Strengthen the Feature Distinguishability of Geo-Object Details in the Semantic Segmentation of High-Resolution Remote Sensing Images

Jie Chen, Hao Wang, Ya Guo, Geng Sun, Yi Zhang, and Min Deng

***Abstract*—Semantic segmentation is one of the hot topics in the field of remote sensing image intelligent analysis. Deep convolutional neural network (DCNN) has become a mainstream technology in semantic segmentation due to its powerful semantic feature representation. The emergence of high-resolution remote sensing imagery has provided massive detail information, but difficulties and challenges remain in the "feature representation of fine geo objects" and "feature distinction of easily confusing geo objects." To this end, this article focuses on the distinguishing features of geo-object details and proposes a novel DCNN-based semantic segmentation. First, the cascaded relation attention module is adopted to determine the relationship among different channels or positions. Then, information connection and error correction are used to capture and fuse the features of geo-object details. The output feature representations are provided by the multiscale feature module. Besides, the proposed model uses the boundary affinity loss to gain accurate and clear geo-object boundary. The experimental results on the Potsdam and Vaihingen datasets demonstrate that the proposed model can achieve excellent segmentation performance on overall accuracy and mean intersection over union. Furthermore, the results of ablation and visualization analyses also verify the feasibility and effectiveness of the proposed method.**

***Index Terms*—Attention mechanism, geo-object details, high-resolution remote sensing imagery, multiscale feature representation, semantic segmentation.**

I. Introduction

IMAGE semantic segmentation aims to assign a semantic category label to each pixel in an image automatically. Semantic segmentation is a research hotspot in the field of remote sensing image analysis, and it remains a challenging task. It has been successfully used in urban planning [1]–[3], precision agriculture [4], [5], environmental monitoring and protection [6], urban change detection [7], land resource survey [8], [9], and other fields.

Traditional semantic segmentation methods considerably rely on handmade low-level visual features to identify and distinguish objects. The low-level visual features require prior knowledge provided by professionals but are unstable when the observation conditions change. For this reason, the authors in [10] used superpixel segmentation to subdivide an image into subregions and alleviate the variability of low-level visual features. The pixels in these subregions have similarity in visual features, which can be used for further semantic segmentation. Researchers have also used high-performance classification algorithms, such as support vector machines [11] and random forests [12], [13], to realize efficient semantic segmentation.

Compared with traditional algorithms, deep convolutional neural networks (DCNNs) have gradually become the mainstream method in the field of semantic segmentation due to the powerful feature representation capability, and solve the dependence of traditional methods on handmade feature. Early DCNNs implemented pixel-level segmentation based on patches [14], [15], but exhibited low training efficiency. The authors in [16] proposed the fully convolutional network (FCN) model to achieve efficient end-to-end pixel-level semantic segmentation. FCN replaces all fully connected layers with convolutional layers and achieves image semantic segmentation of any size by restoring to the input image scale through upsampling operation. As a groundbreaking end-to-end semantic segmentation network, FCN has become the foundation for most of the subsequent works, such as SegNet [46], Unet [20], and PSPNet [37]. These subsequent works have achieved excellent accuracy on big datasets, such as ImageNet scene, PASCAL VOC 2012, and Cityscapes [17]–[19].

With the development of remote sensing technology, remote sensing image data show the characteristics of massiveness, multisource, and high resolution. The resolution increase of remote sensing images has resulted in rich geo-object details, and a large number of easily confusing objects. "Feature representation of fine geo objects" and "feature distinction of easily confusing geo objects" become two challenges in semantic segmentation, which lead to great difficulties in semantic annotation [8]. When the FCN-based model processes high-resolution remote sensing images, the spatial position and boundary information of geo objects are seriously lost due to rough upsampling operation, resulting in incomplete structure and rough boundary. Meanwhile, the feature representation ability of most networks is not enough to deal with easily confusing geo objects. In view of the abovementioned problems, the "feature distinguishability of geo-object details" still needs to be considered in the current semantic segmentation field.

Manuscript received August 30, 2020; revised October 30, 2020 and December 22, 2020; accepted January 13, 2021. Date of publication January 20, 2021; date of current version February 15, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 41671357 and Grant 42071427 and in part by the Natural Science Foundation of Hunan Province, China, under Grant 2020JJ4691. (*Corresponding author: Min Deng*.)

The authors are with the School of Geosciences and Info-Physics, Central South University, Changsha 410083, China (e-mail: cj2011@csu.edu.cn; wastonh@csu.edu.cn; 1203392419@qq.com; 0106170206@csu.edu.cn; zhangyi_csu@csu.edu.cn; dengmin@csu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3053067

Aiming at the feature representation of fine objects, researchers have used an encoder–decoder network with skip connections to capture their detail information. Although using shallow layers to extract low-level visual features helps restore spatial information, inherent semantic differences exist between high-level semantic features and low-level visual features. Directly fusing features of different levels will cause representation errors [1]–[3], [17], [20]–[22], and the single-scale output of the encoder–decoder structure cannot be effectively applied to multiscale geo objects in remote sensing images [2], [20], [23].

Aiming at the distinction of easily confusing objects, researchers have used the attention mechanism to determine the relationship among the feature representation of different geo objects, and enhance the feature difference among easily confusing objects. On the one hand, channel attention is used to focus on the relationship among different channels to enhance feature representation [24]–[26]. On the other hand, spatial attention is used to focus on the relationship among pixels of different objects to enhance the feature similarity of homogeneous geo objects and the feature difference of inhomogeneous geo objects [19], [25], [27]. However, recent studies have generated attention maps through pooling operations, which lead to the loss of spatial information and the difficulty of considering the relationship among objects [24]–[26], [28], [29]. In addition, the outputs of different attention modules are fused mainly by adding or concatenating [19], [21], [29], which makes the relationship among different attention modules insufficiently close.

In this regard, this article proposes a novel feature representation enhancement network to capture and fuse the detailed information of geo objects effectively and pay further attention to the relationship in feature representation. The network includes a multiscale feature representation module, a cascaded attention module, and boundary affinity loss. The learning of the distinguishing feature of geo object details can be realized by coupling the three parts. The main contributions of this study are as follows.

1) High-level semantic features and low-level visual features are fused through feature pyramid and error correction; hence, the model can not only pay further attention to the detailed information in an image but can also extract accurate multiscale features.
2) The coupling degree of the attention module can be enhanced by cascading channel and spatial attention, and the distinction of easily confusing geo objects can be improved.
3) Accurate segmentation of geo-object boundary is promoted using the boundary affinity loss to monitor the relationship between pixels on both sides of the geo-object boundary.

The rest of this article is organized as follows. Section II of this article introduces the related works, section III focuses on the proposed model, section IV presents the experiment and analysis, and finally the article is concluded in Section V.

II. Related Works

This article mainly focuses on the distinguishing features of geo-object details and improves the capability of DCNNs in feature learning and generalization. Therefore, this section mainly reviews related works from two aspects: "Capturing geo-object details" and "using attention mechanisms."

A. Capturing Geo-Object Details

Existing researches on semantic segmentation models are mainly conducted from two aspects: Boundary enhancement and feature fusion. That is, the boundary is input as additional information into the network to enhance the segmentation performance of the network, or low-level visual features and high-level semantic features are fused to supplement the detailed information of geo objects to solve the problem of rough boundaries.

In terms of boundary enhancement, the authors in [30] proposed a boundary semantic awareness network to create the clear boundary, and input the boundary detection results as additional information into the semantic segmentation network to enhance the segmentation performance. The authors in [31] proposed a boundary loss enhancement network that first obtains boundary labels through gradient calculation and then uses multiple weighted edge supervision to prompt the network to retain spatial boundary information. The authors in [23] embedded superpixel algorithm into a segmentation neural network to provide extra rich boundary information for building detection. These methods enable the network to capture boundary details effectively but increase computational complexity.

In terms of feature fusion, the U-net proposed by [20] is one of the earliest DCNNs using an encoder–decoder framework. The encoder extracts semantic features through a continuous convolutional layer and uses a pooling layer to reduce the resolution of the feature map, and the decoder gradually restores the size of the feature map through an upsampling operation. Feature maps with the same scale in the encoder and decoder are fused by skip connection. Thus, low-level features can be used to supplement the missing details in high-level features. In the field of remote sensing, the authors in [21] regarded the residual unit as the basic unit of the encoder to extract robust features and enhance the accuracy of spatial details, then fused low-level visual features and high-level semantic features through information connection to supplement the missing details in the decoder. The authors in [3] proposed a new convolutional neural network composed of RCCN unit, which can better exploit spatial context and visual features to enhance the effectiveness of features in the process of feature fusion. The authors in [1] used wavelet transform to enhance the original image and adopted it as an additional information input to every stage of the encoder, which can promote the encoder to learn effective features and improve the effect of feature fusion. The authors in [22] embedded generative adversarial network into the convolutional neural network. To learn more robust features, the network combines adversarial features of the discriminator with features of the encoder through the feature fusion module, and then fuses it with the high-level features in the decoder to supplement the missing spatial details. The authors in [2] proposed an encoder–decoder network with gating. The gating module can filter low-level features and allow only important and effective features to pass before performing skip connection, reducing unnecessary

feature transmission, and enhancing the effectiveness of feature fusion.

The aforementioned studies have used shallow features to supplement the missed detail information in deep features by adding or concatenating and enhancing the segmentation effect of the network, especially on fine geo objects. Most researches focus on improving the effect of feature fusion by optimizing or filtering features before feature fusion. However, after feature fusion, these models simply use one or several convolutional layers without excessive processing in many cases. These works ignored the inherent semantic differences between features at different levels—visual shallow and semantic deep features. The semantic difference between these two levels of features is large. If they are simply added together, considerable error information will be generated, destroying the semantic consistency in the original high-level features. Moreover, the encoder–decoder structure used in most studies can only output features of a single scale and is unsuitable for geo objects of different scales commonly existing in remote sensing images.

Therefore, this study uses a feature pyramid structure and multiple connections to capture the detailed information of geo objects. When feature fusion is performed, the error correction block is used to eliminate feature errors after the fusion of high- and low-level features. The output terminal of the feature pyramid is connected to an atrous spatial pyramid pooling (ASPP) module. The features of different scales are extracted through the convolutional layers of different atrous ratios to generate effective multiscale feature expressions.

B. Using Attention Mechanisms

The DCNN extracts the features of geo objects, with each feature channel representing the characteristic attributes of different geo objects. Geo objects of the same category have similar feature responses on the same channels, and geo objects of different categories have different feature responses. A semantic segmentation network realizes segmentation in accordance with the feature response differences of different geo objects. In view of the problem of "feature distinction of easily confusing geo objects" in high-resolution remote sensing images, the attention mechanism is often used to overcome the limitation of the local dependence of a neural network and enhance the connection among geo objects by capturing the global long-range dependence and improve the distinguishability of features. The attention mechanism can consider the relationship among features from position or channel and generate corresponding attention weight maps to enhance or suppress the original feature response selectively.

The attention module used in SENet [24] focuses on the relationship among feature channels. The squeeze-excitation module obtains the global features of different feature channels by global average pooling, and then learns the relationship among the channels to obtain the channel weight vector. SENet allows the network to pay attention to important feature channels with rich information and enhance the feature expression of the geo object of interest. EMANet [32] combines the expectation–maximization algorithm with the attention mechanism to execute spatial attention, it can model the relationships between pixels in a low-dimensional manifold, reduces a lot of computation, and then reconstruct higher dimensional features with global information through bases and attention maps. The convolutional block attention module (CBAM) [25] tandem connects spatial and channel attention, in which the output of channel attention is used as the input of spatial attention. The two attention modules use the average and maximum pooling layers to calculate the attention weight map. DANet [19] parallelly processes position and channel attention mechanisms, which are based on convolution and matrix operations, and the outputs of the two attention modules are fused by an addition operation. In the field of remote sensing, the authors in [26] proposed a channel attention module to distinguish the saliency of different spectrum channels. This module generates channel weight vectors through global convolution operations, which can adaptively emphasize useful spectrum channels and suppress useless spectrum channels, resulting in enhance feature representation capability of convolutional neural networks for geo objects of interest. The authors in [34] proposed a multiscale context aggregation network based on HRNet architecture, which executes spatial attention for small-size feature maps by spatial reasoning module. The spatial reasoning module based on convolution and matrix operations can model long-range spatial correlations to aggregate global spatial information and enhance the feature relationship of different geo objects. The authors in [29] proposed a relation-enhanced multiscale convolutional network that includes spatial relationship enhancement block based on convolution and matrix operations and channel relationship enhancement block based on global average pooling. The network can adaptively learn global contextual relations between any two positions or feature maps to enhance feature representations, and apply the spatial relationship enhancement block to the encoder and the channel relationship enhancement block to the decoder.

In summary, the network model focusing on the attention mechanism has a powerful global dependency capture capability and can extract effective global context information. The authors in [25], [26], and [28] calculated the attention weight map via a pooling operation or feature dimension reduction, which can reduce the computational complexity of the model and improve computational efficiency, nevertheless, the effectiveness of the attention map may be affected due to the loss of part of feature information. Furthermore, substantial spatial position information of feature maps may be lost due to pooling operation. The authors in [26]–[27], [33], and [34] used channel or spatial attention alone in the network and achieved good model performance, but a single type of attention module may not be able to consider multiple relation of feature representations. In [19] and [29], channel attention and spatial attention were both used in their models, which can fully capture the relation among different feature representations. However, when both attention modules are used in the model, they are in parallel or are embedded in different parts of the model in most cases. The coupling degree of this combination is relatively low, which ultimately leads to the low efficiency of the attention module.

In this regard, this study aims at the limitation of single relational attention to feature enhancement and uses the dual-attention mechanism of channel and spatial relation attention to enhance feature representation comprehensively. Specifically,

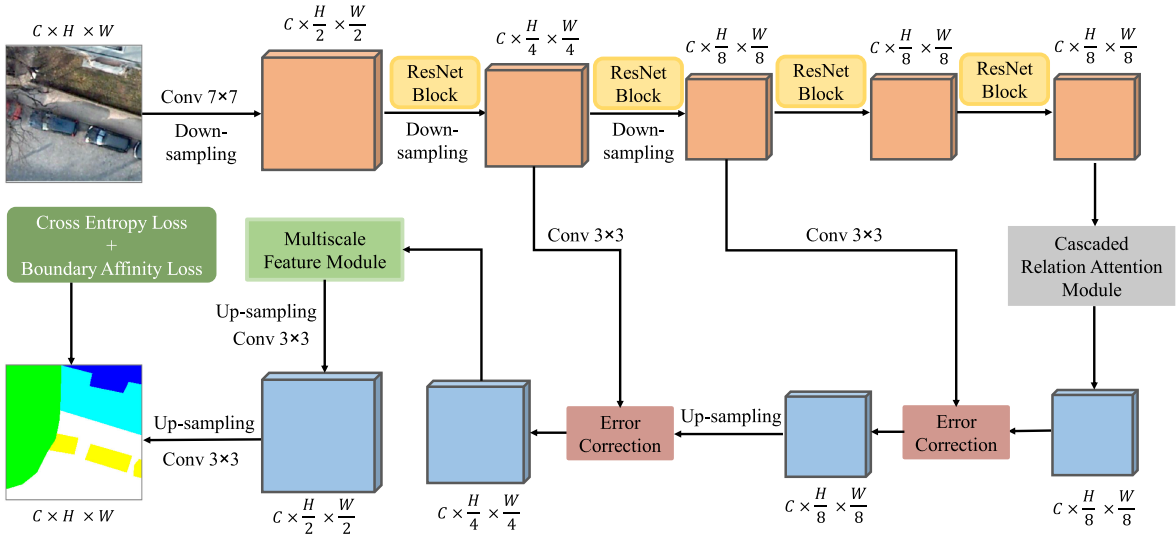

Fig. 1. Overview of the proposed method.

the two relation attention blocks are coupled, with the output of the channel relation attention as the input of the spatial relation attention. Continuous feature relationship enhancement is achieved by enhancing the information interaction of different relation attention blocks. Moreover, the proposed model obtains the attention map through convolution and matrix calculation, which cannot lose the information of feature maps and can utilize the spatial position information to obtain an accurate attention location.

III. Methodology

A. Overview

The proposed model in this article is shown in Fig. 1. The network mainly includes three parts: Multiscale feature representation module, cascaded relation attention module, and boundary affinity loss. The multiscale feature representation module can extract effective high-level semantic features with rich spatial information and efficiently capture features at multiple scales. The cascaded relation attention module can adjust the feature relationship among channels or positions and alleviate the problem of "feature distinction of easily confusing geo objects." The boundary affinity loss can consider the affinity relationship between pixels at the boundary, prompting the network to supervise the boundary of geo objects and helping segment the accurate boundary.

The proposed model embeds the cascaded relation attention module into the multiscale feature representation module, and the cascaded relation attention module is placed between the encoder and decoder. First, before feature fusion, the encoded features undergo the dual attention of channel and spatial relations to enhance the distinguishability of feature representation, reduce the intraclass variance, and increase the interclass variance. Then, after information connection and error correction, the enhanced high-level semantic features fully contain the detailed information of the geo objects. Finally, the boundary affinity loss supervises the object boundary at the output end, which promotes the optimal performance of the two main modules in the model. The three parts in this article interact with one another and jointly enhance the "feature distinguishability of geo object details."

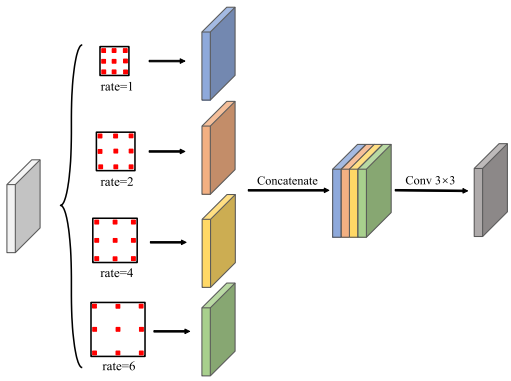

Fig. 2. Multiscale feature module.

B. Multiscale Feature Representation

This module is mainly composed of a feature pyramid and a multiscale feature module. The feature pyramid is used to extract effective high-level semantic features, including a bottom-up encoder and a top-down scale restoration. The encoder uses ResNet101 network as the feature extractor and atrous convolution to keep the feature map with the smallest scale (8 times downsampling). Three feature stages exist in the top-down process of the feature pyramid. The feature map of each feature stage is obtained by fusing the feature map of the previous stage and the corresponding shallow feature map in the encoder. This operation can offset the missing spatial information in the deep feature map. The multiscale feature module is mainly inspired by the ASPP module [17], which uses a convolutional layer with different atrous ratios on the final output image in the feature pyramid to extract multiscale semantic features, as shown in Fig. 2. The difference is that this study removes the global average pooling branch in ASPP to avoid producing invalid noise

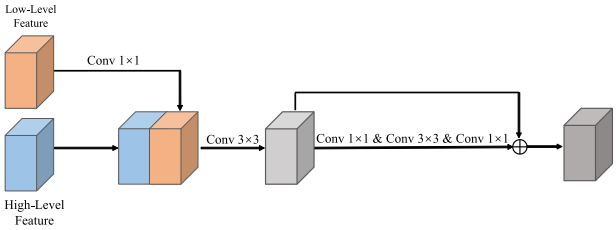

Fig. 3. Error correction block.

information, and the cascaded relation attention module in the model is used to capture the global context information fully. The output of multiscale branches undergoes feature fusion by a concatenation operation, and a 3×3 convolutional layer is used to aggregate multiscale features while decreasing the number of channels.

The shallow feature maps in the network contain rich spatial information. However, inherent semantic differences exist between low-level visual features and high-level semantic features. If the features of the two levels are integrated by adding or concatenating directly, a certain error will be produced in the feature representation of geo objects, which affects the effectiveness and robustness of the fused features. Therefore, this article refers to the idea in [8], in which an error correction block is used to eliminate semantic errors caused by different levels of feature fusion. This module mainly includes a 3×3 convolutional layer and a bottleneck block of ResNet. It eliminates fitting errors caused by the fusion of features of different semantic levels by learning the semantic differences in the feature map. The authors in [8] directly used an addition operation to perform simple feature fusion and then passed through a bottleneck block to eliminate the fusion error, thereby generating effective fusion features. In our study, the high- and low-level features are combined through concatenation operations to achieve fine feature fusion (Fig. 3). Then, the error correction block uses a 3×3 convolutional layer to realize preliminarily rough feature fusion and inputs the rough result into the bottleneck block for detailed error correction. Compared with the addition operation, the concatenation operation more easily captures effective features and achieves a robust feature representation. The concatenation operation can improve the flexibility of feature fusion here.

C. Cascaded Relation Attention

The relationship attention module determines the degree of feature interaction by considering the relationship among objects. This study aims to use the attention module to enhance the feature similarity among homogeneous objects and the feature difference among inhomogeneous objects. The spatial relation attention module considers the relationship among pixels at different spatial positions; thus, any location in the feature map can perceive the feature information of all other positions and capture the spatial dependence between any two positions in the feature map. The features of all positions are optimized and updated through weighted summation, and the weight is determined using the similarity of the corresponding two positions. The channel relationship attention module can consider the relationship among different feature channels; hence, any feature channel can perceive the information in all other feature channels and capture the channel dependency between any two feature channels. Each feature channel is updated and optimized using the weighted sum of all feature channels to produce an effective and robust feature representation.

The spatial relationship attention module mainly has three branches (Fig. 4). The inputs of A, B, and C branches are all obtained from the original input through a convolution operation. Among them, A branch uses the reshape and transpose operations to change the feature map size from $C \times H \times W$ to $N \times C$ (C represents channel, H represents height, W represents width, $N = H \times W$). B branch first reshapes the feature map size to $C \times N$, then implements matrix multiplication with the output result of A branch to obtain an attention weight map of size $N \times N$, and finally uses a softmax operation to normalize the attention weight. C branch transforms the feature map into $C \times N$ through a reshape operation and implements matrix multiplication with the normalized attention map to obtain an optimized feature map after spatial attention processing. The optimized result is matrix-added with the original feature map through an identity-mapping operation to obtain the final output feature map.

The channel relation attention module also has three branches (Fig. 4), which are similar to those in spatial attention. Nevertheless, the matrix multiplication of A and B branches includes a channel attention weight map with size $C \times C$, and a softmax operation is used to normalize the attention weight. The attention map and C branch output are subjected to a matrix multiplication operation to obtain an optimized feature map after channel attention processing. The optimized result is added to the original feature map through identity mapping to obtain the final output feature map.

The model in this article uses two types of relation attention modules to optimize dual relations and compensate for the shortcomings and one-sidedness of single attention modules. The output of the channel relation attention module is used as the input of the spatial relation attention module, which improves the coupling degree among different attention mechanisms and enhances the information flow between different relation attention modules. Compared with the attention calculation based on pooling, the attention map calculation based on convolution and matrix multiplication used in this study will not lose considerable spatial information in the feature map, thereby improving the optimization effect of the attention module.

D. Boundary Affinity Loss

The modeling of the relationship among objects is important in semantic segmentation because it can alleviate the problems of boundary blur and class variance. A clear correspondence should exist among pixels in remote sensing images. This study mainly uses attention modules to extract and adjust this relationship. In addition, a loss function is used at the network output to guide the learning of the relationship between boundary pixels.

Inspired by [35], which used affinity to express the pixel relationship of the geo-object boundary, our study uses affinity loss to capture and match the relationship between boundary pixels on the basis of the label space. Specifically, the K–L divergence

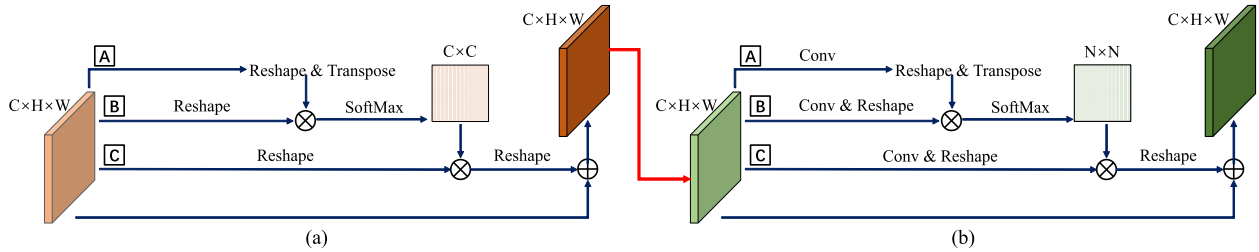

Fig. 4. Cascaded relation attention module. (a) Channel Relation Attention Module and (b) Spatial Relation Attention Module.

is used to express the affinity relationship between the pixels. The larger the K–L divergence is, the weaker the affinity will be. The affinity loss uses the real label as *a priori* knowledge to calculate the boundary affinity loss between the pixels at the boundary with a certain range of neighboring pixels. In formula (1), $\mathcal{L}^i_{BAL}$ represents the boundary affinity loss of category i, F^i represents the prediction probability map of category i after softmax processing, m is the pixel in F^i, m^r represents the neighborhood window with pixel m as the center and a range of r, and n is the pixel in the neighborhood window. $\mathcal{L}^i_{mn}$ is the affinity loss of pixels m and n in category i, K is the threshold affinity, $D_{KL}(.)$ represents the K–L divergence, $y_i(.)$ represents the real probability, $\overline{y_i}(.)$ represents the prediction probability, and the specific calculation is shown in formula (2). The final boundary affinity loss calculation is shown in formula (3), and N represents the number of categories

$$\mathcal{L}^i_{\mathrm{BAL}} = \sum_{m \in F^i} \sum_{n \in m^r} \mathcal{L}^i_{mn} \tag{1}$$

$$\mathcal{L}^i_{mn} = \begin{cases} \max\{0,\ K - D_{KL}(\overline{y_i}(m) \,||\, \overline{y_i}(n))\} & \text{if } y_i(m) \neq y_i(n) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$\mathcal{L}_{\mathrm{BAL}} = \sum_i \mathcal{L}^i_{\mathrm{BAL}} \quad (i = 1, 2, \ldots, N). \tag{3}$$

Compared with the cross-entropy loss that only performs loss supervision for a single pixel, the affinity loss can supervise the relationship between boundary pixels, standardize the affinity relationship between the boundary pixels, and compensate for the lack of cross-entropy loss. The affinity loss encourages two pixels with different labels to make a difference in prediction and realizes the supervision of geo-object boundaries, which helps create clear geo-object boundaries.

IV. Experimental Result Analysis

A. Dataset

Two standard airborne image datasets, Potsdam [36] and Vaihingen dataset [45], are used in our experiment to evaluate the performance of the proposed method in this article. Potsdam dataset shows a typical historical city, with large architectural blocks, narrow streets, and dense settlement structures. The Vaihingen dataset is a relatively small village with many detached buildings and small multistory buildings. Potsdam and Vaihingen dataset have been well-labeled manually and can be used as benchmark dataset in the research community to test the performance of networks [2], [8], [9], [30].

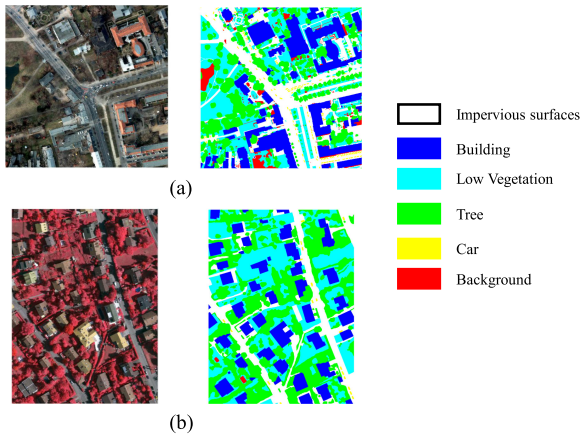

Fig. 5. Examples of Potsdam and Vaihingen dataset. Remote sensing image on the left and labeled ground truth on the right. (a) Potsdam and (b) Vaihingen.

Potsdam and Vaihingen dataset contain six categories: Impervious surfaces, building, low vegetation, tree, car, and background, as shown in Fig. 5. The dataset includes digital orthophotos and digital surface model (DSM) images. However, considering the inaccessibility and nonuniversality of DSM data, we do not use DSM data in the experiments and use only digital orthophotos to participate in the network training and testing.

The Potsdam dataset consists of 38 high-resolution images, each with a size of 6000 $\times$ 6000 and a spatial resolution of 5 cm. It consists of red, green, blue, and NIR bands. Only 24 of the 38 images have their label images. Therefore, this study uses the 24 images with real labels to train and test the proposed model. Six images numbered (2_12, 3_12, 4_12, 5_12, 6_12, and 7_12) are used to test the model performance, and the remaining 18 images are used for training models. The training image is cropped into subimages with a size of 256 $\times$ 256, given that the size of each image is excessively large to train. Finally, 18000 samples are obtained for training.

The Vaihingen dataset consists of 33 high-resolution images, each with an average size of 2494 $\times$ 2064 and a spatial resolution of 9 cm. It consists of green, blue, and NIR bands. Only 16 of the 33 images have their label images. Therefore, this study uses the 16 images with real labels to train and test the proposed model. Four images numbered (30, 32, 34, and 37) are used to test the model performance, and the remaining 12 images are used for training models. The training image is cropped into subimages with a size of 256 $\times$ 256, given that the size of each image is

TABLE I
QUANTITATIVE EVALUATIONS ON POTSDAM DATASET WITH DIFFERENT METHODS. ACCUARCY OF EACH CATEGORY IS PRESENTED IN THE OA/IOU FORM

Model	Backbone	Impervious Surfaces	Building	Low Vegetation	Tree	Car	Clutter/Background	OA	mIOU
FCN8s[16]	VGG16	92.93/82.90	91.29/86.59	87.62/75.49	77.91/65.06	83.99/74.35	22.09/15.48	86.33	66.65
Unet[20]	/	91.77/82.85	93.28/88.68	89.00/76.16	73.60/63.88	85.95/77.70	28.24/17.49	86.81	67.79
SegNet[46]	VGG16	89.70/80.59	93.05/86.57	86.96/75.11	75.78/63.89	82.51/70.00	20.85/12.23	85.53	64.73
PSPNet[37]	ResNet101	91.43/84.27	**94.63**/89.67	88.51/77.10	**80.05**/66.27	84.99/76.38	25.29/17.45	87.45	68.52
ICNet[39]	ResNet101	88.91/78.63	93.38/84.40	86.12/75.07	70.92/59.48	81.67/74.83	18.60/12.96	84.60	64.23
BiseNet[38]	ResNet18	89.14/79.04	91.61/83.68	88.45/75.54	72.96/62.37	80.34/72.97	19.97/13.29	84.81	64.48
DFANet[42]	Xception	91.19/82.21	92.20/86.53	90.07/76.41	72.39/61.81	82.93/75.65	21.48/14.30	86.17	66.15
DeepLabv3+[17]	ResNet101	91.51/83.11	92.60/87.82	88.99/76.94	78.42/65.36	87.17/80.51	30.27/17.46	86.91	68.53
DenseASPP[18]	DenseNet121	90.02/81.06	92.34/87.65	**89.53**/76.42	76.43/65.31	84.22/76.12	29.89/15.25	86.21	66.97
HRNetV2-W48[47]	/	90.83/80.88	91.01/84.79	87.05/75.38	77.31/62.06	82.77/75.15	19.18/11.52	85.27	64.96
OCNet[48]	ResNet101	92.32/84.61	92.21/88.08	89.03/77.64	77.45/65.50	86.90/76.67	29.49/17.32	87.03	68.30
CBAM[25]	ResNet101	92.31/82.68	90.83/86.26	90.28/76.00	74.47/64.14	82.64/75.27	22.97/15.61	86.28	66.66
EMANet[32]	ResNet101	91.42/84.48	94.56/**90.22**	89.36/76.79	76.96/**67.03**	86.96/79.38	**31.80**/17.87	87.63	69.30
DANet[19]	ResNet101	90.25/81.15	92.57/86.38	88.83/76.15	75.10/62.67	84.73/75.96	25.64/16.81	86.01	66.52
Ours	ResNet101	**93.88/85.94**	94.51/89.52	88.17/**78.03**	78.77/66.95	**89.06/81.13**	27.88/**20.37**	**88.20**	**70.32**

excessively large to train. Finally, 12000 samples are obtained for training.

B. Evaluation Metrics

This study uses mean intersection over union (mIOU) and overall accuracy (OA) as evaluation indicators to evaluate the semantic segmentation performance of the model. OA represents the proportion of correctly marked pixels in the total pixels. The larger the value is, the better the segmentation effect will be. mIOU is used to measure the average correlation at the category between the real and predicted results. The higher the correlation is, the better the segmentation effect will be. The specific calculation process is shown as follows:

$$OA = \frac{TP + TN}{FP + FN + TP + TN} \tag{4}$$

$$\text{IOU} = \frac{TP}{TP + FN + FP} \tag{5}$$

where *TP* represents true positive, *TN* represents true negative, *FP* represents false positive, and *FN* represents false negative.

C. Experimental Setting

In the experiments, the backbone of each network is shown in Table I, and the downsampling operation of all models is maintained at 8 times. All models are trained using Adam optimizer, a weight decay of 1e-4, a training epoch of 100, and an initial learning rate of 1e-3. We employ a poly learning rate policy to adjust the learning rate. The batch size of each model is set in the dynamic range of 8 to 16 due to the limitation of computer resources and the number of model parameters. At the time of testing, only the 128 $\times$ 128 prediction result of the central area of each 256 $\times$ 256 image is regarded as the final prediction. All experiments are performed on the PyTorch framework version 0.4.1 with one NVIDIA 1080Ti GPU.

D. Performance Comparison of Potsdam dataset

Table I shows the model test results on the Potsdam dataset. Among all the comparison models, the model with the worst effect is ICNet, whose OA is 84.60% and mIOU is 64.23%; the model with the best accuracy is EMANet, whose OA is 87.63% and mIOU is 69.30%. The proposed model achieves the best accuracy, with OA reaching 88.20% and mIOU reaching 70.32%. In terms of OA, the proposed model shows an increase by 3.6% over ICNet and by 0.57% over EMANet. In terms of mIOU, the proposed model presents an increase by 6.09% over ICNet and by 1.02% over EMANet. Experiments show that the proposed model can achieve excellent segmentation performance on the Potsdam dataset and is superior to state-of-the-art methods on OA and mIOU.

Compared with other models, the proposed model in this article achieves the better performance on all six categories, and all of them have a certain degree of accuracy improvement. The proposed model achieves good segmentation accuracy on impervious surfaces, buildings, and cars. OA reaches more than 89%, and IOU reaches more than 81%. As for impervious surfaces and cars, the performance of this proposed model is better than all comparison models on OA and IOU. As for building, the proposed model achieves the good performance on OA and IOU, which is second only to EMANet and PSPNet. The distinguishability of these three categories is relatively high, and they can easily gain effective feature representations. Therefore, most models can achieve improved accuracy on impervious surfaces, buildings, and cars. However, the proposed model can achieve higher accuracy improvement than the comparison methods because it has advantages in capturing the details of geo objects. The feature fusion and error correction strategies can effectively restore the fine boundaries of the objects, which helps further improve the segmentation accuracy.

As for the low vegetation and the trees, the proposed model is considerably different from the other models. The OA of the proposed model on low vegetation is lower than most comparison models, but the best accuracy is achieved on IOU. The recall rate of the proposed method on low vegetation is low, but the accuracy rate is high. The OA of the proposed method on trees is better than that of all comparison models except PSPNet, and the proposed model is also better than all comparison models except EMANet on IOU (only 0.08% lower than EMANet), which shows that the proposed method can achieve an effective feature representation on trees.

As for low vegetation and trees, the proposed model achieves 83.47% and 72.49% accuracy on the average OA and IOU. The average IOU is better than that of all comparison models. The

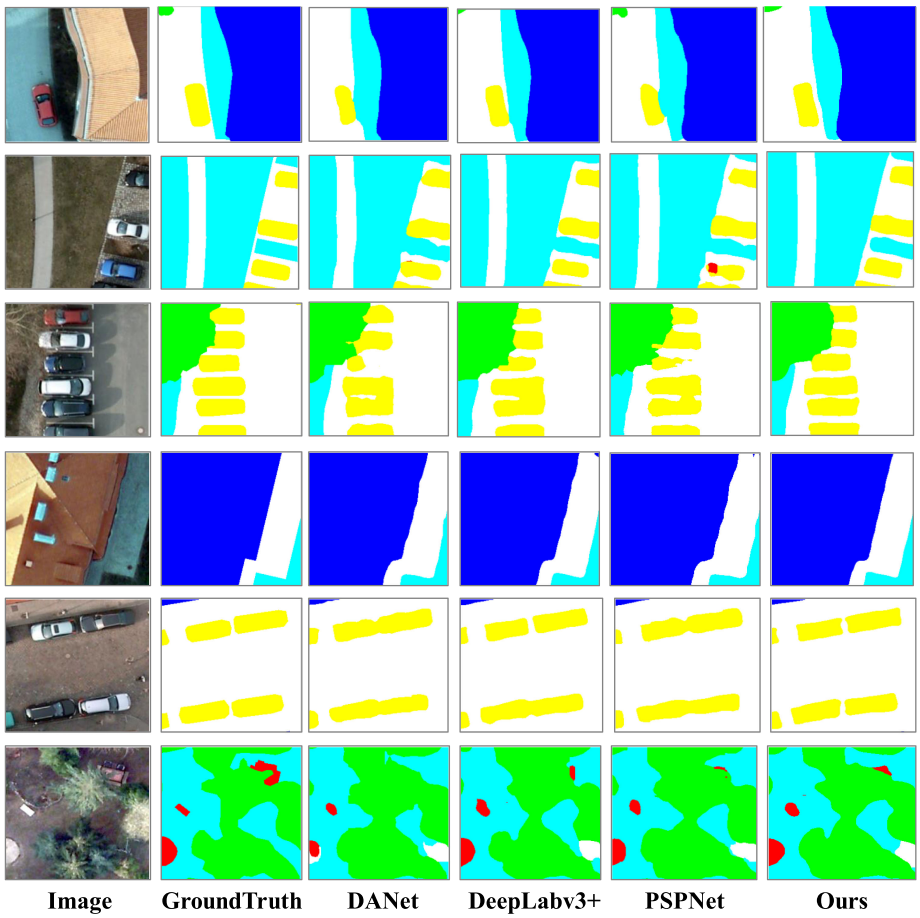

Fig. 6. Visualization results on the Potsdam test dataset.

average OA is better than that of all comparison models except PSPNet and DeepLabv3+. Low vegetation and trees are easily confusing due to the high similarity in spectrum features, which leads to most neural network models having great shortcomings in distinguishing capability on these two categories. However, the cascaded relation attention module in the proposed method can capture global context information and capture the feature correlation on channels or positions. As a result, the feature similarity of homogeneous objects and the feature difference of inhomogeneous objects are improved. Therefore, the proposed model can enhance the ability to distinguish between low vegetation and trees, effectively alleviating the issue of "feature distinction of easily confusing geo objects" and achieving enhanced segmentation performance.

Clutter/background is the most difficult category to segment because the feature representations of this category are complex and most geo objects are meaningless in this category. Consequently, the accuracy of all models is unsatisfactory on this category. Nonetheless, the proposed model can still achieve relatively good performance in general. OA reaches 27.88%, which is better than that of most comparison models; mIOU reaches 20.37%, which indicates the best performance.

The visualization results are shown in Fig. 6. In visualization comparison, DANet, DeepLabv3+, and PSPNet all have a certain degree of competitiveness on the Potsdam dataset; all of them achieve enhanced segmentation performance. All these methods have comparable model structures. For example, DANet has a parallel attention module, DeepLabv3+ conducts a fusion of low- and high-level features, and PSPNet includes a pyramid pooling layer for multiscale feature extraction. Thus, these three models can be used as a visualization comparison benchmark to evaluate the semantic segmentation effect of the proposed model.

The proposed model has a good feature representation capability; hence, it has a good effect in maintaining the integrity and independence of geo objects. For example, in the results of the third and fifth rows of Fig. 6, the proposed model can segment each individual vehicle well and ensure the integrity of the vehicle; in the segmentation results of other models, some vehicles are glued together, or some vehicles are divided into a small part.

Moreover, the proposed model has a good effect in distinguishing objects with confusing features. For example, in the visualization results of the sixth row of Fig. 6, low vegetation and trees are two types of geo objects that are difficult to distinguish. Among the segmentation results of the three comparison models, the segmentation effect of low vegetation and trees is relatively rough. For example, in the upper half of the image in the sixth row of Fig. 6, the confusion of two similar objects is serious. Nevertheless, the proposed model can distinguish the two geo objects, and the segmentation effect is greatly improved compared with that of the three other models. The

TABLE II
QUANTITATIVE EVALUATIONS ON VAIHINGEN DATASET WITH DIFFERENT METHODS. ACCUARCY OF EACH CATEGORY IS PRESENTED IN THE OA/IOU FORM

Model	Backbone	Impervious Surfaces	Building	Low Vegetation	Tree	Car	OA	mIOU
FCN8s[16]	VGG16	92.37/81.50	90.96/83.71	80.61/66.11	77.22/67.98	60.61/50.78	85.82	70.02
Unet[20]	/	93.93/81.43	90.93/83.84	79.08/65.72	75.56/67.63	**73.03/61.79**	85.87	72.08
SegNet[46]	VGG16	92.20/77.97	88.62/80.98	78.42/64.46	75.14/67.13	58.56/46.17	84.28	67.34
PSPNet[37]	ResNet101	92.47/82.66	93.09/85.53	**83.23**/68.29	77.07/69.44	59.64/54.02	86.91	71.99
ICNet[39]	ResNet101	93.16/81.43	90.88/84.72	81.28/66.89	77.43/68.89	62.92/56.42	86.28	71.67
BiseNet[38]	ResNet18	93.18/80.81	89.99/83.98	79.60/65.58	**77.92**/68.53	63.08/56.47	85.79	71.07
DFANet[42]	Xception	93.39/80.85	90.51/84.12	80.22/65.95	75.69/67.45	63.81/53.43	85.72	70.36
DeepLabv3+[17]	ResNet101	93.53/80.50	86.26/81.38	80.85/66.32	76.18/67.62	71.32/52.84	85.05	69.73
DenseASPP[18]	DenseNet121	93.66/82.39	91.24/85.81	82.19/67.51	76.62/68.20	68.72/60.55	86.67	72.89
HRNetV2-W48[47]	/	92.58/82.00	**93.41**/85.50	81.10/66.68	75.68/68.11	67.73/58.91	86.43	72.24
OCNet[48]	ResNet101	92.50/81.92	90.75/83.13	80.23/65.46	76.99/68.52	70.60/57.84	85.84	71.37
CBAM[25]	ResNet101	93.32/80.34	90.14/84.09	79.36/66.28	**77.92**/68.99	63.63/51.82	85.83	70.30
EMANet[32]	ResNet101	**94.35**/81.56	90.35/84.83	81.65/67.41	76.56/**69.52**	68.31/60.54	86.54	72.77
DANet[19]	ResNet101	93.44/80.94	89.20/83.68	82.67/67.40	75.80/68.25	66.66/57.86	86.01	71.63
Ours	ResNet101	93.44/**83.24**	92.14/**86.14**	83.21/**68.54**	76.90/68.92	69.38/60.48	**87.12**	**73.46**

cascaded relation attention module adopted in this model can establish the relationship among geo objects, thereby improving the similarity of homogeneous geo objects and the difference of inhomogeneous geo objects and enhancing the distinguishability of features.

Besides, the proposed model also has an enhanced segmentation effect at the object boundary, and the boundary is relatively smooth. By contrast, the segmentation results of the other comparison models at the boundary are uneven and inadequately flat, and the boundary positioning is insufficiently accurate. For example, in the results of the first and fourth rows of Fig. 6, the segmentation effect of the proposed model at the building boundary is smoother than that of the comparison models, and the positioning is more accurate. In the segmentation result of the second row of Fig. 6, a small low vegetation area exists between the blue and white vehicles which is separated from the low vegetation area in the middle of the image. This area belongs to the detailed information in the image. None of the three other models can separate these two regions well, but the proposed model accurately separates these regions. This result shows that the proposed model has a strong ability to capture geo-object details, and feature fusion and error correction play an important role in segmentation results.

E. Performance Comparison of Vaihingen dataset

Table II shows the model test results on the Vaihingen dataset. The proposed model in this article achieves excellent performance on Vaihingen dataset, and achieves 87.20% on OA and 70.12% on mIOU. Among the comparison models, the model with the best comprehensive performance is DenseASPP, and the model with the worst comprehensive performance is SegNet. In terms of OA, the proposed model shows an increase by 2.84% over SegNet and by 0.45% over DenseASPP. In terms of mIOU, the proposed model presents an increase by 6.12% over SegNet and by 0.57% over DenseASPP.

To be specific, the proposed model also achieves excellent segmentation performance on different categories. On impervious surfaces, buildings, and low vegetation, the proposed model achieves best performance in mIOU and well performance on OA, the OA of the proposed model is higher than that of most comparison models and close to the best OA. In terms of tree, the proposed model achieves 76.90% on OA and 68.92% on mIOU, and the performance is better than that of most comparison models. In terms of cars, the proposed model in this article achieves 69.38% on OA and 60.48% on mIOU. Among all the comparison models, the performance of the proposed model is second only to that of Unet on cars. UNet achieves the best accuracy on cars, which is significantly better than that of other models, because the multiple information connections supplement a large amount of geo-object details in the decorder of UNet. Therefore, the segmentation performance of small scale fine geo objects such as vehicles is greatly improved.

Low vegetation and trees are easily confusing geo objects. The proposed model achieves 80.06% accuracy on average OA and 68.73% accuracy on average IOU on low vegetation and trees. The proposed model achieves excellent accuracy, which is second only to PSPNet with about 0.1% gap. However, on the whole, the segmentation performance of the proposed model is better than that of PSPNet.

In visualization comparison, we also choose PSPNet, Deeplabv3+, and DANet as comparison models, because they can achieve good performance on the Vaihingen dataset. The specific visualization results are shown in Fig. 7. It can be clearly seen that compared with the other three models, the proposed model has great advantages in the integrity and independence of geo objects. For example, as for cars, the proposed model can efficiently segment each car without adhesion, and each car maintains a complete shape and construction. However, the other three models all appear the adhesion and incompleteness on cars.

On the other hand, the proposed model has a well performance in the geo-objects' boundary, especially the artificial geo objects. Compared with the other three models, the proposed model can segment smoother boundary, and the location of the boundary is more accurate. It can be clearly seen from the visualization results in Fig. 7 that the segmentation effect of the proposed model is better than that of the other three models on the boundary of buildings, and the obtained boundary is smooth with accurate position.

In addition, the distinguishing ability of the proposed model among low vegetation and trees is better than that of the other three models. Because of the complex spectrum features, the

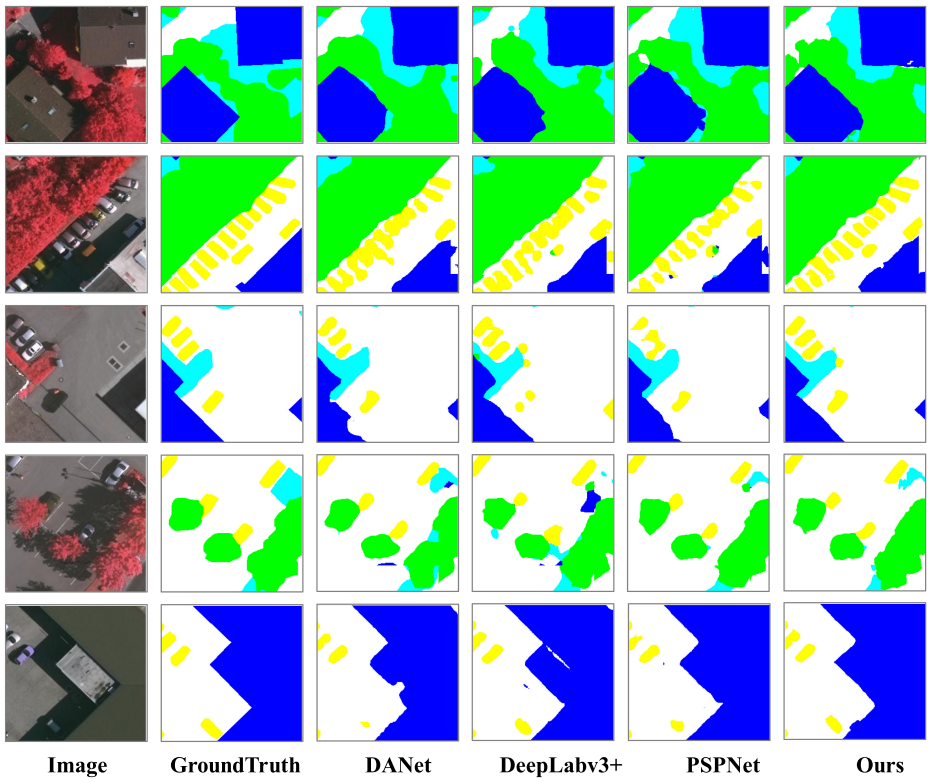

Fig. 7. Visualization results on the Vaihingen test dataset.

TABLE III
ABLATION ANALYSIS ON POTSDAM DATASET

Multiscale Feature Pyramid	Feature Representation Module: Error Correction	Cascaded Relation Attention Module: Channel Relation Attention	Cascaded Relation Attention Module: Spatial Relation Attention	Boundary Affinity Loss	OA	mIOU
√	√			√	86.89	68.65
		√	√	√	87.59	69.62
√		√	√	√	87.84	69.94
√	√	√		√	87.80	70.05
√	√		√	√	87.71	69.54
√	√	√	√		87.17	68.25
√	√	√	√	√	**88.20**	**70.32**

two kinds of geo objects are often confusing together in most cases. In the visualization results of comparison models, the segmentation result of these two types of geo objects is relatively fuzzy. However, the proposed model can effectively improve the segmentation effect among confusable categories, which mainly depends on the cascaded relation attention module in the proposed model.

F. Ablation Analysis

A series of ablation experiments is performed to prove the effectiveness of the proposed method in this article. A two-part ablation experiment is conducted on the Potsdam dataset. One part is to explore the importance of the cascaded relation attention module and the multiscale feature representation module in the network structure; the other part is to evaluate the importance of the internal structure of the two modules.

The results of the ablation experiment are shown in Table III. When the model contains only the multiscale feature representation module, the segmentation accuracy is the worst, OA is 86.89%, and mIOU is 68.65%. When the model contains only the cascaded relation attention module, OA is 87.59% and mIOU is 69.62%. When the model includes these two modules, the segmentation accuracy is the best, OA is 88.20% and mIOU is 70.32%. Experiments show that the accuracy improvement contribution of the cascading relational attention module is greater than that of the multiscale feature representation module. The cascaded relation attention module in the proposed model is the main part of feature representation and can greatly improve the semantic segmentation. The multiscale feature representation module is mainly responsible for capturing the detailed information of the geo objects in the image, further improving and enriching the original high-level semantic features.

Further analysis of the internal structure of the two modules shows that when using cascaded relation attention module and multiscale feature representation module without error correction block, OA is 87.84% and mIOU is 69.94%. Compared with only using cascaded relation attention module, the accuracy of

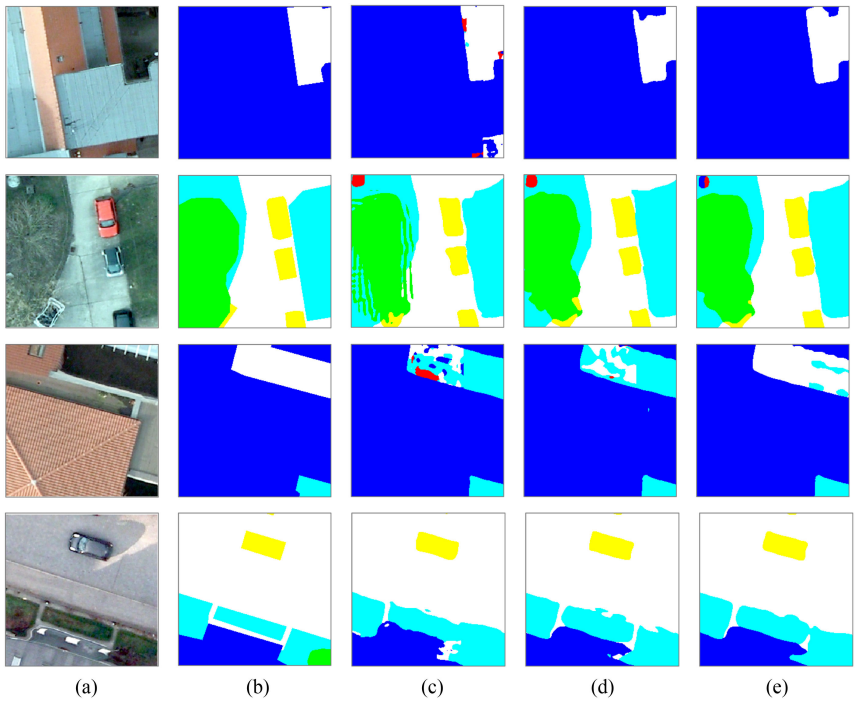

Fig. 8. Visualization results of ablation analysis on the Potsdam test dataset. (a) represents the original image; (b) represents the ground truth; (c) represents the model containing only the multiscale feature representation module; (d) represents the model containing only the cascaded relation attention module; and (e) represents a complete model that includes multiscale feature representation module and cascaded relation attention module. (The above three models all use boundary affinity loss).

the model is not improved much. That is, OA increased by 0.25% and mIOU increased by 0.32%. However, when using cascaded relation attention module and multiscale feature representation module with error correction block, OA is 88.20% and mIOU is 70.32%, the performance is greatly improved, OA increased by 0.61%, mIOU increased by 0.7%. The effect of error correction block is very significant. When the error correction block is not used, the semantic error generated by the fusion of different levels of features may not be effectively eliminated. Therefore, after direct fusion of low-level visual features and high-level semantic features, the low-level visual features may produce a negative effect on the high-level semantic features, interfering with the original feature representation.

When only the channel or spatial relation attention is used in the proposed model, the segmentation accuracy of the model is reduced to a certain extent. This result shows that a single relation attention module cannot achieve the optimal effect. The effect of channel relation attention is better than that of spatial relation attention, but the improvement is limited. The dual relation attention module by coupling the channel and spatial relation attention can achieve an effective feature representation. When the proposed model contains only the cascaded relation attention module, the structure is similar to that of DANet, but the two types of attention modules in DANet are combined in parallel. With reference to Tables I and III, compared with DANet, the proposed model is increased by 1.58% on OA and 3.1% on mIOU, which indicates that the proposed strategy in the attention module contributes to the improvement of model performance.

The boundary affinity loss also plays an important role. When the proposed model does not use boundary affinity loss, the segmentation performance of the model decreases significantly, OA is decreased by 1.03% and mIOU is decreased by 2.07%. Experiments show that the boundary affinity loss can effectively supervise the feature learning of the model and compensate for the cross-entropy loss that cannot consider the relationship among objects, prompting the model to pay further attention to the boundary and enhancing the segmentation effect at the boundary of the geo objects.

To analyze the importance of these two main modules, Fig. 8 shows the visualization of the segmentation results of three models: Using only the multiscale feature representation module, using only the cascaded relation attention module, and using both modules. When the model contains only multiscale feature representation modules, the accuracy of the model is the worst, and the integrity of the geo objects in the segmentation result, such as the buildings in the first row and the trees in the second row of Fig. 8, is poor. Although the multiscale feature representation module can capture detailed information and multiscale features, the features extracted using the encoder are insufficiently effective, and the feature representation capability of the model is limited. Consequently, the segmentation effect of the model is relatively poor.

When the model contains only the cascaded relation attention module, the feature representation capability of the model is enhanced. As a result, the integrity of the geo objects in the segmentation result is well maintained, and the overall segmentation accuracy is improved to a certain extent. However, the

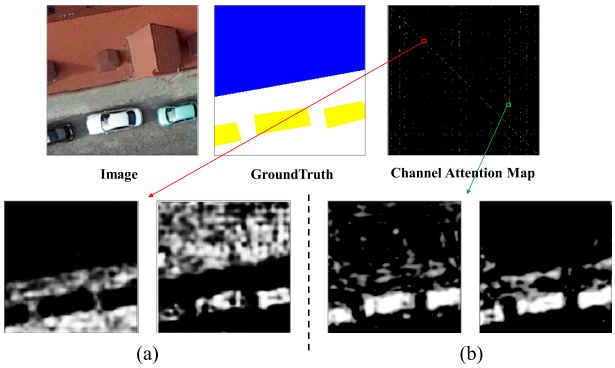

Fig. 9. Visualization analysis of the channel relation attention module. (a) represents an unrelated pair of feature maps; and (b) represents a highly correlated pair of feature maps.

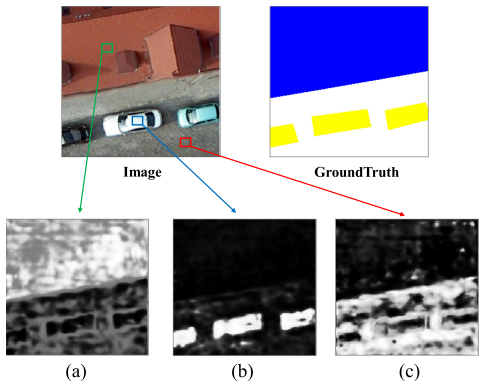

Fig. 10. Visualization analysis of the spatial relation attention module. (a) represents the spatial attention map of the pixel in the green box; (b) represents the spatial attention map of the pixel in the blue box; and (c) represents the spatial attention map of the pixel in the red box.

segmentation at the boundary of the geo objects is rough and inadequately smooth due to the lack of detailed information of the geo objects. For example, the boundary between buildings and low vegetation in the fourth row of Fig. 8 is rough. When both modules are included, the model proposed in this article can achieve the integrity of the geo objects and the smoothness of the boundary of the geo objects. The model can also improve the ability to distinguish the objects with confusing features, such as the accurate segmentation between roads and low vegetation in the upper half of the image in the second row of Fig. 8.

Therefore, the multiscale feature representation module proposed in this article can fully pay attention to the detailed information of geo objects and obtain the accurate boundary of the geo objects. The cascaded relation attention module can effectively enhance the feature representation capability of the proposed model and improve the distinguishability of features.

G. Visualization Analysis

This part presents the visual analysis of the cascaded relation attention module and the error correction block and the study of the internal structure of these two modules and their role in the proposed model.

Fig. 9 shows the visualization analysis of the channel relation attention module. For example, the pixel in the fifth row and the sixth column of the channel attention map represents the correlation weight between the fifth and sixth feature channels. If the feature representations of the two feature channels are similar, the correlation between the two is high, and the relationship is close, then they should appear as a high pixel value in the attention map. The feature channel corresponding to the pixel in the red box is shown in Fig. 9(a). Given that the pixel appears as a dark spot in the image, the correlation between the corresponding feature channels is weak. The specific feature map also depicts that the semantic information conveyed by the feature map on the left in Fig. 9(a) is roads, and the semantic information conveyed by the feature map on the right in Fig. 9(a) is buildings and vehicles. Their feature representations have a low correlation; consequently, the corresponding attention weight value should be low. The pixels in the green box in Fig. 9 are bright; hence, the corresponding feature channel is relevant. The specific feature channel is shown in Fig. 9(b). The semantic information in the left and right feature maps is mainly vehicles. The feature representation similarity of their two feature channels is high, and the corresponding attention weight is also high.

In summary, the channel relation module in the proposed model can fully capture the relationship among different feature channels, thereby enhancing the original feature representation.

Fig. 10 shows the visualization analysis of the spatial relation attention module. The images in the second row represent the spatial attention map corresponding to the pixels at different positions in the original image. The brighter the pixel in the attention map, the higher the correlation between two pixels will be. Therefore, additional attention weights will be allocated at the pixel to achieve the purpose of enhancing the similarity of pixels of the same class and the difference in pixels of different classes. The semantic category of the pixel in the green box is building, the brighter area in the attention map is the building part of the original image, and the other parts in the image are darker. The semantic category of the pixel in the blue box is car, and the area of the car in the corresponding attention image is brighter and will be given more attention. The semantic category of the pixel in the red box is impervious surfaces, and all the area of impervious surfaces in the corresponding attention map are given more attention weight.

In summary, the spatial relation attention module in the proposed model can fully consider the relationship among pixels in an image and capture the information of the global context. Therefore, it can improve the similarity among pixels of the same category and the difference among pixels of different categories and enhance the original feature representation.

Fig. 11 shows the visualization analysis of the error correction block. The last three columns are feature maps that have not undergone feature fusion, without error correction, and undergone error correction, respectively. The feature maps without feature fusion are all high-level semantic information. However, the boundaries of these geo objects are blurry and unclear because the downsampling operation loses substantial position information. In the fused feature map without error correction, the boundaries of the objects become smooth, the positioning is accurate, and the image details are rich due to

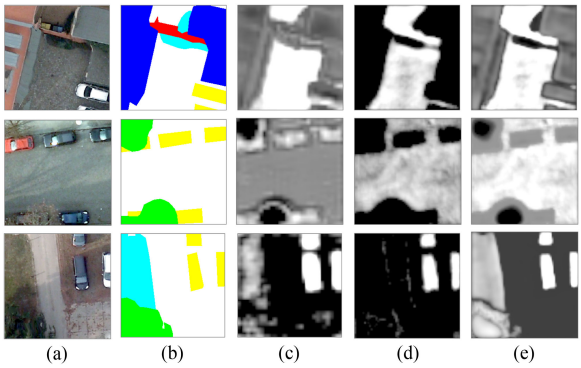

Fig. 11. Visualization analysis of the error correction block. (a) represents the original image; (b) represents the ground truth; (c) represents the high-level semantic feature map without feature fusion; (d) represents the fused feature map without error correction; and (e) represents the fused feature map with error correction.

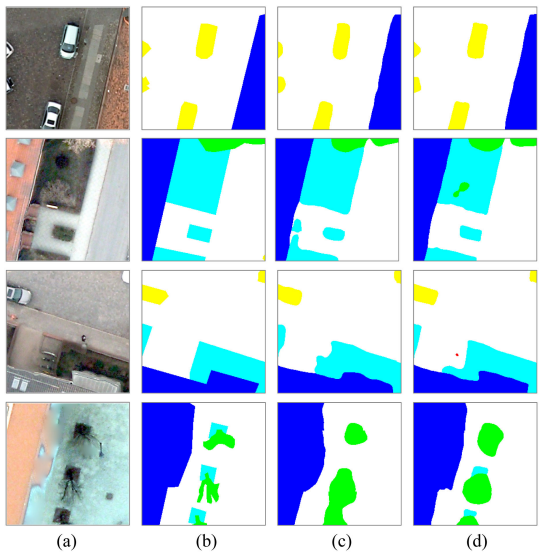

Fig. 12. Visualization analysis of the boundary affinity loss. (a) represents the original image; (b) represents the ground truth; (c) represents model unuse the boundary affinity loss during training; and (d) represents use the boundary affinity loss during training.

the fusion of low-level visual features. However, the original semantic information of objects is destroyed due to the semantic differences among different levels, which affects the feature representation of objects. For example, the original tree and low vegetation information in the feature map in the third row of Fig. 11 is lost after feature fusion. In the fused feature map with error correction, the objects are smooth and clear. The broken semantic information is well restored after error correction, which is more effective than the original semantic information.

In summary, the error correction block in the proposed model can eliminate the feature error generated by the fusion of different levels of features. Therefore, it can achieve effective feature fusion, enhance the original feature representation, and restore the missing geo-object details in an image.

Fig. 12 shows the visualization analysis of the boundary affinity loss. When the proposed model does not use the boundary affinity loss for training, the segmentation effect of the model is relatively rough, the boundary is not smooth enough, and the boundary positioning is not accurate enough. In particular, this situation is more obvious on artificial objects, such as car and building. When the boundary affinity loss is used in the proposed model, it can be clearly seen that the boundary of objects becomes relatively smooth, close to the real situation, and the boundary positioning is more accurate.

In summary, the proposed boundary affinity loss in this article can better monitor the affinity relationship of pixels on both sides of geo-object boundary, the difference between different categories of pixels should be large enough. Thus, the proposed model can create the exact geo-object boundary.

V. Conclusion

In this research, we focus on the "feature distinguishability of geo-object details" and propose a novel semantic segmentation model. It is composed of a multiscale feature representation module, a cascaded relation attention module, and boundary affinity loss. The proposed model can fully pay attention to the relationship among feature representations to improve the feature similarity of homogeneous geo objects and the feature difference of inhomogeneous geo objects. Moreover, it can capture and fuse effective feature information and realize the learning of distinguishable features of geo-object details. Compared with other related segmentation models, the proposed model achieves better segmentation performance. In addition, improved results can be obtained in the boundaries of geo objects and even the easily confusing geo objects.

In future research, we will continue to study different feature fusion methods in-depth and improve and optimize the attention module further to achieve a highly effective feature representation.

References

[1] S. M. Azimi, P. Fischer, M. Korner, and P. Reinartz, "Aerial laneNet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2920–2938, May 2019.

[2] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 91–105, 2019.

[3] X. Yang, X. Li, Y. Ye, R. Y. K. Lau, X. Zhang, and X. Huang, "Road detection and centerline extraction via deep recurrent convolutional neural network U-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7209–7220, Sep. 2019.

[4] J. Dyson, A. Mancini, E. Frontoni, and P. Zingaretti, "Deep learning for soil and crop segmentation from remotely sensed data," *Remote Sens*, vol. 11, no. 16, 2019, Art. no. 1859.

[5] J. D. Sylvain, G. Drolet, and N. Brown, "Mapping dead forest cover using a deep convolutional neural network and digital aerial photography," *ISPRS J. Photogrammetry Remote Sens., Article*, vol. 156, pp. 14–26, Oct. 2019.

[6] S. J. Liu and Q. Shi, "Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China," *ISPRS J. Photogrammetry. Remote Sens.*, vol. 164, pp. 229–242, Jun. 2020.

[7] B. Fang, L. Pan, and R. Kou, "Dual learning-based siamese framework for change detection using bi-temporal VHR optical remote sensing images," *Remote Sens*, vol. 11, no. 11, 2019, Art. no. 1292.

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

[9] 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 J. Photogrammetry Remote Sens.*, vol. 156, pp. 1–13, 2019.

[10] X. Xie, G. Xie, X. Xu, L. Cui, and J. Ren, "Automatic image segmentation with superpixels and image-level labels," *IEEE Access*, vol. 7, pp. 10999–11009, 2019.

[11] Y. Wang, W. Yu, and Z. Fang, "Multiple kernel-based SVM classification of hyperspectral images by combining spectral, spatial, and semantic information," *Remote Sens*, vol. 12, no. 1, p. 120, 2020.

[12] C. Yoo, D. Han, J. Im, and B. Bechtel, "Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 157, pp. 155–170, 2019.

[13] L. Dong *et al.*, "Very high resolution remote sensing imagery classification using a fusion of random forest and deep learning technique—Subtropical area for example," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 113–128, 2020.

[14] D. C. Cirean, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 2852–2860.

[15] C. Farabet, C. Couprie, L. Najman, and Y. Lecun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2015.

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

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

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

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

[21] S. Ghassemi, A. Fiandrotti, G. Francini, and E. Magli, "Learning and adapting robust features for satellite image segmentation on heterogeneous data sets," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6517–6529, Sep. 2019.

[22] J. Chen, J. Zhu, G. Sun, J. Li, and M. Deng, "SMAF-Net: Sharing multiscale adversarial feature for high-resolution remote sensing imagery semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, to be published, doi: 10.1109/LGRS.2020.3011151.

[23] J. Chen, F. He, Y. Zhang, G. Sun, and M. Deng, "SPMF-Net: Weakly supervised building segmentation by combining superpixel pooling and multi-scale feature fusion," *Remote Sens*, vol. 12, no. 6, 2020, Art. no. 1049.

[24] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[25] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis*, 2018, pp. 3–19.

[26] L. C. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens., Article*, vol. 58, no. 1, pp. 110–122, Jan. 2020.

[27] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[28] W. Cheng, W. Yang, M. Wang, G. Wang, and J. Chen, "Context aggregation network for semantic labeling in aerial images," *Remote Sens*, vol. 11, no. 10, 2019, Art. no. 1158.

[29] C. Liu, D. Zeng, H. Wu, Y. Wang, S. Jia, and L. Xin, "Urban land cover classification of high-resolution aerial imagery using a relation-enhanced multiscale convolutional network," *Remote Sens*, vol. 12, no. 2, p. 311, 2020.

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

[31] S. Liu, W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li, "ERN: Edge loss reinforced semantic segmentation network for remote sensing images," *Remote Sens*, vol. 10, no. 9, 2018, Art. no. 1339.

[32] X. Li *et al.*, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis*, 2019, pp. 9166–9175.

[33] B. Zhang, Y. Zhang, and S. Wang, "A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2636–2653, Aug. 2019.

[34] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-Scale context aggregation for semantic segmentation of remote sensing images," *Remote Sens*, vol. 12, no. 4, p. 701, 2020.

[35] T. W. Ke, J. J. Hwang, Z. Liu, and S. X. Yu, "Adaptive affinity fields for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 605–621.

[36] ISPRS potsdam 2D semantic labeling dataset. Accessed: Dec. 10, 2017. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

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

[38] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis*, 2018, pp. 325–341.

[39] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis*, 2018, pp. 405–420.

[40] A. Farooq, X. Jia, J. Hu, and J. Zhou, "Multi-Resolution weed classification via convolutional neural network and superpixel based local binary pattern using remote sensing images," *Remote Sens*, vol. 11, no. 14, 2019, Art. no. 1692.

[41] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[42] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9522–9531.

[43] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, "RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2043–2056, Apr. 2019.

[44] 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 Trans. Geosci. Remote Sens., Article*, vol. 57, no. 10, pp. 7503–7520, Oct 2019.

[45] ISPRS Vaihingen 2D semantic labeling dataset. Accessed: Dec. 10, 2017. [Online]. Available: http://www2.isprs.org/commissions/ comm3/wg4/2d-sem-label-vaihingen.html

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

[47] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2020.2983686.

[48] Y. Yuan and J. Wang, "OCNet: Object context network for scene parsing," 2018, *arXiv:1809.00916*.

[49] Y. Wang, B. Liang, M. Ding, and J. Li, "Dense semantic labeling with atrous spatial pyramid pooling and decoder for high-resolution remote sensing imagery," *Remote Sens*, vol. 11, no. 1, 2019, Art. no. 20.