Feature Guide Network With Context Aggregation Pyramid for Remote Sensing Image Segmentation

Jiaojiao Li, *Member, IEEE*, Yuzhe Liu, Jiachao Liu, Rui Song, *Member, IEEE*, Wei Liu, Kailiang Han, and Qian Du, *Fellow, IEEE*

***Abstract*—In recent years, the deep learning method based on fully convolution networks has proven to be an effective method for the semantic segmentation of remote sensing images (RSIs). However, the rich information and complex content of RSIs make networks training for segmentation more challenging. Specifically, the observing distance between the space-borne cameras and the ground objects is extraordinarily far, resulting in that some smaller objects only occupy a few pixels in the image. However, due to the rapid degeneration of tiny objects during the training process, most algorithms cannot properly handle these common small objects in RSIs with satisfactory results. In this article, we propose a novel feature guide network with a context aggregation pyramid (CAP) for RSIs segmentation to conquer these issues. An innovative edge-guide feature transform module is designed to take advantage of the edge and body information of objects to strengthen edge contours and the internal consistency in homogeneous regions, which can explicitly enhance the representation of tiny objects and relieve the degradation of small objects. Furthermore, we design a CAP pooling strategy to adaptively capture optimal feature characterization that can assemble multiscale features according to the significance of different contexts. Extensive experiments on three large-scale remote sensing datasets demonstrate that our method not only can outperform the state-of-the-art methods for objects of different scales but can also achieve robust segmentation results, especially for tiny objects.**

***Index Terms*—Context aggregation pyramid (CAP), deep learning, edge guide, remote sensing images (RSIs), semantic segmentation.**

I. Introduction

THE semantic segmentation of remote sensing imagery, which assigns a unique category label for each pixel in space-captured images of earth, is a fundamental component for infrastructure planning, territorial change detection, and environmental monitoring [1], [2], [3]. Unlike natural images, the semantic segmentation of subdecimeter aerial images faces many challenges due to the complex remote sensing contents and illumination conditions, making it challenging to obtain fine-grained semantic segmentation results [4], [5].

A major challenge is that the vast imaging range of aerial satellites results in the remote sensing scenes usually being accompanied by high complexity and diversity so that the ground objects in the scene have very great and different geometric shapes, sizes, and textures, bringing about great difficulties to the feature extraction of remote sensing images (RSIs) [6], [7]. In addition, the downsampling operation in deep convolutional neural networks gradually leads to the disappearance of details in the original image, which may lose a lot of pivotal information and result in unsatisfactory segmentation prediction, especially for boundary details and small objects [8]. A generally adopted method to handle this drawback is to fuse the low-level features containing more detailed information and the high-level features with more semantic information by a decoder structure. For instance, fully convolution networks (FCNs) connect the feature maps from shallow layers to the output feature for gaining more spatial location information. UNet combines the features from the corresponding low layers during every upsampling operation in the resolution recovery stage [9]. DeepLabv3+[10], as an outstanding framework, embeds the features from the first residual convolution group into the refined features. However, these methods handle all the information, such as colors, textures, and edges, in the same way, which ignores the disparate contributions of specific information to the task of semantic segmentation.

Another challenge is the severe category imbalance among different classes in the RSI, which is relatively rare in natural images but quite common in space-captured images of Earth. Furthermore, the distinction between interclass is relatively small. Specifically, roads and cars are ordinarily very tiny and

Manuscript received 17 August 2022; revised 13 October 2022; accepted 8 November 2022. Date of publication 14 November 2022; date of current version 23 November 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102702, in part by the Fundamental Research Funds for the Central Universities under Grant JBF220101, in part by the National Nature Science Foundation of China under Grant 61901343, in part by the State Key Laboratory of Geo-Information Engineering under Grant SKLGIE2020-M-3-1, in part by the Science and Technology on Space Intelligent Control Laboratory under Grant ZDSYS-2019-03, in part by the China Postdoctoral Science Foundation under Grant 2017M623124, in part by the China Postdoctoral Science Special Foundation under Grant 2018T111019, in part by by the Open Research Fund of CAS Key Laboratory of Spectral Imaging Technology under Grant LSIT201924W, in part by the Government-Business-University-Research Cooperation foundation between Wuhu and Xidian under Grant XWYCXY-012021002-HT, in part by the National Nature Science Foundation of China under Grant 61571345, Grant 61671383, Grant 91538101, Grant 61501346, and Grant 61502367, in part by the 111 Project under Grant B08038, and in part by the Innovation Fund of Xidian University under Grant 10221150004. *(Corresponding authors: Rui Song; Kailiang Han.)*

Jiaojiao Li, Yuzhe Liu, Jiachao Liu, and Rui Song are with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi’an 710071, China (e-mail: jjli@xidian.edu.cn; 21011210556@stu.xidian.edu.cn; jcliu1217@163.com; rsong@xidian.edu.cn).

Wei Liu is with the State Key Laboratory of Geo-Information Engineering, Xi’an 710054, China (e-mail: 1271772072@qq.com).

Kailiang Han is with the Key Laboratory of Infrared System Detection and Imaging Technology, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China (e-mail: hankailiang@mail.sitp.ac.cn).

Qian Du is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762 USA (e-mail: du@ece.msstate.edu).

Digital Object Identifier 10.1109/JSTARS.2022.3221860

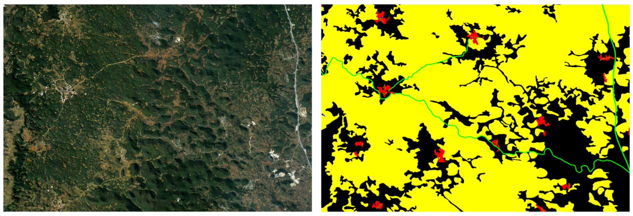


Fig. 1. Semantic segmentation in remote sensing imagery is challenging. On the left-hand side is a remote sensing image, and on the right-hand side is the corresponding semantic label, where yellow represents the vegetation category, green represents roads, red represents settlements, and black represents background. Here, we show a typical example where roads and settlements are few and far between in the scene and are surrounded by dominant vegetation.

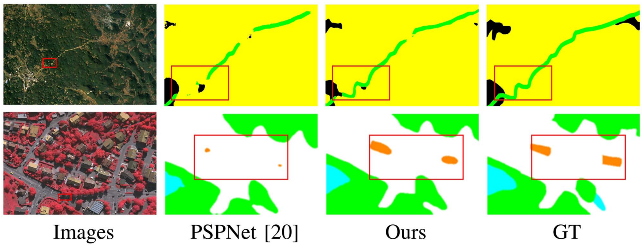


Fig. 2. General semantic segmentation methods hardly can survive in this challenging scenario. Legend of the first line (green: Roads; black: background; and yellow: vegetation); legend of the second line (white: impervious surfaces; cyan: low vegetation; green: trees; and orange: cars). The tiny targets are marked in red box. Predictions of PSPNet lose most of the tiny targets.

occupy only a dozen pixels in the image. Even worse, these tiny objects are surrounded by dominant ground objects, which take up the vast majority of all pixels. Fig. 1 demonstrates an example of these challenges in the RSI scenario. Compared with the vegetation class denoted as a yellow color, the roads are marked with a green color, and settlements denoted as a red color account for an extremely small proportion. In addition, roads and settlements are easily confused with vegetation because they are surrounded by vegetation. The most straightforward strategies for handling class imbalance in semantic segmentation include selective data augmentation and decision level fusion through multiple networks results [11], [12], [13], [14], [15], [16]. These strategies, however, are either overfitting-prone or too heavy for practical application. On the other hand, many current methods combine multiscale representations by relying on the skip connections of typical encoder–decoder network architectures [9], [10], [17], [18], [19]. However, this type of feature extraction does not explicitly enhance the representation of small or especially tiny objects, making them still prone to being easily degraded by large objects with a dominant number of pixels, as shown in Fig. 2. The tiny targets are marked in a red box, and it is evident that the predictions of PSPNet almost lose the tiny targets [20].

To further conquer the abovementioned issues, the present studies propose enhancing the characterization of small or tiny objects explicitly during network training in this work. Specifically, efficient convolution layers are introduced to extract edge and body structure information in the decoder part of a network.

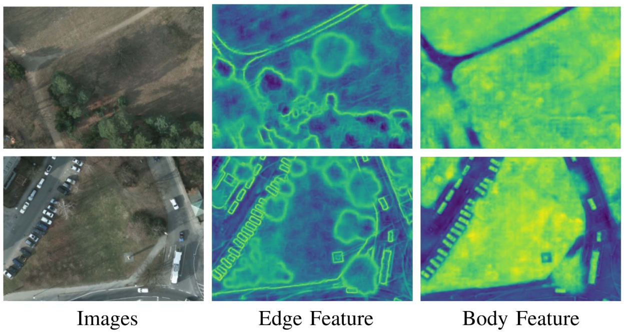


Fig. 3. Visualized body and edge features. Edge and body can enhance nonobvious characteristics in RSIs, which is benefit to the segmentation especially for the tiny structure objects, such as roads and cars.

Fig. 3 depicts the visualized body and boundary features through the designed feature extraction module of the present study, illustrating that the less sharp objects, such as roads, cars, and so on, in the original image are more distinguishable in body and edge features. A practical edge-guided feature transform module (EGFTM) is designed to utilize features to strengthen the edge contour and maintain the internal consistency of a homogeneous area. It operates in a coarse-to-fine manner to progressively enhance the spatial structure of small objects. Furthermore, a context aggregation pyramid (CAP) pooling strategy is devised in the present study to adaptively achieve optimal fusion characterization of multiscale features according to different contexts. The strategies proposed in the present study are general and can be applied in any off-the-shell segmentation network architectures. The proposed network is tested on three large-scale remote sensing datasets, and it is demonstrated that it can consistently improve the accuracy regardless of the network backbones. Moreover, the proposed method outperforms most state-of-the-art (SOTA) methods, especially in segmenting small objects.

The main contributions of the current work can be summarized as follows.

1) A novel RSI semantic segmentation framework feature guide network with context aggregation pyramid (FGN-CAP) based on ResNet backbone is proposed, which can effectively alleviate the degradation of small objects and outperforms other baselines with SOTA results on three RSIs benchmark.
2) An EGFTM is designed, which utilizes the edge and body features extracted in the encoder part to enhance the heterogeneous boundaries and constrain the internal context consistency of homogeneous region, resulting in better parsing of small or tiny targets that are surrounded by dominant objects.
3) CAP is further proposed for improving the segmentation performance, which concentrates on dynamically learning individual superior multiscale context characterizations for each object of varying size. Particularly, a corresponding affine matrix is exploited to modulate the significance of different scale features.

The rest of this article is organized as follows. Section II concisely introduces the related works. Our proposed method is described in detail in Section III. Section IV presents sufficient experiments with analysis. Section V concludes this article.

II. Related Work

A. General Semantic Segmentation

Traditional semantic segmentation methods heavily rely on conditional random fields, which is still a common refinement component for some current methods [21], [22], [23], [24]. These methods, however, are sensitive to scale variations and are less efficient. To overcome these, many methods try to capture contextual information by different strategies [9], [10], [17], [20]. Furthermore, most current methods adopt a nonlocal operator and self-attention mechanism to obtain more detailed context [18], [25], [26], [27], [28]. There are also some attempts by introducing a graph convolutional network to propagate the context information between different regions [19], [29]. These methods, however, do not explicitly enhance the context of tiny objects, making them prone to being easily affected by other larger objects during context propagation.

Recently, aggregating multiscale features based on pyramid network architectures has become a dominant strategy for improving semantic segmentation accuracy [30]. SegNet relies on the common encoder–decoder architecture to combine the low-level and high-level features [31]. PSPNet and Deeplabv3 propose custom modules to capture multiscale contexts [17]. These methods, however, do not explicitly enhance the representation of tiny objects and still cannot prevent their degeneration, especially in the scenario with other dominant categories. In contrast, the authors of the present study propose to explicitly enhance tiny objects' representation with spatial structures, such as edges, to facilitate the aggregation of multiscale features.

Leveraging other tasks by joint learning is also a recent direction for semantic segmentation, such as GSCNN and DSRL [32], [33], which embed into the main network as a subnetwork for shape estimation and super-resolution estimation, respectively. The proposed network conducts standard semantic segmentation tasks while extracting edge body features and using joint loss for supervised learning.

B. Semantic Segmentation in RSIs

Semantic segmentation on RSIs involves classifying houses, roads, vegetation, water, farmland, and more land-cover classes with pixel-level precision. Early research was focused on unsupervised learning based on graphic theory [34], [35], [36]. Recently, semantic segmentation models based on modern deep learning, such as FCNs [37], have dramatically improved the accuracy in most major RSI benchmarks. Volpi and Tuia[38] proposed a CNN architecture based on encoder–decoder to enhance results by deconvolution. Sun et al. [39] proposed ensemble strategies and a residual architecture to target the structural stereotype and insufficient learning in the encoder–decoder framework. However, although the overall accuracy (OA) is increased through the capacity of deep feature representation, the accuracy of some tiny objects is still not satisfactorily improved. In practice, there are many tiny-structured objects in an RSI, including some point objects, such as towers and chimneys, and some linear targets, such as roads and small streams. To further improve the accuracy of the tiny objects in RSIs, the main methodologies following general semantic segmentation are proposed for several specific application scenarios [11], [40], [41], [42], which fuse different context information and extract foreground information and long–short dependence of spatial data for assisting with segmentation accuracy enhancement. Zhang et al. [43] proposed an end-to-end attention-based semantic segmentation network, and a pyramid attention pooling module was designed to introduce the attention mechanism into the multiscale module for adaptive feature refinement. Bai et al. [44] combined multiscale with the attention mechanism, and proposed a multiscale attention module to enhance the fine-grained representation ability of the network and the extraction ability of global context information. However, these existing semantic methods in RSIs handle all the context information, such as color, texture, and edges, in the same way, which ignores the disparate contributions of different information. Furthermore, due to overlapping between categories, the intraclass variance is large, whereas the interclass variance is small. Moreover, the discrimination between the categories is not apparent. The irregularity and complexity of the boundary shapes make it difficult to achieve semantic segmentation near the boundary. In particular, the edge suffers from difficult distinction for the tiny targets often surrounded by different land covers having a large scale. Therefore, to overcome this, the present study proposes an efficient module enhancing the network's sensitivity to the edge structures, which is critical for segmenting tiny objects. The proposed approach considers the body consistency and the edge preservation of the object in images as vital auxiliary information in the segmentation task. With the assistance of body and edge features, the resolution of the feature image is reconstructed from coarse to fine layer by layer in the decoder module. A simple module to extract the body and edge features through supervised learning is adopted to reconstruct the output resolution from coarse to fine with the assistance of body and edge features.

C. Edge Detection

The ideal result of semantic segmentation is the accurate edge division of different objects in the image, so edge detection and semantic segmentation have a certain degree of consistency. Based on the similarity and synergy between edge detection and semantic segmentation, some studies use edge detection tasks to enhance the segmentation results of image edges and improve the overall segmentation accuracy. Bertasius et al. [45] used a "high-for-low" method to predict the boundary using the object level features in the pretrained object classification network. In this method, high-level object features inform the low-level boundary detection process. Kokkinos [46] applied deep convolution neural network training to the boundary detection task. The carefully designed boundary detection loss training and multiresolution architecture improved the boundary detection technology. Bertasius et al. [47] used intermediate features for semantic segmentation, predicted edges through domain transformation, and optimized the quality of target semantic

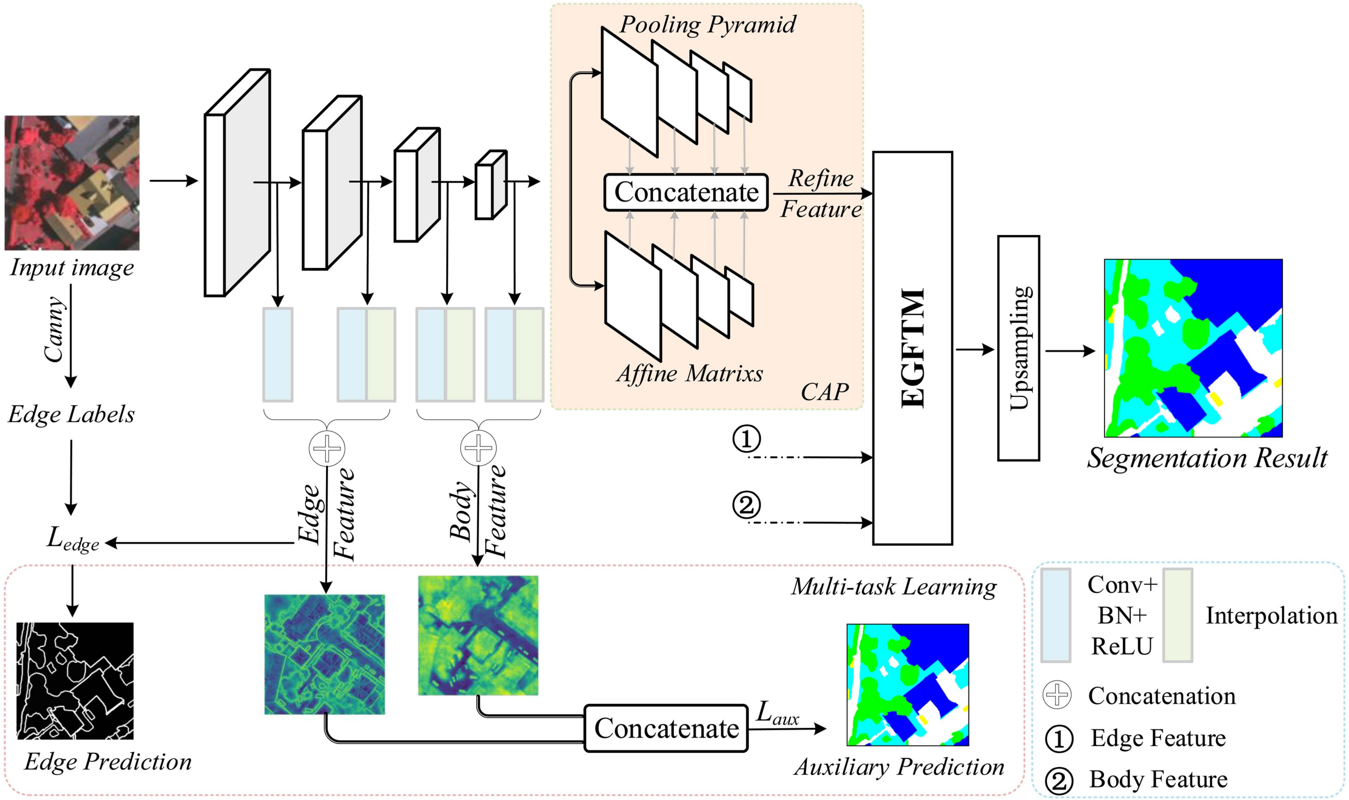


Fig. 4. Our proposed network architecture. The network is based on encoder–decoder structure. In the encoder stage, edge and body information in the feature maps are extracted, respectively, constrained by corresponding loss function. The backbone is followed by a CAP to aggregate multiscale context features adaptively. In the decoder stage, an EGFTM is used to help with upsampling to restore resolution.

segmentation. Marmanis et al. [48] proposed an end-to-end trainable deep convolution neural network structure for semantic segmentation and built in the perception of semantic meaningful boundaries. First, a relatively simple and memory-efficient model is constructed by adding boundary detection to the SegNet encoder–decoder architecture. Second, boundary detection is added to the FCN type model, and a high-end classifier ensemble is established. Combining semantic segmentation and semantic information edge detection to combat this influence clarifies the class boundary in the model. Yang et al. [49] proposed a new end-to-end edge-aware network EANet to obtain accurate buildings from aerial images. Specifically, the architecture comprises image segmentation and edge perception networks, which are responsible for constructing prediction and edge surveys, respectively. The network pays more attention to low-level details, such as edges, rather than emphasizing the multiscale fusion of features, or enhancing more receptive fields to obtain global features. Yu et al. [50] proposed a distinct feature network that consists of two subnetworks: smooth network and boundary network. Specifically, to deal with the problem of intraclass inconsistency, the smooth network selects more discriminative features through a channel attention block and global average pool. The boundary network can distinguish the bilateral characteristics of the boundary through deep semantic boundary supervision. Zhao et al. [7] designed an auxiliary edge detection task to provide edge constraints for semantic segmentation. The present study introduces edge detection tasks into semantic segmentation to take advantage of their synergy. Furthermore, an end-to-end semantic segmentation network structure is proposed. Considering that different features of objects in the image content have other benefits for semantic segmentation tasks, the proposed method explicitly extracts edge and body features in the network and optimizes semantic segmentation results by using edge constraints and internal consistency of the body.

III. Methodology

In this section, the proposed semantic segmentation CNN architecture is proposed. The proposed EGFTM and a CAP module are also described. Then, the loss function exploited in the network training is introduced. Finally, the architecture of the proposed framework is depicted, yielding robust and superior segmentation results.

A. Network Architecture

As depicted in Fig. 4, the proposed network is based on an encoder–decoder structure. The backbone of the encoder module is ResNet [51], which extracts and refines features gradually through multiple groups of residual blocks. A CAP module follows the backbone, which focuses on capturing the objects with multiscales in the image and constructing adaptive contextual representations. Considering the disparity between edge and

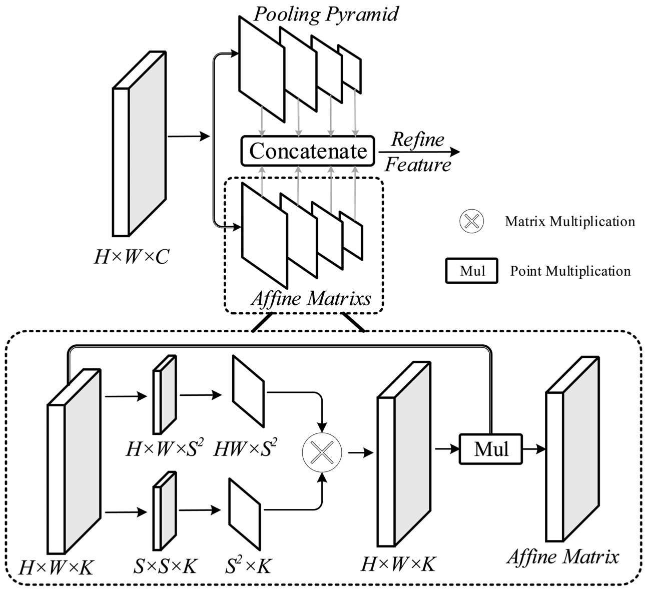


Fig. 5. Context aggregation pyramid model. The feature maps extracted from the backbone network is input into CAP, and a group of feature maps with different scales are obtained through the pooling pyramid. Each feature map corresponds to an affine matrix, which acts with the feature map and concatenate together to obtain the refined features after feature aggregation.

body features, the edge and body features are explicitly extracted at different scales by a light EGFTM. The edge and body features are used in the decoder to enhance the edge and restore the resolution. Specifically, the edge features are extracted from the shallow network of the first two layers while the body features are extracted from the deep network [52].

B. Context Aggregation Pyramid Module

It has been widely demonstrated that context information is critical for scene parsing and semantic segmentation in many works [10], [20]. For complex remote sensing scenes, many objects with uneven scales are included. For example, a telegraph pole occupies only one or two pixels in the image whereas a lake can even take up the whole RSI. Objects of different scales need to aggregate contextual information of either a long or short range. The self-attention mechanism is introduced into the pyramid structure, and a CAP is presented to build a learning model for multiscales objects that ignores irrelevant information and focuses on significant information [18], [26]. Here, the most appropriate context representation is constructed in line with dynamical significance for each scale in the scale pyramid.

Such as PPM in PSPNet, the proposed CAP also constructs a multiscale pooling pyramid to capture contexts. As exhibited in Fig. 5, a pooling features pyramid is first created by different pooling scales s. There is an affine matrix for each scale, where the optimal feature representation at this scale can be calculated by the pooling feature and the corresponding affine matrix. Here, one scale s is taken as an example, and the other scales can be processed similarly. Given an input image I for segmentation, the feature map $F \in \mathbb{R}^{h \times w \times c}$can be calculated by the backbone CNN, where $h \times w$ denotes spatial resolution and c represents channel dimension. For the scale s, F is first processed with a 1×1 convolution to obtain the reduced feature map $x^s \in \mathbb{R}^{h \times w \times k}$. Then, x^s is processed in two parallel branches. In the first branch, the dimension of x, s is first reduced into $h \times w \times s^2$ by 1×1 convolution to acquire the affine matrix, it is then reshaped into $hw \times s^2$. In the second branch, x^s is transformed into $s \times s \times k$ by adaptive pooling, and it is reshaped into $s^2 \times k$. Then, the adaptive weight w^s for every pixel in x^s can be obtained by the matrix product of the outputs of two branches, which can be expressed as the following formula:

$$x^s = \phi(F) \tag{1}$$

$$w^s = x^s \cdot (\varphi(x^s) \otimes \rho(x^s)) \tag{2}$$

where ϕ and φ represent 1×1 convolution and ρ represents adaptive pooling. Finally, the representation z^s can be calculated as follows:

$$z^s = x^s \cdot w^s. \tag{3}$$

The abovementioned steps are also performed at other scales in the scale pyramid. All the z^{s_i} at all the scales and the feature map F are concatenated finally so that the refined representation Z that is adaptive to all objects can be obtained. The formula is as follows:

$$Z = \text{concat}(\cup_{i=0}^{n} z^{s_i}, F). \tag{4}$$

C. Edge-Guided Feature Transform Module

In RSIs, overlapping between categories occurs due to imaging technology, leading to a false alarms problem. Furthermore, for the tiny targets surrounded by different land covers on a large scale, the edge is difficult to distinguish from the irregularity and complexity of the boundary shapes. Thus, it is valuable to propose an efficient decoder structure to support the network, which precisely maintains the edge features and consistently maintains the body feature.

In CNNs, to increase the receptive field of output, the spatial resolution of output will be lost. Therefore, upsampling is needed to restore the feature map to the original spatial resolution. Considering that the pixels inside an object in the scene are similar due to homogeneity in most cases while those distributed near the boundary with possible heterogeneity show differences, an EGFTM is designed in the present study to learn the edge and body feature representations with supervision. These edge and body features are leveraged to restore the output resolution according to their different constraint effects.

Specifically, as demonstrated in Fig. 6, the edge feature F_e is explicitly extracted from lower encoder layers and the body feature F_b from higher encoder layers by a simple convolution operation with supervision [53]. For the body features, F_b, which mainly contain high-level semantic information, is adopted to calculate channel weights and aggregate inner semantic consistency context information with a global average pooling (GAP) operation. The edge features F_e, which have more significant values at the boundary but smaller values elsewhere, are leveraged to reinforce the edges. The modified feature F_m can be

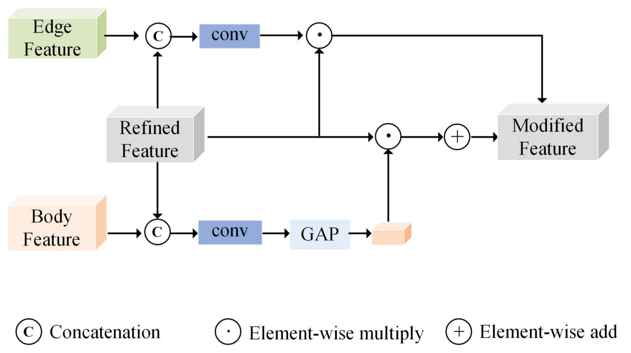


Fig. 6. EGFTM enhances the heterogeneous boundaries by edge features and constrain the internal context consistency of homogeneous region by body features for better parsing of tiny targets that are surrounded by dominant objects.

calculated as follows:

$$F_1 = \mathrm{GAP}(\zeta_1(C(F, F_b))) \cdot F \tag{5}$$

$$F_2 = \zeta_2(C(F, F_e)) \cdot F \tag{6}$$

$$F_m = F_1 + F_2 \tag{7}$$

where ζ_1 and ζ_2 mean two groups of Convolution → Batchnorm → ReLU. F denotes the refine feature output by the CAP module. The semantic predictions can be obtained by up-sampling F_m.

D. Joint Task Learning

During training, we jointly supervise body, edge, and the final semantic segmentation prediction. We use standard binary cross-entropy (BCE) loss for edge prediction and use standard cross-entropy (CE) loss for body and semantic segmentation prediction. The final loss function is expressed as follows:

$$\begin{aligned}\mathcal{L} &= \lambda_1 \mathcal{L}_{\text{seg}} + \lambda_2 \mathcal{L}_{\text{aux}} + \lambda_3 \mathcal{L}_{\text{edge}} \\ &= \lambda_1 \mathcal{L}_{\text{CE}}(y, \hat{y}) + \lambda_2 \mathcal{L}_{\text{CE}}(y, \hat{y}_{\text{aux}}) + \lambda_3 \mathcal{L}_{\text{BCE}}(y_e, \hat{y}_e)\end{aligned} \tag{8}$$

where y represents the ground truth label of final semantic segmentation. y_e represents the real edge label obtained by performing edge detection on y. $\hat{y}$ represents the final semantic prediction result that is the output of the end of the network. $\hat{y}_e$ represents the edge prediction result, which is obtained from the edge feature F_e. $\hat{y}_{\text{aux}}$ represents the auxiliary semantic prediction result, which is obtained by the fusion output of edge features F_e and body features F_b. Here, λ_1, λ_2, and λ_3 are three hyperparameters that control the weighting between the losses. Specifically, λ_1 is the loss of segmentation, λ_2 is the loss of body, and λ_3 is the edge loss. Because of the serious imbalance between edge pixels and body pixels, a coefficient γ is used in the boundary prediction loss $\mathcal{L}_{\text{BCE}}(y_e, \hat{y}_e)$.

IV. Experiments

This section conducts extensive experiments on three challenging RSIs' semantic segmentation datasets, including the Tianzhi, Vaihingen, and Potsdam datasets. An ablation study is first conducted to verify the effectiveness of each module of the proposed framework. Then, the proposed framework is compared with several SOTA baselines, and the comparison results also demonstrate the advancement of the proposed method. In order to highlight the effectiveness of the algorithm, the best result of the comparison algorithm will be shown in bold, and at the same time, the underlined data indicate the second-best result.

A. Datasets

The proposed approach is evaluated with experiments on the ISPRS 2-D semantic labeling benchmark and the dataset released from the Tianzhi cup artificial intelligence challenge.

1) *Vaihingen Dataset:* This dataset contains 33 tiles of different sizes. The average size of the tiles is around 2500 × 2000 pixels with a ground sampling distance (GSD) of 9 cm. Then, 16 tiles were split into a training and validation set, and the rest were for testing.
2) *Potsdam Dataset:* This dataset contains 38 tiles with the same size of 6000 × 6000 pixels and a GSD of 5 cm. Then, 24 tiles are split into a training and validation set and the rest are for testing. Both datasets include six categories: impervious surface, building, low vegetation, tree, car, and clutter (background). Only RGB images are used in the experiments conducted in the present study. The dataset is available at http://www.commission3.isprs.org/wg4.
3) *Tianzhi Dataset:* The Tianzhi dataset consists of 12 pairs of RSIs and corresponding ground truth semantic labels. The size of each image is 7400 × 4950 pixels, and each image contains three channels of red (R), green (G), and blue (B). Five common land cover categories are included: Settlement, Roads, Water, Vegetation, and Backgrounds. To conduct experiments more reasonably, 300 smaller images were obtained by dividing each image into 25 subimages with a size of 1440 × 990 pixels. As for the 25 subimages of each image, 16 are used for training, 7 are used for testing, and the remaining images are set as validation data.
4) *Cityscapes Dataset:* This dataset consist of 5000 images of driving scenes in urban environments across 50 European cities, which has nineteen categories. And the number of training set and validation set is 2975 and 500, respectively, and the rest are testing sets.

B. Evaluation Criteria

The OA and mean intersection over union (mIoU) criterion are employed to assess the overall performance of semantic segmentation results. Besides, the F1 score is exploited to evaluate the performance of each category. It is assumed that there is a total of $k + 1$ categories (from 0 to k, and 0 represents the Backgrounds), and p_{ij} stands for the number of pixels belonging to category i and being predicted as category j.

OA is a straight-forward metric computing a ratio of the amount of correctly classified pixels and the total number of

TABLE I
EFFECTIVENESS OF THE PROPOSED MODULES

DataSet	Method	OA	mIoU
vaihingen	ResNet-101	90.20	78.62
	ResNet-101+CAP	90.47	79.19
	ResNet-101+EGFTM	90.60	80.48
	ResNet-101+CAP+EGFTM	**90.65**	**80.75**
Potsdam	ResNet-101	91.08	86.35
	ResNet-101+CAP	91.17	86.64
	ResNet-101+EGFTM	91.20	**86.83**
	ResNet-101+CAP+EGFTM	**91.24**	86.80
Cityscapes	ResNet-101	94.84	71.23
	ResNet-101+CAP	95.28	73.10
	ResNet-101+EGFTM	95.30	71.62
	ResNet-101+CAP+EGFTM	**95.79**	**76.36**

pixels. It can be calculated as follows:

$$\mathrm{OA}=\frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}}. \tag{9}$$

The intersection over union (IoU) represents a ratio of the intersection of pixels predicted to be of a certain category and the ground truth pixels of that category and their union. The mIoU can be calculated by averaging the IoU for all the categories besides Backgrounds. The mIoU metric can be calculated as follows:

$$mIoU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{i=0}^{k} p_{ij}+\sum_{j=0}^{k} p_{ji}-p_{ii}}. \tag{10}$$

The F1 score harmonic average of precision (OA) and recall. The recall and F1-score can be obtained as follows:

$$\mathrm{Recall}=\frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} p_{ii}+\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ji}} \tag{11}$$

$$F1=2\times\frac{OA\times\mathrm{Recall}}{OA+\mathrm{Recall}}. \tag{12}$$

C. Implementation Details

All the experiments are carried out with the PyTorch framework. The proposed network is trained using Adam optimization with a batch size of 8. Then, the betas of (0.9, 0.999) and a weight decay of 5e-4 are set. The learning rate is initialized to 2e-5, and the "poly" policy is adopted to decay the learning rate by multiplying $(1-\frac{\text{iter}}{\text{total\_iter}})^{0.9}$ after every training iteration. The experiments are conducted on a single NVIDIA GTX 2080Ti GPU.

D. Ablation Study

The following experimental results of the ablation study are conducted on both the Vaihingen and Potsdam datasets to evaluate the effectiveness of each component of the proposed method.

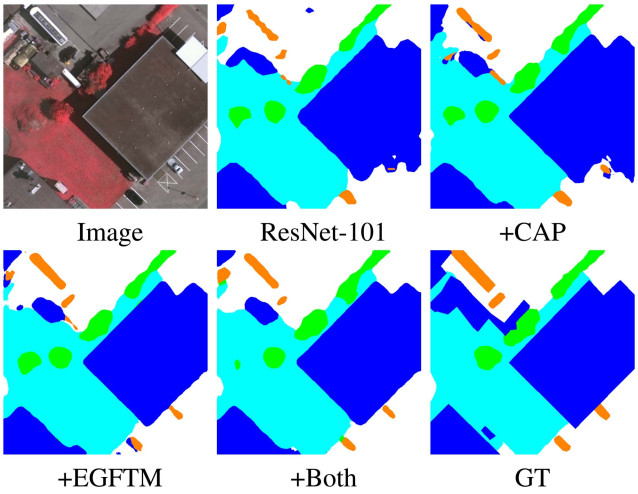


Fig. 7. Visualization comparison of ablations on segmentation results. Legend (white: impervious surfaces noted as Imp.S., blue: buildings noted as Build., cyan:low vegetation noted as Low.V., green: trees, and orange: cars).

TABLE II
ABLATION STUDIES OF CLASSES ON THE VAIHINGEN DATASET

Method	Imp.S.	Build.	Low.V.	Tree	Car
ResNet-101	91.90	95.44	83.65	89.73	78.42
+CAP	92.26	95.56	**84.58**	89.64	78.29
+EGFTM	92.39	**96.03**	84.08	89.76	82.50
+Both	**92.51**	95.90	84.22	**89.82**	**83.24**

TABLE III
ABLATION STUDIES OF CLASSES ON THE POTSDAM DATASET

Method	Imp.S.	Build.	Low.V.	Tree	Car
ResNet-101	93.33	97.04	87.37	89.46	95.58
+CAP	93.14	96.97	87.84	89.49	96.17
+EGFTM	**93.34**	**97.21**	87.67	89.54	**96.36**
+Both	93.06	97.14	**88.01**	**89.58**	96.29

1) Effectiveness of the Proposed Modules: Effectiveness of the proposed modules: The experiments are conducted to evaluate the effectiveness of the proposed modules. The experimental results are tabulated in Table I. For the Vaihingen dataset, first, the ResNet-101 and simple upsampling are selected as the encoder and decoder of the baseline, which achieves 78.82% mIoU and 90.20% OA. Then, the adaptive optimal multiscale representations are simply concatenated after the ResNet-101 part, denoted as the CAP layer, which obtains 79.19% mIoU and 90.47% OA with a slight enhancement. Furthermore, to verify the effectiveness of the proposed EGFTM module, it is inserted before the upsampling layer of the baseline, increasing mIoU to 90.60% and 80.48%, respectively. Finally, by employing CAP and EGFTM, the mIoU is enhanced to 90.65%, with a 0.45% increment better than the baseline. Moreover, the OA result is also improved to 80.75%, which is 1.93% better than the baseline. For the Pstsdam dataset, the baseline combined with CAP and EGFTM attains superior semantic segmentation performance, improving 0.16% OA and 0.45% mIoU. Similar ablation results are depicted in Cityscapes dataset and the results are more pronounced. Specifically, by employing CAP and EGFTM, the mIoU is enhanced to 76.36%, with a 5.13% increment better than the baseline. All these improvements demonstrate that the proposed modules greatly benefit scene parsing. The visualization comparison of the proposed modules is shown in Fig. 7.

For a more detailed analysis, the F1 scores of each class under different settings are tabulated in Tables II and III. From Table II, it can be claimed that the class "Car" with a tiny

TABLE IV
COMPARISON WITH DIFFERENT BACKBONE ON THE VAIHINGEN DATASET

Method	OA	mIoU	Build.	Car
ResNet-50	89.32	76.94	94.66	75.22
+CAP+EGFTM	90.20	76.39	95.55	68.07
ResNet-101	90.20	78.82	95.44	78.42
+CAP+EGFTM	90.65	80.75	95.90	83.24
ResNet-152	90.41	79.46	95.79	79.64
+CAP+EGFTM	**90.71**	**81.28**	**96.15**	**84.61**

TABLE V
COMPARISON WITH DIFFERENT BACKBONE ON THE POTSDAM DATASET

Method	OA	mIoU	Build.	Car
ResNet-50	90.43	84.67	96.47	93.59
+CAP+EGFTM	90.99	86.09	96.81	96.46
ResNet-101	91.08	86.35	97.04	95.58
+CAP+EGFTM	91.24	86.80	97.14	96.29
ResNet-152	91.30	85.55	97.20	95.86
+CAP+EGFTM	**91.34**	**87.14**	**97.39**	**96.59**

structure has been increased by 4.82%, whereas the large classes, named "Imp.S.," "Build.," "Low.V.," and "Tree," improved by 0.61%, 0.46%, 0.57%, and 0.09%, respectively. The F1 scores in the Potsdam datasets observed in Table III demonstrate that although the "Imp.S." class slightly decreased, the other classes achieved better performance than the baseline. In particular, the tiny targets named "Car" can obtain an improvement of 0.71%. Also, it can be observed that CAP can obtain more precise multiscale feature representations for each class with slight accuracy enhancement. Moreover, EGFTM generates more consistent segmentation inside large and tiny objects or along the boundaries, which is why EGFTM obtains a better increment than CAP.

2) Robustness of Different Backbones: To show the generalization capability of the proposed modules, a series of experiments is further conducted to compare exploiting different backbone networks. It is worth noting that the presented modules can be easily inserted into various backbone networks. The ResNet-50, ResNet-101, and ResNet-152 are selected as representatives. Due to limited context space, the OA, mIoU, and F1 scores of the classes "Build." and "Car" are chosen to verify the robustness of the proposed framework. The detailed criterion values of other large objects, such as "Imp.S.," "Low.V.," and "Tree," are also recorded, which have similar comments as the "Build." class. It should be noted that all the deep networks are equally pretrained on ImageNet. The comparison results are reported in Tables IV and V. Both modules combined with ResNet-152 achieve the best performance for irrelevant datasets, indicating that ResNet-152 has the most superior feature representational capability. Furthermore, the backbones exploited with the proposed strategies can obtain considerably better mIoU, OA, and F1 scores on either the Vaihingen or Potsdam datasets than without utilizing the proposed modules, especially for the tiny objects, the increments are larger than in other classes. All these experimental results imply that the EGFTM and CAP modules can enhance the boundary feature of tiny objects, which can effectively extract a more detailed multiscale context.

3) With Different Pyramid Levels: The proposed CAP is constructed in a pyramid manner to extract spatial features from multiple scales. Meanwhile, the best representation can be adaptively obtained by the attentional weight. The models are assessed with different scale rates adopted in CAP to verify the effectiveness of adaptive optimal multiscale features. As given in Table VI, harvesting abundant multiscales of adaptive attentional features outperforms single-scale features, indicating the effectiveness and necessity of employing an adaptive attentional pyramid structure. Moreover, using a scale of 2 has the best OA and mIoU than the performance obtained from other single-scale, and using scales 1, 2, 3, and 6 in combination with the best OA and mIOU. Furthermore, the experimental results indicate that adaptively learning appropriate representations of objects aggregating multiscale features via different contexts is critical for improving the segmentation accuracy of tiny targets.

TABLE VI
COMPARISON WITH DIFFERENT SCALES S

1	2	3	4	5	6	OA	mIoU
✓						89.78	77.74
	✓					90.03	77.98
		✓				89.82	77.88
			✓			89.52	77.29
				✓		89.57	78.36
					✓	89.77	78.27
✓	✓	✓			✓	**90.47**	**79.19**

E. Comparison to State of the Arts.

The proposed method is further compared against several SOTA semantic segmentation baselines on the ISPRS 2-D semantic labeling benchmarks, Tianzhi dataset and cityscapes dataset. Furthermore, the complexity comparison is reported among the SOTA algorithms and the proposed one with the ResNet-101 backbone. The proposed method is moderatamente in terms of calculation complexity. Compared with DeepLabV3+ [10], although the Flops of the proposed method is slightly greater, both the number of parameters and occupied memory are smaller, and the inference time is almost the same.

Tables VII and VIII tabulate the results on typical ISPRS benchmarks, named Vaihingen and Potsdam, respectively, demonstrating that the proposed method achieves the best performance for almost every land-cover class. Specifically, the proposed method obtains significant improvement and outperforms other SOTA methods for tiny objects, such as cars and some small buildings. The values of OA and mIoU achieved by the two tested modules are still better than those of not using, indicating that the EGFTM and CAP significantly enhance the segmentation performance in RSIs, especially for tiny objects. The corresponding comparison segment maps are also provided for visual perception. From Fig. 8, the edges of cars are more finely attained from the proposed method, demonstrating that the proposed modules can help improve the segmentation accuracy of tiny objects. Furthermore, all these results imply that the boundaries of objects are a vital part of precision in RSI semantic segmentation. Besides, the experimental results based on the Tianzhi dataset are also reported in Table IX and Fig. 9. For this typical RSI, existing with the winding and narrow roads surrounded by large-scale vegetation, it can be concluded that

TABLE VII
COMPARISON VERSUS SOTA BASELINES ON THE VAIHINGEN DATASET

Method	Backbone	OA	mIoU	Imp.S.	Build.	Low.V.	Tree	Car
SegNet [31]	N/A	88.53	74.81	90.60	93.33	81.10	88.64	71.77
PSPNet [20]	ResNet-101	90.24	79.24	91.94	95.36	84.12	89.61	79.68
DeepLabv3+ [10]	ResNet-101	90.31	79.29	92.00	95.79	83.72	89.67	79.57
OCNet [54]	ResNet-101	89.81	77.70	91.33	95.32	83.35	89.33	77.70
CCnet [26]	ResNet-101	89.96	79.23	91.88	95.67	82.69	88.93	81.60
Ours	ResNet-101	**90.65**	**80.75**	**92.51**	**95.90**	**84.22**	**89.82**	**83.24**

TABLE VIII
COMPARISON VERSUS SOTA BASELINES ON THE POTSDAM DATASET

Method	Backbone	OA	mIoU	Imp.S.	Build.	Low.V.	Tree	Car
SegNet [31]	N/A	87.22	80.45	90.57	92.18	84.20	85.28	93.03
PSPNet [20]	ResNet-101	91.15	86.49	93.17	96.81	87.83	89.34	96.03
DeepLabv3+ [10]	ResNet-101	91.15	86.70	**93.23**	97.13	87.73	89.38	96.28
OCNet [54]	ResNet-101	91.03	86.27	92.27	96.70	87.77	89.44	95.71
CCnet [26]	ResNet-101	91.03	86.28	93.14	96.93	87.48	89.32	95.71
Ours	ResNet-101	**91.24**	**86.80**	93.06	**97.14**	**88.01**	**89.58**	**96.29**

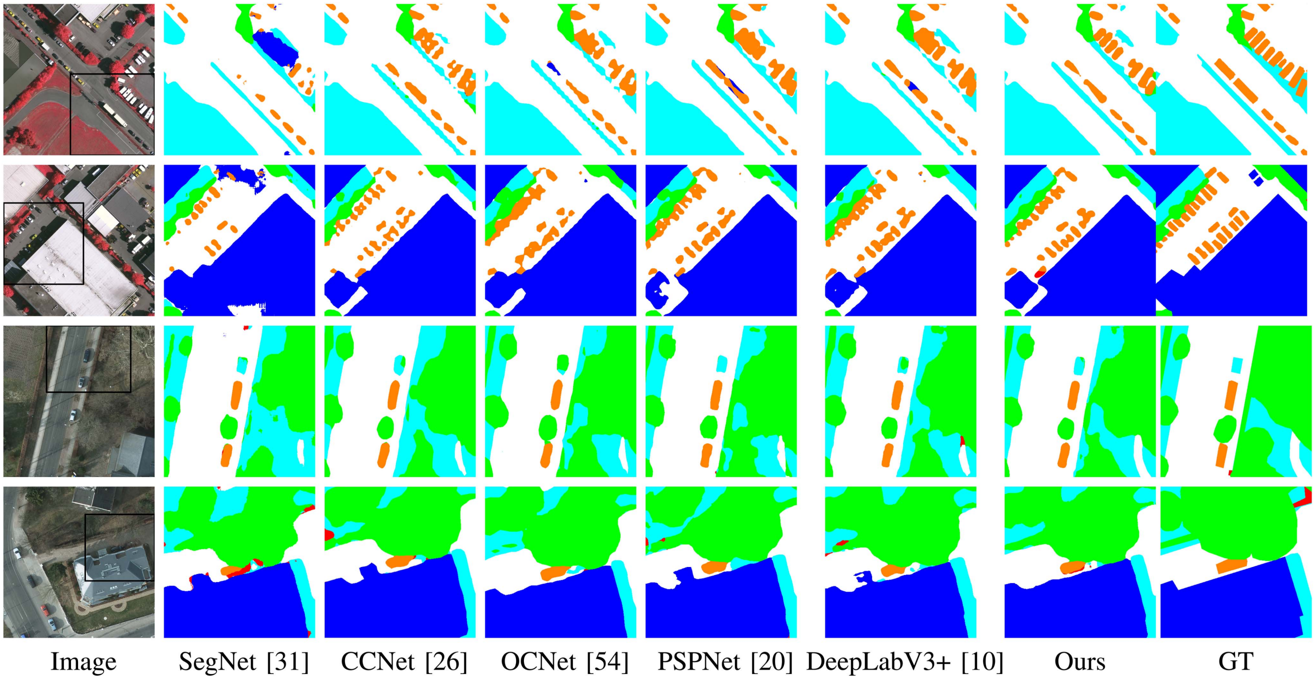


Fig. 8. Visual quality comparison of the proposed model over other different methods.Visualization comparison of ablations on segmentation results. Legend (white: impervious surfaces noted as Imp.S., blue: buildings noted as Build., cyan:low vegetation noted as Low.V., green: trees, and orange: cars, red: clutter).

the proposed modules make an efficient impact on accuracy enhancement. The segmentation performance of all objects in the Tianzhi dataset obtains a positive lift. From the visual comparison in Fig. 9, the roads detected by the proposed method are the most accurate, with relatively complete and acceptable boundaries compared with other methods, indicating that the enhanced edges and adaptive different multiscale contexts work well in the RSI semantic segmentation task with complicated backgrounds. Furthermore, the proposed EGFTM and CAP module demonstrate extraordinary talent in tiny target segmentation. Table X and Fig. 10 show the segmentation results on the Cityscapes dataset. As can be seen from the table, our method also achieves the optimal segmentation result. Compared with the comparison algorithm, the proposed method achieves 3.1%, 0.58%, 1.64%, 3.04%, and 2.96% improvement in mIoU index, respectively.

F. Parameter Tuning

When designing the overall segmentation loss of the network, we simultaneously used three parameters, λ_1, λ_2, and λ_3, to weight the segmentation loss, edge loss and body loss. In order to determine the final weight value, we set the overall segmentation loss as 1.0 unchanged, while changing the body and edge loss to conduct ablation experiments. Fig. 11 shows the influence of different λ_2/λ_3 values on the final segmentation results. As can

TABLE IX
COMPARISON VERSUS SOTA BASELINES ON THE TIANZHI DATASET

Method	Backbone	OA	mIoU	Set.	Roads	Water	V.
SegNet [31]	N/A	73.07	43.52	61.38	0.2	86.15	76.09
PSPNet [20]	ResNet-101	68.35	51.37	73.45	21.21	93.29	55.97
DeepLabv3+ [10]	ResNet-101	73.78	47.83	59.67	17.61	92.88	75.90
OCNet [54]	ResNet-101	73.11	42.13	51.25	13.77	91.20	**79.39**
CCNet [26]	ResNet-101	69.06	51.93	73.50	23.31	**93.78**	58.59
Ours	ResNet-101	**77.99**	**55.68**	**74.82**	**32.39**	92.81	79.33

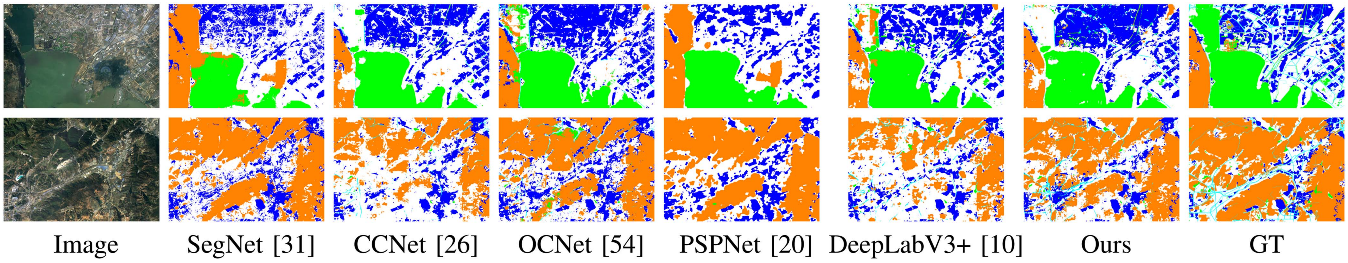


Fig. 9. Visual quality comparison of the proposed model over other different methods on Tianzhi dataset. Legend (blue: Settlement, cyan: Roads, green: Water, orange: Vegetation, and white: Background).

TABLE X
COMPARISON VERSUSS SOTA BASELINES ON THE CITYSCAPES DATASET

No.	SegNet	PSPNet	DeepLabv3+	OCNet	CCNet	Ours
road	97.77	98.00	97.83	97.59	97.70	**98.16**
sidewalk	83.16	84.33	83.06	81.44	82.65	**85.01**
building	91.29	91.65	**91.77**	90.96	91.11	91.54
wall	49.06	48.15	52.45	**53.68**	51.42	49.88
fence	57.65	**59.71**	57.44	58.85	57.87	59.66
pole	49.49	55.72	57.41	49.17	51.37	**58.39**
traffic light	62.55	62.98	65.83	61.21	62.71	**65.96**
traffic sign	72.78	74.55	74.77	72.08	73.56	**76.10**
vegetation	91.71	92.01	91.95	91.61	91.69	**92.13**
terrain class	63.61	64.31	60.55	**65.19**	63.44	63.99
sky	93.62	94.49	**94.60**	93.29	93.71	94.21
person	79.23	80.37	80.49	78.67	78.88	**81.19**
rider	57.75	56.93	58.24	57.41	57.15	**59.94**
car	93.67	**94.78**	94.14	93.58	93.90	94.68
truck	61.95	**80.79**	66.65	70.79	69.30	80.38
bus	80.49	**84.67**	77.65	79.44	81.04	84.00
train	68.19	**75.79**	73.93	59.97	61.25	73.67
motorcycle	63.64	65.25	65.43	64.25	61.37	**66.53**
bicycle	74.33	75.36	**75.55**	73.92	74.41	75.48
OA	95.39	95.72	95.63	95.27	95.37	**95.79**
mIoU	73.26	75.78	74.72	73.32	73.40	**76.36**

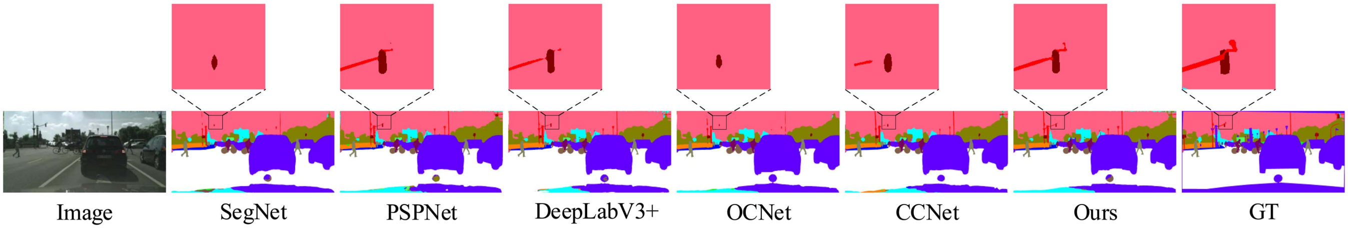


Fig. 10. Visual quality comparison of the proposed model over other different methods on Cityscapes dataset. Legend (white: class1 Road, class2: Sidewalk, class3: Building, class4: Wall, class5: Fence, class6: Pole, class7: Traffic Light, class8: Traffic Sign, class9: Vegetation, class10: Terrain Class, class11: Sky, class12: Person, class13: Rider, class14: Car, class15: Truck, class16: Bus, class17: Train, class18: Motorcycle, and class19: Bicycle).

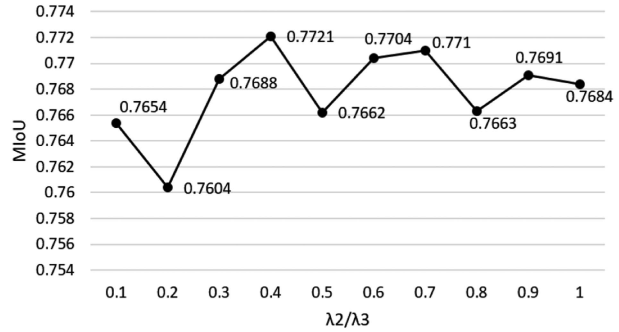


Fig. 11. Influence of different lambda values on segmentation results.

TABLE XI
NUMBER OF PARAMETERS AND THE NUMBER OF COMPUTATIONS FOR DIFFERENT ALGORITHMS

Method	Params	FLOPS
SegNet	49.8	92.94G
PSPNet	46.8	28.38G
OCNet	51.7	23.39G
CCNet	66.5	29.18G
DeepLabV3+	59.4	33.88G
Ours	50.5	55.45G

be seen from the Fig. 11, the best segmentation results can be achieved when the lambda value is 0.4, so we take the weight of 0.4 for both edge and body loss.

G. Computational Cost

Table XI tabulates the number of parameters as well as FLOPS of each compared algorithm. The proposed method has a slight increase in the number of parameters compared with SegNet and PSPNet, but it is still less than OCNet, CCNet, and DeepLabV3+ because the proposed network utilizes both the edge features and body features extracted by the network and uses its loss calculation for backpropagation. At the same time, this network requires a lot of feature matrix operations in the CAP module. Therefore, the FLOPS may be slightly higher than the other networks. However, compared with the final performance improvement, we think the computational cost is acceptable.

V. CONCLUSION

This article proposes a novel semantic segmentation framework with two flexible and effective modules for RSIs, named FGN-CAP. The EGFTM leverages the edge and body information as a guide to enhance edge contours and internal consistency, which can explicitly improve the representation of small or tiny objects. The CAP adaptively aggregates multiscale features in a coarse-to-fine manner through different contexts, enhancing the performance of objects of varying scales in RSIs. Both strategies can be applied in any optional network to boost their performance. The thorough experiments show their effectiveness and robustness, which achieve SOTA on three public remote sensing benchmarks.

REFERENCES

[1] Q. Zhang and K. C. Seto, "Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data," *Remote Sens. Environ.*, vol. 115, no. 9, pp. 2320–2329, 2011.

[2] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, "Transferred deep learning-based change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6960–6973, Sep. 2019.

[3] H. Fan, G. S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, 2015.

[4] 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.

[5] A. Dc, B. Sn, and D. Jhc, "Aerial image semantic segmentation using DCNN predicted distance maps," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 309–322, 2020.

[6] H. Li et al., "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618014.

[7] Q. Zhao, J. Liu, Y. Li, and H. Zhang, "Semantic segmentation with attention mechanism for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5403913.

[8] Y. Wang, W. Ding, R. Zhang, and H. Li, "Boundary-aware multitask learning for remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 951–963, 2020.

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

[10] 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.

[11] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 1–9.

[12] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Comput. Methods Programs Biomed.*, vol. 140, pp. 93–110, 2017.

[13] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[14] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Netw. Vis. Recognit.*, vol. 11, pp. 1–8, 2017.

[15] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. PhD Workshop*, 2018, pp. 117–122.

[16] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 13001–13008.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 6, 2017.

[18] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[19] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8950–8959.

[20] 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.

[21] 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. 1, 2018.

[22] S. Zheng et al., "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1529–1537.

[23] V. Jampani, M. Kiefel, and P. V. Gehler, "Learning sparse high dimensional filters: Image filtering, dense CRFs and bilateral neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4452–4461.

[24] X. He and S. Gould, "An exemplar-based CRF for multi-instance object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 296–303.

[25] H. Zhao et al., "PsaNet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 267–283.

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

[27] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, "Adaptive pyramid context network for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7519–7528.

[28] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9167–9176.

[29] M. Welling and T. N. Kipf, "Semi-supervised classification with graph convolutional networks," *J. Int. Conf. Learn. Representations*, 2016.

[30] 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.

[31] 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.

[32] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2020, pp. 5228–5237.

[33] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3773–3782.

[34] Y. Y. Boykov, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in *Proc. 8th IEEE Int. Conf. Comput Vis*, 2001, pp. 105–112.

[35] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.

[36] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

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

[38] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, Feb. 2017.

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

[40] F. Bastani et al., "Roadtracer: Automatic extraction of road networks from aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4720–4728.

[41] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4096–4105.

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

[43] C. Zhang, W. Jiang, and Q. Zhao, "Semantic segmentation of aerial imagery via split-attention networks with disentangled nonlocal and edge supervision," *Remote Sens.*, vol. 13, no. 6, 2021, Art. no. 1176.

[44] L. Bai, X. Lin, Z. Ye, D. Xue, C. Yao, and M. Hui, "MsanlfNet: Semantic segmentation network with multiscale attention and nonlocal filters for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6512405.

[45] 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.*, 2016, pp. 504–512.

[46] I. Kokkinos, "Pushing the boundaries of boundary detection using deep learning," *Comput. Vis. Pattern Recognit.*, 2015.

[47] 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.*, 2015, pp. 4380–4389.

[48] 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, 2017.

[49] G. Yang, Q. Zhang, and G. Zhang, "EANet: Edge-aware network for the extraction of buildings from aerial images," *Remote Sens.*, vol. 12, no. 13, 2020, Art. no. 2161.

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

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

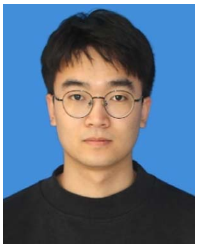
[52] X. Li et al., "Improving semantic segmentation via decoupled body and edge supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2020.

[53] A. Latif et al., "Content-based image retrieval and feature extraction: A comprehensive review," *Math. Problems Eng.*, vol. 2019, 2019, Art. no. 9658350.

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

Jiaojiao Li (Member, IEEE) received the B.E. degree in computer science and technology, the M.S. degree in software engineering, and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2009, 2012, and 2016, respectively.

She was an exchange Ph.D. Student of Mississippi State University supervised by Dr. Q. Du. She is currently an Associate Professor and Master supervisor with the School of Telecommunication, Xidian University. Her research interests include hyperspectral remote sensing image analysis and processing, and pattern recognition.



Yuzhe Liu received the B.E. degree in telecommunications engineering in 2021 from Xidian University, Xi'an, China, where he is currently working toward the M.S. degree in information and communication engineering with the State Key Laboratory of Integrated Service Network.

His research interests include hyperspectral image process, machine learning, and deep learning.

Jiachao Liu received the B.E. degree in telecommunications engineering in 2019 from Xidian University, Xi'an, China, where he is currently working toward the M.S. degree in information and communication engineering with the State Key Laboratory of Integrated Service Network.

His research interests include hyperspectral image process, machine learning, and deep learning.

Rui Song (Member, IEEE) received the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 2009.

He is currently a Professor and Ph.D. advisor with the State Key Laboratory of Integrate Service Network, School of Telecommunications, Xidian University. His research interests include image and video coding algorithms and VLSI architecture design, intelligent image processing, and understanding and reconstruction of 3-D scene.

Wei Liu received the master's degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2011.

She is currently with the State Key Laboratory of Geo-Information Engineering, Wuhan University. Her research focuses on real-time processing of UAV images and satellite images.

Kailiang Han received the Ph.D. degree in physical electronics from the Shanghai Institute of Technical Physics of Chinese Academy of Sciences (SITP of CAS), Shanghai, China, in 2008.

He is currently a Professor of SITP of CAS with major research on space photoelectric detection, remote sensing technique, image processing and machine vision, etc.

Qian Du (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland-Baltimore County, Baltimore, MD, USA, in 2000.

She is currently a Bobby Shackouls Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA. From 2009 to 2013, she was as a Cochair for the Data Fusion Technical Committee of the IEEE GRSS. From 2010 to 2014, she was the Chair with the Remote Sensing and Mapping Technical Committee of the International Association for Pattern Recognition. Her research interests include hyperspectral remote sensing image analysis and applications, pattern classification, data compression, and neural networks.

Dr. Du is a Fellow of SPIE-International Society for Optics and Photonics. She was the recipient of the 2010 Best Reviewer Award from the IEEE Geoscience and Remote Sensing Society (GRSS). She was the General Chair of the fourth IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing held at Shanghai, China, in 2012. She was an Associate Editor for the *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (JSTARS), *Journal of Applied Remote Sensing*, and *IEEE Signal Processing Letters*. Since 2016, she has been the Editor-in-Chief for the IEEE JSTARS.