Boundary-Aware Dual-Stream Network for VHR Remote Sensing Images Semantic Segmentation

Zhixian Nong, *Student Member, IEEE*, Xin Su, *Member, IEEE*, Yi Liu, *Member, IEEE*, Zongqian Zhan, *Member, IEEE*, and Qiangqiang Yuan, *Member, IEEE*

***Abstract*—Semantic segmentation for very-high-resolution remote sensing images has been a research hotspot in the field of remote sensing image analysis. However, most existing methods still suffer from a challenge that object boundaries cannot be finely recovered. To tackle the problem, we develop a dual-stream network based on the U-Net architecture, Instead of the traditional skip connections, a boundary attention module is proposed to introduce the boundary information from the EDN module to the SSN module. Experiments on ISPRS Potsdam and Vaihingen datasets show the effectiveness of the proposed network, especially in man-made objects with distinct boundaries.**

***Index Terms*—Attention module, edge detection subnetwork, semantic segmentation, very high spatial resolution.**

I. Introduction

IN THE field of remote sensing, semantic segmentation of very-high-resolution (VHR) remote sensing images is of significance to many applications, such as land cover mapping, urban change detection, and environmental monitoring [1]–[3]. Automatic semantic segmentation for remote sensing image has thus become a fundamental problem for a long time.

Early methods, including unsupervised classification (such as K-means [5]) and supervised classification (such as SVM [6]), mainly rely on spectral information of each individual pixel and ignore spatial information, which is inapplicable for the case of VHR remote sensing images. As the outstanding performance of convolutional neural networks (CNNs) in computer vision, recent researches have proven CNNs very successful tools for VHR remote sensing images semantic segmentation [4]. FCN [7] is a milestone for semantic segmentation and many FCN-based methods [8]–[10] have also been developed. Despite their promising performance in semantic segmentation tasks, FCN and its extensions may fail on the segmentation of complex objects with intraclass inconsistency and interclass indistinction issues in VHR remote sensing images. To tackle the problem, a larger receptive field should be used to mine rich contextual information, instead of extracting feature in a small region or even isolated pixel. Many methods use complex feature extractors to expand the receptive field and get high-level semantic information [36], such as replacing the feature extractor of VGG with the feature extractor of ResNet. These feature extractors use downsampling operations and complex connections to ensure large receptive fields.

However, downsampling operations lead to the loss of spatial resolution. In order to recover deep heat maps to the same size of the input data, upsampling layers are often used in the semantic segmentation networks. However, the upsampling operation cannot recover the details caused by the pooling layers, which may lead to fuzzy boundaries in the segmentation results. In other words, there is a trade-off between deep feature and fine boundaries [10]. Obviously, the results of the networks with excessive downsampling layers tend to blur object boundaries and ignore small objects. It could worsen for VHR remote sensing images, which include more boundaries.

In this article, we propose a boundary-aware dual-stream network based on U-Net [25], in which an auxiliary edge detection stream is introduced to improve the result of boundaries by explicit supervision of object boundaries. Correspondingly, to better fuse the boundary information with the semantic information and the shallow feature with the deep feature, spatial and channel attention modules (AMs) instead of direct skip connections are used in the proposed dual-stream network. The contributions of this article can be summarized as follows: 1) A boundary-aware dual-stream network is proposed to combine the edge detection network and the semantic segmentation network, which recover finer object boundaries in the results of semantic segmentation; 2) Spatial and channel AMs are used to replace the classic skip connections for U-Net to capture local details and the result shows promising benefit; 3) The experimental results show that the combination of the proposed two modules yields finer boundaries in VHR data semantic segmentation, especially in man-made objects.

The remaining parts of the article are as follows. Section II introduces some related work and the intuition behind our method. A detailed description of the proposed method is then presented in Section III. We later describe the setup of experiments and

Manuscript received February 1, 2021; revised March 10, 2021 and April 1, 2021; accepted April 21, 2021. Date of publication April 27, 2021; date of current version June 1, 2021. This work was supported in part by the National Geomatics Center of China through the National Key Research & Development Program of China under Grant 2016YFB0501403 and in part by the Fundamental Research Funds for the Central Universities under Grant 2042019kf0046. *(Corresponding author: Xin Su.)*

Zhixian Nong is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: zhixiannong@ whu.edu.cn).

Xin Su is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: xinsu.rs@whu.edu.cn).

Yi Liu and Zongqian Zhan are with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: yliu@sgg.whu.edu.cn; zqzhan@sgg.whu.edu.cn).

Qiangqiang Yuan is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China, and also with the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: qqyuan@sgg.whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3076035

evaluate our method in Section IV. Finally, conclusions are drawn in Section V.

II. Related Work

A. Semantic Segmentation

FCN [7] opens new avenues for semantic segmentation by using CNNs. Applications in remote sensing draw on the empirical works in computer vision and combine them with the multiscale characteristic of the objects in remote sensing data. Fu *et al*. [11] improve FCN by using multiscale classification. In order to reduce spatial ambiguities, Liu *et al*. [12] apply skip connections with residual units to feed encoding-stage information to the decoder. Wang *et al*. [13] utilize gate mechanism to automatically select features when merging different-scale feature maps. Considering the scale inconsistency of objects in remote sensing image, Liu *et al*. [14] propose a self-cascaded convolution module. Most of the networks use the high-to-low and low-to-high-resolution framework [15], without reusing spatial information from low-level feature. However, these networks designed for natural images may fail when dealing with VHR remote sensing data with more boundaries.

To alleviate the loss of image details and spatial resolution caused by pooling layers and/or strided convolutions, some methods have been proposed. These methods can be categorized into two groups. One is to use atrous convolutions [23] that expand the receptive field without reducing the spatial resolution. In consideration of global context information, global average pooling is first used in ParseNet [16]. PSPNet [37] applies spatial pyramid pooling [17] to capture multiscale information and then Deeplab v3 [18], [42]–[44] extends it to atrous spatial pyramid pooling. However, since the atrous convolution is a kind of sparse operation, successive atrous convolutions may cause grid artifacts [24] and ignore small objects. The other group of methods is to reintroduce low-level features to the decoder by skip connections, which brings high-resolution information (e.g., details and boundaries) to the decoder. Unpooling operator in SegNet [19] saves the max pooling indices and makes use of the indices in upsampling layers, which help to recover the spatial information. U-Net [25] and its variants have shown their advantages in VHR remote sensing images semantic segmentation [26]. However, the simple skip connection also brings some weak semantic information from low-level features to deep ones, which may lend to some misclassification inside an object.

B. Edge Detection

Semantic segmentation has been coincided with edge detection, since ideal semantic segmentation results are with effective edges. Motivated by the synergies between edge detection and semantic segmentation, some studies make use of edge detection to enhance object boundaries in the result of semantic segmentation. Bertasius *et al*. [20] and Kokkinos [21] utilize the edge learned from CNN to enhance the performance of semantic segmentation, the two tasks however are independent, and they are not end-to-end frameworks. Extended Deeplab [22] employs intermediate features for semantic segmentation to predict edges and optimize the target semantic segmentation quality by domain transform. Marmanis *et al.* [10] make edge detection and semantic segmentation in series and perform a higher semantic segmentation accuracy in VHR remote sensing data. EANet [41] designs an end-to-end edge-aware multitask learning network and EALoss for the extraction of buildings. Yu *et al.* [33] design a network to learn these two tasks simultaneously by sharing weights in the feature extractor (encoder) but using different decoders. However, considering that these two tasks need different level features, the feature for one task may not be suitable for the other one. Single-stream network thus may not be the optimal solution for this multitask learning.

C. Attention Module

AMs have been widely used in various tasks using deep learning methods in recent year [27], since they improve the performance with introducing few network parameters [33]. Channel AM in SENet [28] adaptively enhances beneficial channels and suppresses useless channels by different weights. Based on SENet, CBAM [29] applies AM both in channel domain and spatial domain. Residual attention network [30] proposes residual attention learning, which solves the performance degradation caused by the stack of AMs. PAN [31] combines spatial pyramid structure with attention mechanism, instead of a series of atrous convolution. In the field of semantic segmentation, AMs are usually employed to generate enhanced feature, e.g., local details and boundary information are fed into AMs and attend the origin feature in the semantic segmentation network.

Thus, in order to further take advantage of reweighted low-level feature and alleviate the problems of insufficient semantic information in low-level feature in U-Net, we utilize AMs to enhance local information for low-level feature and to re-weight low-level feature and high-level feature. Besides, in contrast to other studies, we use a dual-stream network to learn the two tasks respectively. In addition, boundary attention modules (BAMs) are introduced into our dual streams to enhance boundary information for the semantic segmentation stream. Boundary feature learned in edge detection stream is treated as boundary weight map to attend feature in semantic segmentation stream. With the two types of AMs, the feature in the semantic segmentation stream is further enhanced by boundary information to get finer boundaries and higher classification accuracy.

III. Methodology

In this section, we first present the overview of the proposed dual-network for semantic segmentation of VHR remote sensing images. Then, the two AMs, i.e., the AMs for the sematic segmentation subnetwork and the BAMs between the edge detection subnetwork (EDN) and the sematic segmentation subnetwork, are introduced in detail.

A. Network Architecture

The structure of our network is shown in Fig. 1, which consists of two subnetworks, i.e., the semantic segmentation subnetwork (SSN) and the EDN. The SSN subnetwork is built upon the

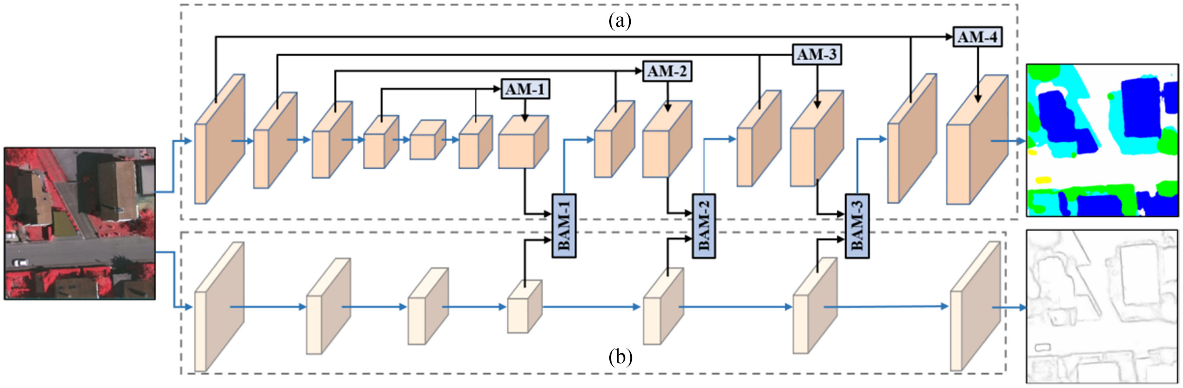

Fig. 1. Overview framework of the proposed method. It consists of two streams: (a) semantic segmentation network (SSN), and (b) edge detection network (EDN).

classic U-Net. As mentioned in ection II, semantic segmentation methods are also effective tools for edge detection. Thus, we design another lighter U-shape network that performs as an auxiliary subnetwork to detect edges. As shown in Fig. 1, each blue arrow represents a pooling layer (or an upsampling layer) and a convolution group (the first and last blue arrows do not include any pooling or upsampling layer). Each convolution group consists of two 3×3 convolution layers, and each convolution layer is followed by a batch normalization layer and a ReLU activation function. Thus, the spatial size of the feature map after the four pooling layers in the encoder is reduced to 1/16 of the input size in our UNet-AM. In the EDN subnetwork, there are three pooling layers in the encoder, so the spatial size of the feature map after encoder is reduced to 1/8 of the input size. In addition, between the decoders of the two subnetwork, three BAMs are utilized ($\mathrm{BAM} - i (i \in \{1, 2, 3\})$) to enhance the boundary information for the SSN subnetwork in different scales. The structure of the proposed BAM is described in detail in Section III-C.

The supervisory label of the edge detection task is generated from the ground truth of the VHR semantic segmentation dataset by traditional edge detection operators, such as Canny [32]. Since boundaries are in the minority in an image, edge detection is a class-imbalanced problem. The binary cross entropy loss tends to submerge the boundaries with fewer samples into the background. Hence, a weighted binary cross-entropy loss is used in the EDN subnetwork, which can be formulated as

$$L_{\mathrm{WBCE}} = \ - \left(\alpha y \log y' + (1 - \alpha) \left(1 - y\right) \log \left(1 - y'\right)\right) \quad (1)$$

where $y \ = \ 0$ or 1 is the true label of boundaries. $y' \in [0, 1]$ is the estimated probability for the corresponding pixel to be boundary category. Parameter α is introduced to deal with the class-imbalanced problem in the edge detection task, which is selected manually according to the experiment performance.

For the semantic segmentation task, we use the traditional cross-entropy loss in the SSN subnetwork. To balance the loss of the two subnetworks, we select a parameter β to balance L_{CE} and L_{WBCE}. Thus, the total loss L of our network can be formulated as

$$L \ = L_{\mathrm{CE}} \ + \beta L_{\mathrm{WBCE}}. \qquad (2)$$

B. Attention Modules for SSN Subnetwork

In the classic U-Net network, skip connections are used to reintroduce local details from the encoder to the decoder. More precisely, the feature map f_e in the encoder is concatenated to its corresponding feature map f_d in the decoder, as shown in Fig. 2(a). However, direct skip connections fusing low-level feature and high-level feature may introduce some confusing information that leads to misclassifications. To further develop the idea of reusing low-level feature in U-Net and alleviate the problem of introducing weak semantic information, an AM, namely AM-s, is introduced to capture local details in the skip connection stage without introducing insufficient semantic information, as shown in Fig. 2(b). Different from the direct skip connections in the original U-Net, we exploit f_e as spatial attention weights for itself by employing a 1×1 convolution layer to generate one-channel spatial weights. Such spatial AM is formulated as

$$f'_e = f_e \, Sig\left(h\left(f_e\right)\right) f_e \qquad (3)$$

$$f'_d = Con\left(f'_e, f_d\right) \qquad (4)$$

where $h(\cdot)$ denotes 1×1 convolution with one kernel, $h(f_e)$ is of shape $1 \times H \times W$. $Sig(\cdot)$ denotes a sigmoid activation function that generates spatial weights. $\otimes$ denotes spatial element-wise multiplication. It applies the element-wise multiplication between $Sig(h(f_e))$ and each channel slice of f_e, where $Sig(h(f_e)) \in [0, 1]$. Different from the traditional AMs, we utilize an element-wise addition to compute the value of f'_e, instead of excessively suppressing f_e by the multiplication. $Con(\cdot)$ donates the concatenation operation between the encoder and the decoder.

Low-level feature f_d is with less semantic information but more local details. In contrary, high-level feature f'_e is with more semantic information but less local details. To balance the contributions of f_d and f'_e for the semantic segmentation task, we use a channel AM to adaptively reweight each channel in

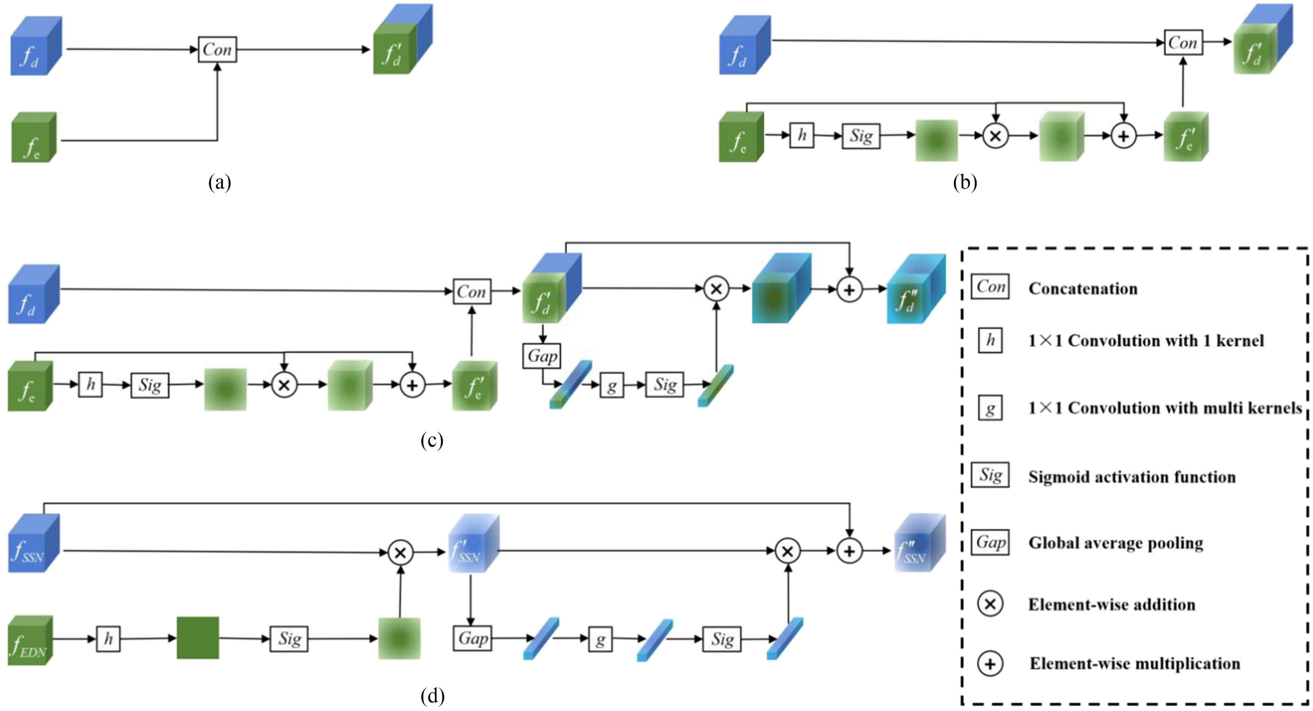

Fig. 2. Attention modules [(a)–(c) are for U-Net, (d) is for the two streams]. Blue cubes represent that the feature maps come from SSN, and green cubes represent that the feature maps come from EDN. (a) Skip connection in U-Net. (b) AM-s attention module. (c) AM-sc attention module. (d) BAM attention module.

f'_d after the concatenation operation, as shown in Fig. 2(c). The proposed AM, namely AM-sc, can be formulated as (5)

$$f''_d = f'_d\ Sig\left(g\left(Gap\left(f'_d\right)\right)\right) \oplus f'_d. \tag{5}$$

$Gap(\cdot)$ denotes global average pooling. $Gap(f'_d)$ is of size $(C_1 + C_2) \times 1 \times 1$. $g(\cdot)$ denotes 1×1 convolution with $C_1 + C_2$ kernels, which is used to generate a channel weight vector. Here, $\oplus$ denotes the multiplication of each element in the channel weight vector and each channel of f'_d. Finally, by spatial and channel attention, we get a refine feature f''_d with sufficient semantic information and local details.

As shown in Fig. 2(b) and (c), the AM-s corresponds (3) and (4), and AM-sc corresponds (3), (4), and (5). As shown in Fig. 1, $\mathrm{AM} - i (i \in \{1, 2, 3, 4\})$ represents the four AMs used in our SSN subnetwork. The effectiveness of the two modules is discussed in detail in Section III.

C. BAMs Between SSN and EDN Subnetworks

As mentioned before, edge learning is used to guide the semantic segmentation learning for finer boundary recovery in our dual-stream network. To further combine the two branches, we treat the feature in the EDN subnetwork as attention maps to attend the origin feature in the SSN subnetwork. The BAM introduces the boundary information from EDN subnetwork to SSN subnetwork, which is formulated as

$$f'_{\mathrm{SSN}} = f_{\mathrm{SSN}}\ Sig\left(h\left(f_{\mathrm{EDN}}\right)\right) \tag{6}$$

$$f''_{\mathrm{SSN}} = f'_{\mathrm{SSN}}\ Sig\left(g\left(Gap\left(f'_{\mathrm{SSN}}\right)\right)\right) f_{\mathrm{SSN}}. \tag{7}$$

The feature map in SSN subnetwork is defined as f_{SSN}, and the feature map in EDN subnetwork as f_{EDN}, where f_{SSN}, f'_{SSN}, and $f''_{\mathrm{SSN}} \in \mathbb{R}^{(C_1+C_2)\times M\times N}$, $f_{\mathrm{EDN}} \in \mathbb{R}^{C_3\times M\times N}$. We use f_{EDN} to enhance boundary information of f_{SSN} and generate boundary-enhanced feature f'_{SSN}. Then, f'_{SSN} is used to generate channel-reweighted feature f''_{SSN}. The structure of BAM module looks like AM-sc module introduced in Section III-B; however, the differences are as follows. First, in (6), we exploit a one-channel feature map produced by f_{EDN} as spatial attention weights for f_{SSN}. Since the attention map for SSN subnetwork comes from the EDN subnetwork, BAM is not a self-AM. Second, as mentioned in [40], the direct concatenation of feature maps from different task streams may reduce the distinguishability of features. So, we design BAM, which is not of the same architecture with AM-sc. The complete structure of BAM is shown in Fig. 2(d).

IV. Experimental Results and Analysis

In this section, we present the implementation details and evaluate the performance of the proposed dual-stream network on the semantic segmentation task of VHR remote sensing images. The results are discussed in the subsequent sections.

A. Datasets and Evaluation Metrics

We carried out experiments on two ISPRS 2D semantic labeling challenge datasets, namely Potsdam and Vaihingen [34]. The Potsdam dataset is composed of 38 images with a spatial resolution of 5 cm. The image size of Potsdam is 6000 × 6000 pixels, and each pixel is manually annotated with a certain category of six categories, including impervious surface, building, low vegetation, tree, car, and clutter/background. Each image has four bands, near infrared (NIR), red (R), green (G), and blue (B). The corresponding digital surface models (DSMs) are provided as complementary data. For the network training and testing, all the 38 images in the Potsdam dataset are divided into train set (17 images, IDs: 2_10, 3_10, 3_11, 3_12, 4_11, 4_12, 5_10, 5_12, 6_8, 6_9, 6_10, 6_11, 6_12, 7_7, 7_9, 7_11, 7_12), validation set (7 images, IDs: 2_11, 2_12, 4_10, 5_11, 6_7, 7_8, 7_10), and test set (the remaining images), which are following the setup in [35] and [36]. The Vaihingen dataset consists of 33 images (with average image size of 2494 × 2064) with a spatial resolution of 9 cm. As same as the Potsdam dataset, each pixel in the Vaihingen dataset is manually annotated with one of the six categories. Each image has three bands, i.e., NIR, R, G, with DSMs as well. Similarly, we follow the setup in [35] and [36], where 33 images divide into train set (11 images, IDs: 1, 3, 5, 7, 13, 17, 21, 23, 26, 32, 37), validation set (5 images, IDs: 11, 15, 28, 30, 34), and test set (the remaining images).

To compute the loss in the EDN sub-network, we evaluate the edge detection accuracy on object boundaries. Note that the boundaries of objects are eroded by three pixels, and the eroded areas are ignored during evaluation to relieve the impact caused by uncertain boundaries. To evaluate the performance of every pixel, overall accuracy (OA) is used, which represents the percentage of correctly classified pixels. In addition, to evaluate the performance of each class, F1-score is tested, which considers both the recall and the precision. Because of the insensitivity for minority classes of OA, we also calculate mean-F1 score among all classes.

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} . \tag{8}$$

B. Implementation Details

In this article, the corresponding DSM data of Potsdam and Vaihingen dataset is normalized and considered as an extra band. Thus, there are five bands in Potsdam dataset and four bands in Vaihingen dataset. The training data is augmented to eight times by flips and rotations. Following the most works on the two datasets, in the case of Potsdam, all six classes are predicted. However, in the case of Vaihingen, the clutter/background class is ignored, due to the lack of training data for this category [8].

The channel numbers of convolution groups in U-Net are 32, 64, 128, 256, 512, 256, 128, 64, 32. In the proposed EDN subnetwork, channel numbers of convolution groups are 32, 64, 128, 256, 128, 64, 32. Adam optimizer is used to train the networks. The pixel resolution of the input image is 512 × 512. The initial learning rate is set as 2e−04, and decays by 0.2 when the validation loss does not decrease in 3 epochs. The batch size is 5. Training is finished when the loss fails to decrease.

Before training the dual-stream network, we separately train the EDN using the loss in (1). The parameter α in (1) is set as a series of $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. When α is set as 0.4, the EDN yields the best boundary IoU. Thus, the parameter α is determined as 0.4 in the following experiments. Finally, we set $\alpha = 0.4$ in (1) and $\beta = 0.2$ in (2) by a series of experiments. The experiments on the effect of different β values are discussed in detail in Section IV-D.

C. Comparing With Existing Works

In this section, the proposed method is compared with some state-of-the-art methods, including the baseline network U-Net [25], U-net with CBAM module for the heat map (UNet-CBAM) [37], FCN [7], FCN with feature rearrangement (FCN-FR) [8], FCN with relation-augmented module (RA-FCN) [34], RotEqNet [36], PSPNet [37] with ResNet101, Deeplab v3+ [38], DST-2 [9], and ONE-7 [38].

The results of the Vaihingen dataset are shown in Table I. Our methods are implemented as U-Net with the AM-sc, namely "UNet-sc," and the dual-stream network with EDN and UNet-sc, namely "BAM-UNet-sc." It is demonstrated that our method outperforms other methods in terms of mean F1-score and OA. The proposed BAM-UNet-sc yields a gain of 3.54% and 2.14% in mean F1 and OA compared to U-Net.

Compared with UNet-CBAM, the proposed UNet-sc reaches an improvement of 2.55% in mean F1-score and 1.11% in OA. It shows that the AM-sc exploits more effect local details from the low-level feature by the AMs than only using AMs for the heat map. Besides, by using the proposed BAMs module and the EDN subnetwork, our method contributes to an improvement of 0.34% in the mean F1-score with respect to the UNet-sc, which shows that our EDN subnetwork and BAMs module further benefit the sematic segmentation subnetwork to recover fine boundaries.

Furthermore, the proposed method has an obvious advantage in man-made objects compared with the other methods. In Fig. 4(a), the proposed method accurately identifies cars and their boundaries. In Fig. 4(b), our method remarkably recovers more accuracy boundaries of buildings.

However, the F1-score in low vegetation and tree has less improvement. The reason is that our network improves object with regular boundaries, e.g., cars, buildings. However, low vegetation and tree have inherent fuzzy boundaries, which may have no underlying patterns that can be learned by the network. In such situation, accurate boundaries cannot be found. Then, inaccurate boundary information is passed to the semantic segmentation subnetwork, which leads to misclassifications. Fig. 3 shows four typical classes of objects and their corresponding ground truth.

As shown in Table II, similar results on the Potsdam dataset are obtained. The proposed BAM-UNet-sc performs the best performance, outperforming the classic U-Net by 3.01% in mean F1-score. As the same as experiments in the Vaihingen dataset, in the case of man-made objects, our method surpasses other methods. While in the case of vegetation, it shows fewer advantages. In Fig. 5(a), the regular road greenbelts are recovered completely. In Fig. 5(b), the problem of the internal inconsistency of buildings is alleviated.

Considering that DSM data is not always available in the reality, we launch experiments on the Vaihingen dataset without using the DSMs, including the original U-Net, UNet-sc, and

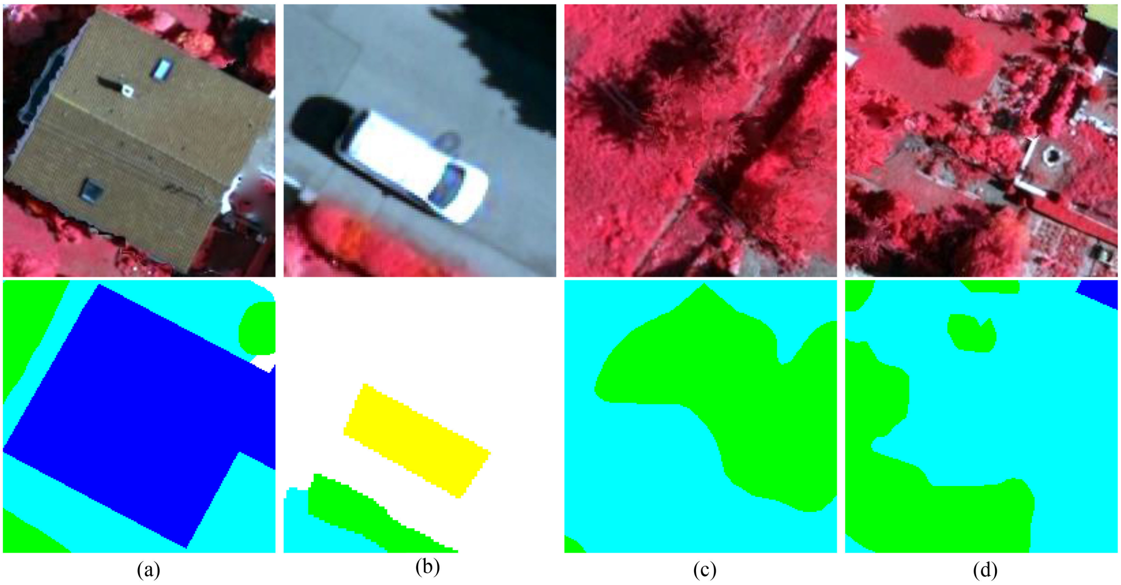

Fig. 3. Images and labels of the typical classes of man-made objects and vegetation. Classes: (a) building (blue), (b) car (yellow), (c) tree (green), (d) low vegetation (cyan). (a) and (b) have clear boundary while (c) and (d) have confused boundary.

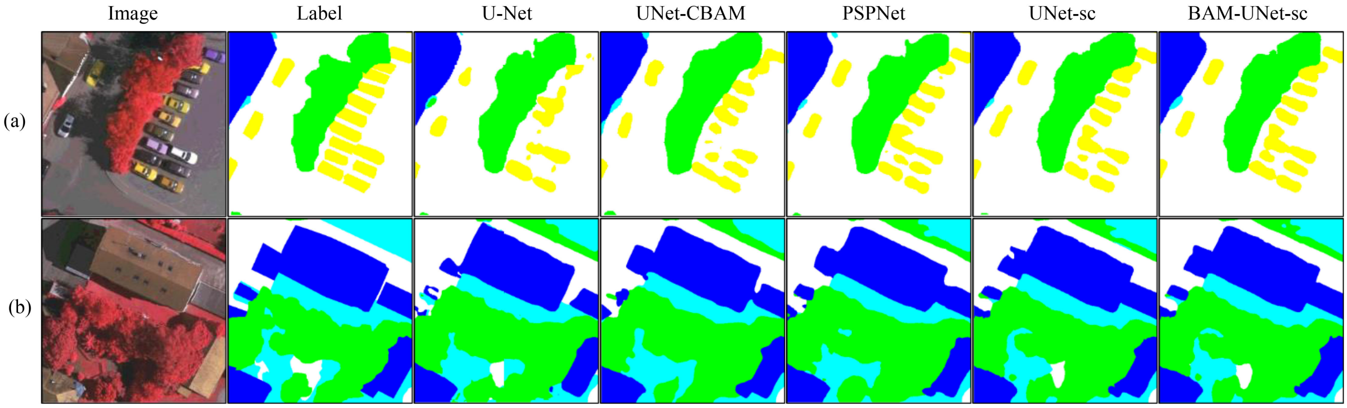

Fig. 4. Experiment results of ISPRS Vaihingen dataset. Classes: impervious surface (white), building (blue), low vegetation (cyan), tree (green), car (yellow), and clutter (red).

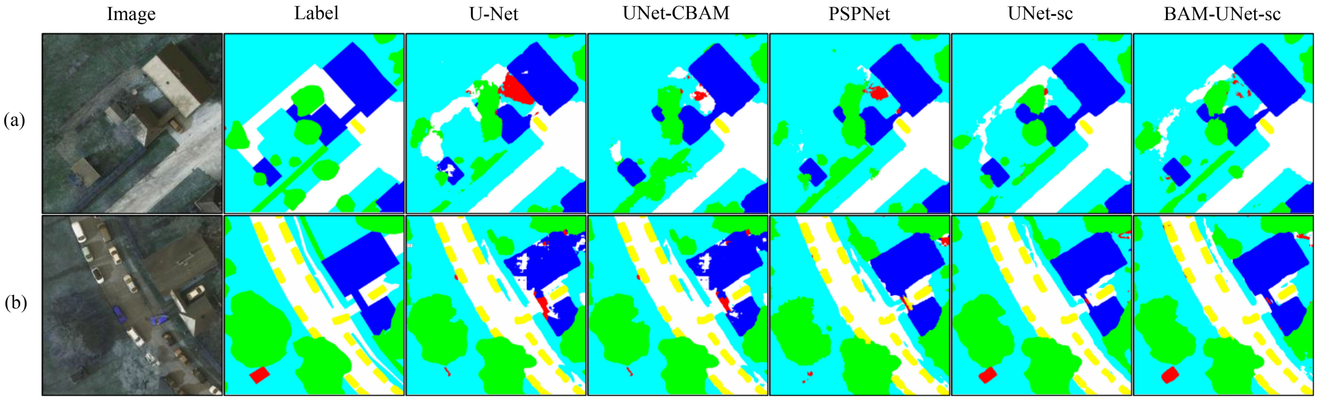

Fig. 5. Experiment results of ISPRS Potsdam dataset. Classes: impervious surface (white), building (blue), low vegetation (cyan), tree (green), car (yellow), and clutter (red).

TABLE I
COMPARISONS BETWEEN OUR METHOD AND OTHER PUBLISHED METHODS ON THE ISPRS VAIHINGEN DATASET

	Imp.surf.	Building	Low veg	Tree	Car	Mean F1	OA
U-Net	90.19	94.81	77.61	87.11	78.11	85.56	87.61
UNet-CBAM	90.62	95.11	78.13	87.17	79.99	86.21	88.45
DST-2	90.50	93.70	83.40	89.20	72.60	85.90	89.10
ONE-7	91.00	94.50	**84.40**	**89.90**	77.80	87.50	**89.80**
RotEqNet	89.50	94.80	77.50	86.50	72.60	84.18	87.50
RA-FCN	91.47	94.97	80.63	88.57	87.05	88.54	89.23
PSPNet	91.69	95.23	80.23	87.86	80.71	87.14	88.87
Deeplab v3+	91.84	95.15	80.72	88.86	79.96	87.31	89.02
UNet-sc	92.18	95.99	80.44	88.27	87.71	88.76	89.56
BAM-UNet-sc	**92.26**	**96.17**	80.36	88.14	**88.55**	**89.10**	89.75

TABLE II
COMPARISONS BETWEEN OUR METHOD AND OTHER PUBLISHED METHODS ON THE ISPRS POTSDAM DATASET

	Imp surf	Building	Low veg	Tree	Car	Clutter	Mean F1	OA
U-Net	88.99	92.68	84.45	81.50	93.75	72.10	85.58	86.25
UNet- CBAM	90.34	93.62	85.34	81.87	94.24	72.64	86.34	86.82
FCN	88.03	93.13	83.72	79.10	92.94	70.67	84.60	85.54
FCN-FR	89.31	94.37	84.83	81.10	93.56	76.54	86.62	87.02
RA-FCN	91.33	94.70	86.81	**83.47**	94.52	77.27	88.01	88.59
PSPNet	90.74	94.24	**87.15**	83.21	93.23	77.83	87.87	88.32
Deeplab v3+	90.99	94.15	86.91	83.11	93.61	78.49	87.88	88.57
UNet-sc	91.18	95.35	86.88	83.43	94.95	77.23	88.19	88.75
BAM-UNet-sc	**91.50**	**95.56**	86.94	83.37	**95.09**	**78.97**	**88.59**	**89.13**

TABLE III
ABLATION EXPERIMENTS FOR UNET-AM AND BAM

	Mean F1	OA	Prediction time	Parameters
U-Net	85.56	87.21	0.0433s	4.32M
UNet-s	87.39	88.02	0.0495s	4.36M
UNet-sc	88.76	89.56	0.0618s	4.45M
SS-UNet-sc	88.51	89.47	0.0620s	4.45M
BN-UNet-sc [33]	88.89	89.68	0.0712s	6.12M
BAM-UNet-sc	**89.10**	**89.75**	0.0632s	5.18M

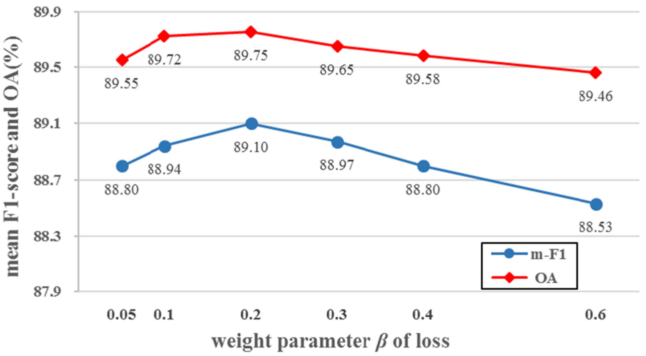

Fig. 6. Influence on different values of weight parameter β (blue: mean F1-score, red: OA).

BAM-UNet-sc. Due to the lack of DSM data, evaluation Metrics decline slightly. However, the conclusion still holds that our method outperforms the other reference ones.

D. Effectiveness of the Proposed UNet-AM and BAM

In Table III, we verify the effectiveness of the two proposed modules: AMs between the encoder and decoder of U-Net (the proposed SSN sub-network), and BAMs between the SSN and EDN subnetwork.

U-Net with AM-s and U-Net with AM-sc are called “UNet-s” and “UNet-sc,” respectively. As shown in Table III, both UNet-s and UNet-sc outperform the original U-Net, with a gain of 3.20% and 1.83% in mean F1-score, respectively. The proposed AM-sc is an effective module to improve the skip connection from the encoder to the decoder in U-Net. Besides, the effectiveness of BAMs and the integration strategy of the two sub-networks is also tested in this section. Our method is compared with two multitask learning methods for semantic segmentation. “**SS***” denotes the learning edge detection task and semantic segmentation task simultaneously in a single-stream network. “**BN***” denotes the sharing weights in the encoding stage for the two tasks, which is proposed in [33]. “**BAM***” denotes the proposed dual-stream network with the EDN network. In this part, the SSN subnetwork is U-Net with AM-sc. α and β are 0.4 and 0.2, respectively, in this experiment.

The proposed AM-sc and BAMs modules are further integrated to fuse the boundary information and the shallow features. We make use of the BAMs module to combine the SSN stream and the EDN stream, named “BAM-UNet-sc,” which outperforms UNet-sc by 0.34% in mean F1-score and 0.19% in OA. It implies that the boundary information is successfully sent to the SSN subnetwork from the auxiliary EDN subnetwork by our BAMs module. SS-UNet-sc fails to get improvement, which indicates that sharing weights for the semantic segmentation task and the edge detection task in a same network does not work. The reason may be that the two tasks need different types of features. Compared with the two multitask learning methods for the semantic segmentation task and the edge detection task, the proposed BAM-UNet-sc gets higher mean F1-score and OA, which implies the effectiveness of the proposed dual-stream network and BAM.

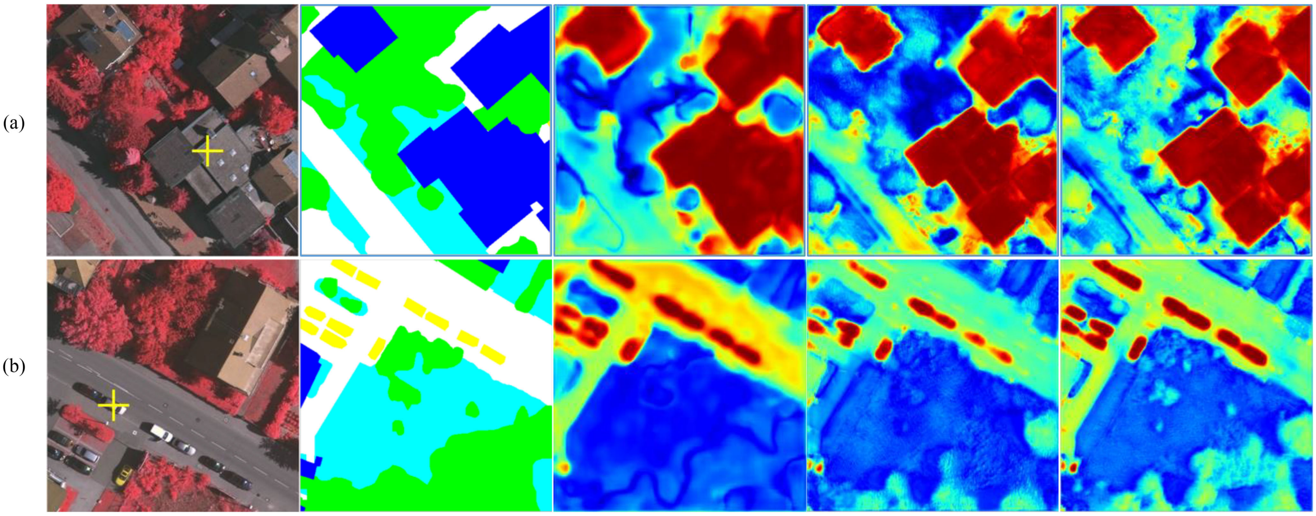

Fig. 7. Feature similarity maps between the selected pixel and other pixels. Hotter color denotes more similar in feature level. (a) A pixel from "building" (yellow cross sign in image) is selected. (b) A pixel from "car" (yellow cross sign in image) is selected.

Furthermore, Table III lists the total parameters of each network tested in the experiment as well, in which the time cost is corresponding to a test image with size of 512 × 512 pixels. The experiment is carried out on a single desktop PC, equipped with an Intel Core$^{\mathrm{TM}}$ i7-8700 CPU, 32GB RAM and NVIDIA RTX 2070. Compared with the original U-Net, our method has an extra 46% time consumption and extra 20% network parameters.

E. Influence on Different β Values

In the proposed dual-stream network, the SSN subnetwork is the major stream, and the EDN is designed as an auxiliary stream. To balance the contributions of the two streams, the weight parameter β is designed in (2). To further analyze the effect of different weight parameter β and to select the proper value of β, we launch a series of experiments with different $\beta = \{0.05, 0.1, 0.2, 0.3, 0.4, 0.6\}$ for β. The result is shown in Fig. 6.

With the increase of β, the evaluation metrics, mean F1-score, and OA all present the trend that first increased and then decreased. And mean F1-score and OA reach a maximum when the value of β is set as 0.2. However, there are some differences between the two metrics. Boundaries are in the minority in an image. Thus, the overall classification accuracy fluctuates within a narrow range. However, mean F1-score is more sensitive to the categories with a small proportion, which mainly consist of small objects in VHR remote sensing data, such as cars. Boundaries occupy a larger proportion in these categories. When the boundaries of these categories are recovered finer with the proposed BAM, there is an obvious improvement in F1-score. We can find similar conclusion from Tables I and II, which shows a great improvement of cars in F1-score. With the further increase of the value of β, the two evaluation metrics decline significantly, since the loss of edge detection task has occupied a large proportion of total loss in this situation, which confuses the primary task and secondary task.

F. Feature Similarity Visualization of the Feature Maps

In order to further show the effect of the proposed AM-sc and BAM modules, we calculate the cosine similarity between a selected point and other pixels in the feature map. The feature similarity maps are shown in Fig. 7. As mentioned in [39], the uncertainty of semantic segmentation is usually higher in the pixels near object boundaries. In the feature similarity map of the original U-Net, there are obvious transition regions between the selected category and its neighboring objects, where the pixels are hard to be classified. It corresponds to the conclusion in [39]. In UNet-sc, the introduction of local details by the AM-sc module makes the pixels near boundaries easier to be classified accurately. Visually, in the feature similarity map of UNet-sc, the boundaries are finer than those in the original U-Net, as shown in the regions of buildings and cars in Fig. 7. The feature similarity of the feature maps confirms that our network learns a more boundary-fine feature by adding the two modules.

V. Conclusion

In this article, we have introduced AMs for U-Net to enhance low-level feature, which helps to recover local details lost caused by the downsampling operations in the encoder–decoder structure semantic segmentation network. In addition, an auxiliary edge detection stream and BAM are proposed to provide semantic segmentation stream with strong boundary information. The experiments on VHR data have validated the effectiveness of recovering object boundaries, especially for man-made objects with clear boundaries.

However, the main limitation of the proposed method relies on the requirement of ground-truth images with fine boundaries. In the future, the combination of color feature, texture feature, and shape feature may be an alternative solution of this limitation.

References

[1] Y. Yuan, J. Lin, and Q. Wang, “Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization,” *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2966–2977, Dec. 2016.

[2] L. Matikainen, and K. Karila, “Segment-based land cover mapping of a suburban area—Comparison of high-resolution remotely sensed datasets using classification trees and test field points,” *Remote Sens.*, vol. 3, no. 8, pp. 1777–1804, Aug. 2011.

[3] Y. Yuan, J. Lin, and Q. Wang, “Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization,” *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2966–2977, Dec. 2016.

[4] X. X. Zhu *et al.*, “Deep learning in remote sensing: A comprehensive review and list of resources,” *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[5] J. Senthilnath *et al.*, “Hierarchical clustering algorithm for land cover mapping using satellite images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 3, pp. 762–768, Jun. 2012.

[6] M. Pal, and G. M. Foody, “Evaluation of SVM, RVM and SMLR for accurate image classification with limited ground data,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 5, pp. 1344–1355, Oct. 2012.

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

[8] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “High-resolution aerial image labeling with convolutional neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7092–7103, Dec. 2017.

[9] J. Sherrah, “Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery,” 2016, *arXiv:1606.02585.*

[10] 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, Jan. 2018.

[11] G. Fu *et al.*, “Classification for high resolution remote sensing imagery using a fully convolutional network,” *Remote Sens.*, vol. 9, no. 5, pp. 498, May 2017.

[12] Y. Liu *et al.*, “Semantic labeling in very high resolution images via a self-cascaded convolutional neural network,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 78–95, Nov. 2018.

[13] H. Wang *et al.*, “Gated convolutional neural network for semantic segmentation in high-resolution images,” *Remote Sens.*, vol. 9, no. 5, pp. 446, May 2017.

[14] Y. Liu *et al.*, “Hourglass-ShapeNetwork based semantic segmentation for high,” *Remote Sens.*, vol. 9, no. 6, pp. 522, May 2017.

[15] J. Wang *et al.*, “Deep high-resolution representation learning for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Apr. 2020.

[16] W. Liu, A. Rabinovich, and A. Berg, “ParseNet: Looking wider to see better,” 2015, *arXiv:1506.04579.*

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[18] L. C. Chen *et al.*, “Rethinking atrous convolution for semantic image segmentation,” 2017, *arXiv:1706.05587.*

[19] 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.

[20] G. Bertasius, J. Shi, and L. Torresani. “High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision,”in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 504–512.

[21] I. Kokkinos, , “Pushing the boundaries of boundary detection using deep learning,” 2015, *arXiv:1511.07386.*

[22] G. Bertasius, J. Shi, and L. Torresani. “Deepedge: A multi-scale bifurcated deep network for top-down contour detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4380–4389.

[23] 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.

[24] P. Wang *et al.*, “Understanding convolution for semantic segmentation,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2018, pp. 1451–1460.

[25] 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.*, Cham, Switzerland: Springer, 2015, pp. 234–241.

[26] A. Ghosh, M. Ehrlich, S. Shah, L. Davis, and R. Chellappa, “Stacked U-Nets for ground material segmentation in remote sensing imagery,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 252–2524.

[27] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[28] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[29] S. Woo *et al.*, “Cbam: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[30] F. Wang *et al.*, “Residual attention network for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3156–3164.

[31] H. Li *et al.*, “Pyramid attention network for semantic segmentation,” 2018, *arXiv:1805.10180*.

[32] J. Canny, “A computational approach to edge detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

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

[34] International Society for Photogrammetry and Remote Sensing. 2018. 2D Semantic labeling contest.

[35] D. Marcos *et al.*, “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, Nov. 2018.

[36] L. Mou, Y. Hua, and X. X. Zhu, “A relation-augmented fully convolutional network for semantic segmentation in aerial scenes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, Jun. 2019, pp. 12416–12425.

[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.

[38] N. Audebert, B. L. Saux, and S. Lefèvre, “Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks,” in *Proc. Asian Conf. Comput. Vis. Cham*, 2016, pp. 180–196.

[39] Y. Gal, and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” 2018, *arXiv: 1506.02142*. [Online]. Available: https://arxiv.org/abs/1506.02142

[40] H. Li, G. Chen, G. Li, and Y. Yu, “Motion guided attention for video salient object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7273–7282.

[41] Y. Guang, Z. Qian, and Z. Guixu, “ EANet: Edge-aware network for the extraction of buildings from aerial images,” *Remote Sens.*, vol. 12, no. 13, pp. 2061, Jul. 2020.

[42] S. Liu, Q. Shi, and L. Zhang, “Few-Shot hyperspectral image classification with unknown classes using multitask deep learning,” *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3018879.

[43] Q. Shi, X. Tang, T. Yang, R. Liu, and L. Zhang, “Hyperspectral image denoising using a 3-D attention denoising network,” *IEEE Trans. Geosci. Remote Sens*, to be published, doi: 10.1109/TGRS.2020.3045273.

[44] Q. Shi, M. Liu, X. Liu, and P. Liu, “Domain adaption for fine-grained urban village extraction from satellite images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1430–1434, Aug. 2020.