# SPANet: Spatial Adaptive Convolution Based Content-Aware Network for Aerial Image Semantic Segmentation

Jianlong Hou, *Member, IEEE*, Zhi Guo, *Member, IEEE*, Yingchao Feng, *Member, IEEE*, Youming Wu, *Member, IEEE*, and Wenhui Diao, *Member, IEEE*

***Abstract*—Semantic segmentation of remote sensing images encounters four significant difficulties: 1) complex backgrounds, 2) large-scale differences, 3) numerous small objects, and 4) extreme foreground–background imbalance. However, the existing generic semantic segmentation models mainly focus on the modeling context information and rarely focus on these four issues. This article presents an enhanced remote sensing image semantic segmentation framework to solve these problems through the hierarchical atrous pyramid (HASP) module and spatial-adaptive convolution-based FPN decoder framework. On the one hand, HASP solved the problem of complex backgrounds and large-scale differences by further enlarging the receptive field of the network through the cascade of atrous convolution with various rates. On the other hand, spatial adaptive convolution is embedded in FPN decoder framework step by step to solve the problems of numerous small objects and extreme foreground–background imbalance. Besides, a boundary-based loss function is constructed to help the network optimize the relevant segmentation results. Extensive experiments over iSAID and ISPRS Vaihingen datasets reflect the superiority of the presented structure to conventional the state-of-the-art semantic segmentation approaches.**

***Index Terms*—Attention module, remote sensing, semantic segmentation, spatial adaptive.**

## I. Introduction

With the fast growth of remote sensing technology, high-resolution remote sensing images can be easily obtained for semantic segmentation. Remote sensing image semantic segmentation has a broad domain of usages in various areas, such as urban planning [1], [2], disaster monitoring [3], meteorological monitoring [4], and environmental modeling [5], [6]. This demonstrates the important academic and application value of the research and development of remote sensing image semantic segmentation.

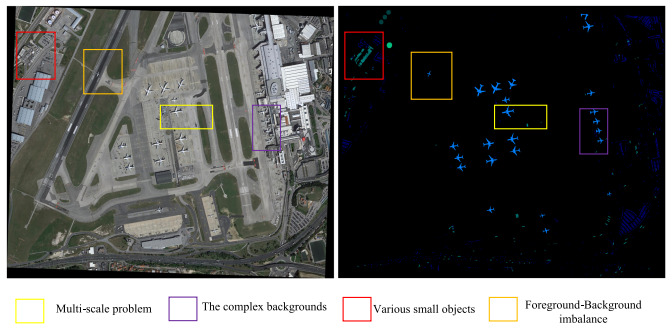


Fig. 1. Illustrations of potential challenges in remote sensing aerial images. (1) Multiscale problem. (2) Complex backgrounds of remote sensing images. (3) Various small objects in remote sensing images. (4) Foreground–background imbalance.

In recent years, deep learning-based methods have brought revolutionary development in semantic segmentation. Although FCN and its variants [7], [8], [9], [10] perform well in general semantic segmentation tasks, their performance in remote sensing images is not satisfactory. This is because remote sensing images have the following challenges, which are described in detail in Fig. 1.

1) Multiscale problem. The scale differences between different categories of objects in remote sensing images and within the same category are significant.
2) The complex backgrounds of remote sensing images can easily lead to false alarm.
3) There are various small objects in remote sensing images, and their dense arrangement leads to the misdetection problem.
4) Foreground–background imbalance. The number of the foreground is much smaller than the background, and the lack of foreground modeling in existing networks will easily cause the network to fall into local optimum.

As presented in Fig. 2, the existing approaches mainly employ the multiscale information to solve the problem of multiscale and complex backgrounds. PSPNet [11] presented a pyramid

Manuscript received 18 November 2022; revised 4 January 2023 and 7 February 2023; accepted 8 February 2023. Date of publication 13 February 2023; date of current version 27 February 2023. *(Corresponding author: Zhi Guo.)*

Jianlong Hou is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China, also with the University of Chinese Academy of Sciences, Beijing 100190, China, and also with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: houjianlong18@mails.ucas.ac.cn).

Zhi Guo, Yingchao Feng, Youming Wu, and Wenhui Diao are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, and also with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: guozhi@mail.ie.ac.cn; fengyc@aircas.ac.cn; youming_wu@yeah.net; diaowh@aircas.ac.cn).

The code will be accessible at: https://github.com/jlhou/SPANet.

Digital Object Identifier 10.1109/JSTARS.2023.3244207

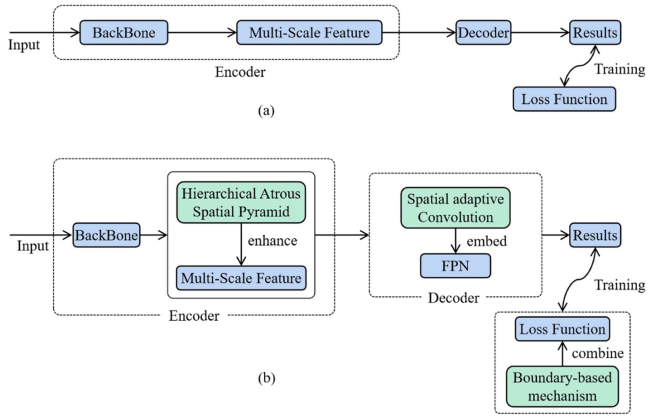


Fig. 2. Description of the segmentation pipeline. (a) Previous methods usually employ a simple encoder–decoder framework for feature encoding. (b) SPANet adopts HASP for promoting the extraction of multiscale information at the encoder stage. In the decoder stage, spatial adaptive convolution is embedded into the FPN framework to enhance the network's perception of the foreground objects and small objects. The boundary-based loss function is introduced into the training process to optimize the results.

pooling module (PPM). Deeplab series [12] proposed atrous spatial pyramid pool (ASPP). However, the performance of these modules for natural scenes in remote sensing images is limited. This is because the scale differences in natural scenes is much smaller than that in remote sensing images, and the receptive fields of the multiscale module proposed for natural scenes are insufficient remote sensing images. Thus, a critical issue is how to extract the features of objects with very large-scale differences. Inspired by the ideas of DenseNet and ASPP, the hierarchical atrous spatial pyramid (HASP) is proposed. It further extends the pyramid module's receptive field range by cascading the output of the atrous convolution with a lower rate into the input of the atrous convolution with a higher rate. Besides, since the existing methods usually adopt concat operation for the properties of various receptive fields, they cannot fully play to the advantages of different receptive field features. Therefore, a scale-aware fusion (SAF) module is proposed for optimization. The SWF module employs the channel-weighted (CW) and spatial-weighted (SW) mechanisms to weigh the properties of various receptive fields and optimize the feature fusion mode of various receptive fields.

In order to solve the problem of foreground–background imbalance and numerous of small objects. The existing methods use FPN-based methods to directly fuse low-level features with high-level ones. However, it is not reasonable to adopt the direct integration method. On the one hand, vanilla convolution is utilized to maintain the matching of dimensions in direct fusion, while vanilla convolution is space sharing. As a result, the gradient is calculated as the mean of the entire image. At the same time, remote sensing data with a more significant proportion of background than foreground are prone to fall into local optimum. On the other hand, the direct fusion method makes the network more advantageous in predicting large objects while making it challenging to predict small objects. Therefore, an FPN framework is proposed based on spatially adaptive convolution to solve the low-level and high-level features' fusion problem. The spatially adaptive convolution is embedded in the decoder framework of FPN step by step to enhance the network's perception of foreground information so that the network performs better on remote sensing images with the foreground–background imbalance and many small objects.

For the segmentation of foreground objects, we can simplify to find the boundary of foreground objects in the images. Thus, a new boundary loss function is designed, integrating with focal loss and dice loss to achieved more competitive performance.

Extensive experiments are performed on two challenging large-scale aerial datasets, i.e., iSAID [13] and ISPRS Vaihingen [14].

Based on the abovementioned considerations, spatial-adaptive convolution-based content-aware network (SPANet) is presented for aerial image semantic segmentation. In summary, our main contributions are as follows.

1) The HASP module is established for solving the large-scale difference problem and complex backgrounds of remote sensing images.
2) A spatial-adaptive convolution-based FPN decoder framework (SPA-FPN) is designed to solve the problem of foreground–background imbalance and numerous small objects.
3) A robust boundary loss is proposed to achieve the competitive performance.
4) To verify the effectiveness of our method, we conduct experiments on two datasets as iSAID and ISPRS Vaihingen. Our approach yields the state-of-the-art results on both datasets.

The rest of this article is organized as follows. Section II briefly introduces some relevant methods. The details of the presented method are given in Section III, and the experiments are presented in Section IV. In the end, Section V includes the conclusions and discussion of the results.

## II. Related Work

### A. Image Segmentation

Traditional image semantic segmentation methods based on machine learning mainly include two processes: first, feature extraction of images, and then classification of image pixels. Image features can be extracted by manually marking [15], [16], [17], [18], but these manually made features depend on the experience of researchers. At the same time, when classifying pixels, only support vector machine [19] and other machine learning algorithms are used [20], [21], [22], [23], even though later algorithms based on probability graph model [24], [25], [26], [27], [28] consider using the values of the pixels around the classified pixels to improve the segmentation edge results. However, generally speaking, the feature expression ability of manually extracted features is limited and has limitations in practical applications [29].

With the development of deep learning, deep learning has achieved the highest accuracy rate in image segmentation [30], [31], [32]. Image segmentation is mainly divided into two categories: semantic segmentation and instance segmentation. In this article, we study the semantic segmentation method based on deep learning. Semantic segmentation is an intensive prediction

TABLE I
OBJECT SEMENTATION MIOU(%) ON ISAID VAL

| Method | Backbone | IoU per category(%) | | | | | | | | | | | | | | | mIoU(%) | mean $F_1$ | Acc.(%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Plane | BD | Bridge | GTF | SV | LV | Ship | TC | BC | ST | SBF | RA | Habor | SP | HC | | | |
| DANet [67] | ResNet-50 | 80.9 | 71.4 | 28.6 | 52.5 | 42.1 | 57.5 | 60.2 | 84.7 | 50.9 | 63.0 | 63.3 | 40.6 | 48.8 | 46.1 | 30.4 | 57.5 | 71.0 | 94.4 |
| CCNet [68] | ResNet-50 | 78.8 | 72.4 | 30.5 | 52.2 | 44.4 | 53.1 | 56.3 | 85.3 | 49.2 | 63.2 | 67.1 | 51.8 | 48.7 | 43.6 | 29.2 | 57.8 | 71.3 | 94.6 |
| EMANet [69] | ResNet-50 | 77.8 | 71.2 | 28.3 | 53.1 | 40.0 | 56.9 | 58.4 | 82.6 | 52.4 | 62.8 | 63.5 | 40.2 | 48.7 | 43.1 | 30.8 | 56.8 | 70.3 | 94.0 |
| DenseASPP [70] | ResNet-50 | 78.2 | 66.9 | 28.4 | 53.5 | 38.2 | 57.4 | 60.8 | 82.4 | 53.7 | 50.1 | 69.9 | 54.8 | 51.2 | 43.4 | 25.7 | 57.1 | 70.6 | 94.2 |
| Non-local [71] | ResNet-50 | 81.0 | 71.9 | 29.1 | 49.8 | 47.8 | 59.1 | 62.4 | 84.9 | 48.6 | 62.2 | 41.6 | 53.5 | 50.9 | 45.1 | 29.9 | 57.3 | 70.8 | 94.3 |
| DeepLab v3+ [72] | ResNet-50 | 82.9 | 71.4 | 32.2 | 52.0 | 48.5 | 60.1 | 64.5 | 81.1 | 55.7 | 68.5 | 61.0 | 60.2 | 53.0 | 47.1 | 30.8 | 60.5 | 73.9 | 94.8 |
| PSPNet [11] | ResNet-50 | 82.8 | 76.8 | 31.2 | 53.4 | 47.3 | 58.2 | 63.9 | 86.5 | 56.7 | 68.3 | 67.4 | 52.6 | 53.0 | 44.0 | 33.3 | 60.9 | 74.2 | 95.1 |
| SETR [73] | T-Large | 81.5 | 74.5 | 30.8 | 52.3 | 48.8 | 58.6 | 63.4 | 86.9 | 57.4 | 67.9 | 68.0 | 61.6 | 52.9 | 46.2 | 32.6 | 61.4 | 74.4 | 95.3 |
| UperNet [74] | Swin-B [75] | 83.0 | 75.3 | 30.5 | 51.8 | 49.6 | 59.2 | 63.1 | 85.7 | 59.2 | 68.2 | 68.1 | 59.5 | 53.3 | 46.8 | 34.3 | 61.6 | 74.7 | 95.5 |
| AlignSeg [76] | ResNet-50 | 82.9 | 76.2 | 28.7 | 52.0 | 50.3 | 60.7 | 67.4 | 86.2 | 62.1 | 68.9 | 71.2 | 56.2 | **54.8** | 45.7 | 31.2 | 62.1 | 75.0 | 95.6 |
| OCR [77] | HRNet-W48 | 83.3 | 73.6 | 33.1 | 47.7 | 49.6 | 61.4 | **67.8** | 87.9 | 63.4 | 70.7 | 72.8 | 59.5 | 53.3 | 48.4 | 30.4 | 62.6 | 75.4 | 95.8 |
| HMANet [55] | ResNet-50 | 83.8 | 74.7 | 29.0 | 54.6 | 50.3 | 59.7 | 65.4 | 88.7 | 60.5 | **70.9** | 70.2 | 62.9 | 51.9 | 51.4 | 32.6 | 62.6 | 75.5 | 95.8 |
| Semantic FPN [78] | ResNet-50 | 81.2 | 71.5 | 33.8 | 52.2 | 45.4 | 60.1 | 63.5 | 87.1 | 57.8 | 61.5 | 60.2 | 59.0 | 51.5 | 46.6 | 31.2 | 60.1 | 73.5 | 94.8 |
| SFNet [79] | ResNet-50 | 83.7 | 77.5 | 34.6 | 53.5 | 50.4 | 60.8 | 64.9 | 88.5 | 63.2 | 70.8 | 73.5 | 59.0 | 53.9 | 46.4 | 31.9 | 63.2 | 76.0 | 96.1 |
| FarSeg [80] | ResNet-50 | 82.0 | 77.7 | 36.7 | 56.7 | 46.3 | 60.6 | 65.4 | 86.4 | 62.1 | 61.8 | 72.5 | **71.4** | 53.9 | 51.2 | 35.8 | 63.7 | - | - |
| baseline | ResNet-50 | 80.7 | 71.3 | 31.9 | 52.0 | 44.9 | 58.7 | 61.6 | 86.6 | 55.3 | 62.2 | 64.5 | 55.7 | 49.6 | 45.0 | 30.3 | 59.4 | 72.4 | 94.7 |
| baseline+HASP | ResNet-50 | 82.7 | 75.2 | 30.8 | 52.0 | 48.2 | 59.2 | 63.3 | 86.6 | 59.0 | 68.2 | 68.1 | 61.3 | 53.0 | 46.6 | 33.7 | 61.6 | 74.8 | 95.5 |
| baseline+SPA-FPN | ResNet-50 | 83.3 | 77.6 | 35.3 | 55.5 | 49.3 | 60.6 | 65.1 | 87.2 | 62.9 | 66.8 | 73.1 | 67.4 | 53.8 | 49.6 | 34.3 | 63.3 | 76.6 | 96.6 |
| **Ours** | ResNet-50 | **85.1** | **78.3** | **40.1** | **56.9** | **50.6** | **61.6** | 66.9 | **89.1** | **64.3** | 70.6 | **74.2** | 70.8 | 54.5 | **53.1** | **39.3** | **65.9** | **78.6** | **97.3** |

Bold entities indicate optimal performance.

TABLE II
ABLATION STUDIES FOR PROPOSED MODULES

| Method | HASP | SPA-FPN | mIOU |
|---|---|---|---|
| baseline | | | 59.4 |
| Ours | ✓ | | 61.6 |
| | | ✓ | 63.3 |
| | ✓ | ✓ | **64.1** |

Bold entities indicate optimal performance.

TABLE III
ABLATION STUDIES FOR THREE LOSS FUNCTIONS

| Method | $L_{fc}$ | $L_{dice}$ | $L_b$ | mIOU |
|---|---|---|---|---|
| Ours | ✓ | | | 65.0 |
| | ✓ | ✓ | | 65.3 |
| | ✓ | | ✓ | 65.5 |
| | ✓ | ✓ | ✓ | **65.9** |

Bold entities indicate optimal performance.

task that classifies every pixel in an image, which is more difficult than image classification [33], [34]. Long et al. [7] proposed a fully convolution network (FCN) to semantically segment images in 2015. FCN modified the architecture of the convolutional neural network VGG16 [35]. FCN can accept images of any size as input and generate segmentation maps of the same size. In order to produce more accurate results, FCN uses jump connections to upsample the last layer of the feature map of the model and fuse it with the shallow feature map to combine semantic and spatial information. FCN uses an end-to-end approach to train deep neural networks to semantically segment images. Because the final result of FCN does not respond well to local features and the localization is inaccurate [36], [37], [38], Chen et al. [39] used the fully connected conditional random fields on the final output to associate any two pixels on the image to produce more refined results. FCN ignores useful scene-level context because it does not obtain global semantic information. In order to obtain a wider range of contextual information, Liu et al. [40] proposed ParseNet, which enhances the features of each location by global average pooling, so that each pixel can obtain global contextual information. PSPNet [11] proposed a PPM. The PPM performs multiple pooling operations on the input features at different scales, and then upsamples all feature maps to the same size to obtain contextual information at different scales.

At present, the mainstream semantic segmentation model is the encoder–decoder structure, the encoder can obtain high-level semantic information, and the decoder can recover the detailed information of the segmentation. The encoder can use a classification network that removes fully connected layers, such as ResNet [41] and Xception [42], and the decoder fuses the feature maps of different stages output by the encoder to obtain the final result. Ronneberger et al. [43] proposed UNet, which is a fully symmetric encoder–decoder structure. Feature maps are downsampled using pooling layers at the encoder stage, and feature maps are downsampled at the decoder stage. The stage is upsampled by transposed convolutions, and shallow features are connected by skip connections. The decoder stage requires

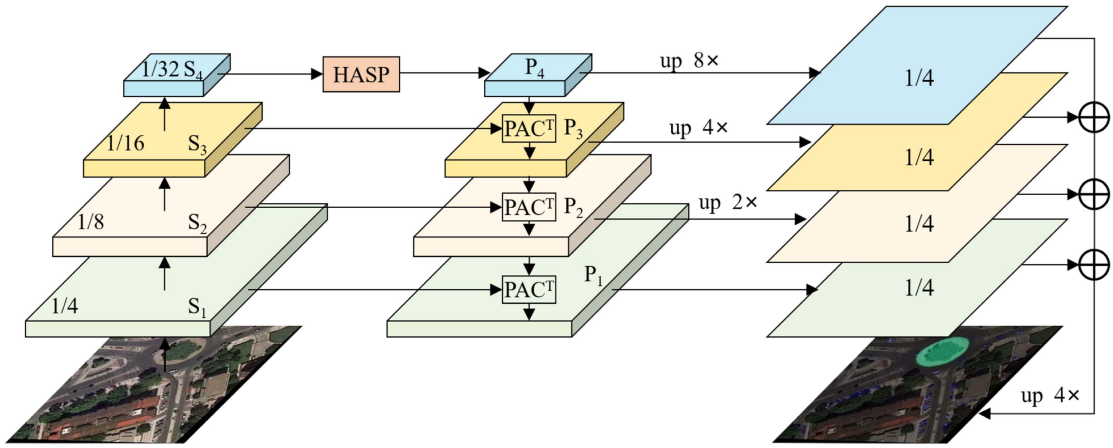


Fig. 3. Overall SPANet framework. The HASP module models the encoded features with multiscale information. PAC convolution is embedded in each upsampling stage of FPN to assist high-level information to predict more refined foreground information.

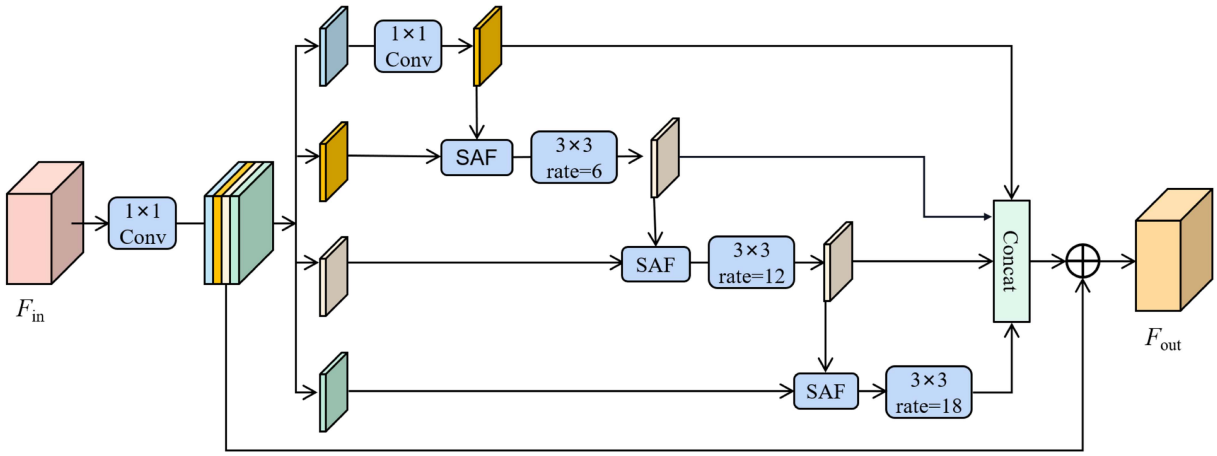


Fig. 4. HASP. In this module, the output of the smaller atrous convolution and the original input are cascaded into the larger atrous convolution kernel to increase the range of receptive field of the network. SAF means the proposed SAF module.

upsampling of high-level features, which is especially critical for subsequent feature fusion. Badrinarayanan et al. [9] proposed SegNet, which uses the index of the encoder stage max-pooling to nonlinearly upsample the feature map, so that upsampling cannot be learned separately. Since the result of upsampling is sparse, dense feature maps can be generated through trainable convolutional layers. After upsampling, shallow features need to be connected to obtain finer segmentation results.

### B. Remote Sensing Imagery Segmentation

Compared with natural images, remote sensing images are rich in high-resolution spatial information, texture details and complex scenes. Traditional remote sensing image semantic segmentation methods only rely on the texture or color difference of pixels in the image to achieve pixel classification, which leads to the segmentation accuracy has been difficult to meet the actual needs [44], [45], [46], [47], [48], [49].

The remote sensing image semantic segmentation method also benefits from the development of depth learning technology. Guo et al. [50] used the full convolution neural network with atrous convolution for remote sensing image segmentation, and used conditional random fields for postprocessing to smooth the prediction results. Diakogiannis et al. [51] proposed a method based on Unet that integrates residual connection, and then finetuned the loss function to obtain more accurate segmentation results. Wang et al. [52] used ResNet101 as a bone net to extract the high-level semantic features of remote sensing images. HMANet [53] introduced the category attention mechanism to calculate the category based correlation to recalibrate the category information. GraFNet [54] proposed a transformer-induced hierarchical graph network for multimodal semantic segmentation in remote sensing scenes, which promotes the exploration of potential intra- and intermodal relations by introducing a new modeling paradigm. GLCNet [55] proposed propose a global style and local matching contrastive learning network for remote sensing image semantic segmentation.

### C. Dynamic Convolution

Spatial adaptive convolution is a kind of dynamic convolution. Ordinary convolutions have many shortcomings, such as only

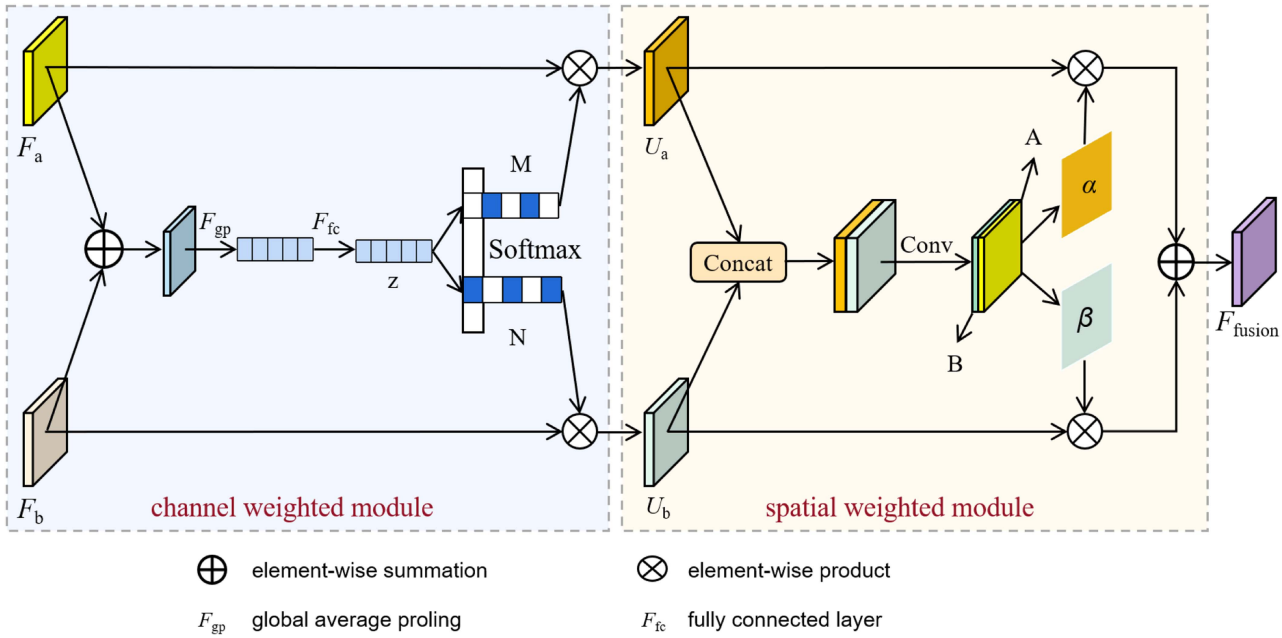


Fig. 5. SAF module. The module is composed of CW module and SW module in series.

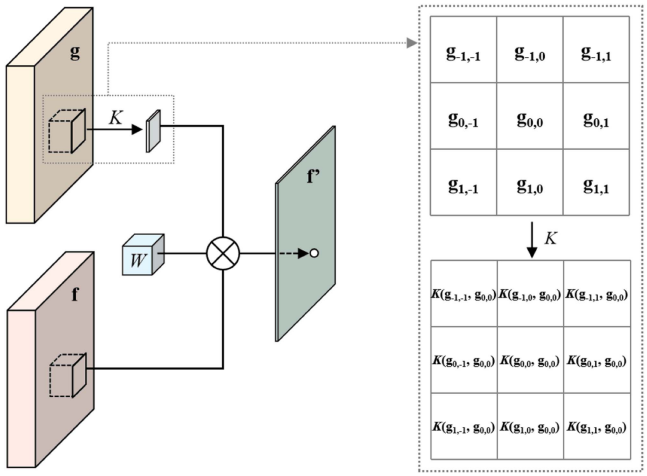


Fig. 6. PAC alerts a typical convolution on an input f by changing the spatially invariant filter W with an adapting kernel K, established via either predetermined or trained. ⊗ describes the elementwise multiplication of matrices followed by a summation. Only one output channel is presented for the description.

establishing local relationships and failing to capture geometric transformations of objects. Yu et al. [56] proposed atrous convolution, which adds an interval between each sampling point to extent the convolutional layer's receptive field. Peng et al. [57] reduced the amount of computation by replacing k × k convolution with k × 1 and 1 × k convolutions, so that larger convolution kernels can be used to extract features without increasing too much computation and parameters.

The abovementioned methods are all static convolution types, that is, the parameters of any input convolution kernel will not be transformed. Nowadays, many dynamic convolution algorithms proposed [58], [59], [60], [61]. For example, Zhang et al. [62] started from the module, and the convolutional layers in a module share the same convolutional kernel aggregation attention vector to reduce the computational amount. CondConv [61] uses the same convolution kernel for each target point on the feature map, while Chen et al. [58] proposed dynamic region-aware convolution (DRConv). DRConv selects different convolution kernels for each target point to extract features, which improves the representation ability and translation invariance of convolution.

## III. Methods

Aiming at the four major challenges in remote sensing images, we propose spatial adaptive convolution-based content-adaptive network for aerial image semantic segmentation. On the one hand, aiming at the characteristics of remote sensing images with large-scale differences and complex backgrounds, the HASP is proposed. On the other hand, aiming at the characteristics of foreground–background imbalance and numerous small objects, a SPA-FPN is proposed. Fig. 3 shows the overall network framework. In the following, the presented approach is illustrated in detail.

### A. HASP Module

There is considerable difference in the objects' size in remote sensing images. For example, cars and buildings are not of the same order of magnitude in size of the objects. However, it is crucial to make the network accurately perceive small objects and fully perceive large objects. Extending the network's receptive field range as much as possible is a friendly solution to deal with object interpretation under complex backgrounds. Inspired by the multiscale modules ASPP and PPM in natural scenes, the HASP module is proposed.

As shown in Fig. 4, for the feature $F_{in}$ obtained by the encoder, 1 × 1 convolution is first utilized for decreasing the dimension (the channel number is a multiple of 4). The features after

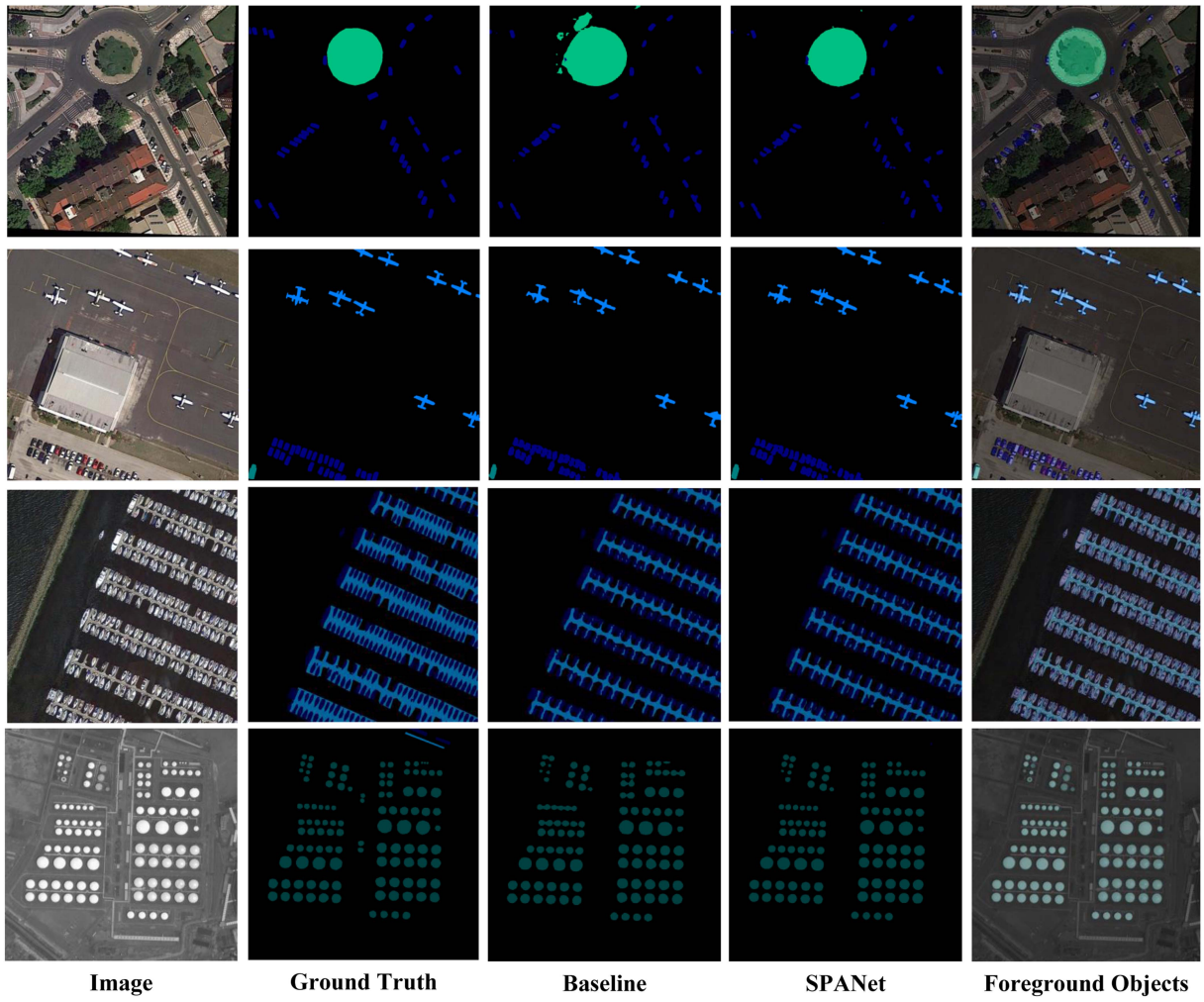


Fig. 7. Comparison of segmentation results between SPANet (ResNet-50) and baseline on the iSAID validation set.

dimensionality reduction are divided into four groups according to the channel dimension. Then, $1 \times 1$ convolution is adpoted for the separated features, and $3 \times 3$ atrous convolutions with rates of 6, 12, and 18 are used. Different from the ASPP module, the output of the $3 \times 3$ atrous convolution at the upper level (such as rate $= 6$) is cascaded with the original input, and then fused with the SAF module, and then input to the atrous convolution at the next level (such as rate $= 12$) for operation. Since the feature output of atrous convolution with a small rate is cascaded to the input of convolution with a large rate, the output feature map of convolution with a large atrous rate has richer receptive field range so that can solve the problem of large size difference of objects and complex backgrounds.

The SAF module is designed to perform feature fusion at different scales, and its structure is shown in Fig. 5. Specifically, two feature maps $U_a$ and $U_b$ are obtained from the features $F_a$ and $F_b$ of two different receptive fields by CW module, and the channel attention weight map are generated as follows:

$$\begin{aligned} U_a &= M \otimes F_a \\ U_b &= N \otimes F_b \end{aligned} \tag{1}$$

where, $\otimes$ stands for the elementwise product. $M$ and $N$ represent the characteristic coefficients, and the specific calculation process is as the following:

$$M = \frac{e^{Qz}}{e^{Qz} + e^{Pz}}, N = \frac{e^{Pz}}{e^{Qz} + e^{Pz}} \tag{2}$$

where, $Q, P \in R^{C \times C}$ denote the learnable transformation matrices, $z$ represents an input feature, and the specific calculation process is as the following:

$$z = \mathrm{FC}\left(\mathrm{GAP}\left(F_a + F_b\right)\right) \tag{3}$$

where, FC stands for full connection layer.

Next, $U_a$ and $U_b$ enter the SW module. Specially, we pass via some convolutions and obtain two feature maps $A, B \in R^{H \times W}$ of the different receptive fields. Then spatial attention weight maps $\alpha, \beta \in R^{H \times W}$ are generated as the following:

$$\alpha_i = \frac{e^{A_i}}{e^{A_i} + e^{B_i}}, \beta_i = \frac{e^{B_i}}{e^{A_i} + e^{B_i}}, i = [1, 2, 3 \ldots, H \times W]. \tag{4}$$

In the end, the $F_{\text{fusion}}$ is attained as the following:

$$F_{\text{fusion}} = \alpha \otimes U_a + \beta \otimes U_b \tag{5}$$

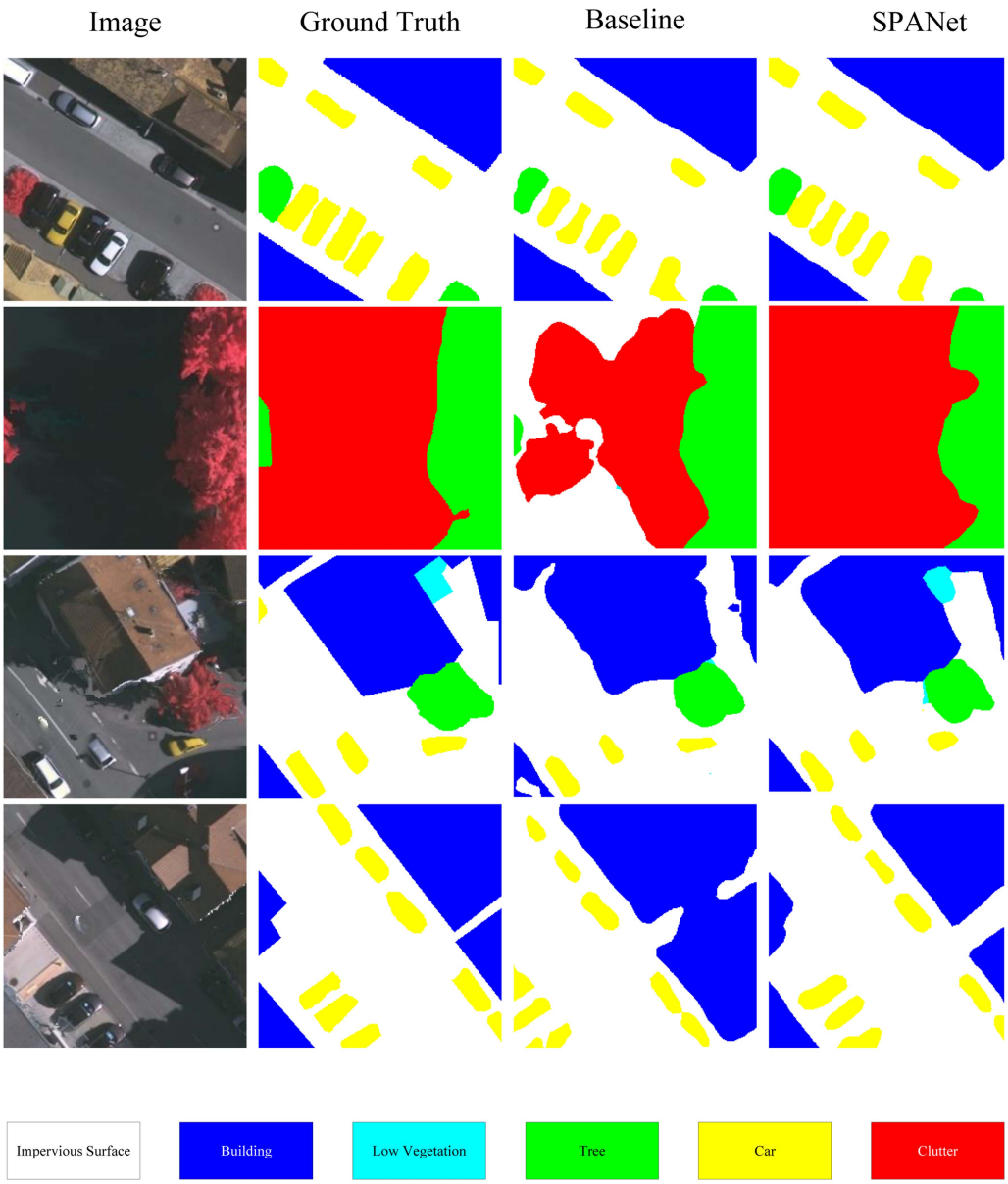


Fig. 8. Qualitative analysis of SPANet (ResNet101) and baseline (ResNet-101+FPN) on Vaihingen test set.

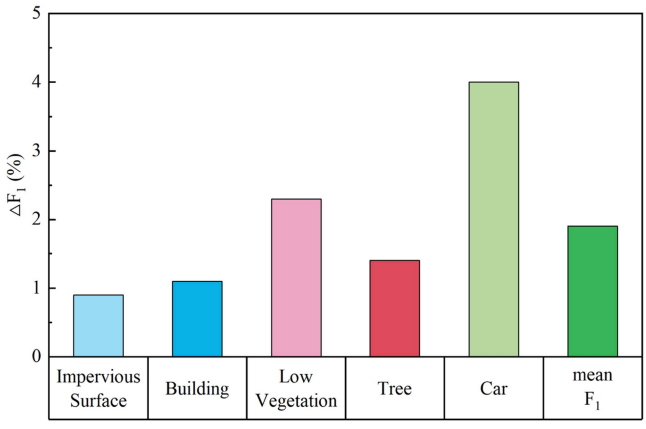


Fig. 9. Comparison of performance improvement at different objects.

where, $\otimes$ denotes the elementwise product operations, which acts between spatial weight maps and feature maps of different receptive fields for obtaining the fused feature map $F_{\text{fusion}}$.

### *B. Spatial-Adaptive Convolution-Based FPN Decoder Framework*

The existing networks utilize FPN framework is used to upsample the encoded features, and features from the encoder are channeled into the decoder to directly fuse to help the network predict more accurate results. However, this method still cannot show good performance for remote sensing images with extreme foreground–background imbalance and numerous small objects. This is because vanilla convolution is used to maintain the matching of dimensions in a direct fusion way, whereas ordinary convolution is spatially shared. As a result, the gradient

is calculated as the mean of the whole image, while remote sensing data with a much larger proportion of background than the foreground is prone to fall into local optimum and tend to predict large objects.

Inspired by spatial adaptive convolution, a hierarchical FPN structure based on pixel-adaptive convolution (PAC) [63] is proposed to solve this problem. PAC is a popular spatial adaptive convolution. As shown in Fig. 3, the low-level feature map in the encoder is adopted to guide the high-level feature map in the decoder step by step. The low-level features in the encoder contain richer foreground information and $\mathrm{PAC}^{\mathrm{T}}$ convolution is a foreground-aware enhanced operator. $\mathrm{PAC}^{\mathrm{T}}$ is the deconvolution of PAC. For clarity, this article first introduces PAC. PAC convolution is a content-adaptive operator, which can be network parameters that pay more attention to foreground information and small objects. This helps the decoder to recover more refined foreground segmentation results.

In order to clearly describe PAC convolution, we first introduce vanilla convolution. The convolution kernel $W \in R^{c' \times c \times r \times r}$ acting on image $f = (f_1, \ldots, f_n), f_i \in R^c$ over n pixels and c channels is denoted as

$$f_i{}' = \sum_{j \in \Omega(i)} W[p_i - p_j] f_j + b \tag{6}$$

where, $p_i = (x_i, y_i)^{\mathrm{T}}$ represents the pixel coordinates, $\Omega(\cdot)$ describes an $r \times r$ window, and $b \in R^{c'}$ describes biases. $[p_i - p_j]$ lists the relative positions of pixels in window r × r. The mentioned convolution operation gives a c-channel output, $f_i{}' \in R^{c'}$, at every pixel $i$. Equation (6) expresses that the weight of the vanilla convolution kernel is spatially shared, that is, image agnostic. Such convolution is performance constrained, so we want to use a content-adaptive convolution kernel.

Further, as shown in Fig. 6, the description of the PAC convolution we used is as follows. We choose to modify the spatially invariant convolution in (6) with a spatially varying kernel that depends on pixel features g

$$f_i{}' = \sum_{j \in \Omega(i)} K(g_i, g_j) W[p_i - p_j] f_j + b \tag{7}$$

where, $K \in R^{c' \times c \times r \times r}$ is a kernel function that has a fixed parametric form. We followed PAC convolution for the value of $K$

$$K(g_i, g_j) = \exp\left(-\frac{1}{2}(g_i - g_j)^{\mathrm{T}}(g_i - g_j)\right) \tag{8}$$

where, $g$ is an adaptive feature, and in this article is a low-level feature that provides detailed information. Similar to the transpose convolution corresponding to vanilla convolution, $\mathrm{PAC}^{\mathrm{T}}$ can be obtained in the same way.

PAC convolution is a content-aware spatial convolution, which can acutely capture the rich foreground information in the encoder to help the decoder recover more accurate segmentation results. The hierarchical progressive FPN framework proposed by $\mathrm{PAC}^{\mathrm{T}}$ convolution can recover more accurate foreground information step by step. Furthermore, enhanced capture of foreground information also leads to enhanced perception of small objects in the foreground.

### C. Loss Function

We optimize our objective from three perspectives: distribution-based, region-based, and boundary-based, and combine them, as shown in the following:

$$L = L_{fc} + L_{\mathrm{dice}} + L_b. \tag{9}$$

*1) Distribution-Based Perspective:* In the distribution perspective we use focal loss, it deals with extreme foreground–background category imbalance and reduces the loss of samples allocated to simple classification. The definition of focal loss is described as

$$L_{fc}(y, p) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \cdot \alpha \cdot (1 - p_i)^{\gamma} \cdot \log(p_{i,c}) \tag{10}$$

where, y describes the ground truth and p describes the predicted value of the network in every pixel. $\alpha$ represents the normalization factor which is set to 0.25, and $\gamma$ is chosen as 2.

*2) Region-Based Perspective:* From the region-based perspective, we use dice loss. As shown in (11), the dice efficient is a set similarity measure,calculate the similarity of two samples in the range [0,1]

$$\text{Dice efficient} = \frac{2\,|X \bigcap Y|}{|X| + |Y|} \tag{11}$$

where, $|X \bigcap Y|$ is the intersection computes between $X$ and $Y$, $|X|$ and $|Y|$ represents the number of elements of $X$ and $Y$. The dice loss is shown as

$$L_{\mathrm{dice}} = 1 - \frac{2\,|X \bigcap Y|}{|X| + |Y|}. \tag{12}$$

The dice loss is to control the gradient of the loss function to make the model pay more attention to samples with larger loss values.

*3) Boundary-Based Perspective:* We propose a robust boundary loss, where the boundary of the label is extracted from the edge by the gradient convolution kernel (GCK) [64]. GCK is a learnable boundary detection operator evolved from Sobel operator, and the cross entropy is calculated separately for the edge part to strengthen the information of the boundary

$$\mathcal{L}_{\mathrm{b}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{c} \mathrm{GCK}\left(y_{i,c}\right) \cdot \log\left(p_{i,c}\right). \tag{13}$$

## IV. Experiments

### A. Datasets and Assessment Metrics

1) iSAID: contains 15 types of foreground objects with spatial resolutions ranging from $\sim 800 \times 800$ to $\sim 4000 \times 13000$. We followed [79], with 1411 images as the training set and 458 images as the evaluating. Note that annotations for 937 images in test sets are not available. The mean $F_1$ score, the mean of categorywise intersection over union (mIoU) and overall accuracy (OA) of foreground classes are used for evaluation.
2) Vaihingen: contains six types of objects whose average spatial resolution is $2494 \times 2064$. We followed [64], [79],

TABLE IV
ABLATION STUDIES FOR EVERY BLOCK OF HASP

| Method | CW | SW | mIOU |
|---|---|---|---|
| baseline | | | 59.4 |
| baseline+PPM | | | 60.5 |
| baseline+ASPP | | | 60.9 |
| Ours | ✓ | | 61.2 |
| | | ✓ | 61.1 |
| | ✓ | ✓ | **61.6** |

*CW means the channel weighted module and SW means spatial weighted module.
Bold entities indicate optimal performance.

TABLE V
ABLATION STUDIES FOR SPA-FPN IN EVERY STAGE

| Method | Stage3 | Stage2 | Stage1 | mIOU |
|---|---|---|---|---|
| baseline | | | | 59.4 |
| Ours | ✓ | | | 60.2 |
| | ✓ | ✓ | | 61.5 |
| | ✓ | ✓ | ✓ | **63.3** |

Bold entities indicate optimal performance.

with 16 images as the training set and 17 images as the test set. The $F_1$ score, OA and mIoU are used for evaluation.

### B. Implementation Detail

ResNet-50/101+FPN is baseline due to its excellent performance. And ResNet-50/100 is our backbone. The samples are cropped with a fixed size of $896 \times 896$ for iSAID and $512 \times 512$ for Vaihingen. SGD optimizer with batch size $= 8$ is employed over four GPUs. We set the momentum and weight decay to 0.9 and 0.0001, respectively. The initial learning rate is 0.007 for iSAID and 0.01 for Vaihingen. We employ the poly learning rate policy $(\text{lr} = \text{base\_lr} * (1 - \frac{\text{iter}}{\text{max\_iter}})^{\text{power}})$, where power $= 0.9$. Each experiment is performed on $4\times$ NVIDIA 20180Ti GPU for 80 k iterations. Horizontal and vertical flip, rotation of $90 \cdot k(k = 1, 2, 3)$ degree are utilized within the learning to perform training data augmentation. It is worth noting that we performed the He normal initialization [80] for the HASP module.

### C. Experiments on iSAID

*1) Comparison to Traditional Methods:* To assess the SPANet, various experiments are accomplished on the iSAID dataset. SPANet was compared with numerous deep learning approaches from typical to state-of-the-art, most of which employed the ResNet-50 as the backbone. SETR [71] and Swin-B [73] utilize transformer [81] as backbone, a more common research area. The quantitative results presented in Table I indicated the superiority of the SPANet to other approach in remote sensing scenario. FarSeg [78] utilized foreground-aware approach to gain better performance. As a contrast, the proposed method achieved 65.9% mIOU, which is a more excellent result. Some examples of the segmentation results of the iSAID dataset is presented in the Fig. 7.

In order to further verify the performance of the proposed HASP and SPA-FPN, we conducted baseline+HASP and baseline+SPA-FPN experiments, respectively.

For the challenge (1) multi-scale problem and (2) complex background problem of remote sensing image, we conducted baseline+HASP experiment. Experiments show that both large-scale baseball diamond (BD) and small-scale small vehicle (SV) have obvious improvement. Furthermore, for plane (PL), which is often accompanied by a complex background, the HASP module is also significantly improves this problem, thus proving the superiority of our approach.

Moreover, experimental results show that the baseline+SPA-FPN can significantly improve the performance of 15 types of foreground objects in iSAID dataset with unbalanced foreground–background. On the other hand, for various small objects, such as SV, the improvement of SPA-FPN can reach 4.4%. This proves the effectiveness of our proposed SPA-FPN module for these challenges (3) and (4) in remote sensing images.

*2) Ablation Study for Presented Method:* Ablation experiments are accomplished to verify the influence of various elements of the presented approach, involving the HASP and SPA-FPN. The ResNet-50+FPN is the baseline. We embed the PAC on each upsampling stage of the FPN and append the HASP at the end of baseline's encoder.

As presented in Table II, the baseline attains 59.4% mIoU. Besides, incorporating the HASP into FPN provides a 2.2% mIoU enhancement. A mIoU of 63.3% is attained using the SPA-FPN. Moreover, the baseline is improved by a large margin by integrating HASP and FPN-PAC into the network, achieving 64.1% mIoU. The experimental results show that HASP and SPA-FPN provide significant improvement in the imbalance scenario and numerous small objects.

*3) Ablation Study for Loss Functions:* In order to make a clear contrast, the common cross entropy of the other ablation experiments is utilized to further verify the proposed losses. The significance of various types of losses is verified and the results are given in Table III. It can be concluded that the focal loss based on distribution considerably enhanced the foreground's segmentation accuracy. Both focal loss and dice loss significantly improved the segmentation accuracy. Finally, it resulted in the most excellent result of 65.9% mIoU via the proposed three loss functions. The proposed boundary-based loss improved mIOU by 0.6%.

*4) Ablation Study for HASP:* In order to assess the performance gain that can be attained via the presented HASP, they are not incorporated into the network and the accuracy of the baseline (ResNet-50+FPN) is evaluated. We find that its mIOU is 59.4%. Directly utilizing the baseline networks cannot fulfill the precision requirements. After attaining the mentioned accuracy, all the presented blocks are introduced into the network and the accuracy is evaluated. The detailed performance gain attained by each block is presented in Table IV. In particular, we first add the CW into the baseline. It can be found that the mIOU scores increase suddenly from 59.4% to 61.2%, demonstrating the importance of the CW. In order to determine the performance gain that can be attained by the SW, this module is added to the

TABLE VI
COMPARISONS WITH THE STATE-OF-THE-ART ON VAIHINGEN TEST SET

| Method | Backbone | $F_1$ per category(%) | | | | | mean $F_1$ | OA(%) | mIoU(%) |
|---|---|---|---|---|---|---|---|---|---|
| | | Impervious surface | Building | Low vegetation | Tree | Car | | | |
| FCN [7] | VGG-16 | 88.7 | 92.8 | 76.3 | 86.7 | 74.2 | 83.7 | 86.5 | 72.7 |
| RoteEqNet [84] | - | 89.5 | 94.8 | 77.5 | 86.5 | 72.6 | 84.2 | 87.5 | - |
| S-RA-FCN [85] | VGG-16 | 91.5 | 95.0 | 80.6 | 88.6 | 87.1 | 88.6 | 89.2 | 79.8 |
| PSPNet [11] | ResNet-101 | 92.8 | 95.5 | 84.5 | 90.0 | 88.6 | 90.3 | 90.9 | 82.6 |
| CASIA2 [86] | ResNet-101 | 93.2 | 96.0 | 84.7 | 89.9 | 86.7 | 90.1 | 91.1 | - |
| V-FuseNet [87] | - | 92.0 | 94.4 | 84.5 | 89.9 | 86.3 | 89.4 | 90.0 | - |
| DLR_9 [88] | - | 92.4 | 95.2 | 83.9 | 89.9 | 81.2 | 88.5 | 90.3 | - |
| DeepLab v3+ [72] | ResNet-101 | 92.4 | 95.2 | 84.3 | 89.5 | 86.5 | 89.6 | 90.6 | 81.5 |
| Semantic FPN [78] | ResNet-101 | 92.6 | 95.1 | 84.5 | 89.5 | 86.6 | 89.7 | 90.7 | 81.8 |
| HMANet [55] | ResNet-101 | **93.5** | 95.9 | 85.4 | 90.4 | 89.6 | 91.0 | 91.4 | 83.5 |
| HUSTW5 [89] | ResNet-101 | 93.3 | 96.1 | 86.4 | 90.8 | 74.6 | 88.2 | 91.6 | - |
| **Ours** | ResNet-101 | **93.5** | **96.2** | **86.8** | **90.9** | **90.6** | **91.6** | **91.8** | **83.8** |

Bold entities indicate optimal performance.

TABLE VII
QUANTITATIVE COMPARISON OF PERFORMANCE OF PARAMETERS SIZE AND FPS (MEASURED ON INPUT IMAGE SIZE OF $3\times896\times896$) ON iSAID VAL

| Method | Params | FPS |
|---|---|---|
| PSPNet [11] | 39.0M | 17.31 |
| DeepLab v3+ [72] | 45.0M | 7.32 |
| CCNet [68] | 34.8M | 19.14 |
| Non-local [71] | 33.6M | 19.43 |
| OCR [77] | 33.3M | 20.41 |
| Semantic FPN [78] | 30.2M | 24.33 |
| SPANet | 35.4M | 11.74 |

baseline, and the performance variation is measured. With the SW, the mIOU scores increase sharply from 59.4% to 61.1%, indicating the importance of the SW. Furthermore, CW and SW modules are integrated into HASP, which can further increase the mIOU to 61.6%.

We further compare the performance with some well-verified context aggregation methods, including ASPP [70] and PPM [11]. For a fair comparison, all modules are only plugged into the FPN architecture's top lateral branch. The details of the results are presented in Table IV. Although both "+ASPP" and "+PPM" attain superior performance, there is still an accuracy gap compared to our method.

*5) Ablation Study for SPA-FPN:* As discussed previously, aerial remote sensing images have the problem of extreme foreground–background imbalance. The current section constructs an ablation test to evaluate the mentioned point. In particular, we concentrate on the foreground objects' mIOU variation. Intuitively, the feature maps from low-level stages have a more significant resolution and can be employed for extracting detailed information.

Thus, we first embed PAC$^T$ convolution from the high stage (i.e., stage 3), and employ the encoder features to guide the upsampling of the features in the highest stage (i.e., stage 4) in the decoder. Subsequently, we embed PAC$^T$ convolution in stage 2 and stage 1 in series. The results are presented in Table V. The foreground's mIOU scores increase more significantly in the low stage. This is due to more detailed information on the low-level features.

### D. Computational Complexity

The results in Table VII provide comparative analysis of the computational complexity of the proposed SPANet, and six representative reference methods: PSPNet [11], DeepLab v3+ [70], CCNet [66], Nonlocal [69], OCR [75], and Semantic FPN [76]. The following attributes are reported: the number of parameters, and FPS for inferencing speed (measured on input image size of $3\times896\times896$). All experiments are conducted on a computer equipped with an Intel Xeon 2.10-GHz CPU, 16-GB RAM, and an RTX 2080TI GPU. Compared to classic PSPNet and DeepLab v3+, our model has less parameters and faster inferencing speed. In terms of the number of parameters and the speed of computation, our method still needs further optimization, which is the focus of our future work.

### E. Experiments on Vaihingen

In order to assess the efficiency of the SPANet, experiments are performed on the Vaihingen. As presented in Table VI, the presented SPANet indicated the optimum performance, exceeding the baseline by 1.9% in the average $F_1$. Fig. 8 presents various results on the Vaihingen dataset. Furthermore, as shown in Fig. 9, the effectiveness of our approach is also demonstrated by the fact that the gains compared to baseline are more significant on small object cars and low vegetation with a complex background.

## V. CONCLUSION

In this article, we focus on the problems of complex backgrounds, large scale differences, numerous small objects, and extreme foreground–background imbalance in remote sensing images. Therefore, we propose SPANet, an enhanced semantic segmentation network that contains two important components:

the HASP and SPA-FPN. The HASP can deal with the problem of large differences and complex backgrounds by cascaded atrous convolution kernels with various rates. SPA-FPN improves the network's perception of numerous small objects and extremely foreground–background imbalance scenes by embedding spatially adaptive convolution kernels in the decoder step by step. Furthermore, to achieve better result, SPANet use three types of loss for optimization. The proposed SPANet achieves better performance than several prevalent methods on two challenging datasets: iSAID and Vaihingen. In the feature, we will concentrate on decreasing the number of parameters and promoting the network's inference rate.

## REFERENCES


[1] M. M. Nielsen, "Remote sensing for urban planning and management: The use of window-independent context segmentation to extract urban features in Stockholm," *Comput., Environ. Urban Syst.*, vol. 52, pp. 1–9, 2015.

[2] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128774–128786, 2019.

[3] K. Nogueira et al., "Exploiting convnet diversity for flooding identification," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1446–1450, Sep. 2018.

[4] L. Lopez-Fuentes, C. Rossi, and H. Skinnemoen, "River segmentation for flood monitoring," in *Proc. IEEE Int. Conf. Big Data*, 2017, pp. 3746–3749.

[5] H. Sheng, X. Chen, J. Su, R. Rajagopal, and A. Ng, "Effective data fusion with generalized vegetation index: Evidence from land cover segmentation in agriculture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 267–276.

[6] L. Mi and Z. Chen, "Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 140–152, 2020.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[8] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[10] A. Li, L. Jiao, H. Zhu, L. Li, and F. Liu, "Multitask semantic boundary awareness network for remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art no. 5400314.

[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[12] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017. [Online]. Available: http://arxiv.org/abs/1706.05587

[13] S. Waqas Zamir et al., "iSAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 28–37.

[14] ISPRS, "ISPRS 2D semantic labeling contest Vaihingen," [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.

[17] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2006, pp. 430–443.

[18] D. Bouchard and N. Badler, "Semantic segmentation of motion capture using Laban movement analysis," in *Proc. Int. Workshop Intell. Virtual Agents*, Springer, 2007, pp. 37–44.

[19] J. Inglada, "Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features," *ISPRS J. Photogrammetry Remote Sens.*, vol. 62, no. 3, pp. 236–248, 2007.

[20] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and $k$-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[21] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 114, pp. 24–31, 2016.

[22] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[23] B. Mičušík and J. Košecká, "Semantic segmentation of street scenes by superpixel co-occurrence and 3D geometry," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops*, 2009, pp. 625–632.

[24] C. Zhang, L. Wang, and R. Yang, "Semantic segmentation of urban scenes using dense depth maps," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2010, pp. 708–721.

[25] G. Passino, I. Patras, and E. Izquierdo, "Aspect coherence for graph-based semantic image labelling," *IET Comput. Vis.*, vol. 4, no. 3, pp. 183–194, 2010.

[26] W. Yang, X. Zhang, L. Chen, and H. Sun, "Semantic segmentation of polarimetric SAR imagery using conditional random fields," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2010, pp. 1593–1596.

[27] J. Feng, X. Wang, and W. Liu, "Deep graph cut network for weakly-supervised semantic segmentation," *Sci. China Inf. Sci.*, vol. 64, no. 3, pp. 1–12, 2021.

[28] R. Fan, H. Wang, Y. Wang, M. Liu, and I. Pitas, "Graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms," *IEEE Trans. Image Process.*, vol. 30, no. 9, pp. 8144–8154, Sep. 2021.

[29] N. Yang and H. Tang, "Semantic segmentation of satellite images: A deep learning approach integrated with geospatial hash codes," *Remote Sens.*, vol. 13, no. 14, 2021, Art. no. 2723.

[30] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1857–1866.

[31] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[32] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding deep learning techniques for image segmentation," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 1–35, 2019.

[33] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, 2021, Art. no. 1224.

[34] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, May 7–9, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[36] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst. 29: Annu. Conf. Neural Inf. Process. Syst.*, Dec. 5–10, 2016, pp. 379–387. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/577ef1154f3240ad5b9b413aa7346a1e-Abstract.html

[37] Y. Lu, Y. Chen, D. Zhao, and J. Chen, "Graph-FCN for image semantic segmentation," in *Proc. Int. Symp. Neural Netw.*, Springer, 2019, pp. 97–105.

[38] B. Singh, H. Li, A. Sharma, and L. S. Davis, "R-FCN-3000 at 30FPS: Decoupling detection and classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1081–1090.

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[40] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," *CoRR*, vol. abs/1506.04579, 2015. [Online]. Available: http://arxiv.org/abs/1506.04579

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[42] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807.

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2015, pp. 234–241.

[44] O. Csillik and M. Belgiu, "Cropland mapping from Sentinel-2 time series data using object-based image analysis," in *Proc. 20th AGILE Int. Conf. Geographic Inf. Sci. Societal Geo-Innov. Celebrating*, Wageningen, The Netherlands, 2017, pp. 9–12.

[45] T. Blaschke et al., "Geographic object-based image analysis–towards a new paradigm," *ISPRS J. Photogrammetry Remote Sens.*, vol. 87, pp. 180–191, 2014.

[46] L. Zhang, A. Li, X. Li, S. Xu, and X. Yang, "Remote sensing image segmentation based on an improved 2-D gradient histogram and MMAD model," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 58–62, Jan. 2015.

[47] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[48] Q. He, X. Sun, Z. Yan, B. Li, and K. Fu, "Multi-object tracking in satellite videos with graph-based multitask modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 2, pp. 1–13, Feb. 2022, Art. no. 5619513.

[49] Q. He, X. Sun, Z. Yan, and K. Fu, "DABNet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, doi: 10.1109/TGRS.2020.3045474.

[50] R. Guo et al., "Pixel-wise classification method for high resolution remote sensing imagery using deep neural networks," *ISPRS Int. J. Geo Inf.*, vol. 7, no. 3, p. 110, 2018, doi: 10.3390/ijgi7030110.

[51] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResuNet-A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020.

[52] J. Wang, L. Shen, W. Qiao, Y. Dai, and Z. Li, "Deep feature fusion with integration of residual connection and attention model for classification of VHR remote sensing images," *Remote Sens.*, vol. 11, no. 13, 2019, Art. no. 1617.

[53] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022, doi: 10.1109/TGRS.2021.3065112.

[54] Q. He, X. Sun, W. Diao, Z. Yan, D. Yin, and K. Fu, "Transformer-induced graph reasoning for multimodal semantic segmentation in remote sensing," *ISPRS J. Photogrammetry Remote Sens.*, vol. 193, pp. 90–103, 2022.

[55] H. Li et al., "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[56] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. 4th Int. Conf. Learn. Representations*, San Juan, Puerto Rico, May 2–4, 2016. [Online]. Available: http://arxiv.org/abs/1511.07122

[57] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in," *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2017, pp. 1743–1751.

[58] J. Chen, X. Wang, Z. Guo, X. Zhang, and J. Sun, "Dynamic region-aware convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8060–8069.

[59] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.

[60] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9300–9308.

[61] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "CondConv: Conditionally parameterized convolutions for efficient inference," *Adv. Neural Inf. Process. Syst. 32: Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 1305–1316. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/f2201f5191c4e92cc5af043eebfd0946-Abstract.html

[62] Y. Zhang, J. Zhang, Q. Wang, and Z. Zhong, "DYNet: Dynamic convolution for accelerating convolutional neural networks," *CoRR*, 2020. [Online]. Available: https://arxiv.org/abs/2004.10694

[63] H. Su, V. Jampani, D. Sun, O. Gallo, E. G. Learned-Miller, and J. Kautz, "Pixel-adaptive convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11158–11167. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Su_Pixel-Adaptive_Convolutional_Neural_Networks_CVPR_2019_paper.html

[64] J. Hou, Z. Guo, Y. Wu, W. Diao, and T. Xu, "BSNet: Dynamic hybrid gradient convolution based boundary-sensitive network for remote sensing image segmentation," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–22, 2022, doi: 10.1109/TGRS.2022.3176028.

[65] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2019, pp. 3141–3149.

[66] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[67] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9166–9175.

[68] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.

[69] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[70] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[71] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6877–6886.

[72] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.

[73] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[74] Z. Huang, Y. Wei, X. Wang, W. Liu, T. S. Huang, and H. Shi, "AlignSeg: Feature-aligned segmentation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 550–557, Jan. 2022.

[75] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 173–190.

[76] H. Liu et al., "An end-to-end network for panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6165–6174.

[77] X. Li et al., "Semantic flow for fast and accurate scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 775–793.

[78] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4095–4104.

[79] R. Niu, X. Sun, Y. Tian, W. Diao, Y. Feng, and K. Fu, "Improving semantic segmentation in aerial imagery via graph reasoning and disentangled learning," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–18, 2022, doi: 10.1109/TGRS.2021.3121471.

[80] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034, doi: 10.1109/ICCV.2015.123.

[81] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.

[82] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 96–107, 2018.

[83] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[84] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 78–95, 2018.

[85] N. Audebert, B. Le Saux, and S. Lefévre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 140, pp. 20–32, 2018.

[86] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 135, pp. 158–172, 2018.

[87] Y. Sun, Y. Tian, and Y. Xu, "Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning," *Neurocomputing*, vol. 330, pp. 297–304, 2019.

**Jianlong Hou** (Member, IEEE) received the B.Sc. degree from Xidian University, Xi'an, China, in 2018. He is currently working toward the Ph.D. degree with the University of Chinese Academy of Sciences, Beijing, China, and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include computer vision, pattern recognition, and remote sensing image processing.

**Zhi Guo** (Member, IEEE) received the B.Sc. degree from Tsinghua University, Beijing, China, in 1998, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2003.

He is a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

**Yingchao Feng** (Member, IEEE) received the B.Sc. degree from Xidian University, Xi'an, China, in 2017, and the Ph.D. degree from Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2022.

He is an Assistant Researcher with Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and deep learning, especially on object detection, semantic segmentation, and remote sensing.

**Youming Wu** (Member, IEEE) was born in Nanchang, Jiangxi, China, in 1990. He received the B.S. degree in electronics engineering from Beihang University, Beijing, China, in 2013, where he also received the Ph.D. degree in signal and information processing, in 2020.

He is currently an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include remote sensing image interpretation, and the improvement of spaceborne SAR image quality.

**Wenhui Diao** (Member, IEEE) received the B.Sc. degree from Xidian University, Xi'an, China, in 2011, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2016.

He is currently an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image analysis.