BAS^{4}Net: Boundary-Aware Semi-Supervised Semantic Segmentation Network for Very High Resolution Remote Sensing Images

Xian Sun, *Senior Member, IEEE,* Aijun Shi, *Student Member, IEEE,* Hai Huang, *Member, IEEE,* and Helmut Mayer, *Member, IEEE*

***Abstract*—Semantic segmentation is a fundamental task in remote sensing image understanding. Recently, Deep Convolutional Neural Networks (DCNNs) have considerably improved the performance of the semantic segmentation of natural scenes. However, it is still challenging for Very High Resolution (VHR) remote sensing images. Due to the large and complex scenes as well as the influence of illumination and imaging angle, it is particularly difficult for the existing methods to accurately obtain the category of pixels at object boundaries—the so-called boundary blur. We propose a framework called Boundary-Aware Semi-Supervised Semantic Segmentation Network (BAS^{4}Net), which obtains more accurate segmentation results without additional annotation workload, especially at the object boundaries. The Channel-weighted Multi-scale Feature (CMF) module balances semantic and spatial information and the Boundary Attention Module (BAM) weights the features with rich semantic boundary information to alleviate the boundary blur. Additionally, to decrease the amount of difficult and tedious manual labeling of remote sensing images, a discriminator network infers pseudo-labels from unlabeled images to assist semi-supervised learning and further improves the performance of the segmentation network. To validate the effectiveness of the proposed framework, extensive experiments have been performed on both the ISPRS Vaihingen dataset and the novel remote sensing dataset AIR-SEG with more categories and complex boundaries. The results demonstrate a significant improvement of accuracy especially on boundaries and for small objects.**

***Index Terms*—boundary-aware, semi-supervised learning, fully convolutional networks, remote sensing images, semantic segmentation**

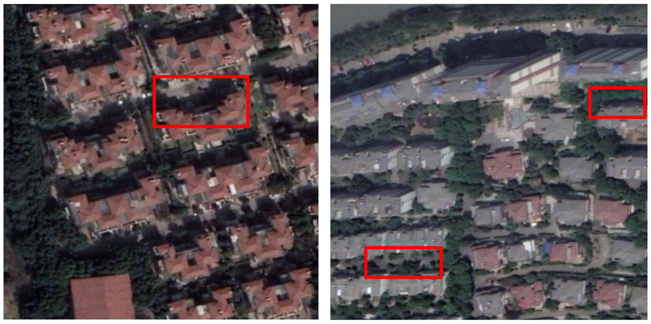


Fig. 1. Visualization of VHR remote sensing images. Particularly in the red boxes, one can see that the object boundary is blurred and there are many scattered small objects.

I. Introduction

NOWADAYS, massive amounts of satellite remote sensing images with Very High Resolution (VHR) are obtained every day. Semantic segmentation is an essential task in remote sensing image understanding to make use of the data. VHR remote sensing images are a significant source of Land Use and Land Cover (LULC) information [1], which have a variety of applications, e.g., in environmental management [2] [3], urban planning [4], and traffic management. While there has been considerable progress in the semantic segmentation of natural scenes, the semantic segmentation of VHR remote sensing images is still challenging because they usually consist of larger and more complex scenes.

Earlier methods had made some progress employing hand-crafted features for machine learning, yet they had a high computational complexity and a poor robustness. With the advent of deep learning, Deep Convolutional Neural Networks (DCNNs) [5] [6] have led to great progress in the semantic segmentation of natural scenes. Most current segmentation methods [7]–[11] are based on Fully Convolutional Networks (FCNs) [7]. Although down-sampling operations in FCNs expand the receptive fields, they also reduce the spatial resolution and degrade the spatial location information in final feature maps. To obtain a high-resolution output, dilated (atrous) convolution [12] is used to generate a high-resolution feature map and to expand the receptive fields. Some methods use a multi-scale context module with dilated convolution to expand the receptive fields and to obtain contextual information, while large receptive fields lead to the reduction of detail. While the above methods have led to great progress on images of natural scenes, they lose much spatial information and cause blur on the boundaries between objects.

VHR remote sensing images consist of large and often complex scenes with heterogeneous objects. Additionally, due to lighting conditions and imaging angles, occlusions and

This work was supported by the National major project on high resolution earth observation system GFZX0404120405. (Corresponding author: Xian Sun.)

X. Sun and A. Shi 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), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China and also with the University of Chinese Academy of Sciences, Beijing 100190, China and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: sunxian@mail.ie.ac.cn; shiaijun18@mails.ucas.ac.cn).

H. Huang and H. Mayer are with the Institute for Applied Computer Science, Universität der Bundeswehr München, Werner-Heisenberg-Weg 39, D-85577 Neubiberg, Germany (e-mail: hai.huang@unibw.de; helmut.mayer@unibw.de).

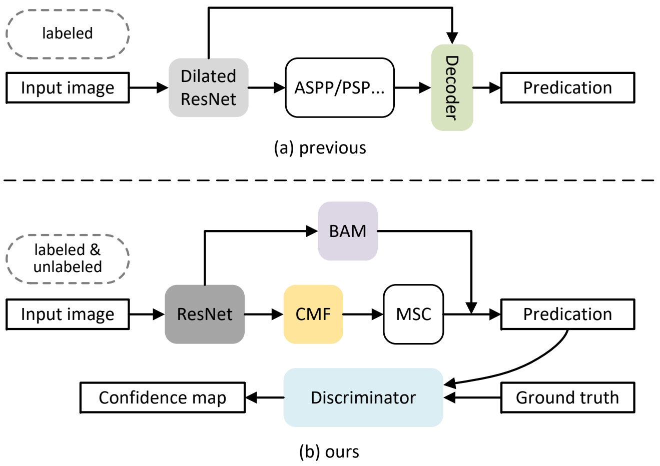


Fig. 2. Pipelines of semantic segmentation. (a) Most previous methods are trained only with labeled data, use dilated convolution in the backbone and a decoder to recover spatial information. (b) The proposed framework can be trained with both labeled and unlabeled data. We use the CMF module to merge feature maps and add the BAM to weight more on the boundaries. Additionally, we use a discriminator as an auxiliary network to train in a semi-supervised manner.

shadows lead to problems at the boundaries of objects. This is especially true, when there are multiple separated objects of different sizes in a VHR remote sensing image, such as individual cars and trees [13], where pixels on the boundary take up a large proportion of the total. Fig. 1 demonstrates boundary blur and that there are many small objects in the VHR remote sensing images. Some approaches use an encoder-decoder structure [9] to recover spatial information, but the boundary accuracy is not very good. To improve the accuracy of the boundary, we fuse multi-scale features weighted by channels inspired by a pyramid structure and an attention mechanism, to balance semantic and spatial information. Another way to address the boundary problem is by adding a new branch to detect the boundary supervised by an additional loss particularly for boundaries [13]–[16]. The human visual system is assumed to pay different attention to each part of an image and to be more sensitive to the shapes of the objects. As the pixels at the boundary are more likely to be misclassified, and considering that the boundary pixels occupy only a small proportion of an image, we think that the model should pay more attention to the boundary information. Unlike previous work, a Boundary Attention Module (BAM) is introduced in this work, which weights the feature map where contains rich boundary information, without supervision based on accurate boundaries.

Another possible solution to improve VHR remote sensing image segmentation is to learn more discriminative feature representations from more finely labeled data through a deeper network. Semantic segmentation requires pixel-wise labeling. However, since VHR remote sensing images consist of complex scenes with an irregular distribution of ground objects, it is difficult and time-consuming to annotate them with dense labels. To reduce the workload of the annotation, some previous work [17]–[19] utilize weakly labeled and unlabeled images to perform weakly-supervised and semi-supervised semantic segmentation. Several approaches employ Generative Adversarial Networks (GANs) [20] for semi-supervised semantic segmentation and have achieved a high effectiveness using a few pixel-wise labeled images and a large amount of unlabeled data. Inspired by this work, we use a discriminator network to obtain a more accurate supervision signal for the unlabeled data, which produces a dense prediction inferring regions close to the ground truth (GT) as pseudo-labels.

In this article, we propose the Boundary-Aware Semi-Supervised Semantic Segmentation Network (BAS^{4}Net) with a focus on the object boundaries in complex scenes. As shown in Fig. 2 (b), a Channel-weighted Multi-scale Feature (CMF) module fuses feature maps, followed by a Multi-Scale Context (MSC) module to capture multi-scale contextual information, and the BAM integrates the semantic information related to the object boundaries. The discriminator network assists by generating pseudo-labels from unlabeled images for semi-supervised learning and further improves the performance of the segmentation. We conduct a series of experiments on the ISPRS Vaihingen dataset to evaluate the performance of the framework. The contributions of this article are summarized as follows.

1. We present a semantic segmentation framework called BAS^{4}Net for VHR remote sensing images, which can effectively semantically segment objects in the image of a complex scene in a semi-supervised manner. Compared with other related approaches, our work focuses on object boundaries. It learns additional information about object boundaries from unlabeled images without extra boundary annotation.
2. To obtain more boundary related semantic information, a CMF module fuses the semantic and spatial information by quantitative calculation, and the BAM weights the feature maps in the spatial position where semantic boundary information is rich, which alleviates the problem of boundary blur especially for small objects.
3. Due to the large effort of annotating VHR images, we designed a discriminator network for semi-supervised learning, which assists the segmentation network to infer trusted regions in the predictions of unlabeled images as pseudo-labels. Additional feature representations can be learned from unlabeled images without complex boundary labeling.

To validate the effectiveness of our framework, we have created a challenging remote sensing semantic segmentation dataset called AIR-SEG with more categories and complex boundaries and extended our experiments with it. The comprehensive experimental results on the public benchmark and our dataset show that our framework has led to considerable improvements. The dataset will be made publicly available.

This article consists of four sections. Related work is presented in Section II. In Section III, we introduce our novel semi-supervised semantic segmentation framework in detail, including the CMF module, the BAM, the discriminator network, and the loss function. Section IV presents the results and the analysis of experiments. Finally, we conclude the article and discuss future work in Section V.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3021098, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

II. RELATED WORK

In this section, we review work related to semantic segmentation based on deep learning concerning three different aspects: FCN-based semantic segmentation, boundary improved semantic segmentation, and semi-supervised semantic segmentation.

A. FCN-based Semantic Segmentation

FCN is an end-to-end trainable neural network that combines appearance information from a shallow layer and semantic information from a deep layer to produce accurate and detailed segmentations, in which the last fully connected layer in DCNNs is replaced with a convolutional layer [7]. Many of the state-of-the-art approaches are based on FCNs. However, downsampling operations in original FCNs lead to the reduction of spatial resolution and spatial location information, which results in the inaccurate prediction of small objects and the boundaries of objects. To obtain a high-resolution output with spatial information also in low-level layers, Badrinarayanan *et al.* [8] use unpooling in the decoder to upsample the feature map to maintain the high-frequency details in the segmentation. Ronneberger *et al.* [9] proposed an encoder-decoder structure with skip-connections between encoder layers with high semantic information and low-level layers with rich spatial information. But these approaches still lose some details. To this end, Chen *et al.* [12] introduced the dilated convolution to expand the receptive fields while maintaining high resolution. However, the dilated convolution has some drawbacks, such as grid effect, less local detail information, and a large computational complexity. Thus Wu *et al.* [21] replaced the dilated convolution in the backbone with a novel joint upsampling module called the Joint Pyramid Upsampling (JPU) to reduce the computation complexity, which formulates the task of extracting high-resolution feature maps into a joint upsampling problem. Subsequently, many improved variants based on dilated convolution have been derived. Atrous Spatial Pyramid Pooling (ASPP) [22] module captures multi-scale contextual information from the final convolutional feature map in the backbone and image-level features encoding global context. Furthermore, Chen *et al.* [11] extend DeepLabv3 [22] by adding a simple and effective decoder module to connect the final feature maps with low-level feature maps, which refines the segmentation results by recovering spatial information, especially along object boundaries. Inspired by the Feature Pyramid Network (FPN) [23] in object detection, some approaches design a feature pyramid structure to combine high-resolution feature maps with rich spatial information and low-resolution feature maps with more semantic information.

In remote sensing, Chen *et al.* [24] proposed a semantic segmentation framework based on FCNs with symmetrical dense-shortcut connections to solve the problems of block effects and ”salt and pepper” noise in large-scale remote sensing images. Liu *et al.* [25] designed an end-to-end self-cascaded network which improves the labeling coherence with sequential global-to-local context aggregation. DCNNs are usually not capable of processing a whole remote sensing image given its huge size, to overcome such limitation, Nogueira *et al.* [26] employ a multi-context paradigm without increasing the number of parameters while defining the best patch size at training time. Sun *et al.* [27] propose ensemble training and inference strategies to suppress the adverse consequences of the structural stereotype in encoder-decoder models. In order to select more discriminative features for classification, Luo *et al.* [28] introduce the channel attention mechanism to a deep FCN for aerial images, which can weigh the semantic and spatial location information in the adjacent-level concatenated feature maps.

The above approaches recover some spatial information, but the accuracy at the boundary is still not good enough.

B. Boundary improved Semantic Segmentation

Several works have made great progress to improve the accuracy of the boundaries between objects. Some previous works improve the architecture of DCNNs by adding a new branch to detect the boundary supervised by accurate boundary information. For example, in natural scenes, Takikawa *et al.* [15] design a Gated Convolution Layer (GCL) to focus on boundary information and add a shaped stream using GCL to capture the feature maps particularly relevant to the boundaries. To tackle the problem of intra-class inconsistency and inter-class indistinction, Yu *et al.* [16] propose a Border Network to make the bilateral features of boundary distinguishable with deep semantic boundary supervision, which uses a bottom-up structure to fuse the features in multi-level layers containing different information. In remote sensing, Li *et al.* [14] employ a superpixel-enhanced region module in the framework to focus on the accuracy and coherence of the boundaries, which utilizes superpixel segmentation to measure the similarities between regions and proposes a region loss to emphasize superpixel segmentation results throughout the task. Marmanis *et al.* [13] integrate a boundary detector into the model by utilizing DSM information to capture semantic boundaries. The above approaches all add an additional loss on the boundary supervised by accurate boundary information. Marin *et al.* [29] design a content-adaptive downsampling operation to extract more pixels from the boundary of the object. Unlike per-pixel loss, the Semantic Encoding Loss (SE-loss) proposed in EncNet [30] uses global context information to treat small objects equally with large objects, which enforces the learning of semantic context and improves the performance of small objects. However, it greatly increases the complexity of the semantic segmentation of remote sensing images due to the large and complex scenes and the heavy manual labeling cost.

C. Semi-supervised Semantic Segmentation

While labeled data is scarce, lots of unlabeled data is available. Consequently, some semi-supervised and weakly-supervised learning methods have emerged using a small amount of pixel-wise labeled data and a large amount of unlabeled data or some weakly-labeled data to reduce the annotation effort. In recent years, some methods have made progress in weakly-supervised object detection for remote sensing

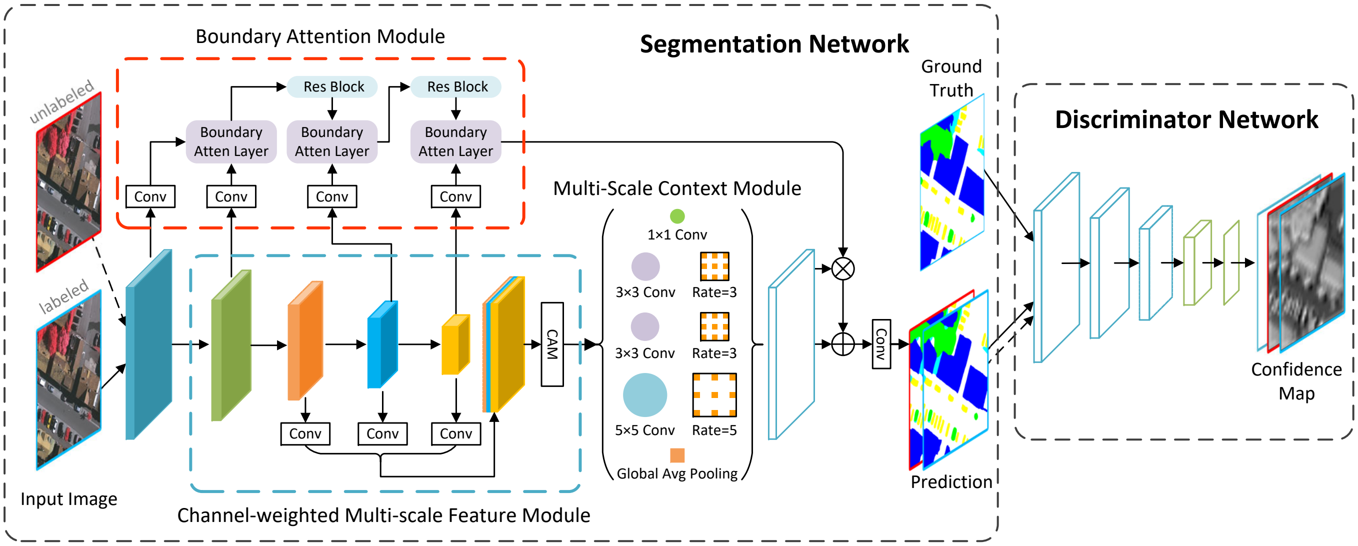


Fig. 3. The overall architecture of the proposed framework consists of the segmentation network and the discriminator network. The segmentation network is mainly composed of the CMF module, the MSC module and the BAM. The CMF module fuses the semantic and spatial information by quantitative calculation, followed by the MSC module to capture multi-scale contextual information, the BAM is designed to focus the feature map, to where it contains rich semantic boundary information.

images. Cheng *et al.* [31] proposed a unified framework to generate and select high-quality proposals, which combines selective search [32] and a Gradient-weighted Class Activation Mapping [33] to generate more proposals with higher quality, and then chooses many confident positive proposals and only class-specific hard negatives to train more effective by up-weighting the losses of discriminative negative proposals. To provide high-quality initial samples and obtain optimal object detectors with only image-level annotations, Yao *et al.* [34] proposed a dynamic curriculum learning strategy with an entropy-based criterion and designed an effective instance-aware focal loss function, which can progressively learn the object detectors by feeding training images with increasing difficulty that matches current detection ability. To avoid selecting only one top-scoring proposal that usually results in learning a suboptimal object detector, Feng *et al.* [35]proposed a novel end-to-end progressive contextual instance refinement method by leveraging both local and global context information for weakly supervised object detection.

With the rapid development of GANs, Luc *et al.* [36] applied adversarial learning by means of a GAN in semantic segmentation, using the segmentation network as a generator. Subsequently, Hung *et al.* [18] designed a discriminator based on FCNs to make a dense prediction, where unlabeled data are trained with a self-taught scheme to further improve the accuracy of semi-supervised segmentation. However, the discriminator network has shallow layers and downsamples five times. Souly *et al.* [19] use GANs in a different manner, employing the discriminator as the segmentation network and using the generator to generate fake images to increase the number of training samples. This ensures that the segmentation network can learn more features. However, the above approaches are designed for natural scenes. They do not perform very well on VHR remote sensing images with complex scenes and ground object distribution.

III. Methods

A. Problem Setup

The task of semi-supervised semantic segmentation can be described as follows: Given an image from a set of images as input,

$$X = \{X_{L_1}, ..., X_{L_m}; X_{U_1}, ..., X_{U_n}\} \tag{1}$$

where X_{L_m} and X_{U_n} represent m labeled and n unlabeled images respectively, the aim is to learn a segmentation $Seg(\cdot)$ model with both labeled and unlabeled images to produce a dense prediction O by assigning a predefined category $c = (c_1, ..., c_k)$ to each pixel in the image. The discriminator network $Dis(\cdot)$ generates a confidence map C trained by labeled data to infer the area close to the ground truth (GT) G, which aims to generate pseudo-labels for unlabeled images.

$$O = Seg(X) \tag{2}$$

$$C = Dis(O, G) \tag{3}$$

The backbone extracts the feature maps from the input image, followed by the CMF module to merge feature maps from different layers weighted by the channel attention matrix α. The output feature map of the CMF module is:

$$P_{Mj} = \sum_{i=1}^{k} \alpha_{ji} P_i + P_j, \tag{4}$$

where P is the feature map concatenated from last three layers, P_i and P_j represent the feature maps of i^{th} and j^{th} channel, respectively, α_{ji} denotes the influence of i^{th} channel on j^{th} channel, and k represents the total number of channels. Next, a MSC module (like ASPP [22]) $MSC(\cdot)$ captures multi-scale contextual information on the feature map P_M and

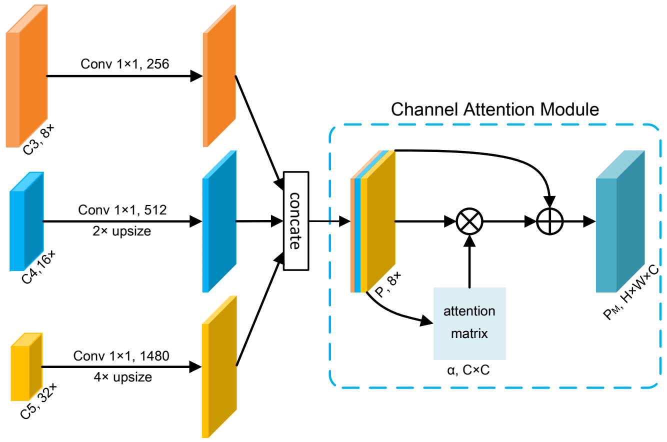


Fig. 4. Channel-weighted multi-scale feature module.

generates the feature map $F \in \mathbb{R}^{C\times H\times W}$ with rich contextual information.

$$F = MSC(P_M) \tag{5}$$

The novel BAM generates a spatial boundary attention map β focusing on the semantic and spatial information on the object boundaries. We perform an element-wise multiplication $\odot$ and sum operation with the feature map F at each pixel (i, j) to obtain the final feature map $F_o \in \mathbb{R}^{C\times H\times W}$ as follows:

$$F_o^{(i,j)} = F^{(i,j)} \odot \beta^{(i,j)} + F^{(i,j)}, \tag{6}$$

where $\beta^{(i,j)}$ represents the weight value at pixel (i, j).

Next, a softmax layer generates a class probability map O as the output of segmentation network, which contains the probability for each pixel (i, j) if it belongs to a certain semantic category c.

$$O = \underset{c,i,j}{\arg\max}\left(\text{softmax}\left(F_o^{(i,j)}\right)\right) \tag{7}$$

The confidence map C generated by the discriminator network aims to infer the area close to the GT G. From it we can get the pseudo-labels L_U^* for unlabeled images for semi-supervised learning by

$$L_U^* = O^{(i,j)} * I\left(C^{(i,j)} \geq \tau\right), \tag{8}$$

where $I(\cdot)$ is a indicator function with a threshold τ.

We present a semi-supervised semantic segmentation framework for VHR remote sensing images, which can learn more information from labeled and unlabeled images than previous work, especially related to object boundaries. As shown in Fig. 3, our proposed framework consists of two networks: The main network is the segmentation network composed of two modules. The channel-weighted multi-scale feature module is the first module. It aims to balance the semantic and spatial information by merging feature maps from different layers of the backbone weighted by channel attention. The second module is the boundary attention module. It captures additional information for the object boundaries. Finally, the auxiliary network is the discriminator network designed for semi-supervised learning, which further improves the performance of the segmentation network by inferring pseudo-labels for unlabeled images.

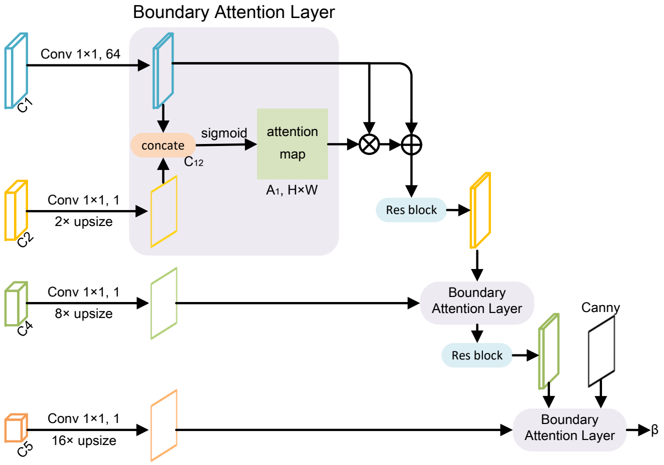


Fig. 5. Boundary attention module.

B. CMF Module

The shallow layers of DCNNs extract rich spatial information but not so much semantic information, while the higher layers contain much semantic information due to larger receptive fields but have a lower resolution. To balance semantic and spatial information and to enlarge the resolution, we use the CMF module to fuse feature maps, as shown in Fig. 4. The backbone of the segmentation network generates five feature maps $\{C1, C2, C3, C4, C5\}$ from layers named Conv1, Res1, Res2, Res3 and Res4. They range from low- to high-level with strides of $\{2, 4, 8, 16, 32\}$ pixels with respect to the input image. In CMF, we merge three feature maps: $\{C3, C4, C5\}$. We use a convolution layer with a stride of 1 to reshape their channels, where the $Conv_{1\times1}$ conducts dimension reduction on the channels. $C5$ and $C4$ are upsampled to adapt the resolution to $C3$ using bilinear interpolation. We concatenate the three feature maps and thus, obtain a feature map P with rich spatial location information.

In existing work, different features are merged through direct concatenation or summation. It is, however, difficult for dense prediction to balance the high-level semantic information and low-level spatial information. Therefore, we attach the channel attentional module (CAM) proposed in [37]. It generates a channel attention matrix α to determine the importance of different channels of the feature map P, which is different from DeepLab V3+ [11] that only employs the last feature map.

C. Boundary Attention Module

As VHR remote sensing images show a complex scene and distribution of ground objects, information related to boundaries is important for an accurate segmentation. For an improved handling of the object boundaries, we propose a boundary attention module. As shown in Fig. 5, the structure of the BAM mainly contains residual blocks (RBs) [38], boundary attention layers, $Conv_{1\times1}$, concatenation and upsampling operations. It takes the feature maps $\{C1, C2, C4, C5\}$ as input. $C1$ contains rich spatial location information, which is reshaped by $Conv_{1\times1}$ and then concatenated with $C2$ reshaped

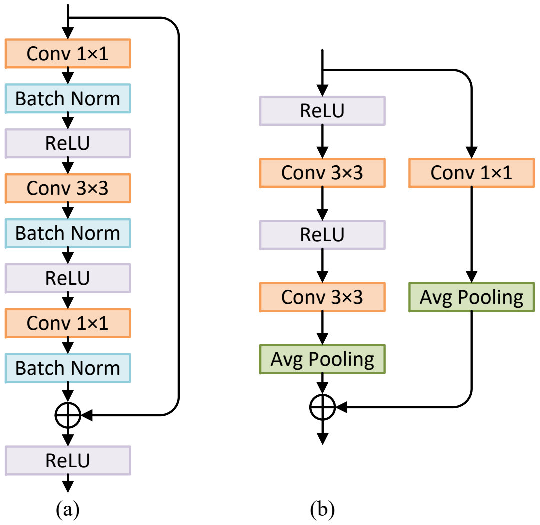


Fig. 6. Residual blocks (RBs): (a) The RB in the segmentation network. (b) The RB in the discriminator network. Conv 1×1 and Conv 3×3 denote the convolution layer with a kernel size of 1×1 and 3×3, respectively. Batch Norm means the batch normalization layer and Avg. Pooling is the average pooling operation.

by $Conv_{1\times1}$ to obtain a feature map C_{12}. Next, we perform the sigmoid operation resulting in the spatial boundary attention map $A_1 \in \mathbb{R}^{H\times W}$:

$$A_1 = \frac{1}{1+\exp(C_{12})} \tag{9}$$

The attention map computed from low-level features pays more attention to texture information. To obtain a map with more attention on the boundary rather than on the texture, we use the GCL [15]. This feature map is reshaped by the RB to join the calculation of the second boundary attention map. The structure of the RB used in the segmentation network is shown in Fig. 6 (a). The image gradient feature map obtained by the Canny edge detection operator [39] is input to the third boundary attention layer. The BAM uses three boundary attention layers to produce the final spatial boundary attention map β, which weights the regions with significant boundary information in the feature map with rich contextual information. The BAM makes the network pay more attention to the information related to the object boundaries and improves the segmentation accuracy at the boundaries.

D. Discriminator Network

The discriminator network takes the class probability map $O = \{O_l; O_u\}$ of labeled and unlabeled data predicted by the segmentation network and one-hot encoded GT G as input. The final output is a confidence map $C \in \mathbb{R}^{H\times W}$, in which the value of each pixel represents the probability that the pixel in the input image comes from GT or the segmentation prediction, that is, the similarity between the segmentation prediction and GT. The greater the value of the pixel point, the closer the segmentation prediction of the point is to the GT. This is used to infer the regions close to the GT as supervision for the unlabeled image to assist the learning of the segmentation network.

$$C = \underset{i,j}{\text{sigmoid}}\, F_d^{(i,j)}, \tag{10}$$

where $F_d \in \mathbb{R}^{H\times W}$ represents the final feature map of the input.

The discriminator network is designed based on an FCN, which consists of three residual blocks [40], followed by two convolution layers. The structure of the residual block in the discriminator is shown in Fig. 6 (b). It contains a spectral normalization convolution layer with kernel size of 3×3 and 1×1, a rectified linear unit (ReLU) [41] activation function and average pooling. The first convolution layer is in the 4th layer with 3×3 kernel size and followed by an average pooling operation for downsampling, as well as a ReLU activation function. The last convolution layer reshapes channels from 512 to 1 with $Conv_{1\times1}$. It is followed by an upsampling layer to make the resolution of the output feature map equal to the input image. Finally, we use the sigmoid layer to generate the confidence map. With it we can infer regions in the prediction sufficiently close to the GT.

E. Loss Function

We train the segmentation network and discriminator network together with labeled data, while unlabeled data is used together with the trained discriminator to assist the training of the segmentation network. To this end, we use the binary cross-entropy (BCE) loss L_D on the discriminator network and a multi-task loss L_S on the segmentation network:

$$L_D = -\sum_{h,w}\left((1-y)\log\left(1-C^O_{h,w}\right) + y\log C^G_{h,w}\right), \tag{11}$$

where $C^O_{h,w}$ and $C^G_{h,w}$ represent the confidence maps at location (h,w) obtained by the discriminator network of the prediction O and the ground truth G of the labeled image, respectively, and y denotes the label of each pixel in the input image.

$$L_S = L_{label} + \lambda_{unlabel} L_{unlabel}, \tag{12}$$

where L_{label} denotes the standard cross-entropy (CE) loss on the predicted semantic segmentation of the labeled images and $L_{unlabel}$ is the loss of the unlabeled images. Here, $\lambda_{unlabel}$ is the weight of $L_{unlabel}$ and is set to 1.

$$L_{label} = -\sum_{h,w,c} l_c(h,w)\log p_{l,c}(h,w), \tag{13}$$

where l_c is the label for class c in the one-hot encoded GT; $p_{l,c}$ is the probability of the segmentation outputs for class l_c.

$L_{unlabel}$ consists of a masked CE loss L_{S-CE} and an adversarial loss L_{adv} as in [18].

$$L_{unlabel} = L_{S-CE} + \lambda_{adv} L_{adv} \tag{14}$$

$$L_{S-CE} = \begin{cases} -\sum\limits_{h,w,c} u_c \log p_{u,c}(h,w), & \text{if } C(h,w) \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

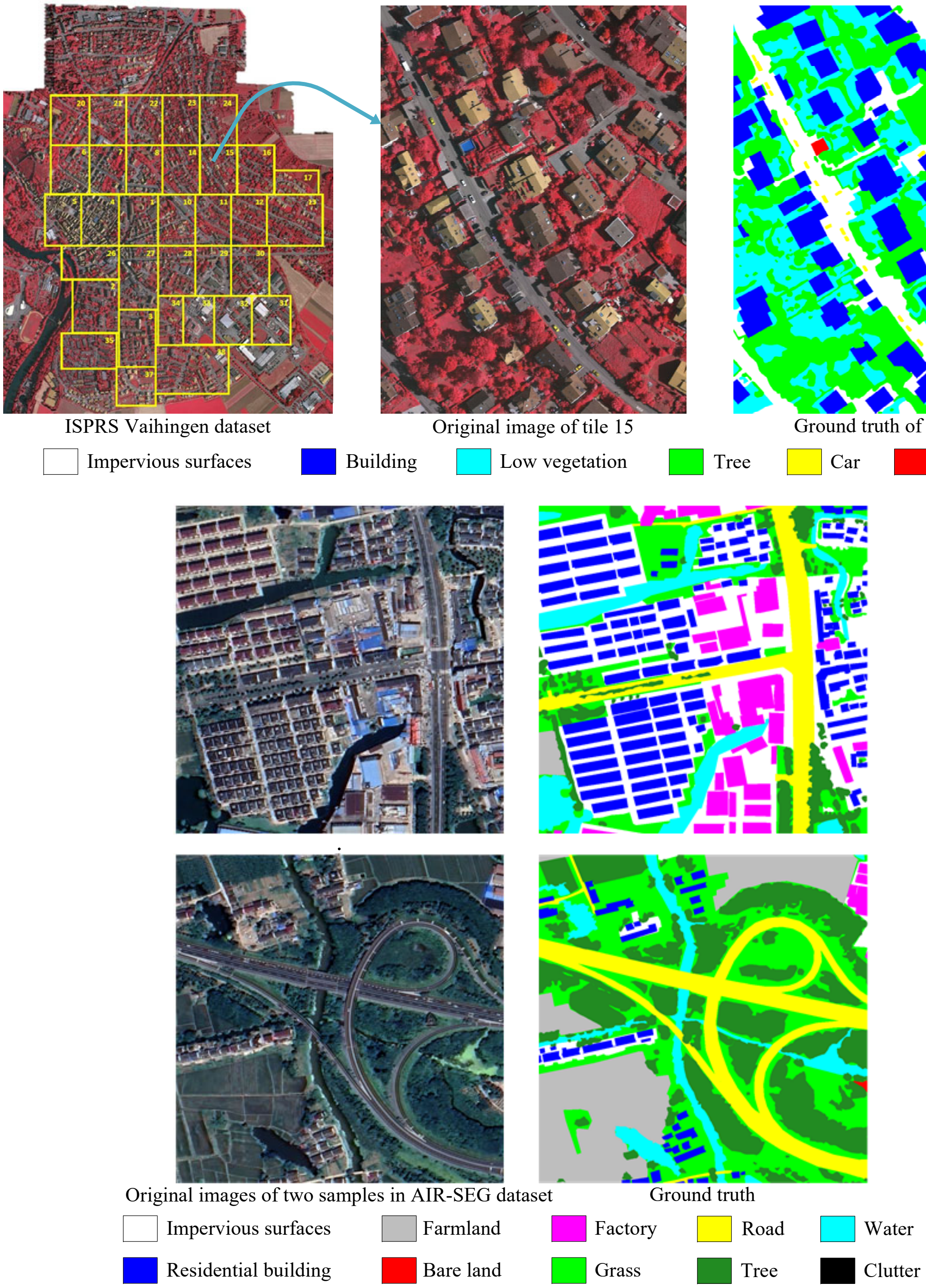


Fig. 7. ISPRS Vaihingen (top) dataset and AIR-SEG (bottom) dataset.

$$L_{adv} = -\sum_{h,w} \log C^{ou}_{h,w}, \tag{16}$$

where λ_{adv} is the weight of L_{adv} and is set to 0.01. u_c is the predicted label from the segmentation output $p_{u,c}$ for class l_c, and $C^{ou}_{h,w}$ represents the confidence map at the location (h, w) generated by the discriminator of the prediction of the unlabeled image. L_{S-CE} is only calculated on the regions where the probability value in the confidence map $C(h, w)$ is larger than the threshold τ, which is set to 0.5.

IV. Experiments

A. Datasets

Many public datasets have been published to advance semantic segmentation in remote sensing. While the INRIA Aerial Image Labeling [42] and Massachusetts Buildings [43] datasets are used to extract Buildings, the Massachusetts Roads dataset [43] only consists of the class of road. We, thus, chose the ISPRS Vaihingen dataset [44] since it contains more categories and is more difficult. We additionally built a

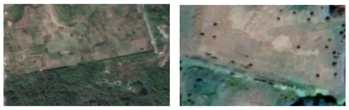
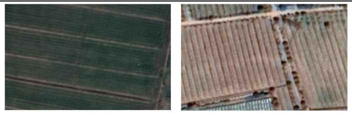
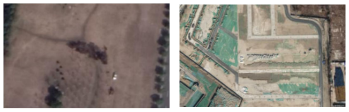
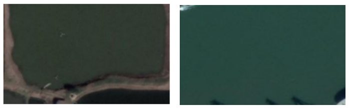
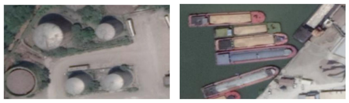
Class	Ratio (%)	Samples	Class	Ratio (%)	Samples
Impervious surfaces	16.37		Residential building	6.46	
Factory	6.57		Grass	16.05	
Farmland	14.49		Tree	15.60	
Road	7.58		Bare land	1.28	
Water	15.34		Clutter	0.26	

Fig. 8. AIR-SEG dataset, including the ratio of each class of the total pixels and image examples corresponding to each category.

challenging dataset called AIR-SEG with even more categories and consisting of complex scenes. To evaluate the effectiveness of our framework, experiments are conducted on the latter two VHR remote sensing image datasets.

ISPRS Vaihingen Dataset: The public 2D semantic labeling benchmark Vaihingen dataset is provided by the International Society for Photogrammetry and Remote Sensing [44]. It consists of high resolution true orthophoto (TOP) tiles and corresponding digital surface models (DSMs) as well as ground truth labels. As shown in Fig. 7, it contains 33 TOP tiles of size 2494×2064 with ground sample distance (GSD) of 9 cm. 16 of them are used for training and the rest for testing. Each TOP tile includes 3 spectral bands: Near Infrared (NIR), Red (R) and Green (G). The pixels are classified into 6 categories: 1. Impervious surfaces (Imf-surf), 2. Building, 3. Low vegetation (Low-veg), 4. Tree, 5. Car, and 6. Clutter. The competition for the benchmark has ended in the summer of 2018 and all training and test data are publicly available. In this work, we only use NIR-R-G images and the DSMs are neglected.

AIR-SEG Dataset: We introduce a more challenging large dataset for semantic segmentation of remote sensing images named AIR-SEG. It contains 8 bit satellite images collected from Google Earth over different geographic locations in South China. The images have different visual appearances and are composed of Red, Green, and Blue spectral bands. The AIR-SEG dataset contains 57 labeled images of size 2000×2000, and 72 unlabeled images of size 2000×2000 with a spatial resolution of 0.27 m. Since the resolution is lower than that of the ISPRS Vaihingen dataset, there are more ground objects in the image at the same image size, especially more individual small objects with less clear object boundaries. To better capture LULC information and facilitate the analysis of urban expansion and urban planning in practical application in the future, according to [45]–[47] ten categories are chosen and annotated, including 1. Impervious surfaces, 2. Factory, 3. Residential building, 4. Road, 5. Water, 6. Farmland, 7. Bare land, 8. Grass, 9. Tree, and 10. Clutter. Details and samples of the AIR-SEG dataset are shown in Fig. 8. The rules and standards of category labeling are as follows, 1. Impervious surfaces: concrete ground such as parking lot and square; 2. Factory: simple buildings with large scales, and regular shapes; 3. Residential building: roofed buildings except for factories, relatively small in size, and diverse in shape; 4. Road: asphalt cement land and other roads, such as artificial pavement, unpaved road, main road, branch road, and airport runway; 5. Water: lakes, oceans, rivers, ponds, swimming pools, etc.; 6. Farmland: the land with regular shape and texture features with or without crop coverage; 7. Bare land: bare mountain, river beach, wasteland, etc.; 8. Grass: artificial grass and natural grass, generally green; 9. Tree: shrubs, trees, forests, and other tall vegetation; 10. Clutter: objects not belonging to the above categories. Considering the resolution of the image and the type and distribution of ground objects in the area covered by the image, objects in the image with pixel size less than 10×10 are not labeled. All images are annotated at pixel-level by experts. We annotated using the Adobe Photoshop tool, first outlining the boundary of each object, and then filling in the corresponding color to the object according to the predefined category and color mapping relationship. To make sure that the training data and test data distributions approximately match, we select 29 labeled images of 57 as the training set and the rest as the test set. Finally, there are 29 labeled images and 72 unlabeled images for training and 28 labeled images for testing.

B. Evaluation Metrics and Implementation Details

We use three overall accuracy assessment metrics to quantitatively evaluate the performance of our framework, Pixel overall accuracy (OA), mean intersection over union (mIoU), and F1 score.

TABLE I
ABLATION STUDY OF THE PROPOSED FRAMEWORK ON THE ISPRS VAIHINGEN DATASET

Method	**F1 score**					**OA**	**mIoU**
	Impervious surfaces	**Building**	**Low vegetation**	**Tree**	**Car**		
Baseline	91.37	94.86	83.61	88.65	68.23	89.60	75.51
+CMF	92.67	95.49	84.07	89.28	85.99	90.57	81.26
+MSC	92.69	95.56	84.32	89.28	86.56	90.68	81.54
+BAM	92.95	95.54	84.29	89.30	88.49	90.74	82.23

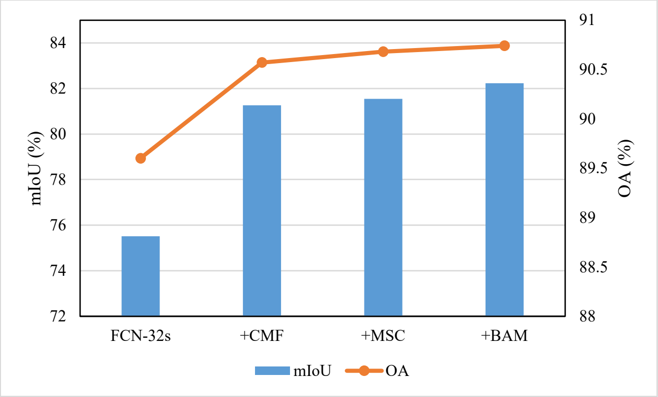


Fig. 9. Segmentation results on the ISPRS Vaihingen dataset. The accuracy is improved by the proposed modules.

TABLE II
COMPARISONS OF SEMI-SUPERVISED LEARNING WITH 1/8, 1/4, 1/2 AND FULL LABELED DATA ON THE ISPRS VAIHINGEN DATASET

Metric	**Method**	**Labeled data**			
		1/8	**1/4**	**1/2**	**Full**
OA	Fully supervised	86.40	88.78	89.62	90.74
	Semi-supervised	87.22	89.30	89.9	-
mIoU	Fully supervised	69.45	76.22	79.47	82.23
	Semi-supervised	74.19	77.53	79.62	-

Our proposed framework is implemented using the PyTorch deep learning framework. All experiments are performed on a single NVIDIA GeForce GTX 1080 Ti GPU with a memory of 12GB. Like most existing segmentation methods we use ResNet-101 pre-trained on the ImageNet [48] dataset as the backbone of the segmentation network. To update the parameters of the segmentation network, we use stochastic gradient descent (SGD), where the momentum is set to 0.9 and the weight decay to 1×10^{-4}. We initially set the learning rate of the backbone to 7×10^{-3}, which gradually decreases to 0 following the cosine decay strategy. The rest of the layers are weighted of 10 times the backbone. To update the parameters of the discriminator network, we use the Adam optimizer [49], where the learning rate is 1×10^{-4}, using a polynomial decay strategy with a power of 0.9.

C. Experiments on the ISPRS Vaihingen Dataset

1) Quantitative Analysis: We conducted a series of experiments on the ISPRS Vaihingen dataset to evaluate the performance of our framework. We first demonstrate the effectiveness of each module of our framework through ablation experiments and then show comparisons with state-of-the-art approaches.

a) Ablation Experiments: Since all training and test data are publicly available, we use the whole training images for training and the test images for validation. Because the size of the original tiles is larger than 2000×2000 pixels, they are too large for the current GPU memory. Therefore, we split all training and test images into 513×513 patches to cover a reasonable area making sure that there is enough contextual information to properly infer the category of each pixel. Training and validation tiles have 320×320 and 50×50 pixels of overlap between neighboring patches. In total, there are 1622 patches in the training dataset and 456 patches in the validation dataset. We augment the training dataset by randomly scaling (from 0.5 to 2.0) and horizontally flipping the input images. We train the models 60 epochs with a batch size of 4. After 30 epochs training with labeled data, we start semi-supervised learning, which randomly interleaves labeled data and unlabeled data. The discriminator network is updated only with labeled data. Inferences are done on a single-scale for all models. Since only less than 1% pixels in the Vaihingen dataset are labeled as clutter, we ignore this class in the experiments on this benchmark. For fairness, all models are trained with the same set of hyperparameters.

Baseline setup. We choose FCN-32s as the baseline for the ablation study to evaluate the effectiveness of each component of our framework. FCN-32s is widely used in semantic segmentation. However, the performance for small objects is not satisfying, the F1 score of car is only 68.23% (cf. Table I). The segmentation results can be seen in the third column of Fig. 14.

Contribution of CMF module. We add the CMF module after the backbone to fuse the multi-scale features with channel weights. As shown in Table I, the segmentation results for all classes have been improved by around 5.7% in terms of mIoU. Small objects have even larger improvements, for example, the car improves around 18% concerning the F1 score. It is obvious from Fig. 9 that the CMF module improves the performance of segmentation by a large margin. This shows that the CMF module balances semantic and spatial information better and recovers lost details.

Contribution of BAM. We weight the feature maps passing through the MSC module with the BAM. We can observe in Table I, that the BAM further improves the performance of the segmentation network by 0.7% concerning mIoU. Especially for small objects such as the cars, it results in an additional

TABLE III
COMPARISION WITH OTHER POPULAR METHODS ON THE ISPRS VAIHINGEN DATASET

Method	**F1 score**					**OA**	**mIoU**
	Impervious surfaces	**Building**	**Low vegetation**	**Tree**	**Car**		
FCN-32s [7]	91.37	94.86	83.61	88.65	68.23	89.60	75.51
FCN-8s [7]	92.70	95.36	83.46	88.79	87.41	90.32	81.32
DeepLab V3+ [11]	92.73	95.53	84.03	**89.44**	85.45	90.63	81.21
BAS^{4}Net (Ours)	**92.95**	**95.54**	**84.29**	89.30	**88.49**	**90.74**	**82.23**

TABLE IV
COMPARISION AT DIFFERENT THRESHOLDS OF THE BOUNDARY QUALITY ON THE ISPRS VAIHINGEN DATASET

Threshold	**Method**	**F1 score**					**OA**	**mIoU**
		Impervious surfaces	**Building**	**Low vegetation**	**Tree**	**Car**		
3px	DeepLab V3+ [11]	66.66	70.11	**55.31**	60.70	66.08	62.67	47.02
	BAS^{4}Net (Ours)	**67.23**	**71.28**	54.83	**61.01**	**68.87**	**63.15**	**48.04**
6px	DeepLab V3+ [11]	71.72	75.57	**59.72**	65.45	73.07	67.59	53.08
	BAS^{4}Net (Ours)	**72.36**	**76.45**	59.48	**65.72**	**76.03**	**68.11**	**54.24**
9px	DeepLab V3+ [11]	75.17	79.12	**63.03**	69.14	77.06	71.10	57.44
	BAS^{4}Net (Ours)	**75.88**	**79.85**	62.93	**69.38**	**79.96**	**71.64**	**58.64**
12px	DeepLab V3+ [11]	77.67	81.62	**65.53**	71.93	79.15	73.68	60.56
	BAS^{4}Net (Ours)	**78.37**	**82.21**	65.53	**72.14**	**81.97**	**74.20**	**61.77**

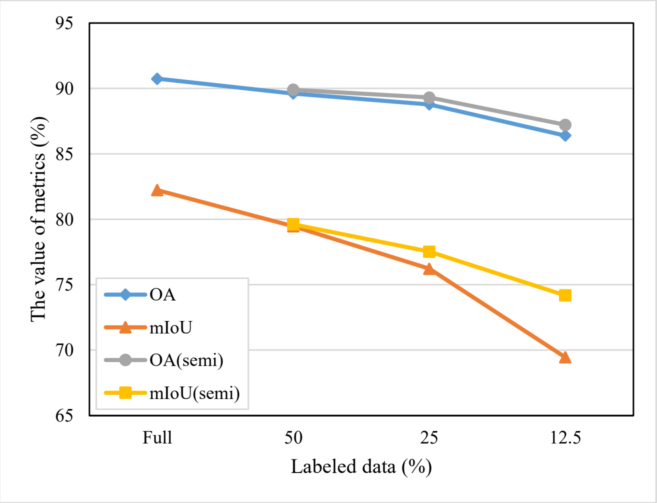


Fig. 10. Semi-supervised learning on the ISPRS Vaihingen dataset with different ratios of labeled and unlabeled data.

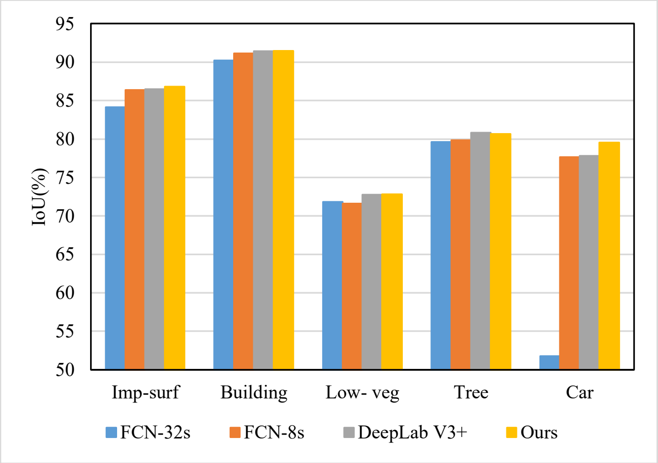


Fig. 11. Results of different methods for the different classes on the ISPRS Vaihingen dataset.

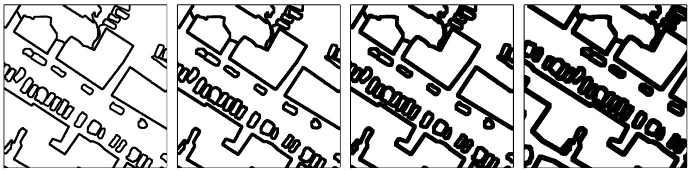

3 pixels 6 pixels 9 pixels 12 pixels

Fig. 12. Boundaries with different thresholds of widths.

improvement of around 2% for the F1 score.

Contribution of semi-supervised learning method. To assess the influence of the semi-supervised learning method in the proposed framework, we produce three datasets by randomly sampling 1/8, 1/4, 1/2 of the images in the training dataset as labeled data and the rest as unlabeled data. We conducted two experiments for each dataset: With labeled data only in a fully supervised manner and both with labeled and unlabeled data in a semi-supervised manner. Table II and Fig. 10 show that the semi-supervised learning method improves the performance significantly particularly for few labeled data and much unlabeled data without extra labeling effort.

b) Comparison with Popular Methods: We also trained the popular FCN-8s and DeepLab V3+ [11] models to evaluate the effectiveness of our framework. For fairness, the backbone of both is ResNet-101 [38]. It can be seen in Table III that our framework outperforms others. Especially the result for cars is improved by 3% F1 score compared to DeepLab V3+ [11]. It is obvious from Fig. 11 that our framework has a better performance concerning IoU especially for small objects

TABLE V
PART OF THE LIST OF ONLINE RESULTS OF THE ISPRS VAIHINGEN TEST DATASET

Method	**F1 score**					**OA**
	Impervious surfaces	**Building**	**Low vegetation**	**Tree**	**Car**	
UPB [50]	87.5	89.3	77.3	85.8	77.1	85.1
RIT_L8 [51]	89.6	92.2	81.6	88.6	76.0	87.8
CVEO [24]	90.5	92.4	81.7	88.5	79.4	88.3
ITC_B2 [52]	90.1	93.5	82.1	88.3	77.1	88.4
UFMG_4 [26]	91.1	94.5	82.9	88.8	81.3	89.4
RIT_7 [53]	91.7	95.2	83.5	89.2	82.8	89.9
V-FuseNet [54]	92.0	94.4	84.5	89.9	86.3	90.0
DLR_9 [13]	92.4	95.2	83.9	89.9	81.2	90.3
TreeUNet [55]	92.5	94.9	83.6	89.6	85.9	90.4
BKHN11 [44]	92.9	96.0	84.6	89.9	88.6	91.0
CASIA2 [25]	93.2	96.0	84.7	89.9	86.7	91.1
NLPR3 [44]	93.0	95.6	85.6	90.3	84.5	91.2
BAS^{4}Net (Ours)	93.3	95.8	85.0	90.1	**90.1**	91.3
HUSTW5 [27]	93.3	96.1	86.4	90.8	74.6	91.6
SDNF [14]	**93.4**	**97.6**	**87.4**	**91.1**	85.3	**92.2**

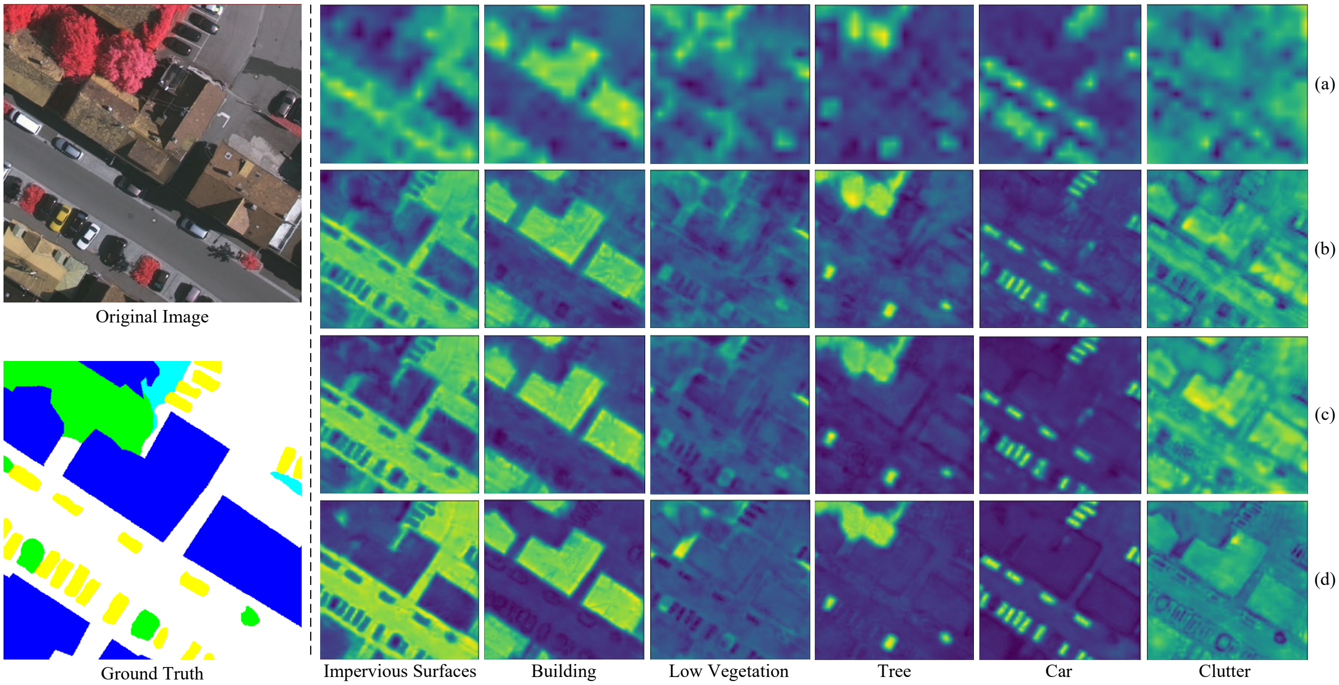


Fig. 13. Visualization results of feature maps for different methods on the ISPRS Vaihingen dataset. (a) Results of FCN-32s. (b) Results when adding the CMF module. (c) Results with additional MSC module. (d) Results when adding the BAM.

TABLE VI
DIFFERENT INFERENCE STRATEGIES ON THE ISPRS VAIHINGEN DATASET

Method	**MS**	**Flip**	**Whole**	**OA**	**mIoU**
BAS^{4}Net				90.74	82.23
BAS^{4}Net	✓			91.11	83.10
BAS^{4}Net	✓	✓		91.18	83.19
BAS^{4}Net	✓	✓	✓	91.25	83.41

like cars, because the pixels on the boundary of the car occupy a large proportion of its total pixels compared to other categories.

To better quantify the effect of our framework on the improvement at the object boundaries, we compare it with DeepLab V3+ [11] concerning the performance at the boundary on a narrow band called trimap [11] with different widths. As presented in Fig. 12, we used the thresholds 3, 6, 9, and 12 pixels. Table IV shows that we achieve around 1.2% improvement in performance in terms of mIoU and around 3% of IoU for the class car. Yet the experiments do not confirm the theoretical expectation that the narrower the trimap width is, the more the metric improves. This is probably due to labeling errors in the dataset, especially at the boundary. Overall, the experiments show that the proposed framework improves the performance at the boundaries.

c) Comparison with the State of the Art: For a deeper understanding of the performance of the proposed framework,

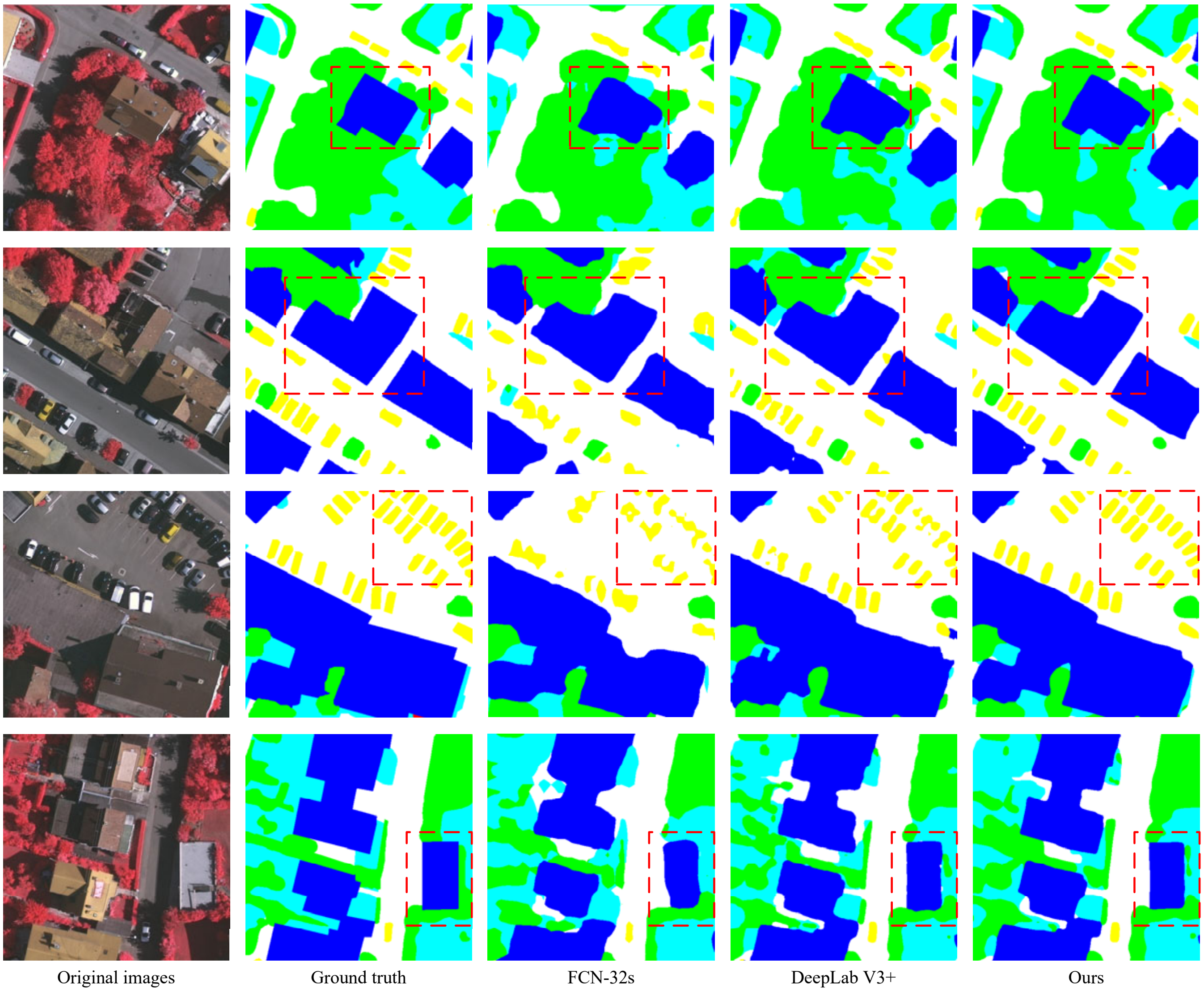


Fig. 14. Example predictions from various methods for the ISPRS Vaihingen dataset. The red dashed boxes are used to mark the regions which have been improved obviously by our method.

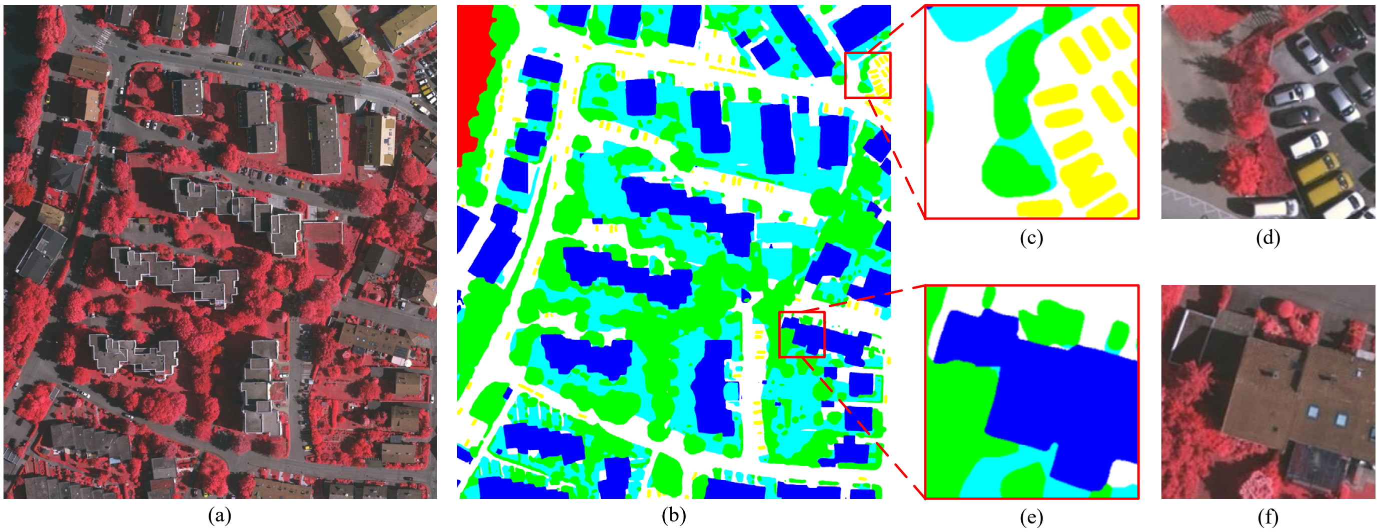


Fig. 15. Example predictions of our framework on the ISPRS Vaihingen dataset: (a) original image of tile 2, (b) prediction, (c and e) enlarged patches of (b), and (d and f) the original imagery corresponding to (c and e).

TABLE VII
COMPARISION OF OTHER METHODS ON THE AIR-SEG TEST DATASET

Method	IoU									OA	mIoU
	Imp-surf	Res-building	Factory	Road	Water	Farm land	Grass	Tree	Bare land		
FCN [7]	60.06	68.85	73.63	69.62	83.64	82.98	52.86	76.88	16.09	82.71	64.96
PSPNet [10]	61.75	69.73	73.92	70.93	84.61	83.38	53.31	79.41	17.63	83.37	66.07
DeepLab V3+ [11]	62.13	69.01	**74.35**	71.02	84.63	83.69	54.66	79.65	20.10	83.56	66.58
BAS^{4}Net (full)	61.92	**70.62**	72.91	71.76	84.44	83.25	54.76	80.38	**25.57**	83.83	67.29
BAS^{4}Net (semi)	**62.34**	70.35	72.62	**72.76**	**84.85**	**84.13**	**55.69**	**80.42**	24.68	**84.02**	**67.54**

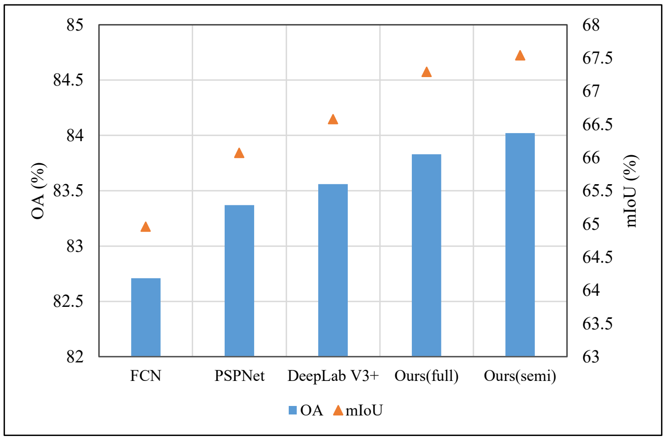


Fig. 16. Comparison of different methods on the AIR-SEG dataset.

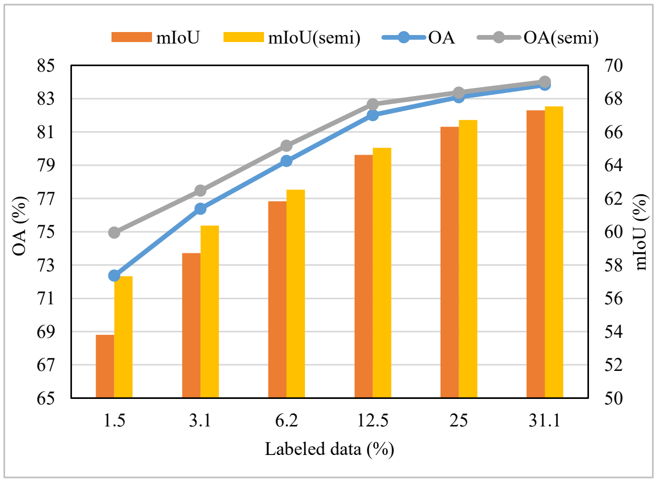


Fig. 17. Results of semi-supervised learning on the AIR-SEG datasets with different ratios of labeled and unlabeled data.

we also compare it with those methods submitted to the online leaderboard. We use different inference strategies on the test data. As Table VI shows, multi-scale (MS) and horizontal flip (Flip) strategies can led to around 0.4% improvement in terms of OA. After all test patches are classified, they are combined back with an overlay fusion (Whole) strategy to get the original size classification maps. Evaluation indicators are calculated on the full test tiles without cropping. Parts of the results for the ISPRS Vaihingen 2-D Semantic Labeling Challenge are listed in Table V, including OA and F1-score metrics. Our approach achieves state-of-the-art performance and outperforms all the other approaches on cars.

2) Qualitative Analysis: Fig. 13 gives a qualitative comparison of the different methods on the ISPRS Vaihingen dataset. Fig. 13 (a) shows that FCN-32s captures few spatial information and cannot accurately segment the objects. It is obvious from Fig. 13 (b) that the CMF module learns much more spatial information and balances semantic and spatial information better. From the buildings and cars, it can be seen that the intra-class gap becomes smaller and the inter-class gap larger, with clear boundaries and regular shapes. Fig. 13 (c) shows that the MSC module captures multi-scale contextual information. Fig. 13 (d) demonstrates that the BAM produces a stronger feature expression, captures more boundary-related information, and enhances intra-class consistency and inter-class difference. From the final segmentation results of Fig. 13 (d), we can see that the boundary is clearer and the segmentation performance of small objects is improved. With the contribution of our framework, we can capture additional information relevant to object boundaries. Fig. 14 shows a comparison of different methods on 513×513 patches. Compared with the baseline and DeepLab V3+ [11], our method performs better, especially on the regions in the red dashed boxes. For example, close aligned cars are segmented well and the boundaries of buildings look clearer. Fig. 15 presents the result of BAS^{4}Net on tile 2 of the test data. Fig. 15 (c and e) show the good result for small objects and boundaries, respectively in the form of the compact arrangement of cars and the weak boundary between buildings and other objects. Therefore, we can state that BAS^{4}Net improves the accuracy and coherence of boundaries, especially for small objects.

D. Experiments on the AIR-SEG Dataset

In this section, we extend the experiments to the AIR-SEG dataset to further evaluate the effectiveness of BAS^{4}Net. As for the experiments on ISPRS Vaihingen dataset, we ignore the category clutter in the experiments, because it makes up less than 1% of the total pixels. We preprocess the dataset and obtain 2438 labeled and 5202 unlabeled patches in the training dataset as well as 700 patches in the test dataset. Again, we select three popular methods namely FCN-32s [7], PSPNet [10] and DeepLab V3+ [11] for comparison. An output stride of 16 is used in the latter two methods and the results are shown in Table VII. Our framework performs better than all other, especially for the category bare land, where the F1 score is 5% higher than for other methods. Our framework also performs well for both the categories tree and residential building. As shown in Table VII and Fig. 16, when trained in a semi-supervised way together with the unlabeled data, the performance is further improved. This demonstrates that the semi-supervised method predicts more accurate pseudo-labels for unlabeled images and learns more significant features. To further validate the semi-supervised learning method, we randomly add different amounts of labeled images to the training dataset and evaluate the results. Fig. 17 shows that semi-supervised learning method improves the performance significantly, particularly for less labeled data and much unlabeled data, which can help to reduce the labeling effort.

Fig. 18 presents the results on patches of size 513×513. The third column is the result for the baseline. There exist block effects and the boundary is blurred. Additionally, scattered small objects, such as trees and residential buildings are not segmented. Our framework segments small objects accurately and produces clearer boundaries. This demonstrates that it can balance semantic and spatial information very well, and capture additional boundary related information. Fig. 19 presents the results of our framework on two images of the test data. Fig. 19 (d) shows that we obtain accurate boundaries and small objects like trees are well segmented. Therefore, our framework has a good generalization capability for the semantic segmentation of VHR remote sensing images.

V. Conclusion

In this work, we have proposed BAS^{4}Net for the semantic segmentation of VHR remote sensing images. It can learn additional information related to object boundaries in a semi-supervised manner. Our framework uses the CMF module to balance semantic and spatial information of multi-scale feature maps. The BAM weights the feature maps with rich semantic boundary information to alleviate the boundary blur. A discriminator network infers pseudo-labels for unlabeled images to assist semi-supervised learning. To validate the effectiveness of our framework, we have conducted experiments on the ISPRS Vaihingen dataset and propose the even more challenging AIR-SEG dataset for extended experiments. The experimental results demonstrate that our framework achieves a state-of-the-art performance on the ISPRS Vaihingen dataset. Furthermore, BAM shows a special advantage for the class car, which contains relatively small objects and is supposed to be more sensitive to boundary errors. Concerning future work, we want to improve the semi-supervised learning method and further increase the accuracy of the model.

References

[1] 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 journal of photogrammetry and remote sensing*, vol. 145, pp. 96–107, 2018.

[2] X. Liu, J. He, Y. Yao, J. Zhang, H. Liang, H. Wang, and Y. Hong, “Classifying urban land use by integrating remote sensing and social media data,” *International Journal of Geographical Information Science*, vol. 31, no. 8, pp. 1675–1696, 2017.

[3] O. Tasar, Y. Tarabalka, and P. Alliez, “Incremental learning for semantic segmentation of large-scale remote sensing data,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3524–3537, 2019.

[4] M. Volpi and D. Tuia, “Dense semantic labeling of subdecimeter resolution images with convolutional neural networks,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, 2016.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[7] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[9] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[12] 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 transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[13] 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 Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158–172, 2018.

[14] L. Mi and Z. Chen, “Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 140–152, 2020.

[15] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, “Gated-scnn: Gated shape cnns for semantic segmentation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5229–5238.

[16] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “Learning a discriminative feature network for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1857–1866.

[17] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, “Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1742–1750.

[18] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, “Adversarial learning for semi-supervised semantic segmentation,” *arXiv preprint arXiv:1802.07934*, 2018.

[19] N. Souly, C. Spampinato, and M. Shah, “Semi supervised semantic segmentation using generative adversarial network,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5688–5696.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[21] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Y. Fastfcn, “Rethinking dilated convolution in the backbone for semantic segmentation,” *arXiv preprint arXiv:1903.11816*, vol. 2, no. 5, p. 6, 2019.

[22] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv preprint arXiv:1706.05587*, 2017.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[24] G. Chen, X. Zhang, Q. Wang, F. Dai, Y. Gong, and K. Zhu, “Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 5, pp. 1633–1644, 2018.

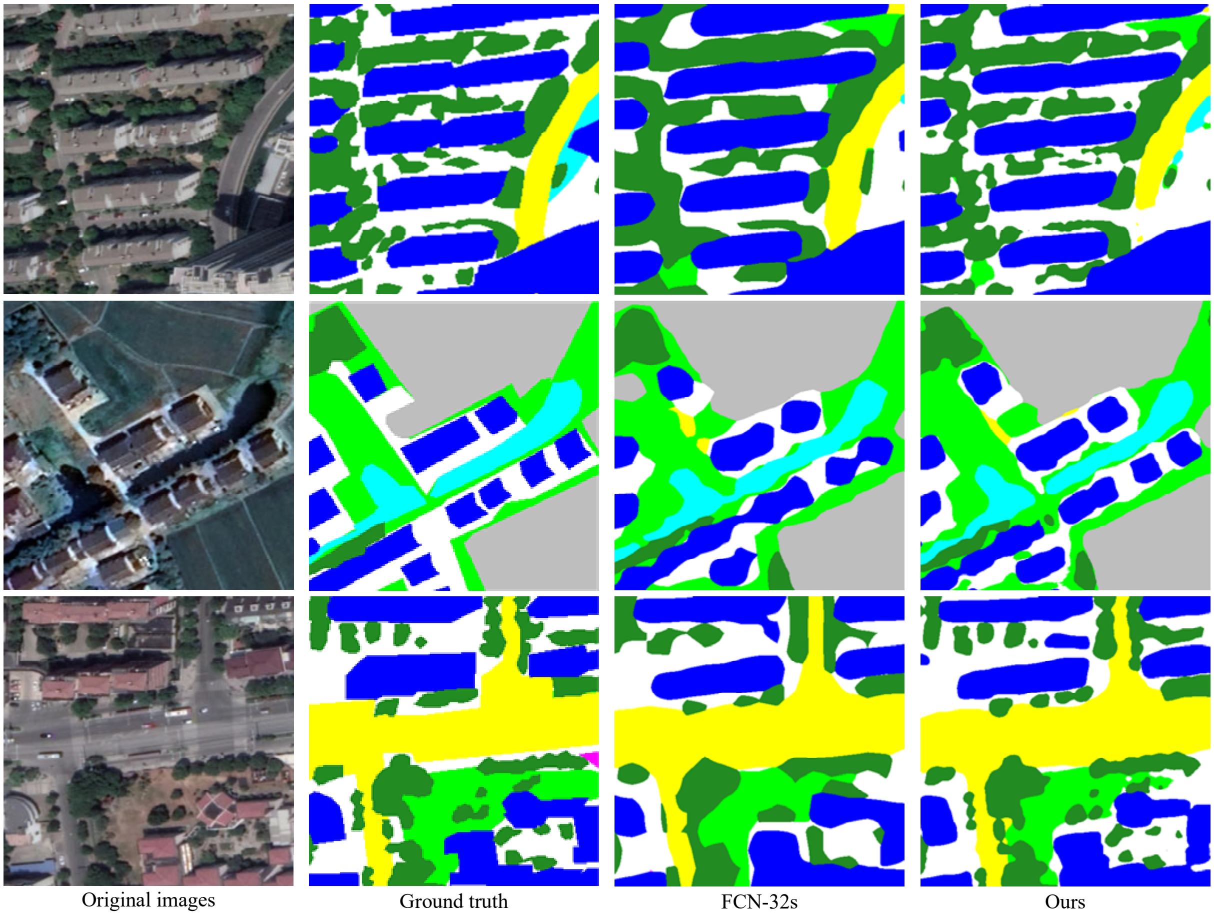


Fig. 18. Comparison of different methods on the AIR-SEG dataset.

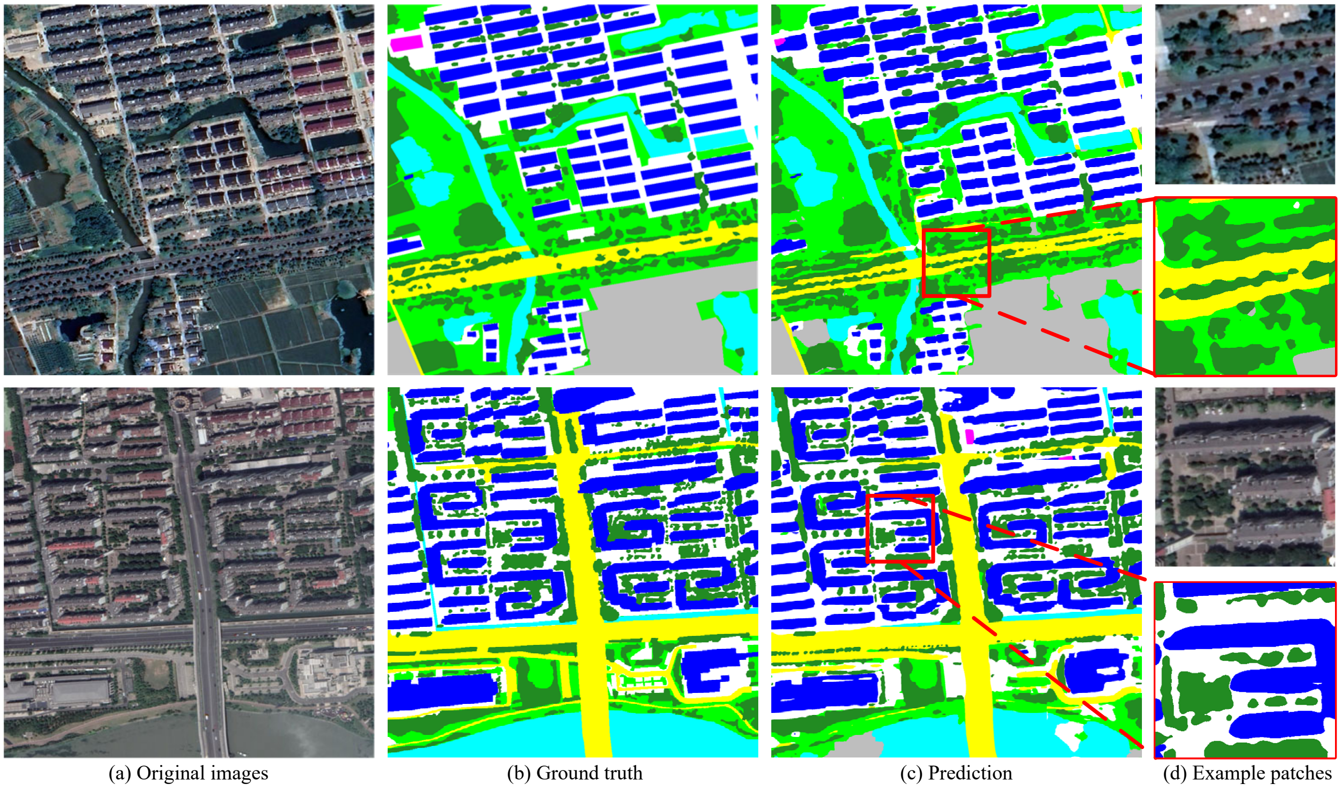


Fig. 19. Experiments on the AIR-SEG dataset.

[25] 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 journal of photogrammetry and remote sensing*, vol. 145, pp. 78–95, 2018.

[26] K. Nogueira, M. Dalla Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7503–7520, 2019.

[27] 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.

[28] H. Luo, C. Chen, L. Fang, X. Zhu, and L. Lu, "High-resolution aerial images semantic segmentation using deep fully convolutional network with channel attention mechanism," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3492–3507, 2019.

[29] D. Marin, Z. He, P. Vajda, P. Chatterjee, S. Tsai, F. Yang, and Y. Boykov, "Efficient segmentation: Learning downsampling near semantic boundaries," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2131–2141.

[30] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[31] G. Cheng, J. Yang, D. Gao, L. Guo, and J. Han, "High-quality proposals for weakly supervised object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 5794–5804, 2020.

[32] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[34] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[35] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervised object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[36] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *arXiv preprint arXiv:1611.08408*, 2016.

[37] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[40] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[42] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017, pp. 3226–3229.

[43] V. Mnih, *Machine learning for aerial image labeling*. Citeseer, 2013.

[44] ISPRS. "2d semantic labeling contest". 2014.

[45] Y. Qin, X. Xiao, J. Dong, B. Chen, F. Liu, G. Zhang, Y. Zhang, J. Wang, and X. Wu, "Quantifying annual changes in built-up area in complex urban-rural landscapes from analyses of palsar and landsat images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 124, pp. 89–105, 2017.

[46] X. Huang, D. Wen, J. Li, and R. Qin, "Multi-level monitoring of subtle urban changes for the megacities of china using high-resolution multiview satellite imagery," *Remote sensing of environment*, vol. 196, pp. 56–75, 2017.

[47] L. N. Kantakumar, S. Kumar, and K. Schneider, "Spatiotemporal urban expansion in pune metropolis, india using remote sensing," *Habitat International*, vol. 51, pp. 11–22, 2016.

[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[50] A. Marcu and M. Leordeanu, "Dual local-global contextual pathways for recognition in aerial imagery," *arXiv preprint arXiv:1605.05462*, 2016.

[51] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order crfs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 76–85.

[52] J. R. Bergado, C. Persello, and A. Stein, "Recurrent multiresolution convolutional networks for vhr image classification," *IEEE transactions on geoscience and remote sensing*, vol. 56, no. 11, pp. 6361–6374, 2018.

[53] S. Piramanayagam, E. Saber, W. Schwartzkopf, and F. W. Koehler, "Supervised classification of multisensor remotely sensed images using a deep learning framework," *Remote Sensing*, vol. 10, no. 9, p. 1429, 2018.

[54] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 20–32, 2018.

[55] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "Treeunet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 156, pp. 1–13, 2019.

Xian Sun (SM'19) received the B.Sc. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, China, in 2009.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, China. His research interests include computer vision, geospatial data mining, and remote sensing image understanding. He is the IEEE Senior Member, and the guest editor of JSTARS, co-editor of Current Chinese Science.

Aijun Shi (S'20) received the B.Sc. degree from Shandong University, Shandong, China, in 2018. She is currently pursuing the master's degree with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

Her research interests include computer vision and deep learning, especially on semantic segmentation, object detection, and remote sensing image understanding.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3021098, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Hai Huang (M'20) received the B.Sc. degree in Aerospace from Beijing University of Aeronautics and Astronautics, Beijing, China, in 1998 and the Dipl.-Ing. degree in Aerospace from Technische Universität München (TUM), Munich, Germany, in 2004. In 2010 he received the Ph.D. degree in Photogrammetry and Remote Sensing from Universität der Bundeswehr München (UniBw), Neubiberg, Germany. He received the Habilitation in Photogrammetric Computer Vision at Leibniz University Hannover, Hannover, Germany, in 2018 and in Computer Vision at UniBw in 2019, respectively.

He is currently a senior researcher with the Institute for Applied Computer Science at UniBw. His research interests include image understanding, spatial data interpretation and 3D urban modeling.

Helmut Mayer (M'94) received the Dipl.-Ing. degree in Surveying from Technische Universität München (TUM), Munich, Germany, in 1990 and the Ph.D. degree and Habilitation in 1993 and 1997, respectively.

In 1999 he became full Professor for Photogrammetry and Remote Sensing at Universität der Bundeswehr München (UniBw), Neubiberg, Germany. In 2008 he moved to the Computer Science department where he is currently Professor for Visual Computing. Since 1997 he has chaired several working groups and been vice president of Commission III of the International Society for Photogrammetry and Remote Sensing (ISPRS). From 2007 to 2011 he has been Editor-in-chief of the German photogrammetric journal PFG. Since 2009 until 2018 he has been a member of the technical committee of the German Society for Pattern Recognition (DAGM).