p = \text{softmax}\left(K_p^T Q_p\right) \tag{1}$$

where A_p measures the influence between the two positions, and the more similar two pixel features are, the larger the value of A_p is. Then, V_p and the transpose of A_p are multiplied, and the resulting product is reshaped to $R^{C\times H\times W}$. Finally, to obtain the output $N_p \in R^{C\times H\times W}$, the feature map is multiplied by the scale coefficient α and sum with the feature map M.

$$N_p = \alpha V_p A_p^T + M \tag{2}$$

where α is a learnable parameter, which is initialized to 0.

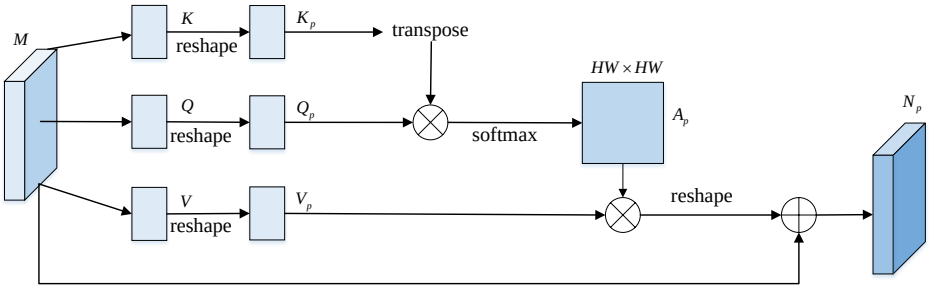

Figure 2. The framework of the position self-attention module (PAM).

2.2.2. Sparse Position Self-Attention Module

The proposed SPAM is based on PAM. Instead of standard PAM operating on the entire image, SPAM implements the sparse mechanism by performing PAM operations on sub-regions. SPAM can not only capture global context information, but can also reduce redundant information. Specifically, we divide the inputs to small regions along the position dimension and perform PAMs in sub-regions. The details of SPAM are shown in Figure 3. Given a feature map X with the spatial size of $H \times W$, the feature map X is grouped along the H and W dimensions and the spatial size of each group is $H\_n \times W\_n$. The feature map X is divided into $H/H\_n \times W/W\_n$ groups, named $\{X_1, X_2, X_3 \ldots\}$. Figure 3 illustrates the details of SPAM by taking $H, W = 8$ and $H\_n, W\_n = 2$ as an example. Then, the pixels at the same relative positions in each group are reorganized into new regions. The number of new regions is $H\_n \times W\_n$, and the pixels of each new region are $H/H\_n \times W/W\_n$. Meanwhile, PAMs are operated in new regions, and the feature map Y_1 is obtained, which is then set as the global operation. Finally, the pixel position of the feature map Y_1 is restored to the original combination, and PAMs are carried on $\{X_1, X_2, X_3 \ldots\}$. The feature Z_1 is obtained. The input of SPAM is S_4 in our SAANet. SPAM efficiently captures the long-range context information and models the pixel-wise relationship. The information propagation process of SPAM is shown in Figure 4. Take regions R_1, R_2 and pixels M, N, P,

Q as examples to illustrate details of the information propagation. Specifically, pixel M in the region R_1 and pixel N in the region R_2 first operate PAM, while pixel P in the region R_1 and pixel Q in the region R_2 operate PAM. Then, the region R_1 and the region R_2 continue PAM, respectively. Finally, pixels M, N, P, and Q aggregate the local and global contextual information. The above operations complete the information propagation between regions R_1 and R_2.

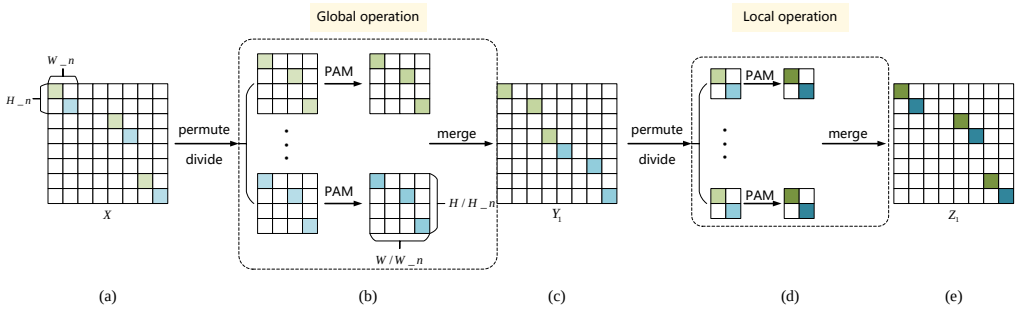

Figure 3. The structure of the proposed sparse position self-attention module (SPAM). (**a**) The input image is divided along position dimension. (**b**) PAMs are performed in rearranged small regions. (**c**) The pixel position of the feature map is restored to the original combination. (**d**) PAMs are performed in restored sub-regions. (**e**) The output of SPAM is obtained.

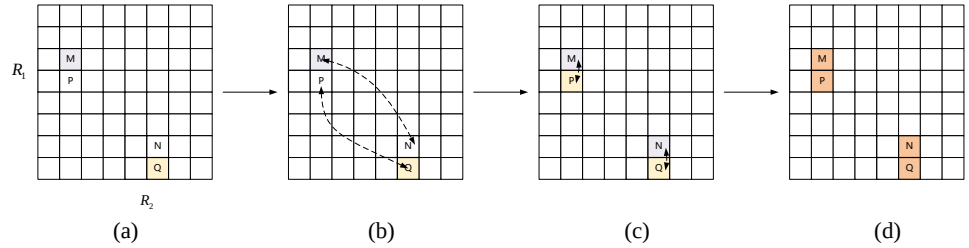

Figure 4. The information propagation process of the proposed SPAM. (**a**) M and P are the two pixels in region R_1; and N and Q are the two pixels in region R_2. (**b**) During the first PAM operation, the information is transmitted between M and N and between P and Q, respectively. (**c**) During the second PAM operation, the information is transmitted between M and P and between N and Q, respectively. (**d**) M, N, P, and Q contain global and local information.

2.3. Sparse Channel Attention Module

Due to the complexity of HRSI, there are large intra-class differences and small inter-class differences. Therefore, operating a standard self-attention module (CAM) on all channels introduces redundant information and causes category confusion. To model interdependent information between channels more efficiently and suppress redundant information, this paper proposes SCAM, which is based on CAM.

2.3.1. Channel Self-Attention Module

The architecture of CAM is shown in Figure 5. We first reshape the local feature map $M \in \mathbb{R}^{C\times H\times W}$ to $M_c \in \mathbb{R}^{C\times HW}$. Then, the matrix multiplication between M_c and the transpose of M_c is performed for the softmax layer, and the attention feature map $A_c \in \mathbb{R}^{C\times C}$ is obtained as follows:

$$A_c = \text{softmax}\left(M_c {M_c}^T\right) \tag{3}$$

where A_c measures the influence of different channels. Then, M_c is multiplied by the transpose of A_c, and the multiplication is reshaped to $\mathbb{R}^{C\times H\times W}$. Finally, the product is

multiplied by the scale coefficient β and added to the original feature M to obtain the final feature map $N_c \in \mathbb{R}^{C\times H\times W}$, as follows:

$$N_c = \beta {A_c}^T M_c + M \tag{4}$$

where β is a learnable parameter and is initialized to 0.

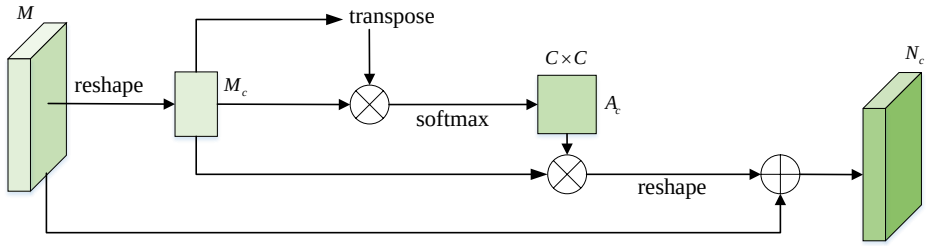

Figure 5. The architecture of the channel self-attention module (CAM).

2.3.2. Sparse Channel Self-Attention Module

The details of SCAM are shown in Figure 6. First, the feature map X is divided into C_n groups $\{C_1, C_2, C_3 \ldots\}$ in the channel dimension. The channel number of each group is C/C_n. Figure 6 illustrates the details of SCAM, taking $C\_n = 2$ as an example. Then, the groups $\{C_1, C_2, C_3 \ldots\}$ are further divided into C_n sub-groups, named $\{C_{11}, C_{12}, C_{13} \ldots\}$, $\{C_{21}, C_{22}, C_{23} \ldots\}$, $\{C_{31}, C_{32}, C_{33} \ldots\}$ .... The channel number of each sub-group is C/C^2_n. Next, for each channel group, sub-groups in the same relative position (for instance, C_{11}, C_{21}, C_{31} ...) are taken out to rearrange and generate new channel groups $\{C_{11}, C_{21}, C_{31} \ldots\}$, $\{C_{12}, C_{22}, C_{32} \ldots\}$, $\{C_{13}, C_{23}, C_{33} \ldots\}$ .... The feature map Y_2 is obtained by operating CAMs in C/C_n new groups. Finally, Y_2 is restored to the original channel arrangement, and Z_2 is acquired by performing CAMs in original groups $\{C_1, C_2, C_3 \ldots\}$.

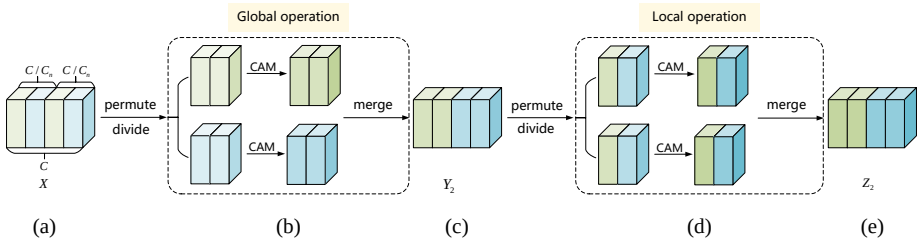

Figure 6. The structure of the proposed sparse channel self-attention module (SCAM). (**a**) The input image is divide along the channel dimension. (**b**) CAMs are performed in rearranged sub-channels. (**c**) The channels of the feature map are resorted to the original combination. (**d**) CAMs are performed in resorted sub-channels. (**e**) The output of SCAM is obtained.

2.4. Feature Alignment Module

Several methods are proposed to refine semantic segmentation results. However, the misalignment of features is ignored. To align features with different resolutions and refine semantic segmentation representations, this paper proposes an FAM. Specifically, the feature map ${S'}_4$ in Figure 1 from the last stage of ResNet fuses global context information and possesses enriched semantic information. However, the feature map ${S'}_4$ with coarse resolution lacks fine details. The proposed SAANet uses an FPN to fuse different resolution features from different stages. The FPN gradually fuses lower-level features with the details and higher-level features with abundant semantic information in a top-down pathway via 2× bilinear upsampling. However, the feature maps with different resolution are misaligned, which causes confusion in edges and small object segmentation. The misalignment has a great influence on the accuracy of semantic segmentation, especially on HRSI with

complex scenes. After a series of operations in SAANet, such as downsampling, residual connection, self-attention, etc., the misalignment of the feature maps is more complicated. In the upsampling process, using bilinear interpolation alone fails to achieve better semantic segmentation results. The proposed SAANet develops a feature alignment module, which utilize convolutions to obtain a learnable offset map for feature alignment.

The details of the FAM are shown in Figure 7. The FAM is structured within the FPN framework. The inputs of the FAM are two feature maps with different spatial resolutions. It is assumed that F^l and F^{l-1} are the two input features of FAM, where $F^l \in R^{H_l \times W_l \times C}$ and $F^{l-1} \in R^{H_{l-1} \times W_{l-1} \times C}$. F^l is first upsampled via the standard regular grid sampling based bilinear interpolation. Then, the upsampled F^l and F^{l-1} are concatenated, and the feature map F' is obtained. The feature F' is passed through a 1×1 convolution, BN, and 3×3 convolution to predict an offset $\Delta_f \in R^{H_{l-1} \times W_{l-1} \times 2}$. Finally, the offset map is used to correct the upsampled F^l, which obtains the output feature map aligned with F^{l-1}. Mathematically, the above operations can be written as:

$$F' = concat(F^{l-1}, upsample(F^l)) \tag{5}$$

$$\Delta_f = conv_{3\times3}(BN(conv_{1\times1}(F'))) \tag{6}$$

where the $upsample$ denotes the bilinear interpolation function, and Δ_f denotes offsets in horizontal and vertical directions. The FAM also involves less computation. SAANet uses three FAMs for the alignment of F_2, F_3, F_4, and F_1, respectively. FAM alleviates the feature misalignment and enhances the performance of semantic segmentation, especially for small objects and boundary regions.

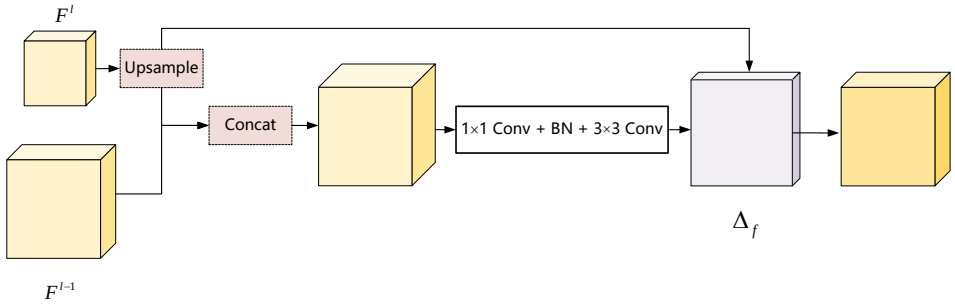

Figure 7. The framework of the proposed feature alignment module (FAM).

3. Experiments

We first introduce the datasets, evaluation metrics, and implementation details and then conduct ablation studies to validate the effectiveness of our framework. Finally, we compare the proposed network with several state-of-the-art methods on ISPRS Vaihingen, Potsdam [32], and LoveDA Urban [33] datasets.

3.1. Datasets and Evaluation Metrics

ISPRS Vaihingen dataset: ISPRS Vaihingen is a high-resolution remote sensing dataset used for semantic segmentation, which is composed of 33 images. The ground sampling distance (GSD) is 9 cm, and the average size of the images is 2496 × 2046 pixels. All images have corresponding semantic segmentation labels. The training and testing sets contain 17 and 16 images, respectively. There are six categories: impervious surface, building, low vegetation, tree, car, and clutter/background.

ISPRS Potsdam dataset: ISPRS Potsdam contains 38 images. The GSD is 5 cm, and the size of each image is 6000 × 6000 pixels. All images have corresponding semantic segmentation labels. The number of images in the training and testing sets is 21 and 17, respectively. As with the Vaihingen dataset, there are six categories.

LoveDA Urban dataset: The LoveDA dataset is constructed by Wang et al. [33]. The historical images were obtained from the Google Earth platform. LoveDA Urban dataset

was obtained from urban areas in Wuhan, Changzhou, Nanjing, and other places in China. The size of each image is 1024 × 1024 pixels, and the GSD is 0.3m. The dataset was divided into three parts: a training set, a val set, and a test set, among which the training set and val set have semantic labels. In our experiment, 1156 training images were used as our training set, and 677 val set images were used as our test set. There are seven categories: background, building, road, water, barren, forest, and agricultural.

Evaluation Metrics: To evaluate the performance of semantic segmentation, this study sets the mean intersection over union (mIoU), F1-score (F1), and overall pixel accuracy (OA) [34] as its evaluation metrics. The aforementioned metrics are as follows.

$$mIOU = \frac{1}{N}\sum_{k=1}^{N}\frac{TP_k}{TP_k + FP_k + FN_k} \tag{7}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN} \tag{8}$$

$$OA = \frac{\sum_{k=1}^{N} TP_k}{\sum_{k=1}^{N} TP_k + FP_k + TN_k + FN_k} \tag{9}$$

where TP, FP, TN, FN, and k indicate the true positive, false positive, true negative, false negatives, and category, respectively.

3.2. Implementation Details

Due to limited computing resources, we cropped all images into 512 × 512 pixels. All experiments were implemented with PyTorch on a single NVIDIA GeForce RTX 2080Ti GPU, and the optimizer was set as standard the stochastic gradient descent (SGD). For different data sets, different learning rates were selected. The learning rates of the Vaihingen, Potsdam, and LoveDA Urban datasets were 0.001, 0.0008, and 0.0007, respectively. For all methods, cross-entropy loss is set as the loss function. For all datasets and networks, the maximum iteration period is 100 epochs.

3.3. Comparison to State-of-the-Art

To verify the superiority of our SAANet, we perform comparisons with several existing semantic segmentation methods, including self-attention-based and other general semantic segmentation networks. Aside from HRNet, whose backbone network is W48, other networks use the dilated ResNet-101 as the backbone. The experimental results on ISPRS Vaihingen, Potsdam, and LoveDA datasets are shown in Tables 1–3, respectively. The proposed SAANet achieves the best mIOU on ISPRS Vaihingen, Potsdam, and LoveDA Urban datasets.

Results on the Vaihingen dataset: Compared with the typical segmentation network FCN, our SSANet obtains 1.72%, 1.33%, and 0.76% improvement and achieves 68.50%, 80.22%, and 86.72% for mIOU, mF1, and OA, respectively. Moreover, the mIOU/F1/OA of our SAANet surpasses 0.92%/0.72%/0.37% by the network based on self-attention DANet. Thanks to SPAM, SCAM, and FAM, SSANet achieves more precise semantic segmentation results in all classes, especially on small objects. For example, SSANet outperforms the previous best one by 1.62% in the car category.

Results on the Potsdam dataset: Compared with a typical segmentation network based on self-attention CCNet, our SSANet obtains 1.20%, 1.02%, and 0.61% improvement and achieves 73.79%, 83.57%, and 88.22%, on mIOU, mF1, and OA, respectively. Moreover, the mIOU, mF1, and OA of our SAANet surpasses 1.07%/0.75%/0.76% by the typical network with multi-scale aggregation PSPNet. Meanwhile, SSANet achieves more precise semantic segmentation results in all classes, with the most significant improvement in the car category.

Table 1. Comparisons of different networks on ISPRS Vaihingen dataset. Note that we chose the IOU as the metric of each category. The best results are shown in boldface.

Method	Imp. Surf.	Building	Low Veg.	Tree	Car	Background	mIOU	mF1	OA
DeepLabv3+	77.15	85.40	61.22	74.54	57.46	26.44	63.70	75.90	85.30
HRNet	**79.12**	85.78	62.46	75.69	60.46	25.87	64.90	76.70	86.10
EMANet	77.83	85.73	62.61	75.37	60.95	32.18	65.78	77.87	85.88
PSPNet	77.94	85.77	62.90	75.65	60.40	34.59	66.21	78.34	85.97
CCNet	77.73	85.64	62.35	75.42	61.61	36.22	66.50	78.66	85.83
FCN	77.99	85.82	62.84	75.39	61.86	36.76	66.78	78.89	85.96
DANet	78.48	86.67	63.19	75.74	63.73	37.67	67.58	79.50	86.35
SAANet	79.00	**87.52**	**63.79**	**76.16**	**65.35**	**39.21**	**68.50**	**80.22**	**86.72**

Table 2. Comparisons of different networks on ISPRS Potsdam dataset. Note that we choose the IOU as the metric of each category. The best results are shown in boldface.

Method	Imp. Surf.	Building	Low Veg.	Tree	Car	Background	mIOU	mF1	OA
DeepLabv3+	81.87	90.32	71.75	73.58	83.12	34.19	72.47	82.50	87.45
CCNet	82.43	90.64	71.72	73.31	83.47	33.97	72.59	82.55	87.61
PSPNet	81.64	89.96	71.43	74.45	82.61	36.21	72.72	82.82	87.46
HRNet	82.65	89.99	72.17	74.16	83.58	35.06	72.94	82.87	87.76
DANet	82.80	**90.94**	72.23	74.42	83.70	33.87	72.99	82.80	87.96
FCN	82.30	90.66	71.62	74.37	83.55	36.03	73.09	83.03	87.73
EMANet	82.54	90.49	71.92	73.73	83.31	37.16	73.19	83.18	87.77
SAANet	**83.40**	90.78	**72.46**	**74.53**	**84.12**	**37.46**	**73.79**	**83.57**	**88.22**

Table 3. Comparisons of different networks on LoveDA Urban dataset. Note that we choose the IOU as the metric of each category. The best results are shown in boldface.

Method	Background	Building	Road	Water	Barren	Forest	Agricultural	mIOU	mF1	OA
DeepLabv3+	35.31	59.73	56.23	54.95	19.45	42.05	31.17	42.70	58.46	57.99
FCN	34.13	59.60	54.99	**68.42**	26.91	**47.90**	23.27	45.03	60.39	58.38
HRNet	37.96	60.03	**59.83**	68.33	25.07	44.63	30.59	46.63	62.11	60.87
PSPNet	38.72	58.80	53.00	60.30	23.18	44.36	48.13	46.64	62.64	63.41
DANet	38.67	**62.04**	58.93	66.52	23.26	43.92	34.37	46.82	62.33	61.54
CCNet	38.83	60.31	56.04	63.89	**39.74**	46.96	29.61	47.91	63.92	61.62
EMANet	40.46	60.02	58.18	64.55	30.36	47.74	46.22	49.65	65.58	65.19
SAANet	**42.09**	61.25	57.26	63.64	33.14	44.32	**48.38**	**50.01**	**66.03**	**65.45**

Results on the LoveDA Urban dataset: Compared with the ISPRS Vaihingen and Potsdam datasets, the LoveDA Urban dataset with lower GSD has more complex scenes, which makes semantic segmentation more challenging. Nevertheless, our SAANet still achieves the best mIOU, mF1, and OA. Particularly, for more challenging classes, the background class with greater intra-class variation, and the agricultural class with a small number of pixels, the proposed SAANet achieves the highest IOU. Specifically, compared with typical segmentation network FCN, our SSANet obtains 4.98%, 5.64%, and 7.07% improvement and achieves 50.01%, 66.03%, and 65.45% for mIOU, mF1, and OA, respectively. Moreover, the mIOU/F1/OA of our SAANet surpasses 2.10%/2.11%/3.83% for the network based on the self-attention CCNet.

Overall, our method achieves state-of-the-art semantic segmentation performance on the ISPRS Vaihingen, Potsdam, and LoveDA datasets. To qualitatively validate the effectiveness, several visualization results are shown in Figures 8 and 9. It is observed that the overall visual effect of our method outperforms other methods. Specifically, for large objects, our method contributes to the intra-class consistency. In the first group in Figure 8, for large buildings, other methods incorrectly predict that several pixels representing

buildings are low vegetation or tree class. In the first group in Figure 9, other methods incorrectly predict pixels inside impervious surfaces. On the contrary, our SAANet can maintain category consistency. Additionally, for small objects, our method achieves more refined semantic segmentation results. For example, in the second group in Figures 8 and 9, several pixels representing cars are incorrectly predicted or have rough edges in the visual results of other methods. However, our SAANet obtains more accurate pixel classification and more precise edges. This suggests that our SAANet can obtain superior semantic segmentation performance and visual effects.

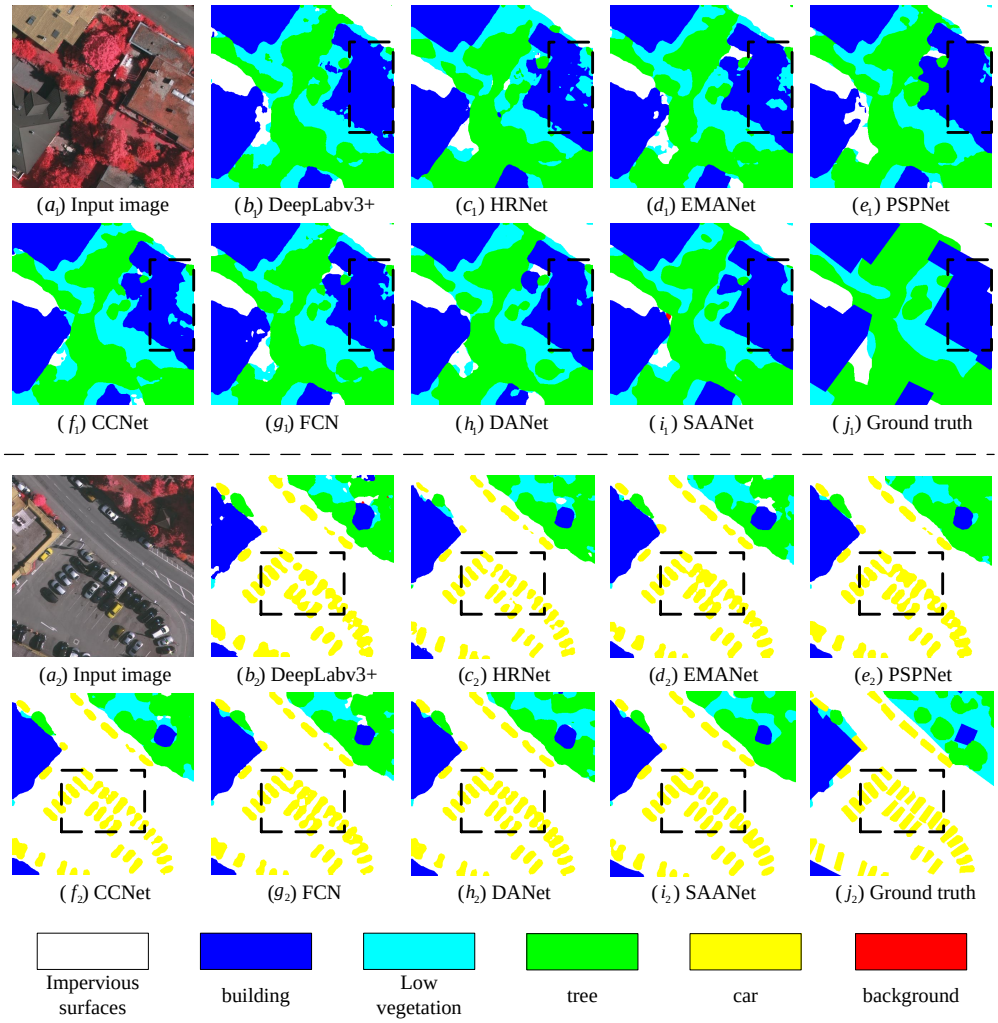

Figure 8. Visual results achieved by different networks on ISPRS Vaihingen dataset. For the first group, other methods incorrectly predict that several pixels representing buildings are low vegetation or tree class. For the second group, several pixels representing cars have rough edges. However, our SAANet can maintain category consistency and obtain more precise edges.

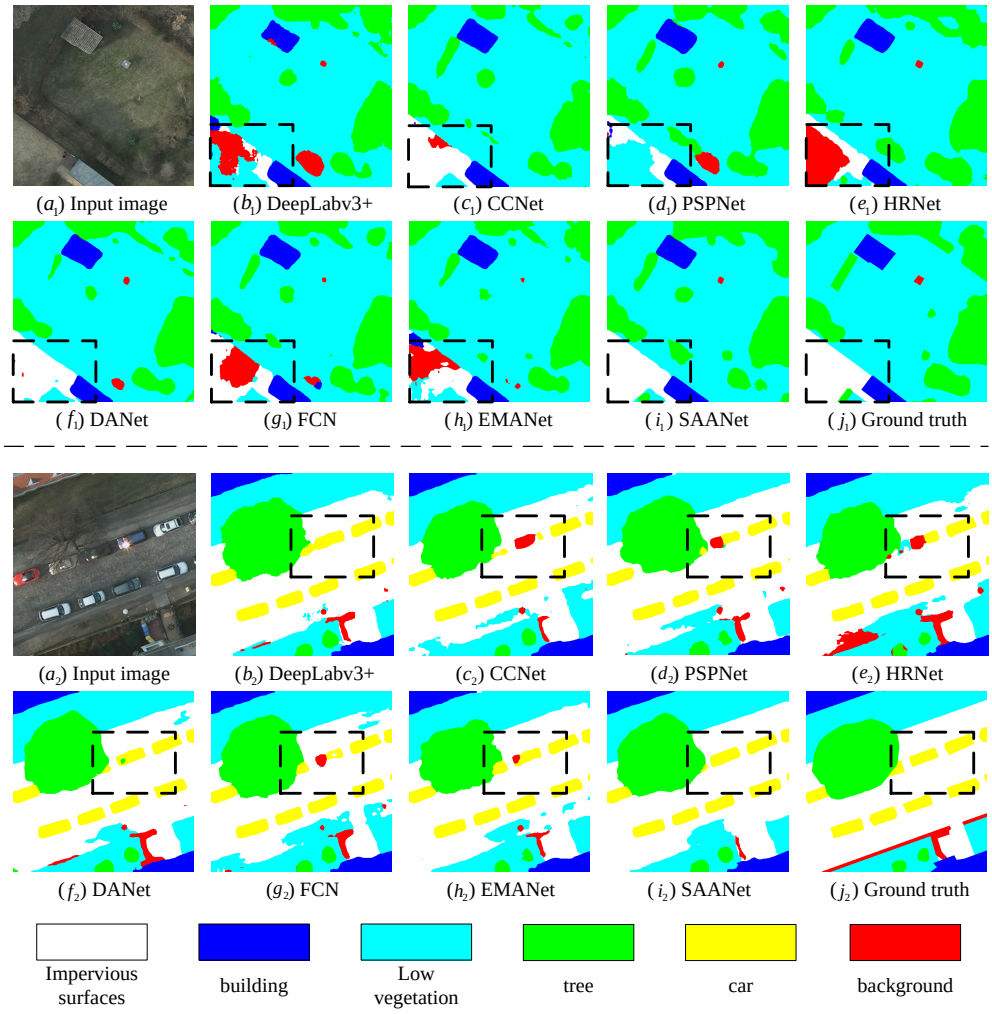

Figure 9. Visual results achieved by different networks on ISPRS Potsdam dataset. For the first group, other methods incorrectly predict pixels inside impervious surfaces. For the second group, several pixels representing cars are incorrectly predicted. However, our SAANet can maintain category consistency and obtain more accurate pixel classification.

3.4. Evaluation in Efficiency

We not only evaluate the segmentation accuracy and visualization results of different methods, but also measure the computational complexity and model parameters, in terms of giga floating-point operations per second (GFLOPs) (G) and the number of parameters with millions (Params) (M). All models are calculated with an input image size of $512 \times 512 \times 3$. The results are shown in Table 4. HRNet uses HRNetv2_W48 as the backbone network and has the lowest computational complexity. The backbone network of other methods is ResNet-101 with dilated convolution strategy. Compared with the self-attention-based networks DANet and CCNet, our method only increases the computational complexity by about 1.76% and the number of parameters by 0.6% to obtain better segmentation accuracy. Although our SAANet achieves better performance, it has a more complex structure and provides a small amount of computational complexity and parameters. We will focus on balancing the relationship between accuracy and complexity in future work.

Table 4. Comparison with other networks on GFLOPs and Params. The best results are shown in boldface.

Method	GFLOPs (G)	Params (M)
DeepLabv3+	254.56	60.21
HRNet	**93.73**	65.85
EMANet	246.63	**58.71**
PSPNet	256.63	65.60
CCNet	278.57	66.45
FCN	275.88	66.12
DANet	277.26	66.45
SAANet	283.46	66.85

4. Discussions

Previous work has focused on the fact that contextual information is important for semantic segmentation. PSPNet [8] uses the pooling operation of a pyramid structure to model the context information of different scales. Deeplabv3+ [11] combines the pyramid structure with the dilated convolution to capture the context information. In addition, several works [13–17] have proved that the self-attention mechanism is an effective way to model global context information. The self-attention mechanism captures context information through a sequence of matrix operations, which improves the accuracy of semantic segmentation. However, HRSI have complex scenes and rich details. The implementation of standard self-attention will introduce excessive redundant information and interfere with semantic segmentation. In this paper, SPAM and SCAM are proposed to model local and global context information, while avoiding the introduction of redundant information. In addition, FAM is proposed to improve the segmentation accuracy of edge regions and refine the semantic segmentation results. To better discuss and validate the effectiveness of each module of our SAANet, extensive ablation studies are conducted on the ISPRS Vaihingen and Potsdam datasets.

4.1. Sparse Position and Channel Attention Module

Both local and global context information is indispensable for the semantic segmentation task. In general, a larger receptive field can fuse a wider range of information, which is conducive to obtaining better feature representation. The standard self-attention operation is equivalent to fusing the information of each pixel of the image indistinguishably, which models long-range context information. The proposed SPAM and SCAM can capture local context information, as well as model long-range context information, and does so both sparsely and efficiently. To acquire a balance between local and global context, different H_n, W_n in SPAM and C_n in SCAM are set. Extensive experiments are conducted on the ISPRS Vaihingen and Potsdam datasets, and the results are shown in Table 5.

The pretrained ResNet-101 with the dilated strategy is adopted to initialize the backbone. The output of the last stage of ResNet-101 is used for semantic segmentation. Baseline based on ResNet101 obtains an mIOU of 65.19%, an mF1 of 77.32%, an OA of 85.60% on the ISPRS Vaihingen dataset. Baseline obtains an mIOU of 72.03%, an mF1 of 82.04%, an OA of 87.43% on the ISPRS Potsdam dataset. Compared with other H_n, W_n in SPAM and C_n in SCAM, SPAM with $H\_n, W\_n = 4$ and SCAM with $C\_n = 2$ achieves the best mIOU of 68.19% on the Vaihingen dataset and mIOU of 73.65% on the Potsdam dataset. The larger H_n and W_n are, the wider the region of capturing local information in the spatial dimension is, which will introduce more redundant information. Each channel map of high-level features is related to the category. By dividing channels into more groups (i.e., the larger C_n is), several channels with strong associations may be dispersed and rearranged, which is adverse to obtaining a better feature representation of each class. Therefore, H_n, W_n, and C_n are set as 4, 4, and 2, respectively, in follow-up experiments.

Table 5. Comparisons of different *H_n*, *W_n*, and *C_n* on ISPRS Vaihingen and Potsdam datasets. The best results are shown in boldface.

Dataset	H_n	W_n	C_n	mIOU	mF1	OA
Vaihingen	/	/	/	65.19	77.32	85.60
	4	4	2	**68.19**	**79.99**	**86.64**
	4	4	4	67.82	79.67	86.47
	8	8	2	67.73	79.71	86.24
	8	8	4	67.82	79.75	86.40
	16	16	2	67.67	79.52	86.42
	16	16	4	68.04	79.75	86.40
Potsdam	/	/	/	72.03	82.04	87.43
	4	4	2	**73.65**	**83.49**	88.09
	4	4	4	73.38	83.28	87.96
	8	8	2	73.44	83.22	**88.15**
	8	8	4	73.11	83.04	87.89
	16	16	2	73.54	83.45	87.98
	16	16	4	72.57	82.39	87.89

Sparse Position Attention Module: In order to efficiently model spatial long-range context information, SPAM is introduced to enhance the output of the backbone. The results are shown in Table 6. Compared with the baseline, SPAM provides an mIOU of 0.54% and 1.15% improvement and achieves an mIOU of 65.73% and 73.18%, an mF1 of 78.06% and 83.12%, and an OA of 85.61% and 87.88%, respectively, on the Vaigingen and Potsdam datasets. It is obvious that SPAM can effectively capture global context information and achieves a better segmentation performance.

Table 6. Comparisons of different versions of our network on ISPRS Vaihingen and Potsdam datasets. The best results are shown in boldface.

Dataset	SPAM	SCAM	FPN	FAM	mIOU	mF1	OA
Vaihingen					65.19	77.32	85.6
	✓				65.73	78.06	85.61
		✓			65.38	77.41	85.85
	✓	✓			68.19	79.99	86.64
	✓	✓	✓		66.82	78.88	86.11
	✓	✓	✓	✓	**68.50**	**80.22**	**86.72**
Potsdam					72.03	82.04	87.43
	✓				73.18	83.12	87.88
		✓			72.80	82.80	87.73
	✓	✓			73.65	83.49	88.09
	✓	✓	✓		73.63	83.45	88.07
	✓	✓	✓	✓	**73.79**	**83.57**	**88.22**

Sparse Channel Attention Module: In order to efficiently capture the interdependencies between channels, SCAM is introduced to enhance the output of the backbone. The results are shown in Table 6. Compared with the baseline, SCAM provides mIOU of 0.19% and 0.77% improvement and achieves an mIOU of 65.38% and 72.80%, an mF1 of 77.41% and 82.80%, and an OA of 85.85% and 87.73%, respectively, on the Vaigingen and Potsdam datasets. SCAM is of great significance when it comes to modeling the dependencies between channels.

We integrate SPAM and SCAM into the baseline to generate a network. Compared with the baseline, the SPAM models long-range context information, and SCAM efficiently captures the interdependencies between channels. SPAM and SCAM provide an mIOU of 3%, an mF1 of 2.67%, and an OA of 1.04% improvement and obtain an mIOU of 68.19%, an F1 of 79.99%, and an OA of 86.64% on the Vaihingen dataset. Additionally, SPAM and SCAM provide an mIOU of 1.62%, an mF1 of 1.45%, and an OA of 0.66% improvement

and obtain an mIOU of 73.65%, an F1 of 83.49%, and an OA of 88.09% on the Potsdam dataset. Extensive experiments demonstrate that SPAM and SCAM enhance the semantic segmentation performance of HRSI.

4.2. Feature Alignment Module

We compare the segmentation results of the method with SPAM and SCAM with labels. Several results are shown in Figure 10, demonstrating that most of the regions with inaccurate segmentation are boundary regions. In this paper, the FPN structure is used to integrate the high-level features with semantic information and the low-level features with detail information to obtain a finer semantic segmentation result. However, feature maps with different resolutions are misaligned. Utilizing the FPN structure to fuse features from shadow layers and deep layers fails to obtain better results. Therefore, it is vital that the FAM aligns and fuses features with different resolutions. The results are shown in Table 6. The network with the FPN obtains an mIoU of 66.82%, an mF1 of 78.88%, and an OA of 86.11% on the Vaigingen dataset. The network with FAM achieves the best mIoU of 68.5%, mF1 of 80.22%, and OA of 86.72%. Meanwhile, the network with FAM achieves the best mIoU of 73.79%, an mF1 of 83.57%, and an OA of 88.22% on the Potsdam dataset. The results prove that feature alignment is essential and the proposed FAM is effective. Additionally, FAM is beneficial, as it refines the boundaries. To prove the effectiveness of FAM for boundary regions, the mIOU, mF1, and OA are calculated on the edge region. Since there is no standard edge region, a neighborhood in which different classes are connected is selected as the edge region in this paper. Specifically, we first extract the boundary of different objects in the label and then perform the dilation operation in morphology operations to obtain the edge region. Note that the pixels closer to the object boundary are more likely to be confused, and the pixels closer to the object interior are more likely to be classified. As the dilation kernel increases, the edge region expands and the pixels grow closer to the interior of the object. The mIOU, mF1, and OA in a larger area cannot fully highlight the improvement in the edge region. Therefore, in this paper, a kernel of 3×3 is selected for the dilation operation to obtain the edge region. The results are shown in Table 7. The network without FAM obtains an mIoU of 35.68%, an mF1 of 51.88%, and an OA of 55.19% on the Vaigingen dataset. FAM provides an mIOU of 1.04%, an mF1 of 0.84%, and an OA of 0.23% improvement. Meanwhile, the method without FAM obtains an mIoU of 38.32%, an mF1 of 54.34%, and an OA of 56.48% on the Potsdam dataset. The method with FAM obtains an mIoU of 38.81%, an mF1 of 54.82%, and an OA of 56.90%. The results demonstrate that FAM refines the edge regions of semantic segmentation results and further explains the necessity and effectiveness of feature alignment.

In general, the experiments and visual results illustrate that SPAM, SCAM, and FAM achieve better semantic segmentation results. As shown in Tables 1–3, the proposed method achieves optimal OA, mF1, and mIOU on the ISPRS Vaihingen, Potsdam, and LoveDA Urban datasets. Specifically, the accuracy of small object cars is significantly improved. Additionally, as shown in Figures 8 and 9, other networks incorrectly the predicted pixels inside large objects, such as impervious surfaces and buildings. For small objects, such as cars, incorrect pixel classifications occur, as well as inaccurate edges. In contrast, our SAANet can maintain intra-class consistency for large objects and accuracy for small objects. Meanwhile, the experimental results show that global context information enhancement on HRSI with complex backgrounds introduces redundant information. The researchers further explore more adaptive global context information fusion methods to suppress redundant information as much as possible.

Table 7. Quantitative results achieved by different variants of our network on boundaries. The best results are shown in boldface.

Dataset	Method	mIOU	F1	OA
Vaihingen	baseline + SPAM + SCAM	35.68	51.88	55.19
	baseline + SPAM + SCAM + FAM	**36.52**	**52.92**	**55.42**
Potsdam	baseline + SPAM + SCAM	38.32	54.34	56.48
	baseline + SPAM + SCAM + FAM	**38.81**	**54.82**	**56.90**

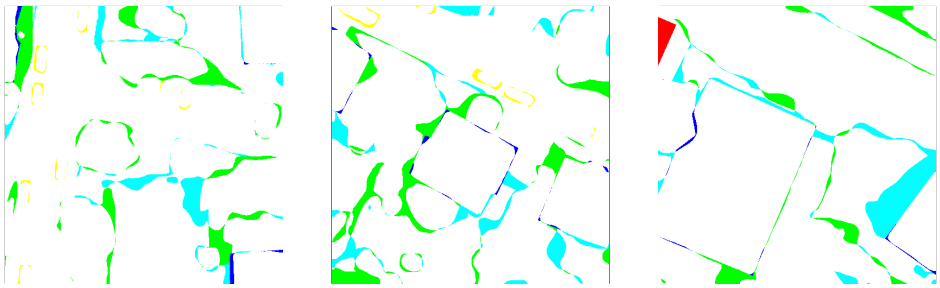

Figure 10. Visualization results of the difference between predictions and labels. (**a**–**c**) from the test set of the Vaihingen dataset. Note that most of the regions with inaccurate segmentation are boundary regions.

5. Conclusions

In this paper, we present a network based on sparse self-attention and feature alignment for semantic segmentation of HRSI. Specifically, SPAM is developed to capture long-range context information. SCAM is adopted to model interdependencies between channels more efficiently, while FAM is introduced to align features with different resolutions and refine semantic segmentation results. Moreover, extensive ablation experiments demonstrate the effectiveness of our method on the ISPRS Vaihingen and Potsdam datasets. Comparative experiments on the ISPRS Vaihingen, Potsdam, and LoveDA Urban datasets demonstrate that our SAANet obtains finer semantic segmentation results and achieves outstanding performance. Although our SAANet enhances the context information and details, there are still problems in the field of semantic segmentation of HRSI, such as large intra-class differences and small inter-class differences. For example, trees and low vegetation in the Vaihingen and Potsdam datasets are easily confused. In subsequent research, we will use comparative learning in the semantic segmentation of HRSI to obtain better feature embedding space and more easily distinguished feature representation.

Author Contributions: H.Z. determined the research direction and revised the expression of the article; L.S. came up with innovative ideas, developed the SAANet, conducted experiments, and completed the manuscript. J.W. helped to modify the conception and provided suggestions for expression. X.C., S.H., M.L. and S.L. checked out the article's writing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of China under grant 62071474.

Data Availability Statement: Not applicable.

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

References

1. 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.* **2014**, *53*, 280–295. [CrossRef]
2. Tang, Y.; Zhang, L. Urban change analysis with multi-sensor multispectral imagery. *Remote Sens.* **2017**, *9*, 252. [CrossRef]

3. 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]
4. Liu, M.Y.; Tuzel, O.; Ramalingam, S.; Chellappa, R. Entropy rate superpixel segmentation. In Proceedings of the The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011.
5. 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]
6. Thanh Noi, P.; Kappas, M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors* **2017**, *18*, 18. [CrossRef] [PubMed]
7. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
8. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Computer Society, Las Vegas, NV, USA, 26 June–1 July 2016.
9. 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] [PubMed]
10. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
11. 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 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
12. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
13. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
14. 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.
15. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 9167–9176.
16. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 603–612.
17. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. OCNet: Object context for semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 2375–2398. [CrossRef]
18. Shi, H.; Fan, J.; Wang, Y.; Chen, L. Dual attention feature fusion and adaptive context for accurate segmentation of very high-resolution remote sensing images. *Remote Sens.* **2021**, *13*, 3715. [CrossRef]
19. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
20. 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.
21. 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] [PubMed]
22. Ronneberger, O.; Fischer, P.; Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*; Springer: Munich, Germany, 2015.
23. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
24. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Wang, J. High-Resolution Representations for Labeling Pixels and Regions. *arXiv* **2019**, arXiv:1904.04514.
25. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 5229–5238.
26. Yuan, Y.; Xie, J.; Chen, X.; Wang, J. Segfix: Model-agnostic boundary refinement for segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 489–506.
27. Liu, S.; Ding, W.; Liu, C.; Liu, Y.; Wang, Y.; Li, H. ERN: Edge loss reinforced semantic segmentation network for remote sensing images. *Remote Sens.* **2018**, *10*, 1339. [CrossRef]
28. Zheng, X.; Huan, L.; Xia, G.S.; Gong, J. Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss. *ISPRS J. Photogramm. Remote Sens.* **2020**, *170*, 15–28. [CrossRef]
29. Li, X.; Li, T.; Chen, Z.; Zhang, K.; Xia, R. Attentively learning edge distributions for semantic segmentation of remote sensing imagery. *Remote Sens.* **2021**, *14*, 102. [CrossRef]

30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
31. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
32. Available online: https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/ (accessed on 1 March 2021).
33. 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.
34. Zhao, Q.; Liu, J.; Li, Y.; Zhang, H. Semantic Segmentation with Attention Mechanism for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]