Multiscale Feature Weighted-Aggregating and Boundary Enhancement Network for Semantic Segmentation of High-Resolution Remote Sensing Images

Yingying Zhao, Guizhou Zheng, Zhangyan Xu, Zhonghang Qiu, Student Member, IEEE, and Zhixing Chen

Abstract—High-resolution remote sensing images (HRRSIs) play an important role in large area and real-time earth observation tasks. However, HRRSIs typically comprise heterogeneous objects of various sizes and complex boundary lines, which pose challenges to HRRSI segmentation. Despite the fact that deep convolutional neural networks dramatically boosted the accuracy, several limitations exist in standard models. Existing methods, mainly concatenate multiscale information to extract the various sizes of objects. However, these methods ignore differentiating information, making it difficult to take advantage of them and completely extract small objects. In addition, there remain some difficulties in extracting boundary information with positions of uncertainty in previous works. In this article, we propose a novel multiscale feature weighted-aggregating and boundary enhancement network (MFBE-Net) for the segmentation of HRRSIs. ResNet-50, possessing a strong ability to extract features, is employed as the backbone. To fully utilize the information that was extracted, we propose a multiscale feature weighted-aggregating module, which aims to weight-integrate deep features, shallow features, and global information. The boundary enhancement module is designed to solve the blurry boundary information problems and locate its positions. Coordinate attention is also applied in the framework to coherently label size-varied ground objects from different categories and reduce information redundancy. Meanwhile, a mixed loss function is used to supervise the network training process. Finally, MFBE-Net was verified on two public HRRSI datasets, and the experimental results show that the proposed framework outperformed other existing mainstream deep learning methods and could further improve the accuracy of HRRSI segmentation.

Index Terms—Boundary enhancement, deep learning, feature weighted-aggregating, high-resolution remote sensing images (HRRSIs), semantic segmentation.

I. INTRODUCTION

HIGH-RESOLUTION remote sensing images (HRRSIs) are an important information source with higher image quality, wider coverage, and more structured data in comparison to traditional images [1]. HRRSI segmentation is to divide each pixel unit in HRRSIs into corresponding categories, which holds great significance in applications, such as land use and land cover [2], [3], ecological system [4], [5], inversion of water depth [6], [7], and other industries. However, visual interpretation of HRRSIs is a method with low efficiency and strong dependence on knowledge and experience. In recent years, automatic segmentation of HRRSIs has become an active research topic [8].

Various methods for HRRSI segmentation have been proposed, which fall into three main methods: 1) pixel-based segmentation, 2) object-oriented segmentation, and 3) deep learning segmentation. The pixel-based method [9] takes pixels as the basic unit with which to extract information, and classifies pixels according to the spectral information depending on prior knowledge, such as the maximum likelihood classifier [10], decision tree analysis [11], and K-means clustering [12]. The object-oriented method [13] takes objects as the basic unit, and an object is an entity composed of a group of adjacent homogeneous pixels, such as nearest-neighbor pattern classification [14] and membership function [15]. However, traditional HRRSI segmentation methods cannot capture more detailed features due to the more complex geometrical structures and richer spectral features in HRRSIs [16]. The complex classification and small objects make the traditional segmentation method difficult to complete the segmentation of HRRSIs.

Deep learning can segment HRRSIs and obtain high accuracy through automatically learning useful features from images [17]. Deep learning methods used in HRRSI segmentation are based mostly on convolutional neural networks (CNNs) and fully convolutional networks (FCNs) [18]. Related works [19], [20] have proven that deep learning methods are more accurate than traditional machine learning in the segmentation of HRRSIs. The FCN proposed by Long et al. replaces the fully connected layers of CNNs with fully convolutional layers [21], allowing input of any size, and is the first true work of semantic segmentation. Later, based on FCNs, semantic segmentation models, such as U-Net [22], DeepLab [23], and PSPNet [24] were successively proposed. Motivated by their remarkable improvements in comparison with traditional machine learning approaches, many scholars utilize these models in HRRSI segmentation.
Diakogiannis et al. combined a U-Net encoder-decoder backbone in combination with residual connections, atrous/dilated convolutions, pyramid scene parsing pooling, and multitasking inference to segment HRRSIs [25]. Based on DeepLab v3+, Liu et al. designed a decoder by adding more skip connections and convolution layers to obtain more detailed information to improve segmentation results [26]. Although the methods based on CNN and FCN show dramatic performance, the incomplete identification of small objects and the difficulty in extracting boundary information have invariably been problems that need to be solved urgently in HRRSI segmentation. To alleviate the problem of incomplete identification and low accuracy of small objects, the main solution is to fuse multiscale information. Zhang et al. applied the atrous spatial pyramid pooling (ASPP) module and the residual module to the proposed models to segment HRRSIs, showing that information fusion of different scales can restore some detailed features to a certain extent [27]. Liu et al. designed ScasNet, where multiscale contexts are captured on the output of a CNN encoder, and then are successively aggregated in a self-cascaded manner to progressively refine the target objects [28]. To a certain extent, multiscale information fusion can restore some detailed features and improve the rate of small object segmentation accuracy. However, previous works have mainly simply concatenated multiscale information, while ignoring the differences in multiscale information. Therefore, how to differentially fuse multiscale information in HRRSI segmentation is of great significance. Different from previous works, we consider the difference in the multiscale information coefficient, so we use not simply concatenation, but a weighting aggregate (which is more beneficial in segmenting small targets). Postprocessing is often used to solve the problem of poor localization near object boundaries, such as conditional random fields (CRF) [29], Markov random fields (MRF) [30], and DenseCRF [31], but these methods require additional parameters and low-level features, greatly affecting the running speed of a computer. In recent years, some studies have designed boundary detection modules as a branch of the model to improve the performance of boundary positions and certainty. Zhen et al. [32] combined semantic segmentation and boundary detection using the iterative pyramid context module. Takikawa et al. [33] think that color, shape, and texture contain very different types of information relevant for recognition, so they proposed a two-stream CNN. That is to say, the shape stream runs parallel to the classical stream as a separate branch to process information. Joint boundary detection and semantic segmentation have achieved certain results, but supervised training of boundary detection greatly increases the burden of a computer, which is secondary training. It is better if one uses one network and one train to extract body and boundary information at the same time. The multilevel structure of CNNs could solve this problem. Differently introducing an additional boundary detection network, we enhance boundary information by fusing multilevel information.

To address the aforementioned problems in the extraction of small objects, as well as blurry and uncertain position boundaries, a multiscale feature weighted-aggregating and boundary enhancement network (MFBE-Net) is proposed for HRRSI segmentation in this article. The MFBE-Net uses Resnet-50 as the backbone, in combination with the multiscale feature weighted-aggregating module (MFW), boundary enhancement (BEM), and coordinate attention (CA). The MFBE-Net adopts the encoding and decoding structure. In the decoding, MFW weighted aggregate of different scale information, CA is added between two decoding blocks, and boundary enhancement (BEM) utilizes multilevel information to enhance the boundary feature. Finally, we construct a mixed loss function to learn local and global contextual information and supervise the feature generation. The main contributions of this article are as follows.

1) A multiscale feature weighted-aggregating and boundary enhancement network is proposed for the segmentation of HRRSIs. The MFBE-Net can alleviate detail recovery and blurry boundary information problems to achieve a better performance of body and boundary.

2) To sufficiently leverage the differences in multiscale features, we design a multiscale feature weighted-aggregating module to model the relationship between multiscale features. MFW uses shallow features to generate the weight matrix of deep features because deep features lose some details related to salient targets, which is beneficial in recovering detail features.

3) We propose BEM to retrieve lost boundary information in the encoding process by adding the difference between low-level information and global information. This has never been experienced, as far as we know.

The efficiency of the proposed framework was verified with two HRRSIs datasets: the “Potsdam Data Set” and “GID.” By comparing the results to those of other state-of-the-art (SOTA) semantic segmentation algorithms, it was found that the MFBE-Net manifests superior performance in both visualization and quantitative evaluation.

The rest of this article is organized as follows. Section II reviews related works’ segmentation of HRRSIs. Section III introduces the proposed framework. The experiment details and the extensive experimental results are presented in Section IV. The detailed discussion is presented in Section V. Finally, Section VI concludes this article.

II. RELATED WORK

A. Segmentation in Computer Vision

Deep learning has been used to great effect in the field of computer vision. VGGNet [34], GoogleNet [35], and ResNets [36] are three deep neural networks designed for feature extraction on which many later image-processing model developments have been built [37]. It is pertinent to mention ResNets. As the network deepens, training the model gets increasingly challenging. ResNets utilize residual modules to allow the construction of very deep networks without gradient vanishing, producing results that are noticeably superior to earlier networks. The main ideas of these networks also support the development of semantic segmentation. Long et al. presented an FCN [21] which uses the fully convolutional layer to replace the fully connected layer of CNNs, allowing the network to allow input of any size. However, the FCN has the problem that the receptive field is fixed and the target details are easily lost or smoothed. Therefore, U-Net was proposed by Ronneberger et al. [22]. U-Net
uses the encode–decode framework. Using the concatenation operation to combine shallow features and deep features could form thicker features, which perform well in medical image segmentation. However, both downsampling and pooling layers in deep learning are accompanied by a loss of information, resulting in reduced spatial resolution, and the segmenter itself has spatial insensitivity. Hence, DeepLab is designed [23], which uses atrous/dilated convolutions to expand the range of filters to incorporate image contextual information into a larger neighborhood, thus being able to explicitly control the resolution of feature responses. Later, PSPNet proposes a pyramid pooling module to aggregate global contextual information, and uses an additional loss function to further improve the robustness and accuracy of the model [24].

B. HRRSI Semantic Segmentation

HRRSI segmentation is divided into three stages: 1) visual interpretation, 2) traditional machine learning, and 3) deep learning. The visual interpretation method has a strong dependence on researchers’ experience and knowledge, and segmentation results are different due to personal experience. Visual interpretation efficiency is low, which is suitable for fewer data. Traditional machine learning predicts the results of target variables by setting appropriate model parameters through train datasets [38]. Traditional machine learning methods include decision tree analysis [39], the maximum likelihood classifier [10], support vector machines (SVMs) [40], and so on. However, there are some problems with traditional machine learning algorithms. Decision tree analyses are prone to overfitting, and SVMs have a higher missing and wrong probability when the number of samples is large. Prior knowledge is very important because these methods usually need feature generation and selection steps before the segmentation process. With the development of segmentation in computer vision, HRRSI segmentation based on deep learning has also been greatly developed. A considerable amount of literature has been published on HRRSI tasks. Early research focused on learning features from the local and global information of images and then designed a supervised classifier to identify the learned features to label. Compared to natural images, HRRSIs have more complex classifications and objects with various sizes. The hierarchical structure of CNNs could solve the problem of various-sized-object segmentation, since the semantic information of small objects exists in the shallow layer, while the semantic information of large objects is situated in the deep layer [41]. Zhao et al. proposed an end-to-end multisource remote sensing image semantic segmentation network (MCENet) aiming at the problems of intraclass inconsistency and interclass indistinguishability in HRRSIs [42]. Dong et al. designed DenseUNet to connect convolutional neural network features through cascade operations and used its symmetrical structure to fuse the detail features in shallow layers and the abstract semantic features in deep layers [43]. Lee et al. presented a segmentation network trained with a small object mask to separate small and large objects in the loss function [44]. In recent years, the self-attentive transformer-based model which models global semantic information through self-attention has developed rapidly. Gao et al. proposed the STransFuse model that combines the benefits of Swin-Transformer with CNN to improve the segmentation quality of HRRSIs [45]. Ma et al. introduced a crossmodal multiscale fusion network (CMFNet) by exploiting both CNN and the transformer architecture to capture long-range dependencies across multiscale feature maps of remote sensing data in different modalities [46].

C. Optimization and Post-processing Strategy

In optimization strategies, attention mechanisms are generally proposed to solve data redundancy and coherently label by explicitly modeling interdependencies between spatial positions or channels. Squeeze-and-excitation (SE) [47] is one of the most popular attention mechanisms, but SE only considers channel information, ignoring the significance of location information. Therefore, Hou et al. [48] put forth coordinate attention (CA), which embeds location information into channel attention and is more suitable for vision tasks with dense predictions, such as semantic segmentation. Li et al. [49] proposed a new semantic segmentation network, i.e., network with spatial and channel attention (SCAttNet), to improve the semantic segmentation accuracy of HRRSIs. Li et al. proposed a multitention network (MANet) to address the underuse of information by extracting contextual dependencies through multiple efficient attention modules [50]. To improve spatial contiguity and sharpen borders in output label maps, a variety of postprocessing techniques have been explored. Chen et al. [29] combined the responses in the final DCNN layer with a fully connected conditional random field (CRF) for HRRSI segmentation, considering convolution scale and spatial positioning characteristics. Pan et al. [51] used a CRF as a postprocessing method to further improve segmentation accuracy. However, the combination of CNNs and a probability graph model is time-consuming with limited precision improvement, and a CRF lacks spatial consistency, at least during the training period, which is generally used in the case of paired or high-order models with few trainable parameters [52]. In order to alleviate the problem of feature maps with large receptive fields losing high-frequency information and causing blurred boundaries, many scholars construct a comparatively simple model by adding boundary detection to existing multi-level architecture. Peng et al. [53] proposed an encoder–decoder convolution network that extracts a set of feature maps shared by the detection branch and the segmentation branch to jointly carry out object detection and semantic segmentation. Sistu et al. [54] presented a joint multitask network design for learning object detection and semantic segmentation simultaneously and demonstrating the efficiency of the joint network. Li et al. [55] explicitly modeled the object’s body and edge, with the body feature and the residual edge feature being further optimized under decoupled supervision.

III. Method

This section is organized as follows. First, in Section III-A, an overview of the proposed model for HRRSIs is illustrated. Thereafter, the main modules, backbone, MFW, and BEM are introduced as following. For feature maps from the backbone,
MFW is utilized to model the relationship of multiple scales of feature maps during decoding. For the boundary with uncertain positions and blurry information, designing a special module to extract boundary information is necessary, and BEM is proposed to enhance the boundary information. Finally, we introduce the loss function used to supervise the training of the network.

A. Pipelines of Proposed Model

For the HRRSI segmentation, the multiscale and multilevel structures are beneficial to the segmentation of challenging objects and boundary extraction. Combining the advantages of multiscale and multilevel structures, the MFBE-Net is proposed in this article to solve the problems of low recognition accuracy of small objects and blurry boundary information. The main components of the proposed framework are as follows.

1) The MFBE-Net adopts the encoder–decoder structure, with one main advantage being that the skip-layer mechanism contributes to extracting the feature maps conveying both high-level and low-level information.

2) For the HRRSI segmentation, the multiscale structure is essential, since it generates feature maps with relatively large receptive fields, which are beneficial to the segmentation of challenging objects, especially for objects of varied sizes [56]. Pretrained Resnet-50 is used as an encoding machine of the backbone network to extract multiscale features, where dilated convolution is used to capture long-range information and stop excessive downsampling. The number of channels continues to increase, and the image size does not decrease after eight times of downsampling.

3) In the decoder part, MFW is used to weight-aggregate high-level, low-level, and global features. The feature-fusing process is gradually completed by the three MFWs. We finally obtain the resulting map to achieve end-to-end segmentation.

4) Multilevel methods can recover the resolutions of feature maps, refine the relatively coarser prediction, and precisely segment the boundaries of finely structured objects [41]. Therefore, in the process of feature extraction, BEM is proposed to enhance the boundary information.

5) Meanwhile, we add a CA module after each MFW in the decoding to coherently label ground objects of varied sizes from different categories and reduce information redundancy. Moreover, we construct a mixed loss function to supervise the local and global feature generation.

The overall framework of the MFBE-Net is shown in Fig. 1.

B. Backbone

As the network deepens, the features are more abstract and the information is richer, but the resulting gradient explosion and disappearance make the network optimization worse. Thus, we use U-Net as the backbone, which a residual module is introduced in the U-Net encoding process to achieve consistent training in this article. The residual module in Resnet-50 is shown in Fig. 2. Its main idea can be defined as follows:

$$F(x) = H(x) - x$$  \hspace{1cm} (1)

where \(x\) is the input map, \(F(x)\) is the stacked nonlinear layer fit, and \(H(x)\) formally denotes the desired underlying mapping.

After the residual module is introduced, the network highlights small changes, making the mapping more sensitive to changes in the output. However, too many downsampling operations in Resnet-50 cause a loss of detailed information and a reduction in resolution, which is not conducive to semantic segmentation tasks. Downsampling is done to broaden the receptive
Therefore, it is necessary to expand the receptive field without reducing the size of an image. Atrous/dilated convolution comes into being. Based on this, this article adjusts Resnet-50 to make it more suitable for semantic segmentation. In the adjusted Resnet-50, block1 and block2 remain unchanged. The residual module in block3 and block4 is shown in Fig. 3. We set the convolution step size to 1, and add the expansion coefficient within each residual building block so that the network can increase the receptive field without downsampling.

C. Multiscale Feature Weighted-Aggregating Module

For the segmentation of objects of varied sizes, feature maps from multiple scales could be helpful, since the information from high-resolution feature maps of CNNs can provide semantic information of small or threadlike objects, and information from low-resolution feature maps contains semantic information of relatively large objects. Meanwhile, the shallow layers of the network include low-level features, such as color, boundary, and spatial information, etc., [57], but they also contain a large amount of noise. The deep layer can provide rich semantic information and suppress noise, but its resolution is low, which is not conducive to localization. The global context feature can reduce the dilution of high-level features. But different spatial scales and global contextual information have great connections and differences. The conventional fusion strategy achieved by concatenation does not have different hierarchical feature maps, which might account for the challenge that exists in labeling objects of varied sizes. Based on the above, this article proposes MFW, which is an aggregation strategy for global features, high-level features, and low-level features, as illustrated in Fig. 4.

Different from previous works, we consider the difference in the multiscale information coefficient. We design MFW which not simply concatenates multiscale features, but a weighting aggregate. In this article, we use shallow features to generate the weight matrix of deep features, which is more radical and more efficient. As background noise is suppressed, the multiplication process can enhance the response of important targets. Specifically, first, the low-level feature \( f_{tl} \) and the high-level feature \( f_{th} \) are fed into a 1 × 1 convolution layer conv1, which compresses the features to have the same number of channels as \( f_{th} \). Thereafter, the 3 × 3 convolution layer conv2 is applied to the compressed low-level feature layer \( \tilde{f}_{tl} \) to generate the weight matrix \( W_{tl}^{t} \) of the compressed high-level feature layer \( \tilde{f}_{th}^{t} \). Finally, we multiply \( W_{tl}^{t} \) to upsample \( \tilde{f}_{th}^{t} \). Finally, the fusion features of the \( t \) stage \( f_{hl}^{t} \) are obtained through the RELU activation function

\[
W_{tl}^{t} = \text{conv}_2 \left( f_{tl}^{t} \right)
\]

\[
f_{hl}^{t} = \delta \left( W_{tl}^{t} \circ \text{upsampling} \left( f_{th}^{t} \right) \right)
\]

where \( t \) is the stage index, \( \tilde{f}_{tl}^{t} = \text{conv}_1 \left( f_{tl}^{t} \right) \) represents the compressed low-level features, \( \tilde{f}_{th}^{t} = \text{conv}_1 \left( f_{th}^{t} \right) \) is the compressed high-level features, \( \circ \) denotes multiplication, \( \delta \) denotes the RELU activation function, and \( \text{upsampling} \) is the upsampling operation through bilinear interpolation.

We apply the same fusion strategy to the global feature \( f_g \). The same method is different from the aforementioned in that the weight matrix \( W_{th}^{t} \) is generated by \( f_{th}^{t} \) through a 3 × 3 convolution
layer, and then the mask $W^t_2$ is multiplied to upsample the compressed global features $f^t_y$. The final weighted-aggregating features $f^t_w$, which can be characterized as the following, are created by combining these three level features and then passing them through a $1 \times 1$ convolution layer $\text{conv}_3$

$$f^t_w = \text{conv}_3 \left( \text{concat} \left( f^t_1, f^t_h, f^t_{gh} \right) \right). \quad (4)$$

D. Boundary Enhancement Module

Boundary information is considerably more ambiguous than body information during the feature extraction process, and there are uncertain positions [58], [59]. Therefore, it is unreasonable not to consider boundary. Some previous works have directly fused the multiscale and multilevel features and then upsampled them to obtain the final result map, and some works have trained the boundary as a single branch, which increases the burden on a computer. We concluded that neither approach was reasonable. For the segmentation of boundaries, feature maps from multiple levels could be helpful, since features from the shallow layer of CNNs can provide spatial information threadlike objects, such as the boundary, and information from the deep layer contains semantic information, such as position information. Thus, this article proposes a BEM based on multilevel layers to enhance the boundary information.

The BEM is mainly used to decrease lost boundary information by finding the differences between the global features extracted from the deep network and the low-level features extracted from the shallow network. BEM adds these differences to the final segmentation map. Here, we use upsampling of global feature $f_g$ to obtain the body information, which gathers contextual information from within objects to create a distinct body for each object. We consider low-level features $f^t_1$ to contain all of the information. There, boundary information $y_b$ is obtained by subtracting $y_g$ from $f^t_1$. Therefore, we can obtain lost boundary information in the encoding and decoding process and add it to the prediction map to achieve the purpose of information enhancement. The process can be written as follows:

$$y_b = \gamma \left( \text{conv}(f^t_1) - \text{upsampling} \left( \text{conv}(f^t_g) \right) \right) \quad (5)$$

where $\gamma$ is a convolution layer and upsampling operation, and $\text{conv}$ denotes a convolution layer with BatchNorm and RELU activation function. $\text{upsampling}$ is the upsampling operation through bilinear interpolation.

The final prediction map is jointly obtained by the $y_b$ and the fusion feature information $f^t_w$, which can be calculated as follows:

$$\hat{y} = \text{conv}_4 \left( y_b + f^t_w \right) \quad (6)$$

where $\text{conv}_4$ denotes a convolution layer with BatchNorm and RELU activation function.

E. Mixed Loss Function

This article uses a mixed loss function, which not only supervises the final segmentation graph but also jointly supervises $f^t_w$ obtained by the MFW, which allows us to efficiently train with the training samples. To train the MFBE-Net end to end, the total loss function of this article can be computed as follows:

$$L_{\text{total}} = \sum_{i=1}^{3} l_w + l_y \quad (7)$$

where $\sum_{i=1}^{3} l_w$ denotes the weighted sum of the losses between the label map $y$ and the segmentation map $f^t_w$ ($t = 1, 2, 3$) obtained via upsampling after MFW, and $l_y$ is the loss between the label map $y$ and the final segmentation map $\hat{y}$. The loss functions of both parts are DiceLoss [60]. DiceLoss was proposed by Milletari et al., which does not need to give weights to samples of different classes to establish the ideal balance between foreground and background pixels. Furthermore, this article chooses IoU and F1-score as the evaluation value indicators. Therefore, it is more appropriate to choose DiceLoss as the loss function in this article, which can be formulated as follows:

$$L_{\text{dice}} \left( y, \hat{y} \right) = 1 - \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{N} \left[ 2 \sum_{i=1}^{N} y_{ij} \hat{y}_{ij} \right] \quad (8)$$

where $m$ is the number of categories, $N$ is the number of pixels, and $y_{ij}$ is the true category of the prediction $\hat{y}_{ij}$.

IV. EXPERIMENTS AND ANALYSIS

To assess the effectiveness of the proposed framework, experiments were conducted on two HRRSI datasets. In this part, we first introduce the dataset, followed by the implementation details and accurate measurement. Finally, we compare our results with SOTA methods in Sections D and E.

A. Datasets

1) ISPRS Potsdam Dataset: Potsdam is a typical historic city with large building blocks, narrow streets, and dense settlement structures. The Potsdam 2D semantic labeling challenge data are provided in the framework of the 2D semantic labeling contest organized by the International Society for Photogrammetry and Remote Sensing (ISPRS) Commission III.4 (http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html) [61]. The Potsdam dataset has six categories [im pervious surfaces (imp. surf.), building, low vegetation (low veg.), tree, car, and background] and 38 tiles with a size of 6000 × 6000 pixels, whose spatial resolution is 5 cm with uniform color and texture distributions. The near infrared, red, green, and blue channels, the DSM, and the normalized DSM are all included in this dataset. We employ red, green and blue channels as input for our networks (three dimensions). The 38 images were cropped into a patch of 521 × 521 pixels, and a total of 18392 images were obtained, of which 12874 were used for training, 3678 were used for validation, and 1840 were used for testing according to the ratio of 7:2:1.

2) GaoFen Image Dataset: The GaoFen image dataset (GID) was published by Tong et al. [62] for semantic segmentation of HRRSIs, which contains 150 HRRSIs of five categories (built-up, farmland, forest, meadow, and waters) and 15 HRRSIs of fifteen categories captured from more than 60 cities in China.
Fig. 5. Image sample and corresponding ground truth of the Potsdam dataset and GID.

| Table I | Detailed Information of Setting on Experimental Datasets |
|---------|--------------------------------------------------------|
|          | Train | Val  | Test | Resolution | Category |
| Potsdam  | 12,874| 3,678| 1,840| 5m    | 6        |
| GID      | 13,230| 3,780| 1,890| 4m    | 6        |

Each original image is $7200 \times 6800$, and the panchromatic band resolution is 4 m. Twenty-eight images of five categories with good quality were selected as an experience dataset and cropped into a patch of $521 \times 521$ pixels, with a total of 18,900 images being obtained, of which 13,230 were used for training, 3,780 were used for validation, and 1,890 were used for testing according to the ratio of 7:2:1.

Table I summarizes the detailed information of the datasets illustrated above, and Fig. 5 shows some examples of these datasets. As Fig. 5 shows, for GID, there are confusing objects and more high fragmentation, which pose an extra challenge for the labeling task.

B. Implementation Details

Our model is implemented using the PyTorch. The semantic segmentation model was optimized by Adam [63] with a beta1 of 0.9 and a beta2 of 0.999. The ResNet-50 used in the network was pretrained on ImageNet [64] to avoid overfitting. We trained the models with a total batch size of eight for 100 iterations. The initial learning rate is set at le-5 and the “poly” policy in which the initial learning rate is multiplied by $(1 - (\text{iter} / \text{max}\_\text{iter}))^{\text{power}}$ with the power = 2 is employed to change the learning rate. On a server with NVIDIA GeForce RTX 3090 GPU accelerators (with 24GB GPU memory), all of the tests were conducted.

C. Accuracy Measurement

To assess the quantitative performance, two common and widely accepted metrics are utilized here. We adopt the accuracy evaluation algorithm of pixel-by-pixel labeling: mean intersection over union (MIoU), and the F1-score as the model performance evaluation criteria. We assume that there are $K$ categories in HRRSIs to be segmented, where $K$ includes the defined segmentation category and background. TP, FP, and FN represent the number of true positives, false positives, and false negatives, respectively.

Intersection over union (IoU) is the ratio of the intersection and union of the prediction result and the real value of a certain category. MIoU is obtained by averaging based on IoU, as shown in the following formula:

\[
\text{IoU} = \frac{TP}{FN + FP + TP} \tag{9}
\]

\[
\text{MIoU} = \frac{1}{K} \sum_{k=1}^{K} \text{IoU}_k \tag{10}
\]

The harmonic mean of recall and precision is called the F-score. We obtain the F1-score when setting the precision and recall to have the same weight. The F1-score has a range of 0 to 1, with a number closer to 1 indicating a superior model. The calculation formula is as follows:

\[
F1\_\text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{11}
\]

\[
\text{precision} = \frac{TP}{TP + FP} \tag{12}
\]

\[
\text{recall} = \frac{TP}{TP + FN} \tag{13}
\]

To evaluate the performance, we calculate the aforementioned metrics in each dataset.

D. Experiments on the Potsdam Dataset

In this part, we conduct experiments comparing our model with the SOTA network on the Potsdam dataset. We have chosen FCN, U-Net, PSPNet, Deeplabv3+, ScasNet, and HRCNet as benchmark approaches. FCN, U-Net, PSPNet and Deeplabv3+ are the popular network in natural scene imagery segmentation and are transferred learning to various fields, and ScasNet and
HRCNet have also achieved success in HRRSI segmentation. ScasNet and HRCNet are as follows.

1) ScasNet: This model was proposed by Liu et al. in 2018 [28]. ScasNet was proposed to solve confusing, manmade objects’ and intricate, finely structured objects’ problems regarding coherent and accurate labeling, as well as improving labeling coherence with sequential global-to-local context aggregation. Technically, multiscale contexts are captured in the output of a CNN encoder, and then are successively aggregated in a self-cascaded manner. Meanwhile, for finely structured objects, ScasNet boosts labeling accuracy with a coarse-to-fine refinement strategy.

2) HRCNet: This model was proposed by Xu et al. in 2021 [65]. In order to solve the imbalance of category scale and uncertain boundary information, HRCNet contains high-resolution network (HRNet) lightweight dual attention (LDA) and boundary awareness (BA). HRNet was adopted to retain spatial information. LDA was designed to obtain global contextual information in the feature extraction stage, and the feature enhancement feature pyramid (FEFP) structure is promoted and employed to fuse the contextual information of different scales. BA is combined with the boundary-aware loss function to achieve boundary information.

Using MIoU and the F1-score as the reference, Table II presents the semantic segmentation performance of our model and benchmark approaches, and it is clear that our model outperforms them all in terms of MIoU and F1-score. Our model achieves an MIoU of 0.956 and an F1-score of 0.978. Since FCN ignores multiscale feature fusion, the extraction of small objects (such as cars) is poor. We also can find that U-Net cannot benefit from semantic supervision efficiently. Convolutional kernels have fixed receptive fields, which makes it difficult for FCN and U-Net to effectively collect visual context information. PSPNet and DeepLabv3+ ignore the differences in features and lack consideration of boundary information. Although the overall segmentation effect is better than with FCNs and UNet, the improvement in small target segmentation is not obvious. In general, a multiscale structure is necessary to extract objects of varying sizes. As shown in Table II, ScasNet is beneficial to small objects, especially buildings and cars, which can achieve competitive performance on the Potsdam dataset. However, multilevel information is just as important as multiscale information, and it is also necessary to process the boundary information separately. Our MFW considers the above, while ScasNet only uses multiscale information within the models. Therefore, overall ScasNet is inferior to our model. Similar to HCRNet, our model contains the MFW and BEM to aggregate multiscale features and enhance the boundary, respectively. By contrast, the performance of HRCNet is not so fine, because it only supervises the training through the boundary loss function, and its effect is not even as good as that of ScasNet.

To assess the qualitative performance, we also visualized the benchmark and our proposed MFBE-Net results, as shown in Fig. 6. To zoom in on a more understandable comparison with the other methods, four 1000 × 1000 pixel regions are randomly chosen. For HRRSIs, U-Net does not perform as well as in the field of medical images, and there are many misclassifications, even fragmentation (first row). In the first row, the performance of each network is not good. There are discontinuities in all comparing models, and the boundaries are not extracted completely. In contrast, MFBE-Net can produce images with better continuity (first row, third row), which also shows that multiscale feature weighted aggregating is more efficient in detail recovery and image smoothing. Although our model does not fully recover the boundary loss, our model outperforms them in comparison. Among the six groups of control experiments, ScasNet is the most outstanding, and the details are handled well on the whole in the remaining sets of maps, but there are some ambiguities and uncertainties boundary extraction. HRCNet, which is dedicated to solving the problem of boundary ambiguity, is slightly better than other models in boundary processing, but the overall accuracy is not ideal. Our model has added a boundary enhancement module to generate sharper boundaries. For example, the building edge (second row, fourth row) is very easy to confuse and bend. However, the MFBE-Net succeeds in segmenting it. Overall, our model has better segmentation than that of other models, with clear boundaries and complete extraction of small objects. The quantitative and qualitative results both support the effectiveness of the MFBE-Net at the system level.

### E. Experiments on GID

We also performed benchmark experiments on another significant HRRSIs GID to further assess the proposed framework. Since GID has more complex and various-sized objects, the difficulty of the segmentation task has increased. The experimental results are provided in Table III. Our model achieved the best performance with an MIOU of 0.950 and an F1-score of 0.974, exceeding the performance of other models. We can find the same evidence that U-Net still has the lowest accuracy, probably because there are more categories of HRRSIs in comparison with medical images. Compared with ScasNet, which is the best among other models, the MFBE-Net further improves
Fig. 6. Visualization of results of compared methods in the Potsdam dataset. (a) Original Image. (b) GT. (c) FCN. (d) U-Net. (e) PSPNet. (f) Deeplabv3+. (g) ScasNet. (h) HRCNet. (i) Ours.

Fig. 7. Visualization of results of compared methods in GID. (a) Original Image. (b) GT. (c) FCN. (d) U-Net. (e) PSPNet. (f) Deeplabv3+. (g) ScasNet. (h) HRCNet. (i) Ours.

MIoU by 4.28% and the F1-score by 3.29%. Experimental results confirm the effectiveness of the MFBE-Net again. In addition, compared with the Potsdam dataset, GID has larger coverage and greater intraclass variation, which poses a great challenge for HRRSI segmentation. For example, the FCN, DeepLabv3+, PSPNet, and ScasNet are not as effective in GID as the Potsdam dataset. However, our model significantly exceeds other models and maintains robustness. Even for complex backgrounds, our model is also well segmented. Overall, our model is superior to these models as always, which also shows the robustness of our model and the importance of MFW and BEM.

The six meticulous visualization results are displayed in Fig. 7 to exhibit a microlevel visual performance. We also select the FCN, U-Net, DeepLabv3+, PSPNet, ScasNet, and HRCNet for comparison. We can find that on the whole our model and ScasNet detail spatial information more accurately, and obtain results closer to the real truth. For narrow targets, our model achieved the best segmentation results (first row, second row). For boundary information that is difficult to extract, our model can also handle it well (third line). In the extraction of the built-up, our model and ScasNet achieved good results, which also proves the importance of multiscale information for small object extraction. In contrast, the FCN, U-Net, and DeepLabv3+
all face the problems of blurry boundary information and small object extraction.

V. DISCUSSION

A. Computational Cost Analysis

Model accuracy should be measured, but also model size and inference speed should be taken into account. It is important to consider the optimal tradeoff between accuracy, computational efficiency, and the number of operations measured by the floating point operations per second (FLOPs) and the number of parameters. To evaluate the required amount of calculation resources of the compared models, the number of parameters, inference time for a single image, and Flops of our method and comparative methods are shown in Table IV.

As we can see from Table IV, our proposed network has low parameters, medium FLOPs, and inference time, which are 170.78G, 51.32M, and 54 ms in Potsdam, 52 ms in GID, respectively. For inference time of a single figure, the same model performs slightly less well than GID on the Potsdam dataset. The reason is the resolution of remote sensing images in the Potsdam dataset is higher than that of GID. In addition, combing Table IV with Tables II and III, although the FCN showed much better inference speed, its accuracy is not appropriate for semantic segmentation. ScasNet showed better accuracy results, but it required more than 321.11G FLOPs. HRCNet is slow in inference speed and not appropriate for real-time semantic segmentation. It can be observed that our method can achieve better performances with other models on most objects, while performing slightly worse on computational cost. In contrast, our model can better meet the higher requirements of semantic segmentation on model deployment and application.

B. Ablation Experiment

In this part, we conduct an ablation experiment to confirm the efficiency of each crucial element of the proposed MFBE-Net. The ablation study is conducted on the Potsdam dataset. As a baseline, we train U-Net, which uses Resnet-50 as the encoder. Thereafter, each module is added progressively. We conduct experiments on those architectures and report the performance via MIoU and the F1-score. As shown in Table V, we can find that MIoU increased by 11.1% and the F1-score by 11.4% after MFW replaced the sample concatenation operation. Meanwhile, as we expected, the MFW is beneficial in the extraction of small objects, such as low vegetation, trees, and cars, which show a noticeable boost in performance. In addition, the addition of the BEM increased MIoU from 0.840 to 0.956, and the F1-score by 8.91%. The accuracy improvement for each category proves the effectiveness of the BEM in the MFBE-Net for HRRSI segmentation. Finally, the combination of the MFW and BEM has already produced the best results.

We visualize the result maps generated by the ablation study in Fig. 8 to better understand the effect of the MFW and BEM, where we can see that with the addition of the MFW, the recognition of small targets is easier than the baseline, such as small-scale impervious surfaces (fourth row), and the identification of false positives could be reduced with the proposed modules (third row). Moreover, the BEM makes the object boundary smoother and more accurate (first row, second row). This implies that the BEM has improved the occurrence of wrong segmentation and image connectivity problems. To sum up, our proposed modules can better extract small targets and boundaries. To obtain the best segmentation of HRRSIs, each component in the proposed model is necessary.

C. Model Analysis

Compared with other semantic segmentation methods, our model outperforms other methods in each category and overall segmentation performance. This is mainly attributable to the following points. Multiscale feature weighted aggregating and boundary enhancement are both key factors that influence how well segmentation performs. These factors are taken into account by the MFBE-Net, resulting in successful semantic segmentation. Additionally, our model improves the learning process by utilizing intermediary loss functions, where intermediate loss supervises the backpropagation process, by defining how poorly the network performs at the intermediary layers of the network, rather than using a single loss function at the end of the network.

The attention mechanism supports the network’s ability to concentrate on key discriminative regions in the images by giving such areas higher weights, while suppressing redundant and unimportant regions, such as backgrounds. However, the attention mechanism greatly increases the complexity of the model, which leads to extra computational costs and more training times. In the future, we will carry out further research to reduce the computational complexity of MFBE-Net for better prospects.
Fig. 8. Visualization of results of ablation study on the Potsdam dataset. GT means ground truth. (a) Original Image. (b) GT. (c) Baseline. (d) Baseline+MFW. (e) Baseline+MFW+BEM.

Fig. 9. Partial visualizations of the results obtained from uncertain areas. (a) Original Image. (b) GT. (c) Ours.

D. Generalizability to Uncertainties

The generalizability of the model when the image quality is poor is discussed in this part. HRRSI segmentation becomes more challenging when some key objects are invisible, or suppressed due to their size, shadow, occlusion from the surrounding objects, or where the background suppresses the objects of interest. The results shown in Fig. 9 come from some areas where there are some quality issues (e.g., shadow, mosaic, image distortion) either on image or ground truth. In most instances, our model can correctly predict the category of the area with poor imaging quality. For instance, in the first row of Fig. 9, cars are occluded by building shadows, but our model correctly discriminates shadows, buildings, and low vegetation. In the second row of Fig. 9, there is a huge region covered with mosaic, and the region is not labeled as impervious surface or background but building. Although our network prediction result is consistent with the label, it should not be the building according to its true category.

Although the proposed model attains competitive accuracy results, shadow detection, alignment, and correction for HRRSI segmentation remain a great area of interest requiring further attention to mitigate shadow-prone errors. Finally, blending HRRSIs multiband information and making full use of spectral features requires further attempts to obtain optimum results.

VI. Conclusion

Most previous HRRSI segmentation works have ignored the difference in multiscale features and have especially not considered boundary information, which makes it difficult to extract small targets and make boundary positioning accurate and information complete. In this article, we propose a multiscale and multilevel semantic segmentation network for HRRSI segmentation, which performs excellently in small object extraction and boundary refinement. To fully utilize multiscale features and their differences, we design an MFW to use low-level features to weigh and fuse high-level features, which improves detail recovery and small object extraction by establishing the relationship between multiscale features. The MFW is more effective than the general concatenation. Aiming at the problem of blurred boundaries and uncertain positioning in the encoding process, a BEM is introduced in our network to enhance boundary features which utilize multilevel layer features to retrieve a loss of boundary information to generate more accurate boundaries. The detailed ablation study suggests that the MFW and BEM are of significant importance for semantic segmentation. The effectiveness and
superiority of the MFBFE-NET are thoroughly evaluated in comparison to two different HRRSIs, the Potsdam datasets, and GID. Compared to existing algorithms, our network achieved the best results locally and globally. Future experiments are necessary to validate the proposed framework for HRRSIs segmentation in other datasets and confirm the effectiveness of the proposed MFW and BEM in different networks.

REFERENCES

[1] M. Wang, H. Zhang, W. Sun, S. Li, F. Wang, and G. Yang, “A coarse-to-fine deep learning based land use change detection method for high-resolution remote sensing images,” Remote Sens., vol. 12, no. 12, Jun. 2020, Art. no. 1933.

[2] S. Kaliraj, N. Chandrasekar, K. K. Ramachandran, Y. Srinivas, and S. Saravanan, “Coastal land use and land cover change and transformations of Kanyakumari coast, India using remote sensing and GIS,” Egyptian J. Remote Sens. Space Sci., vol. 20, no. 2, pp. 169–185, Dec. 2017.

[3] I. Ahmad, A. Singh, M. Fahad, and M. M. Waqas, “Remote sensing-based framework to predict and assess the interannual variability of maize yields in Pakistan using Landsat imagery,” Comput. Electron. Agriculture, vol. 178, Nov. 2020, Art. no. 105732.

[4] Z. Li et al., “Urban landscape extraction and analysis in the mega-city of China’s coastal regions using high-resolution satellite imagery: A case of Shanghai, China,” Int. J. Appl. Earth Obs., vol. 72, pp. 140–150, Oct. 2018.

[5] K. Yuan, X. Zhuang, G. Schaefer, J. Feng, L. Guan, and H. Fang, “Deep-learning-based multispectral satellite image segmentation for water body detection,” IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens., vol. 14, pp. 7422–7434, 2021.

[6] G. Zheng, X. Le, H. Wang, and W. Hua, “Inversion of water depth from worldview-02 satellite imagery based on BP and RBP neural network,” Earth Sci., vol. 42, no. 12, pp. 2345–2353, Dec. 2017.

[7] G. Zheng, Z. Pan, Y. Meng, and H. Wang, “Inversion of sea surface flow field in Southern South China Sea based on satellite remote sensing data,” Earth Sci., vol. 46, no. 01, pp. 341–349, Jan. 2021.

[8] M. D. Hossain and D. Chen, “Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective,” ISPRS J. Photogrammetry Remote Sens., vol. 150, pp. 115–134, Apr. 2019.

[9] C. Zhang et al., “Joint deep learning for land cover and land use classification,” Remote Sens. Environ., vol. 221, pp. 173–187, Feb. 2019.

[10] M. Z. Ali, W. Qazi, and N. Aslam, “A comparative study of ALOS-2 PALSAR and landsat-8 imagery for land cover classification using maximum likelihood classifier,” Egyptian J. Remote Sens. Space Sci., vol. 21, pp. 520–535, Jul. 2018.

[11] R. M. D. S. Vieira et al., “Land degradation mapping in the MATOPIBA region (Brazil) using remote sensing data and decision-tree analysis,” Sci. Total Environ., vol. 782, Aug. 2021, Art. no. 146900.

[12] M. S. Navin and L. Algelandesswari, “Land use land cover change detection using k-means clustering and maximum likelihood classification method in the javadi hills, Tamil Nadu, India,” Int. J. Eng. Adv. Technol., vol. 9, no. 153, pp. 51–56, Dec. 2019.

[13] S. W. Myint, P. Gober, A. Braelz, S. Grossman-Clarke, and Q. Weng, “Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery,” Remote Sens. Environ., vol. 115, no. 5, pp. 1145–1161, May 2011.

[14] T. Cover and P. Hart, “Nearest neighbor pattern classification,” IEEE Trans. Inform. Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[15] S. S. Jamsandekar and R. R. Muddolkar, “Fuzzy classification system by self generated membership function using clustering technique,” BVICA M’s Int. J. Inf. Technol., vol. 6, no. 1, pp. 697–704, Feb. 2014.

[16] Q. Zhu et al., “A global context-aware and batch-independent network for road extraction from VHR satellite imagery,” ISPRS J. Photogrammetry Remote Sens., vol. 175, pp. 353–365, May 2021.

[17] Q. Zhu, Y. Zhong, L. Zhang, and D. Li, “Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 10, pp. 6180–6195, Oct. 2018.

[18] Y. Gao, Y. Liu, T. Georgiou, and M. S. Lew, “A review of semantic segmentation using deep neural networks,” Int. J. Multimedia Inf. Retrieval, vol. 7, no. 2, pp. 87–93, Jun. 2018.
Zhixing Chen received the B.S. degree from the School of Resource and Environmental, Hubei University, Wuhan, China, in 2020. He is currently working toward the master’s degree in cartography and geographical information engineering with the China University of Geosciences, Wuhan, China.

His research interests include remote sensing image processing, image super-resolution reconstruction, and deep learning.

Zhongyang Xu received the B.S. degree in geographic information science from Henan Polytechnic University, Jiaozuo, China, in 2021. He is currently working toward the master’s degree in cartography and geographical information engineering with the China University of Geosciences, Wuhan, China, in 2021. He is currently working toward the Ph.D. degree in cartography and geographic information engineering from the China University of Geosciences, Wuhan, China.

She is the author of five books, more than 50 journal papers and five software copyrights. His research interests include marine remote sensing, environmental remote sensing, deep learning, and spatio-temporal data mining.

Yingying Zhao received the B.S. degree in human geography and urban planning from the College of Sciences, Shihzei University, Shihzei, China, in 2021. She is currently working toward the master’s degree in cartography and geographical information engineering with the China University of Geosciences, Wuhan, China.

Her research interests include remote sensing image semantic segmentation, the processing of remote sensing image data, and deep learning.

Guizhou Zheng received the Ph.D. degree in cartography and geographic information engineering from the China University of Geosciences, Wuhan, China, in 2008. Since 2005, he has been a Professor with the Faculty of Information Engineering, China University of Geosciences. He is the author of five books, more than 60 journal papers and five software copyrights. His research interests include marine remote sensing, environmental remote sensing, deep learning, and aerosols.

Guizhou Zheng received the Ph.D. degree in cartography and geographic information engineering from the China University of Geosciences, Wuhan, China, in 2008. Since 2005, he has been a Professor with the Faculty of Information Engineering, China University of Geosciences. He is the author of five books, more than 50 journal papers and five software copyrights. His research interests include marine remote sensing, environmental remote sensing, deep learning, and aerosols.

Zhongyang Xu received the B.S. degree in geographic information science from Henan Polytechnic University, Jiaozuo, China, in 2021. He is currently working toward the master’s degree in cartography and geographical information engineering with the China University of Geosciences, Wuhan, China, in 2021. He is currently working toward the Ph.D. degree in cartography and geographic information engineering from the China University of Geosciences, Wuhan, China.

His research interests include marine remote sensing, environmental remote sensing, deep learning, and spatio-temporal data mining.

1 [43] R. Dong, X. Pan, and F. Li, “DenseU-net-based semantic segmentation of small objects in urban remote sensing images,” IEEE Access, vol. 7, pp. 65347–65356, 2019.
2 [44] J. Lee, C. Kim, and S. Sull, “Weakly supervised segmentation of small buildings with point labels,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 7406–7415.
3 [45] G. Ma et al., “STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 10990–11003, 2021.
4 [46] G. Xu, X. Zhang, and M. -O. Pun, “A crossmodal multiscale fusion network for semantic segmentation of remote sensing data,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 3463–3474, 2022.
5 [47] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7132–7141.
6 [48] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 13713–13722.
7 [49] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, “SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 18, no. 5, pp. 905–909, May 2021.
8 [50] R. Li et al., “Multitattention network for semantic segmentation of fine-resolution remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 560713.
9 [51] X. Pan, L. Gao, B. Zhang, F. Yang, and W. Liao, “High-resolution aerial imagery semantic labeling with dense pyramid network,” Sensors-Basel, vol. 18, no. 11, Sep. 2018, Art. no. 3774.
10 [52] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, “Semantic segmentation using adversarial networks,” 2016. [Online]. Available: https://arxiv.org/abs/1611.08408
11 [53] J. Peng, Z. Nan, L. Xu, J. Xin, and N. Zheng, “A deep model for joint object detection and semantic segmentation in traffic scenes,” in Proc. IJCNN, 2020, pp. 1–8.
12 [54] G. Sistu, I. Leang, and S. Yogamani, “Real-time joint object detection and semantic segmentation network for automated driving,” 2019. [Online]. Available: https://arxiv.org/abs/1901.03912
13 [55] X. Li et al., “Improving semantic segmentation via decoupled body and edge supervision,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 435–452.
14 [56] X. Ma, L. Wang, K. Qi, and G. Zheng, “Remote sensing image scene classification method based on multi-scale cyclic attention network,” Earth Sci., vol. 46, no. 10, pp. 3740–3752, 2021.
15 [57] A. Li, L. Jiao, H. Zhu, L. Li, and F. Liu, “Multi-task semantic boundary awareness network for remote sensing image segmentation,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5400314.
16 [58] G. Bertasius, J. Shi, and L. Torresani, “Semantic segmentation with boundary neural fields,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3602–3610.
17 [59] 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. 115, pp. 158–172, Jan. 2018.
18 [60] F. Milletari, N. Navab, and S. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in Proc. 4th Int. Conf. 3D Vis., 2016, pp. 565–571.
19 [61] 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.
20 [62] X. Tong et al., “Learning transferable deep models for land-use classification with high-resolution remote sensing images,” 2018. [Online]. Available: https://arxiv.org/abs/1807.05713
21 [63] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014. [Online]. Available: https://arxiv.org/abs/1412.6980
22 [64] O. Russakovsky et al., “Imagenet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Apr. 2015.
23 [65] Z. Xu, W. Zhang, T. Zhang, and J. Li, “HRCNet: High-resolution context extraction network for semantic segmentation of remote sensing images,” Remote Sens., vol. 13, no. 1, Dec. 2020, Art. no. 71.