Multilateral Semantic With Dual Relation Network for Remote Sensing Images Segmentation

Weiheng Zhao, Jiannong Cao, and Xueyan Dong

***Abstract*—Semantic segmentation of remote sensing images is an extensively employed and demanding task. Although deep convolutional neural networks have significantly increased the accuracy of semantic segmentation, the problems of losing detailed features in segmentation and ignoring rich contextual information of images still exist. To solve these challenges, we propose a multilateral semantic with dual relation network (MSDRNet) for remote sensing images segmentation. The proposed MSDRNet consists of two parallel modules, the detail semantic module and the global semantic module, for extracting image detail and global features, respectively. Subsequently, improved spatial relation block and channel relation block are introduced in two separate parallel modules to further enhance the contextual connection of the images. Finally, a feature refinement module is added to balance the multilateral features between the features extracted from the two branches. We display the robustness and effectiveness of the proposed MSDRNet on the publicly available ISPRS Potsdam and Vaihingen datasets. We further experimented with the Gaofen image dataset, which contains information on larger scale features, to demonstrate the validity of our model. The results of extensive experiments conducted on the aforementioned three datasets show that the proposed approach outperforms several state-of-the-art semantic segmentation methods.**

***Index Terms*—Attention mechanisms, deep learning, remote sensing image, semantic segmentation.**

I. Introduction

SEMATIC segmentation plays a crucial role in recognizing and extracting information from remote sensing images, making it an essential tool for remote sensing analysis. Therefore, it has become a prominent research topic in the field of remote sensing, including land cover classification [1], building extraction [2], road extraction [3], forest stands identification [4], urban land-use classification [5], and environmental monitoring [6].

Manuscript received 27 June 2023; revised 11 September 2023 and 1 October 2023; accepted 1 November 2023. Date of publication 8 November 2023; date of current version 23 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 41571346, in part by the Open Fund for Key Laboratory of Degraded and Unused Land Consolidation Engineering, and in part by the Ministry of Natural Resource under Grant SXDJ2017-10 and Grant 2016KCT-23. *(Corresponding author: Xueyan Dong.)*

Weiheng Zhao and Xueyan Dong are with the School of Earth Science and Resources, Chang'an University, Xi'an 710064, China (e-mail: 2019027008@chd.edu.cn; 2019027009@chd.edu.cn).

Jiannong Cao is with the School of Geological Engineering and Surveying, Chang'an University, Xi'an 710064, China, and also with the Ministry of Natural Resources, Key Laboratory of Degraded and Unused Land Consolidation Engineering, Chang'an University, Xi'an 710064, China (e-mail: caojiannong@126.com).

Digital Object Identifier 10.1109/JSTARS.2023.3330731

In the emerging research on semantic segmentation of remote sensing images, deep convolutional neural networks (DCNNs) [7] have shown their strong capability in feature learning and target extraction. In particular, the introduction of fully convolutional network (FCN) [8] has provided innovative ideas for image semantic segmentation, which is essentially a pixel-level classification task. Subsequently, advanced end-to-end semantic segmentation models have drawn significant inspiration from the FCN concept and embraced an encoder–decoder architecture. The encoder component extracts information from input images and generates high-level feature maps. On the other hand, the decoder component utilizes these feature maps to reconstruct masks and achieve pixel-level segmentation through upsampling operations. In contrast to natural images, remote sensing images display extensive variations on a large scale, with a broader imaging range and more complex and diverse backgrounds. Consequently, the diverse nature of objects within remote sensing images poses significant challenges for semantic segmentation. Specifically, in remote sensing images, some classes of objects, particularly buildings, are commonly composed of various materials. Conversely, different objects such as forests and grasslands often exhibit striking visual similarities. The presence of global intraclass variance and local interclass variance in remote sensing images presents significant difficulties in achieving consistent labeling for semantic segmentation. To address the aforementioned challenges, He et al. [9] introduced edge information into FCN to refine the segmentation results based on remote sensing imagery. Tian et al. [10] proposed an end-to-end class-level segmentation framework called class-wise FCN, which utilizes a shared encoder to process class-specific features, thereby improving segmentation accuracy. Fu et al. [11] employed the FCN as the backbone network and replaced the fully connected layer with convolutional layers to enhance network efficiency. Mo et al. [12] utilized FCN to generate directional elevation gradients between adjacent pixels and established the relativity between the altitude system of the training data and the output. DCNNs have also achieved excellent results in hyperspectral image segmentation. Zhang et al. [13] proposed SOT-Net to redesign the information complementary collaboration approach and redundancy exclusion operator to enhance the semantic relevance. Yu et al. [14] proposed an FADCNN with a feedback attention mechanism, and for the first time, a feedback attention module was developed, which utilizes semantic knowledge at the higher level of the dense model to enhance the attention mapping. However, these methods often neglect the problem of missing details in semantic segmentation and fail to effectively integrate global and local features. As a result, they may not capture the intricate characteristics of objects in remote sensing images, leading to potential limitations in segmentation accuracy and boundary delineation.

In remote sensing images, objects exhibit strong spatial dependencies in the geographic space. Objects of the same class, particularly those at larger scales such as forests and water bodies, tend to have highly consistent pixel-level features with surrounding pixels of the same class, while significantly differing from pixels belonging to different classes. In addition, remote sensing images also contains a considerable number of dispersed objects with intricate structures, such as cars and small buildings. These detail objects exhibit long-range dependencies. However, due to the limitations imposed by the FCN-based methods, the relationships and dependencies between objects cannot effectively contribute to their role in semantic segmentation. In recent years, transformers have demonstrated advanced performance in various natural language processing tasks, offering new insights for semantic segmentation in the field of computer vision [15]. Yue et al. [16] proposed TreeUNet based on the framework of deep semantic modeling, utilizing the TreeCutting algorithm to transmit the feature layer, fusing the multiscale features and learning the optimal weights of the model, which achieves better results in the semantic segmentation work. Hang et al. [17] proposes a multiscale progressive segmentation network that utilizes a three-level subnetwork to segment image objects layer by layer, and a scale-guided module for the internetwork feature learning. Niu et al. [18] introduced channel attention modules and region attention modules into the network to reduce feature redundancy during semantic segmentation. Liu et al. [19] proposed a novel hybrid attention semantic segmentation network consisting of intraclass attention branch and interclass attention branch, resulting in significant improvements in Mars terrain segmentation. Shi et al. [20] utilized a high-resolution network as the backbone network and integrated region attention modules and context fusion modules to enhance the relationship between local and global pixels. Li et al. [21] introduced a multiattention network that addresses the insufficient correlation of multiscale feature information by extracting contextual information through attention modules. Cheng et al. [22] achieved higher object localization accuracy in the semantic segmentation process by aggregating multiscale features through a context fusion module.

Global features and local features play a pivotal role in semantic segmentation tasks. Despite the integration of spatial and channel relationships into image processing through attention mechanisms in these methods, certain challenges persist. On one hand, these two types of features inherently demand distinct network architectures. Detail features are situated in the initial stages of the network and necessitate a broader channel dimension. Conversely, high-level global information relies on abstraction from deep-level features. Managing these two types separately guarantees the precision of semantic segmentation. However, simply fusing the acquired global features and local features leads to a network constrained in exploring the intricate associations among relation modules, feature map attributes (size, depth, and abstraction level), and receptive fields. Consequently, this constrains the overall performance of the semantic segmentation model.

To tackle these aforementioned problems, a novel multilateral semantic with dual relation network (MSDRNet), which takes into consideration both relational and contextual features, is proposed for remote sensing images segmentation. The backbone structure of the MSDRNet is based on ResNet [23]. In order to fully capture the relationship features and contextual features required for semantic segmentation, we establish two parallel branches: the detail semantic module (DSM) and the global semantic module (GSM). Furthermore, to enhance feature aggregation, we introduce the feature refinement module (FRM) to reduce the disparities among semantic information.

Intuitively, remote sensing images are rich repositories of spatial information. To enhance the representation of intricate details, we introduce a spatial relation block (SRB) within the first branch's DSM. This leverages a broader channel width and a shallower branch depth to extract meticulous semantic features. Furthermore, global contextual information plays a pivotal role in resolving the issue of inconsistencies within and between classes in images, a vital consideration in determining object categories during semantic segmentation. Hence, in the second branch's GSM, we incorporate a channel relation block (CRB). This module utilizes multiple residual blocks to extract deep-level features effectively, thereby capturing global information in the image and ensuring the precision of semantic segmentation. Finally, considering that the distribution and positioning of local and global features differ, directly merging them inevitably introduces background noise from detail features. This noise can compromise feature robustness and potentially lead to the loss of details. Therefore, we introduce the FRM to harmonize semantic feature discrepancies between the two branches and to comprehensively assess the aggregation capability of context information, facilitating multiscale feature encoding.

The contributions of this article are summarized as follows.

1) We introduce MSDRNet, a multilateral network that merges global and detail semantic information through the global semantic module (GSM) and detail semantic module (DSM), addressing the problem of neglecting small-scale features.
2) We consider the pixel-channel relationship vital for semantic segmentation. Thus, we introduce SRB and CRBs to boost attention mechanism performance and minimize information loss.
3) To enhance multilateral features, we introduce the feature refinement module (FRM), which reduces semantic differences, distinguishes similar features, and integrates features from various levels to boost prediction accuracy.

The rest of this article is organized as follows. Related works on DCNNs-based semantic segmentation methods and Attention mechanisms are reviewed in Section II. Section III describes the details of MSDRNet and the structure of each module. The experimental results and discussion are presented in Section IV. Finally, Section V concludes this article.

II. Related Work

A. Semantic Segmentation for Remote Sensing Imagery

Semantic segmentation aims to assign labels to each pixel in remote sensing imagery, dividing the image into meaningful and distinct regions with homogeneous attributes. Deep learning-based approaches leverage the training and learning between multiple layers of networks, enabling effective capture of both detail and deeply abstract features in remote sensing imagery. As a result, these approaches have been widely applied in semantic segmentation of remote sensing imagery [24]. The FCN was the first to apply image classification models to semantic segmentation in natural images. The end-to-end FCN effectively addresses the inefficiency and limited feature utilization issues commonly found in traditional methods.

Building upon the foundation of the FCN, several prominent encoder–decoder semantic segmentation models have been developed. The U-Net [25] introduced skip connections between the encoder and decoder, which significantly improved the accuracy of medical image segmentation. SegNet, proposed by Badrinarayanan et al. [26], employed indexing in the encoder and utilized it for efficient upsampling in the decoder, reducing the number of trainable parameters. The Deeplab series of networks, pioneered by Chen et al. [27], [28], [29], have made remarkable contributions to deep-learning-based semantic segmentation. DeepLab_v1 [27] introduced dilated convolutions into the network to expand the receptive field, enabling the preservation of multiscale feature information. DeepLab_v2 [28] further advanced the model by introducing the atrous spatial pyramid pooling (ASPP), which incorporated multiple parallel branches of dilated convolutions to extract multiscale features and effectively address the challenge of small-scale object filtering by deep networks. DeepLab_v3 [29] built upon the ASPP by incorporating a global pooling layer, allowing for better integration of both local and global information.

There are also some state-of-the-art methods that have been proposed. Adversarial learning [30] can enhance the accuracy of 3-D semantic segmentation by reducing the domain gap between the source and target domains. Gao et al. [31] combines adversarial complementary learning (ACL) with CNN to obtain complementary information for multisource data classification. Qiu et al. [32] introduce the adversarial semantic guidance network (ASGN) for image segmentation, addressing the issue of pixel distribution similarity. Furthermore, the introduction of cross-scale mixing attention [33] aims to handle long-term dependencies among large-scale features in multispectral data. There is also a research [34] that utilizes a cross-scale mixed attention network to segment smoke, which is fundamentally inspired by attention methods. Depthwise separable convolution networks are also applicable to image segmentation. Huang et al. [35] propose an end-to-end depthwise separable U-shaped convolution network for medical image segmentation.

However, the methods mentioned earlier frequently disregard the challenge of missing details in semantic segmentation, making it difficult to seamlessly integrate both global and local features. Consequently, these approaches may struggle to accurately depict the intricate attributes of objects within remote sensing images, ultimately leading to potential shortcomings in segmentation precision and boundary delineation.

While numerous efforts have been dedicated to leveraging graph neural networks to tackle this issue, their effectiveness is significantly contingent on their ability to learn long-term dependencies, which may not always yield optimal results in the context of semantic segmentation with remote sensing imagery.

B. Attention Mechanisms

Attention mechanisms, as effective signal processing mechanisms, have been widely employed in deep learning [36], [37]. They can rapidly filter out the relevant information from a large volume of visual signals, making them highly applicable in various deep learning tasks. In the context of semantic segmentation of remote sensing images, attention models play a crucial role in enhancing the segmentation capability [38]. By assigning different weights to the scale features of pixels in different positions, attention models focus on important features, effectively improving the performance of semantic segmentation in remote sensing imagery.

Hu et al. [39] introduced the squeeze-and-excitation networks (SENet), which incorporated attention mechanisms into the segmentation network for natural images, effectively improving both the efficiency and accuracy of the models. Building upon SENet, Ding et al. [40] proposed a lightweight network architecture that concentrated on the useful features using the squeeze-and-excitation (SE) block, thereby reducing network overhead. Liang et al. [41] integrated the SE block into a dual-encoder model and achieved excellent segmentation results in medical imaging. However, SE attention, which computes channel attention using global average pooling, may lead to the loss of object localization in remote sensing imagery. To address this issue, researchers have explored various methods to incorporate attention models into neural networks. Fu et al. [42] proposed the dual attention network, which employed two attention modules to enhance the connections between features in an FCN backbone. Huang et al. [43] introduced the criss-cross network, which captured full-image dependencies by acquiring contextual information for all pixels in the image.

In addition, other studies have also incorporated attention mechanisms to enhance segmentation results by focusing on global features and suppressing local features and image noise [44], [45].

While these methods leverage attention mechanisms to incorporate spatial and channel relationships in the imagery, they often fail to deeply explore the intricate connections between relation modules and the size, depth, abstraction level, and receptive field of feature maps. As a result, there may be missed opportunities to fully exploit the rich contextual information and capture long-range dependencies, potentially limiting the overall performance of the semantic segmentation models.

III. Multilateral Semantic With Dual Relation Network (MSDRNET)

In this work, we present an MSDRNet for remote sensing images segmentation, which is illustrated in Fig. 1. In the following

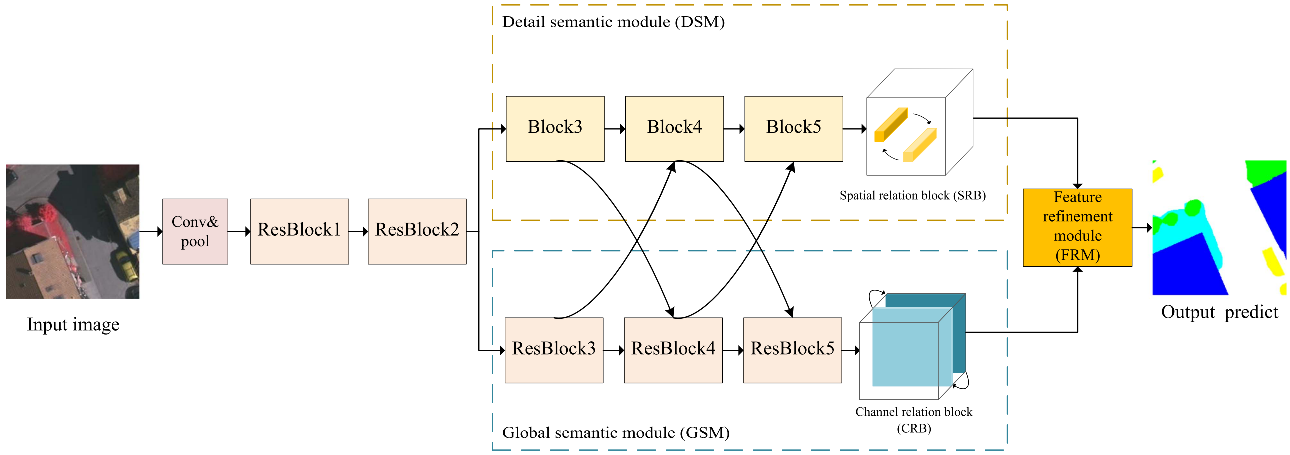


Fig. 1. MSDRNet architecture. The structure primarily consists of two parallel modules, the DSM contains three non-downsampled blocks and an SRB for extracting detailed features, and the GSM includes three residual blocks and a CRB for extracting global features. The FRM is introduced to refine and integrate the feature maps generated by the DSM and the GSM.

sections, we will describe four significant parts of MSDRNet, including MSDRNet backbone, relation blocks, and FRM.

A. MSDRNet Backbone

The backbone of MSDRNet is built upon the ResNet50 architecture, which is known for its multiscale design. ResNet50 consists of 50 residual units, and its skip connections effectively address the problem of network degradation, reducing errors in semantic segmentation models. The ResNet50 used in this study comprises five modules, allowing for effective separation of image features at different levels and extraction of essential information for semantic segmentation. The decision to use ResNet50 instead of ResNet101 or ResNet152 was driven by the aim to achieve competitive segmentation results with a simpler network. In training process, we use a pretraining strategy based on transformation learning, which can reduce overfitting to some extent. Since the output is a feature map, only the convolution operation is retained during pretraining, and the fully connected layer and downsampling operations are removed.

For the input remote sensing images of MSDRNet, we initially utilize a 7×7 convolutional layer and a max-pooling layer to generate a feature map layer that is 1/4 the size of the input image. Subsequently, we pass this layer through two residual blocks, resulting in a feature map layer that is 1/8 the size of the input image. Following that, the network structure branches out into two distinct modules: the DSM and the GSM.

The DSM aims at extracting spatial detail information in low-level features. Remote sensing images have a large amount of spatial detail information due to the complex distribution of features and different scale sizes. As the convolution depth increases with the introduction of residuals, a lot of detailed information is lost, such as edges between features and small-scale targets. Therefore, we need to design a shallow structured module that can retain rich detailed information. The overall design principle of DSM is to use wide channel sizes and shallow layers, i.e., larger branch widths and smaller branch depths, to handle spatial details. In order to preserve more spatial details, pooling operations with residual connections are not used in the module. The feature map after the DSM extraction contains rich spatial detail information, and at the same time, contains a lot of spatial relationship context information due to less image size reduction. Therefore, the SRB is added to the DSM extraction to enhance the detail extraction effect.

The SRB updates the features of each channel by weighting the sum of the features of each channel, where the weights are defined by the similarity of the features at the corresponding positions. Specifically, the entire DSM consists of three convolutional blocks and the SRB. The first block comprises two sets of 3×3 convolutional layers, followed by batch normalization and activation functions. The following two blocks consist of three sets of 3×3 convolutional layers, also followed by batch normalization and activation functions. Since there is no downsampling pooling layer, the output feature map remains at 1/8 the size of the original input after passing through these three blocks. Finally, the feature map undergoes the SRB to obtain the detail semantic feature map layer.

We use the GSM for extracting the deep global segmentation information in the RSIs. The distribution of features in RSIs has certain continuity and regularity, and the extraction of deep semantic features helps to obtain advanced features by integrating global information on a higher perception field. Due to the small size and large number of features of the extracted high-level semantic information, it contains a large amount of channel information. Therefore, the CRB is added to the GSM extraction to enhance the global information extraction effect. The CRB updates the parameters by weighting the features of each layer of channels with the features of other layers. This design enables effective integration of a large number of feature parameters in the global semantics. The GSM consists of three res-blocks and the CRB. Resblock3 and Resblock4 correspond to the third and fourth residual blocks in ResNet, resulting in output feature map sizes of 1/16 and 1/32 of the original input image, respectively. Resblock5 shares the same structure and number of convolutional layers as Resblock4, but it doesn't further reduce the feature map size. Instead, it compresses the

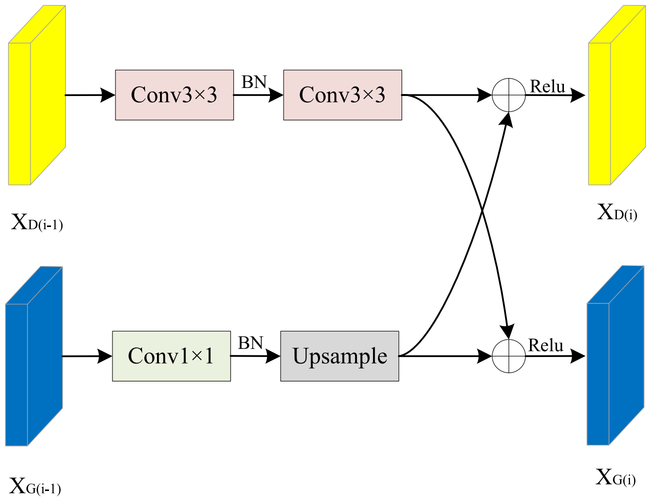


Fig. 2. Cross-fusion operation. The yellow rectangle X_D represents detail semantic feature, and the blue rectangle X_G represent global semantic features. ⊕ denotes matrix addition.

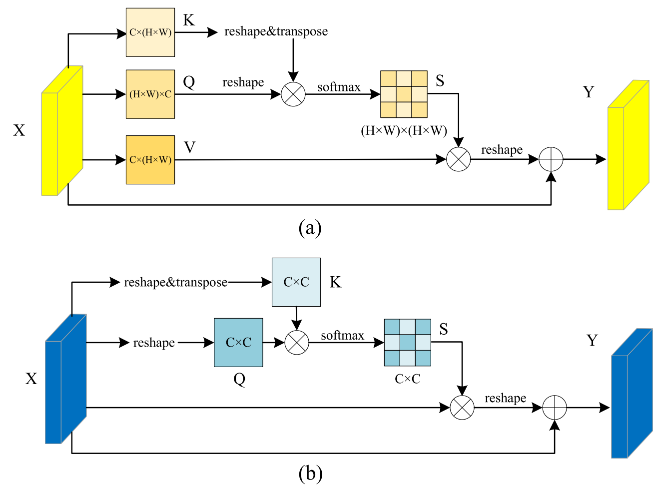


Fig. 3. Relation blocks. (a) Spatial relation block (SRB). (b) Channel relation block (CRB). Here, ⊕ and ⊗ represent matrix addition and matrix multiplication, respectively.

feature dimensions, yielding a feature layer that is 1/4 of the size of the previous block. Finally, after passing through the CRB, deep-level and multifeature high-level semantic information is obtained.

Simultaneously, to enhance the connection between the DSM and the GSM, we introduce cross-fusion operations in two parallel network modules. As shown in Fig. 2, the detail feature map $X_{D(i\text{-}1)}$ undergoes downsampling with 3×3 convolutions using a stride of 2. This downsampling adjusts the feature map size and feature dimensions to match those of the global feature map $X_{G(i\text{-}1)}$ and produces the next layer of global feature map $X_{G(i)}$. On the other hand, the global feature map $X_{G(i\text{-}1)}$ undergoes 1×1 convolutions to modify the feature dimensions and perform upsampling to adapt to the detail feature map $X_{D(i\text{-}1)}$. Through fusion, the next layer of detail feature map $X_{D(i)}$ is generated. The downsampling frequency of $X_{D(i\text{-}1)}$ and the upsampling factor of $X_{G(i\text{-}1)}$ are determined by the corresponding feature map layer sizes.

B. Relation Blocks

In DSM, we introduced SRB with the aim of preserving rich detail information. In addition, for global features within the images, we enhance the effectiveness of global information extraction in the GSM by introducing CRB. The CRB update parameters by weighting the features of each channel layer with features from other layers. This design effectively integrates a vast number of feature parameters in global semantics, resulting in deep, multifeature, and high-level semantic information.

1) Spatial Relation Block (SRB): The SRB updates the target pixel parameters by considering the feature similarity between the target pixel and its weighted sum of the neighboring pixels. In the remote sensing images, the surface objects are distributed in complex and different sizes, in which there are a large number of spatial relations. Therefore, the use of spatial relation for remote sensing images is effective for reducing intraclass errors, enhancing interclass differences, and identifying small targets on the surface.

As shown in Fig. 3(a), the input to the SRB is the feature map X $\in \mathbb{R}^{C\times H\times W}$, where C represents the number of channel layers, and H and W represent the height and width of the feature map, respectively. We first perform the convolution operation on the feature map X and accomplish reshape and transpose operations on the result to get the feature layer Q. We perform the convolution operation on the feature map X, and then, perform reshape operation only to obtain the feature map K. Q and K characteristic sizes are $\mathbb{R}^{C\times N}$, where $N = H \times W$ is the number of pixels.

Subsequently, matrix multiplication is performed on Q and K, and a softmax layer for computing the spatial relation map S $\in \mathbb{R}^{N\times N}$

$$s_{ij} = \frac{\exp\left(Q_i \cdot K_j\right)}{\sum_{j=1}^{N} \exp(Q_i \cdot K_j)} \tag{1}$$

where s_{ij} measures the jth pixel's impact on the ith pixel. We then perform a convolution operation on the feature map X to obtain a new feature map V and reshape it to the same shape as K and Q. Afterwards, S is subjected to reshape and transpose operations and the result is matrix multiplied with V to obtain a fusion feature map. Finally, we use the scaling parameter α to multiply with the fusion feature map and perform the element-wise addition of the obtained features with to obtain the final result $Y \in \mathbb{R}^{C\times H\times W}$ as follows:

$$Y_i = \alpha \sum_{j=1}^{N} (s_{ij} V_j) + X_i. \tag{2}$$

According to (2), the new feature maps at different positions are obtained by applying multilayer convolutions to the original feature X. These new feature maps are then merged with the original feature maps using weighted fusion, resulting in the output feature maps Y at each layer. The parameter α, initially set to 0, allows the weights to be learned and updated continuously.

This process effectively aggregates contextual information from the relationship maps that is beneficial for achieving accurate segmentation. By sharing the acquired information among pixels, semantic consistency is enhanced, reducing the likelihood of misclassifications between objects.

2) Channel Relation Block (CRB): The CRB integrates the information of individual channels in the high-level features into integrated features, and updates the individual channel parameters by the weighted sum of the features of each channel. It can be used to enhance the feature discrimination in the channel domain and improve the feature representation for specific semantics. With the high-dimensionality features and large channel differences of remote sensing images, the interdependencies between channels can be explicitly modeled using the CRB. It also distribution but differing in channel dimensions.

As shown in Fig. 3(b), the input to the CRB is the feature map $X \in \mathbb{R}^{C\times H\times W}$, we first perform the reshape operation on the feature map X to obtain the feature $Q \in \mathbb{R}^{C\times N}$. At the same time, transpose and reshape the feature map X on another path to obtain the feature $K \in \mathbb{R}^{C\times N}$. We then use matrix multiplication to merge features Q and K, and use the softmax layer to compute the channel relation map $E \in \mathbb{R}^{C\times C}$ as

$$e_{ij} = \frac{\exp(K_i \cdot Q_j)}{\sum_{j=1}^{C} \exp(K_i \cdot Q_j)} \tag{3}$$

where e_{ij} is to calculate the degree of influence of the jth channel on the ith channel. The result $\mathbb{R}^{C\times H\times W}$ is obtained by transposing E and after matrix multiplication with the initial feature X. Similar to α of spatial relation, we propose a scale parameter β for the weighting operation. And the obtained features are element-wise added with X to obtain the final result $Y \in \mathbb{R}^{C\times H\times W}$ as follows:

$$Y_i = \beta \sum_{j=1}^{C} (e_{ij} X_j) + X_i \tag{4}$$

where β is gradually updated with weight values starting from 0. As shown in (4), the final semantic feature layer is obtained by weighting and summing the features of each channel with the original features.

C. Feature Refinement Module (FRM)

The introduction of the FRM aims to refine and integrate the feature maps generated by the DSM and the GSM. Due to the different feature extraction methods, making a large semantic difference between the two feature maps. Thus, a module that can combine uniform feature size and reduce semantic differences is needed to merge detailed and global semantic features. The FRM leverages the contextual information from the integration feature maps to guide the feature responses of both the DSM and the GSM. By introducing residual connections, it effectively considers features at different scales, allowing for the encoding of multiscale information.

As shown in Fig. 4, after the extraction from the DSM and the GSM, we obtain the detail semantic feature and the global semantic feature, respectively. To integrate these two feature sets, the global semantic feature is first upsampled by a factor of 4 to match the feature size of the detail semantic feature. Then, the two features are combined through element-wise summation, resulting in integration features. Subsequently, the integration features undergo convolutional operations and are element-wise multiplied with the detail semantic feature and the global semantic feature obtained after the ASPP operation.

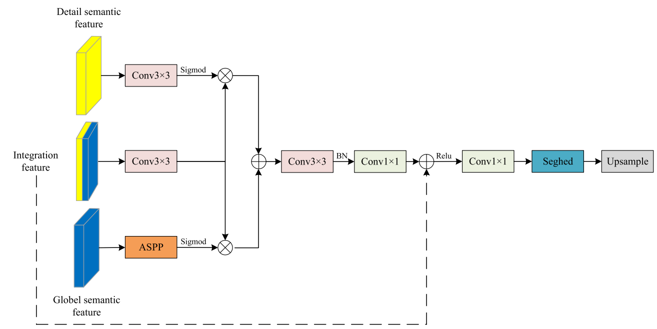


Fig. 4. FRM. The yellow rectangle represents detail semantic feature, and the blue rectangle represents global semantic feature. The mixed color rectangle represents the integration feature of the aforementioned two features. Here, $\oplus$ and $\otimes$ represent matrix addition and matrix multiplication, respectively.

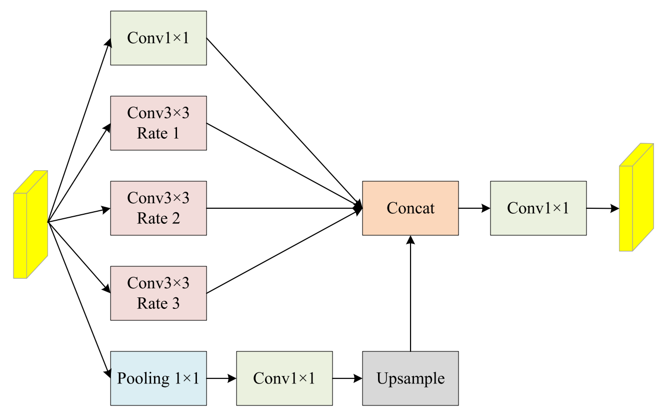


Fig. 5. Illustration of the ASPP.

Finally, the resulting features are further integrated for subsequent processing. The aforementioned ASPP was proposed in [28] and has shown excellent segmentation capabilities in practical studies. The ASPP combines a spatial pyramid pooling module and separable convolutional layers to extract features at different scales, where multiple parallel convolutional layers can effectively expand the range of the received domain and ensure more effective information output. The ASPP uses different dilation rates for different scales to capture multiscale information, and each scale is a separate branch, which effectively avoids the acquisition of redundant information and focuses directly on the correlation between objects. Thus, the ASPP can effectively reduce the information redundancy of global semantic features and enhance the spatial multiscale characteristics and feature diversity of the feature map, while increasing the number of feature layers allows the network to encode rich spatial details, as shown in Fig. 5 . The ASPP consists of a 1×1 convolution, a pooling pyramid with three different expansion factors, and ASPP, so its output feature map combines the information of the aforementioned five layers.

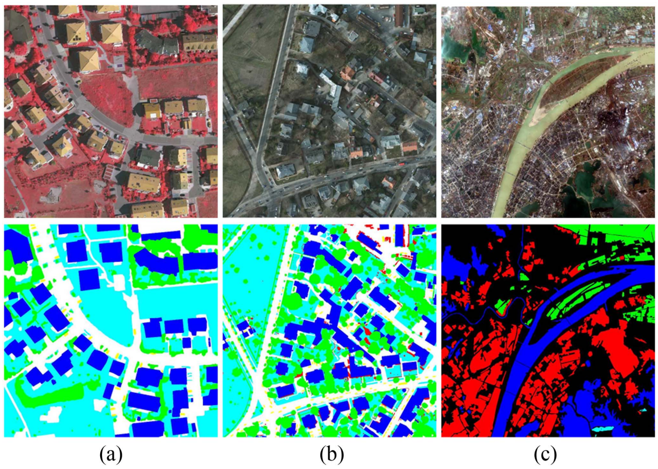


Fig. 6. Datasets display. (a) ISPRS Vaihingen challenge dataset images and labels. (b) ISPRS Potsdam challenge dataset images and labels. (c) GID images and labels.

Finally, the reintegrated features undergo a sequence of convolutional operations and are combined with the initially integrated features through summation. The resulting feature maps are then subjected to further convolution and upsampling operations, leading to the final segmentation results.

IV. Results

In this section, we first introduced the ISPRS 2-D Semantic Labeling Challenge datasets for Vaihingen and Potsdam. We further experimented with the Gaofen image dataset (GID), which contains information on larger scale features, to demonstrate the validity of our model the above three datasets are displayed as shown in Fig. 6. Then, we delve into the specific implementation details of MSDRNet, along with a comprehensive description of the evaluation metrics and comparison methods employed. Finally, we evaluate the effectiveness of the proposed network architecture in semantic segmentation by utilizing these three cutting-edge remote sensing image pairs as benchmarks.

A. Datasets

1) The ISPRS Vaihingen challenge dataset comprises 33 aerial images in the IRRG (infrared, red, green) format. Each image has a spatial resolution of 9 cm and dimensions of 2494 $\times$ 2064 pixels. Among these images, 17 are primarily allocated for online testing, while the remaining 16 annotated images are designated for evaluating our proposed method. Specifically, 11 images are randomly assigned for training data, 3 images for testing data, and 2 images for validation data. It is worth noting that the digital surface model and normalized digital surface model data corresponding to each image in the dataset were not utilized in our experimental analysis.
2) The ISPRS Potsdam challenge dataset comprises 38 aerial images in the IRRGB (infrared, red, green, blue) format. Each image has a spatial resolution of 5 cm and dimensions of 6000 $\times$ 6000 pixels. This dataset offers a larger coverage area compared to the Vaihingen dataset and provides higher resolution. However, the labels remain consistent with six predefined categories. Random partitioning of the dataset resulted in 24 training images, 8 testing images, and 6 validation images. Similar to the Vaihingen dataset, the digital surface model and normalized digital surface model data associated with each image were not employed in our experiments.
3) The GID is a large-scale land-cover dataset with Gaofen-2 (GF-2) satellite images. GID consists of two parts: a large-scale classification set and a fine land-cover classification set. The large-scale classification set contains 150 pixel-level annotated GF-2 images, and the fine classification set is composed of 30 000 multiscale image patches coupled with 10 pixel-level annotated GF-2 images. Note that we have only used large-scale classification set, which involves five classes, to validate the performance of our model for segmentation of large scale RSIs. The large-scale classification set was randomly divided into a training set (100 images), test set (30 images), and a validation set (20 images).

B. Implementation Details

We employed three methods for image augmentation: random scaling, random cropping, and random horizontal flipping. The input image size was set to 512 $\times$ 512. In the first ten epochs, we utilized the warm-up policy [23] to smoothly adjust the learning rate from 0.001 to 0.01, which helps maintain model stability at a deeper level. Afterwards, we implemented a polynomial learning rate policy that reduces the learning rate by a factor of 0.1 every 20 epochs. The optimizer used was AdamW, with an initial learning rate of 0.01. Training was conducted for 110 epochs on an NVIDIA 2070 GPU, with a batch size of 8 and 6 threads.

C. Evaluation Metrics

To assess the quantitative performance, three benchmark metrics were used: overall accuracy (OA) derived from the pixel-based confusion matrix, F1 score (F1), and intersection over union (IoU). OA is defined as

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{N} \tag{5}$$

where TP and TN are the number of true positives and true negatives, respectively, and N is the total number of pixels. F1 is defined as

$$\mathrm{F1} = 2\frac{\mathrm{Pre} \times \mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}} \tag{6}$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{7}$$

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{8}$$

in which FP and FN are the number of false positives and false negatives, respectively. IoU is defined as

$$\mathrm{IoU}\left(\mathcal{P}_m, \mathcal{P}_{\mathrm{gt}}\right) = \frac{\left|\mathcal{P}_m \cap \mathcal{P}_{\mathrm{gt}}\right|}{\left|\mathcal{P}_m \cup \mathcal{P}_{\mathrm{gt}}\right|} \tag{9}$$

TABLE I
ABLATION STUDY FOR ASPP PARAMETERS ON THE ISPRS VAIHINGEN TEST DATASET

Atrous rates	None	(2,4,6)	(3,6,9)	(4,8,12)
OA (%)	88.16	89.58	**90.35**	89.64
mIoU (%)	77.98	80.05	**80.21**	80.19

The best values are indicating in bold.

where $\mathcal{P}_{gt}$ is the set of ground-truth pixels, and P_m is the set of prediction pixels; "$\cap$" and "$\cup$" denote the intersection and union operations, respectively.

D. Comparing Methods

To verify the performance, the proposed MSDRNet is compared with six state-of-the-art deep models on the two datasets above. The main information regarding these models is summarized as follows.

1) *FCN [8]:* Shelhamer et al. [8] propose the FCN for semantic segmentation, which is the first pixel-level segmentation network. There are three versions of FCN models: FCN-32s, FCN-16s, and FCN-8s. We use the best performance model FCN-8s as comparison.
2) *U-Net [25]:* Ronneberger et al. [25] propose U-net for biomedical image segmentation. The main structure of U-Net consists of two parts, upsampling and downsampling parts, and incorporates a jump connection operation between the encoder and decoder.
3) *SegNet [26]:* SegNet is characterized by an FCN architecture and follows an encoder–decoder paradigm. It consists of two parts, an encoder and decoder, and is capable of segmenting the region where the object is located in the image at the pixel level.
4) *PSPNet [46]:* Zhao et al. [46] propose PSPNet for pixel-level prediction tasks, which proposes a pyramid pooling module that integrates global contextual information, and this global priori information is effective in obtaining high quality results in scene semantic analysis.
5) *DeepLabv3+ [47]:* DeepLabv3+ is proposed by Chen et al. [47] for semantic segmentation, which proposes an improved Xception as a backbone network and uses an ASPP structure to solve the multiscale problem.
6) *DDCM-Net [48]:* The dense dilated convolutions' merging network (DDCM-Net) consists of densely dilated image convolutions merged at different dilation rates, expanding the sensory field of the network with fewer parameters and features.

E. Ablation Experiments

1) ASPP Parameters: In the proposed FRM, we have introduced the ASPP operation. To verify the effect of different settings of atrous rate in ASPP on the segmentation results, we designed a set of ablation experiments on Vaihingen dataset presented in Table I.

TABLE II
ABLATION STUDY FOR RELATION BLOCKS AND FRM ON THE ISPRS VAIHINGEN TEST DATASET

Method	CRB	SRB	FRM	OA (%)	mIoU (%)
MSDRNet				85.76	66.44
MSDRNet	√			88.62	76.38
MSDRNet		√		86.18	71.55
MSDRNet			√	87.46	73.63
MSDRNet	√	√		89.24	78.61
MSDRNet	√	√	√	**90.35**	**80.21**

CRB represents the channel relation block, and SRB represents the spatial relation block.
The best values are indicating in bold.

It can be seen that when the atrous rate is (3, 6, 9), OA and mean IoU (mIoU) have the highest values of 90.28% and 81.12% respectively.

2) Relation Blocks and FRM: In the proposed MSDRNet, two relation blocks SRB and CRB are employed in DSM and GSM, respectively. The FRM aims to refine and integrate the feature maps generated by the DSM and the GSM.

To further verify the performance of these blocks and module, we conduct extensive experiments with different settings on Vaihingen dataset in Table II. As shown in Table II, the proposed relation blocks and FRM bring a significant improvement compared with the baseline MSDRNet.

The OA and mIOU of the network with all the relation blocks and FRM added are 90.35% and 80.21%, respectively, which are 4.59% and 3.77% higher than the baseline structure. Meanwhile, we found that CRB has the most significant effect on accuracy improvement among these blocks and module. In the case of adding only CRB and no other blocks OA and mIOU are improved by 2.86% and 9.94%, respectively, over the baseline results.

F. Comparison With Deep Models

To perform a comprehensive evaluation of the proposed MSDRNet, we compare it against six existing methods: FCN, U-Net, SegNet, PSPNet, Deeplabv3+, and DDCM-Net. We employ three accuracy assessment metrics: F1 score, IoU, and OA.

1) ISPRS Vaihingen Challenge Dataset: The comparison results on the ISPRS Vaihingen dataset are shown in Table III. Our proposed MSDRNet demonstrates an average IoU of 79.21%, surpassing the other six comparative methods. Specifically, it achieves a 3.07% improvement over the highest average IoU achieved by U-Net. Furthermore, MSDRNet exhibits an average F1 score of 88.32%, outperforming the other five methods. In particular, it shows a 1.89% enhancement compared to the highest average F1 score obtained by Deeplabv3+.

The OA value of MSDRNet is 90.35%, which is 0.66% higher than U-Net's 89.69%. This result indicates the effectiveness of the connection between the detail and GSMs in MSDRNet. In terms of vehicle segmentation, MSDRNet exhibits an IoU

TABLE III
QUANTITATIVE COMPARISON (%) WITH THE STATE-OF-THE-ART DEEP MODELS ON ISPRS VAIHINGEN CHALLENGE VALIDATION SET

Model	Imp surf		Building		Low veg		Tree		Car		Avg.		OA
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	MIoU	MF1	
FCN	82.13	90.09	87.59	88.47	56.82	76.27	73.47	84.29	65.52	78.41	73.11	83.62	87.50
U-Net	84.37	91.5	90.95	90.44	62.18	80.76	76.96	86.69	66.21	78.94	76.14	85.78	89.69
SegNet	79.74	90.55	90.21	92.02	56.47	77.96	74.34	86.9	48.71	65.83	69.90	82.85	87.45
PSPNet	81.18	89.61	87.90	93.56	57.28	72.84	72.29	83.92	67.05	80.28	73.14	84.16	88.08
Deeplabv3+	83.15	92.74	90.01	91.9	57.77	79.09	74.50	87.01	66.58	81.22	74.40	86.43	88.49
DDCM-Net	78.85	88.18	85.19	92.01	53.88	70.03	69.73	82.17	60.47	75.36	69.63	81.62	86.31
MSDRNet	**85.95**	**93.28**	**91.26**	**94.88**	**75.95**	**86.33**	**78.19**	**88.52**	**69.71**	**82.57**	**80.21**	**89.12**	**90.35**

The best values are indicating in bold.

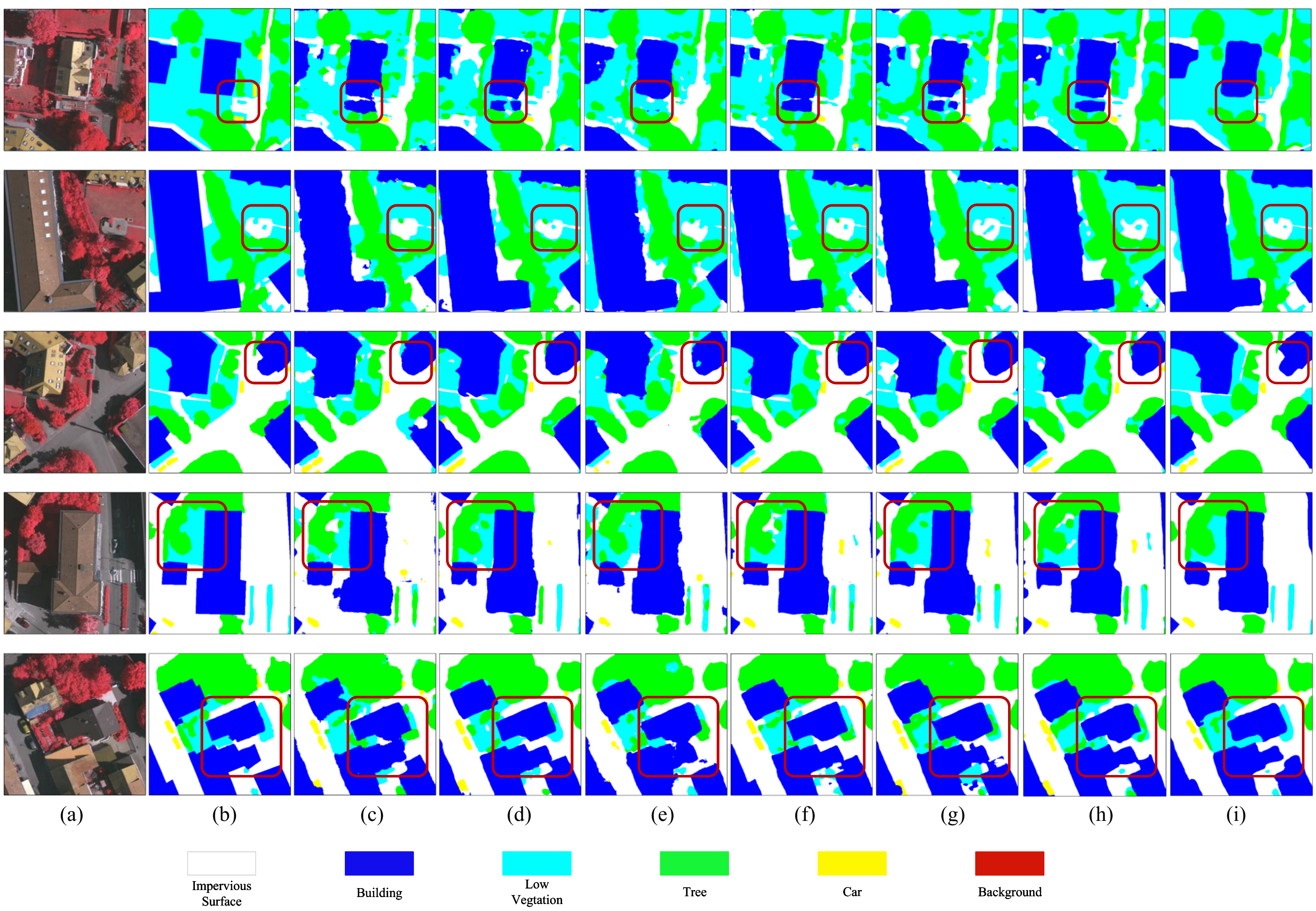


Fig. 7. Qualitative comparison with the state-of-the-art deep models on ISPRS Vaihingen challenge dataset. The label includes six categories: impervious surface (imp surf, white), building (blue), low vegetation (low veg, cyan), tree (green), car (yellow), and clutter/background (red). (a) input image, (b) ground truth, (c) FCN results, (d) U-Net results, (e) SegNet results, (f) PSPNet results, (g) Deeplabv3+ results, (h) DDCMNet results, and (i) MSDRNet results.

that is 2.34% lower and an F1 score that is 1.71% lower than PSPNet. The reason for this outcome is that PSPNet focuses on addressing the problem of predicting small objects by enhancing the representation capability of information at different scales. However, PSPNet is primarily designed for scene parsing applications, which is why its performance in segmenting impervious surface, for example, is inferior to that of MSDRNet. MSDRNet also demonstrates excellent segmentation performance in building segmentation, achieving IoU of 91.26% and F1 score of 94.88%. Particularly noteworthy is its exceptional performance in shrubbery segmentation, where MSDRNet outperforms other models with a remarkable increase of 13.77% in IoU and 5.57% in F1 score. Moreover, the MIoU, MF1, and OA values of MSDRNet are 10.58%, 7.50%, and 4.04% higher than those of the DDCM-Net, respectively.

The qualitative comparison experimental results on the Vaihingen dataset are depicted in Fig. 7. The results clearly indicate that our proposed MSDRNet achieves segmentation results

TABLE IV
QUANTITATIVE COMPARISON (%) WITH THE STATE-OF-THE-ART DEEP MODELS ON ISPRS POSTDAM CHALLENGE VALIDATION SET

Model	Imp surf		Building		Low veg		Tree		Car		Avg.		OA
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	MIoU	MF1	
FCN	77.05	87.04	88.29	93.78	79.46	88.56	77.96	87.61	76.75	86.84	79.90	88.77	88.06
U-Net	82.97	90.69	89.82	94.64	81.05	89.53	79.06	88.31	82.11	90.17	83.0	90.67	89.41
SegNet	71.19	83.17	82.77	90.58	75.01	85.72	71.83	83.61	70.48	82.68	74.26	85.15	84.49
PSPNet	74.42	85.33	87.61	93.4	76.96	86.98	77.19	87.13	76.15	86.46	78.45	87.86	87.03
Deeplabv3+	81.57	89.85	89.82	94.64	**81.56**	**89.85**	80.09	88.94	68.50	81.83	80.31	88.92	89.66
DDCM-Net	65.32	79.02	90.17	95.49	65.65	79.27	63.29	77.52	66.85	80.13	70.49	82.28	82.71
MSDRNet	**83.43**	**90.72**	**90.25**	**95.51**	81.17	89.4	**80.30**	**89.89**	**83.30**	**92.64**	**83.69**	**91.23**	**90.57**

The best values are indicating in bold.

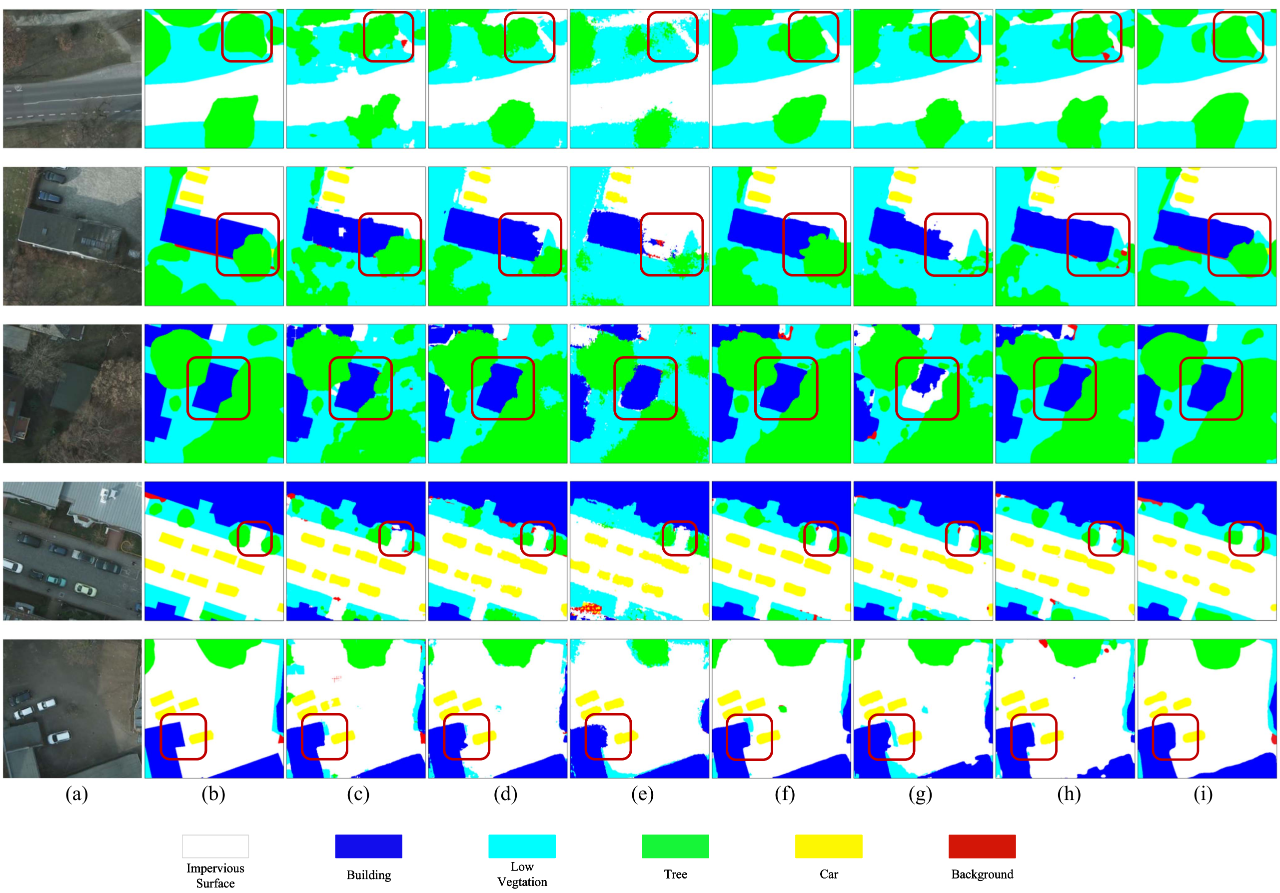


Fig. 8. Qualitative comparison with the state-of-the-art deep models on ISPRS Potsdam challenge dataset. The label includes six categories: impervious surface (imp surf, white), building (blue), low vegetation (low veg, cyan), tree (green), car (yellow), and clutter/background (red). (a) input image, (b) ground truth, (c) FCN results, (d) U-Net results, (e) SegNet results, (f) PSPNet results, (g) Deeplabv3+ results, (h) DDCMNet results, and (i) MSDRNet results.

that closely resemble the ground truth. In comparison to other segmentation models, MSDRNet excels in preserving complete segmentation boundaries and accurately identifying impervious surfaces (as shown in the second row). On the other hand, FCN, SegNet, and Deeplabv3+ are more prone to being affected by shadows, resulting in fragmented object boundaries. When it comes to building segmentation, U-Net and SegNet exhibit holes in their segmented buildings, requiring further refinement. In contrast, MSDRNet demonstrates no significant shortcomings in segmenting various land features and effectively captures intricate details, such as building corners.

2) ISPRS Postdam Challenge Dataset: Table IV presents the numerical results obtained on the Postdam dataset. The results indicate that MSDRNet achieves notable performance with average IoU of 83.69%, average F1 score of 91.23%, and OA of 90.57%. Due to the larger dataset size and the block-like distribution of labels in the Postdam dataset, segmentation accuracy tends to be higher compared to the Vaihingen dataset. Notably,

TABLE V
QUANTITATIVE COMPARISON (%) WITH THE STATE-OF-THE-ART DEEP MODELS ON GID VALIDATION SET

Model	Water body		Residential area		Woodland		Plow land		Background		Avg.		OA
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	MIoU	MF1	
FCN	64.38	78.33	91.79	95.72	37.14	54.17	48.38	65.21	75.11	85.78	63.36	77.19	84.24
U-Net	62.05	76.59	**92.33**	**96.01**	53.48	69.69	57.71	73.19	77.94	87.61	68.71	81.09	85.80
SegNet	63.52	77.69	90.55	95.04	39.36	58.72	46.44	63.43	75.48	86.03	63.07	76.18	84.26
PSPNet	69.47	81.98	87.61	93.4	45.06	62.13	56.15	71.91	78.93	88.23	68.24	80.62	87.05
Deeplabv3+	58.49	73.81	88.62	93.97	39.04	57.28	48.57	65.67	71.77	83.57	61.30	74.86	82.49
DDCM-Net	59.98	74.99	91.49	95.56	51.06	67.6	56.19	71.97	75.31	85.92	66.81	79.94	84.19
MSDRNet	**72.18**	**86.68**	89.58	94.50	**65.04**	**78.82**	**60.54**	**77.14**	**80.66**	**92.81**	**73.60**	**85.99**	**89.12**

The best values are indicating in bold.

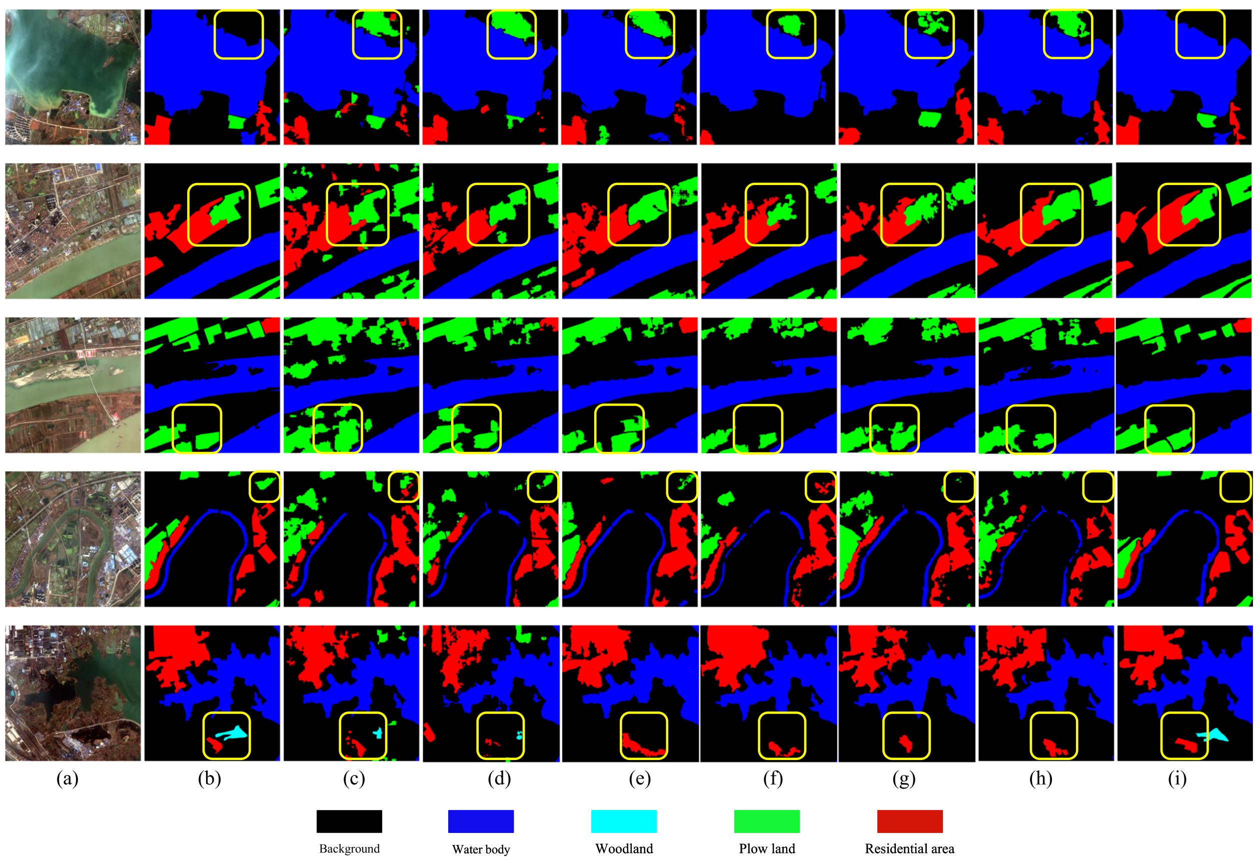


Fig. 9. Qualitative comparison with the state-of-the-art deep models on GID. The label includes five categories: background (black), water body (blue), wood land (cyan), plow land (green), and residential area (red). (a) input image, (b) ground truth, (c) FCN results, (d) U-Net results, (e) SegNet results, (f) PSPNet results, (g) Deeplabv3+ results, (h) DDCMNet results, and (i) MSDRNet results.

MSDRNet demonstrates significant improvements in vehicle segmentation accuracy on the Postdam dataset, exhibiting an impressive 18.59% increase in IoU and a remarkable 14.07% increase in F1 score. MSDRNet achieves an OA that is 7.86% higher than DDCM-Net. In terms of segmentation performance in the impermeable layer, MSDRNet achieves an IoU of 83.43% and an F1 score of 90.72%, both of which significantly surpass the results obtained by DDCMNet. This improvement can be attributed to the larger number of vehicle samples available in the Postdam dataset. Furthermore, in terms of building segmentation, MSDRNet achieves IoU of 90.25% and F1 score of 95.51%. Compared to U-Net, MSDRNet shows an improvement of 0.43% in IoU and 0.87% in F1 score. While MSDRNet's segmentation performance in shrubbery is slightly inferior to that of Deeplabv3+, it demonstrates relatively accurate predictions across all four land cover categories.

Fig. 8 displays the visualized segmentation results on the Postdam dataset. It is evident that U-Net and SegNet struggle with handling segmentation boundaries, resulting in inconsistent edges between different land cover categories. The FCN performs well in segmenting small-scale vehicles, but tends to produce holes in large buildings and continuous shrubbery areas, leading to incomplete segmentation. Conversely, Deeplabv3+ exhibits poor sensitivity to buildings, resulting in noticeable misclassification and omissions in building segmentation. Compared with other methods, our proposed MSDRNet can accurately distinguish the building area sandwiched between shrubs and trees (as shown in the third row), and the segmentation between the building boundary and impervious surface is more complete (as shown in the fifth row). These experimental results show that MSDRNet fully takes into account the local and global semantic information of the features, and effectively avoids the easy misclassification and omission in the segmentation process.

3) Gaofen Image Dataset (GID): Table V shows the accuracy evaluation results of the GID. The results show that the average IoU, average F1, and OA of MSDRNet are 73.60%, 85.99%, and 89.12%, respectively. Since the spatial scale of the GID dataset is more macroscopic and the intraclass attributes are more different, the segmentation accuracy of the GID dataset is overall worse compared to the previous two datasets. For the segmentation of water bodies, the IoU and F1 of MSDRNet are 72.18% and 86.68%, respectively, which is an improvement of 2.71% in IoU and 4.7% in F1 compared to PSPNet. Since residential areas have the most samples in training, each model achieves the highest accuracy here, and although our proposed MSDRNet has a slightly lower segmentation performance than U-Net for residential areas, it makes relatively accurate predictions for segmentation of all four of its landcover classes. On the contrary, Plow land and Plow land have relatively fewer training samples, and hence, poorer validation accuracy.

The visualization results of GID segmentation are shown in Fig. 9. It can be seen that FCN and U-Net have the worst overall segmentation effect, and some of the features are misclassified and fragmented during the segmentation process. SegNet and PSPNet have better overall morphological segmentation of features, but there are still fragmentation and discontinuity in spatial details, which leads to incomplete segmentation. Compared with other methods, our proposed MSDRNet can effectively reduce the omission and mis-segmentation of features (shown in the first and fourth rows), and the segmentation between feature boundaries is more complete, without a large number of broken features (shown in the second and third rows). These experimental results show that MSDRNet fully considers the local and global semantic information of features and effectively enhances the completeness of feature segmentation.

V. Conclusion

In this article, we present an end-to-end MSDRNet for semantic segmentation in remote sensing images. MSDRNet is designed to effectively capture both local details and global features in the image, enabling accurate segmentation results. Our proposed method addresses this challenge through the following three main aspects.:

1) A network consisting of two parallel modules was adopted in this article, which fully considered the details and global semantic features of the image. Shallow and depth information are accurately extracted as a result.
2) SRB and CRB are introduced separately in the two parallel modules to further enhance the contextual relationship of the image.
3) To balance the semantic differences between the extracted features from the two modules, a FRM is incorporated.

These specific designs allow MSDRNet to effectively incorporate image features from both global and local perspectives, enabling semantic segmentation from coarse to fine scales. This greatly improves its performance and accuracy in the segmentation task.

However, MSDRNet has longer processing times compared to other networks. This is mainly because the DSM requires pixel-wise calculations and comparisons for detecting fine-grained semantic features, which consumes a significant amount of time. And due to the limitations of the experimental platform and considering the complexity of the model, this article did not effectively incorporate multiscale features into the model. Therefore, in the future we intend to develop a lightweight network to address the issue of longer processing times and focus on proposing an improved DSB that incorporates multiscale features and optimizing and enhancing the CRB in GSM.

Acknowledgment

The authors would like thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

References

[1] B. Chen, M. Xia, and J. Huang, "MFANet: A multi-level feature aggregation network for semantic segmentation of land cover," *Remote Sens.*, vol. 13, no. 4, p. 731, 2021.

[2] Z. Li, Q. Xin, Y. Sun, and M. Cao, "A deep learning-based framework for automated extraction of building footprint polygons from very high-resolution aerial imagery," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3630.

[3] A. Abdollahi, B. Pradhan, N. Shukla, S. Chakraborty, and A. Alamri, "Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1444.

[4] C. Dechesne, C. Mallet, A. Le Bris, and V. Gouet-Brunet, "Semantic segmentation of forest stands of pure species combining airborne LiDAR data and very high resolution multispectral imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 126, pp. 129–145, Apr. 2017.

[5] F. Fang, X. Yuan, L. Wang, Y. Liu, and Z. Luo, "Urban land-use classification from photographs," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 12, pp. 1927–1931, Dec. 2018.

[6] X. Yuan and V. Sarma, "Automatic urban water-body detection and segmentation from sparse ALSM data via spatially constrained model-driven clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 73–77, Jan. 2011.

[7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[8] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[9] C. He, S. Li, D. Xiong, P. Fang, and M. Liao, "Remote sensing image semantic segmentation based on edge information guidance," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1501.

[10] T. Tian, Z. Chu, Q. Hu, and L. Ma, "Class-wise fully convolutional network for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3211.

[11] X. Fu and H. Qu, "Research on semantic segmentation of high-resolution remote sensing image based on full convolutional neural network," in *Proc. 12th Int. Symp. Antennas, Propag. Electromagn. Theory*, 2018, pp. 1–4.

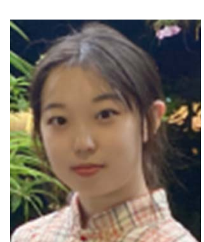
[12] D. Mo, C. Fan, Y. Shi, Y. Zhang, and R. Lu, “Soft-aligned gradient-chaining network for height estimation from single aerial images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 538–542, Mar. 2021.
[13] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, “Hyperspectral and LiDAR data classification based on structural optimization transmission,” *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 3153–3164, May 2023.
[14] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, “Feedback attention-based dense CNN for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5501916.
[15] A. Vaswani et al., “Attention is all you need,” in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.
[16] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, “TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 156, pp. 1–13, 2019.
[17] R. Hang, P. Yang, F. Zhou, and Q. Liu, “Multiscale progressive segmentation network for high-resolution remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5412012.
[18] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, “Hybrid multiple attention network for semantic segmentation in aerial images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5603018.
[19] H. Liu, M. Yao, X. Xiao, and H. Cui, “A hybrid attention semantic segmentation network for unstructured terrain on Mars,” *Acta Astronautica*, vol. 204, pp. 492–499, 2023.
[20] W. Shi, W. Qin, Z. Yun, P. Ping, K. Wu, and Y. Qu, “Attention-based context aware network for semantic comprehension of aerial scenery,” *Sensors*, vol. 21, no. 6, Mar. 2021, Art. no. 1983.
[21] R. Li et al., “Multiattention network for semantic segmentation of fine-resolution remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5607713.
[22] W. S. Cheng, W. Yang, Y. Q. Pan, H. W. Guo, and Y. Cheng, “Context aggregation network for semantic labeling in aerial images,” in *Proc. 26th IEEE Int. Conf. Image Process.*, 2019, pp. 4484–4488.
[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
[24] 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.
[25] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, vol. 9351, pp. 234–241.
[26] 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.
[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. J. C. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” 2015, *arXiv:1412.7062*.
[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. J. A. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, *arXiv:1706.05587*.
[30] W. Liu et al., “Adversarial unsupervised domain adaptation for 3D semantic segmentation with multi-modal learning,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 176, pp. 211–221, 2021.
[31] Y. Gao, M. Zhang, W. Li, X. Song, X. Jiang, and Y. Ma, “Adversarial complementary learning for multisource remote sensing classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Mar. 2023, Art. no. 5505613.
[32] S. Qiu, Y. Zhao, J. Jiao, Y. Wei, and S. Wei, “Referring image segmentation by generative adversarial learning,” *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1333–1344, May 2020.
[33] Y. Gao, M. Zhang, J. Wang, and W. Li, “Cross-scale mixing attention for multisource remote sensing data fusion and classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Mar. 2023, Art. no. 5507815.
[34] F. Yuan, Y. Shi, L. Zhang, and Y. Fang, “A cross-scale mixed attention network for smoke segmentation,” *Digit. Signal Process.*, vol. 134, 2023, Art. no. 103924.
[35] T. Huang, J. Chen, and L. Jiang, “DS-UNeXt: Depthwise separable convolution network with large convolutional kernel for medical image segmentation,” *Signal, Image Video Process.*, vol. 17, no. 5, pp. 1775–1783, 2022.
[36] L. Chen et al., “SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6298–6306.
[37] G. Lin, C. Shen, A. van den Hengel, and I. Reid, “Efficient piecewise training of deep structured models for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 3194–3203.
[38] L. C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, “Attention to scale: Scale-aware semantic image segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.
[39] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[40] P. Ding, H. Qian, Y. Zhou, and S. Chu, “Object detection method based on lightweight YOLOv4 and attention mechanism in security scenes,” *J. Real-Time Image Process.*, vol. 20, no. 2, p. 34, 2023.
[41] B. Liang, C. Tang, W. Zhang, M. Xu, and T. Wu, “N-Net: An UNet architecture with dual encoder for medical image segmentation,” *Signal, Image Video Process.*, pp. 1–3, 2023.
[42] J. Fu et al., “Dual attention network for scene segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 3141–3149.
[43] Z. L. Huang et al., “CCNet: Criss-cross attention for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.
[44] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, “Building extraction in very high resolution imagery by dense-attention networks,” *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1768.
[45] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, “Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning,” *Remote Sens.*, vol. 11, no. 1, p. 83, Jan. 2019.
[46] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017.
[47] 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.
[48] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, “Dense dilated convolutions merging network for land cover classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6309–6320, Sep. 2020.

Weiheng Zhao received the B.S. degree in geographic information science from the Xi’an University of Science and Technology, Xi’an, China, in 2017. He is currently working toward the Ph.D. degree in geo-information systems with Chang’an University, Xi’an, China.

His main research interests include high-spatial-resolution remote sensing image segmentation and classification based on deep learning.

Jiannong Cao received the B.S. degree in aerial photogrammetry and remote sensing from the Wuhan Science and Technology University of Surveying and Mapping, Wuhan, China, in 1987, the M.S. degree in geographic information systems from Northwest University, Xi’an, China, in 1999, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2005.

He is currently a full Professor with Chang’an University, Xi’an, China. His research interests include image understanding, image pattern recognition, and remote sensing imagery processing.



Xueyan Dong received the B.S. degree in remote sensing science and technology from the Xi’an University of Science and Technology, Xi’an, China, in 2017. She is currently working toward the Ph.D. degree in geo-information systems with Chang’an University, Xi’an, China.

Her main research interests include high-spatial-resolution remote sensing image building extraction and edge detection.