Improved Land Cover Classification of VHR Optical Remote Sensing Imagery Based Upon Detail Injection Procedure

Qianbo Sang, Yin Zhuang, *Member, IEEE*, Shan Dong, Guanqun Wang, *Student Member, IEEE*, He Chen, and Lianlin Li, *Senior Member, IEEE*

***Abstract*—Development of very-high-resolution (VHR) remote sensing imaging platforms have resulted in a requirement for developing refined land cover classification maps for various applications. Therefore, aiming at exploring the accurate boundary and complex interior texture retrieval in VHR optical remote sensing images, a novel detail injection network (DI-Net) is proposed in this article, which is composed of three aspects. First, the decoupling refinement module embedded with a multiscale representation is designed to improve the feature extraction capabilities that precede the encoding-to-decoding process. Second, we pay attention to the hard examples of boundary and complex interior texture in land cover classification and design two detail injection attention modules to solve the feature inactivation phenomenon in gradually convolutional encoding-to-decoding process. Third, a specific stage grading loss is proposed to adaptively regulate the structural-level weights of the encoding and decoding stages, which facilitates the details retrieval and produce refined land cover classification results. Finally, various datasets [*incl.* International Society for Photogrammetry and Remote Sensing (ISPRS) and Gaofen Image Dataset (GID)] are employed to demonstrate that the proposed DI-Net achieves better performance than state-of-the-art methods. DI-Net provides more accurate boundaries and more consistent interior textures, and it achieves 86.86% *PA* and 68.37% *mIoU* on ISPRS dataset as well as 77.04% *PA* and 64.38% *mIoU* on GID dataset, respectively.**

***Index Terms*—Encoding-to-decoding, land cover classification, optical remote sensing, refinement module, unmanned aerial vehicles (UAVs), very high resolution (VHR).**

I. Introduction

LAND cover classification is an important application for very high resolution (VHR) optical remote sensing image retrieval, and it could be used for environmental monitoring, urban planning, precision agriculture, forest vegetation survey, and disaster response [1]–[9]. With advancements in optical remote sensing technology, a large quantity of VHR commercial images from unmanned aerial vehicles (UAVs) and satellite platforms have been released and could be accessed. These VHR optical remote sensing images all contain various land covers with clear appearance, which create a challenge for refined land cover classification. The clear appearance could present large intraclass differences for complex interior textures and high requirements of accurate boundary predictions. Facing this challenge, many semantic segmentation methods are available to generate refined land cover classification results. Nevertheless, because of the limited-feature extraction ability of the traditional methods, deep-learning-based convolution neural network (CNN) semantic segmentation methods [10]–[19] are widely used for VHR optical remote sensing refined land cover classification [20]–[24]. Wang *et al.* [20] used CaffeNet to achieve better land cover classification performance than traditional back propagation neural network. Hu *et al.* [21] also proposed a deep convolutional neural network (DCNN) for land cover mapping of Qinhuangdao in China, and their experiments showed that DCNN could provide better mapping results compared with previous methods using the support vector machine and maximum likelihood classification. In relation to [20] and [21], these works all demonstrated that the CNN-based methods could achieve better performance than traditional land cover classification methods.

Recently, to improve the land cover classification performance of CNN-based methods in VHR optical remote sensing images, encoding-to-decoding frameworks and skip connections were widely employed in [25]–[34]. Liu *et al.* [27] proposed the multiscale full convolution network (FCN) to implement maritime semantic segmentation from optical remote sensing images. They proposed a multiscale structure to assemble global-local comprehensive features, and then, they utilized shortcuts to connect low-level and high-level features to generate finer results. Mou *et al.* [28] proposed a recurrent network in FCN (RiFCN) to obtain refined land cover classification results for VHR optical remote sensing images. They proposed a recurrent connection structure to fuse coarse features from deep layers into appearance features from shallow layers. Wurm *et al.* [29] employed an FCN to enlarge

Manuscript received May 13, 2020; revised July 18, 2020 and August 18, 2020; accepted October 5, 2020. Date of publication October 21, 2020; date of current version January 6, 2021. This work was supported in part by the Chang Jiang Scholars Program under Grant T2012122, in part by the Hundred Leading Talent Project of Beijing Science and Technology under Grant Z141101001514005, in part by the China Postdoctoral Science Foundation under Grant 2019M650345, and in part by "the Fundamental Research Funds for the Central Universities CUC200D052." *(Corresponding author: Yin Zhuang.)*

Qianbo Sang, Yin Zhuang, Guanqun Wang, and He Chen are with the Beijing Key Laboratory of Embedded Real-Time Information Processing Technology, Beijing 100081, China (e-mail: sangqianbo@126.com; zhuangyin640829@163.com; wgq@bit.edu.cn; chenhe@bit.edu.cn).

Shan Dong is with the Engineering Center of Digital Audio and Video, Communication University of China, Beijing 100024, China (e-mail: dongshan@cuc.edu.cn).

Lianlin Li is with the School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: lianlin.li@pku.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.3032423

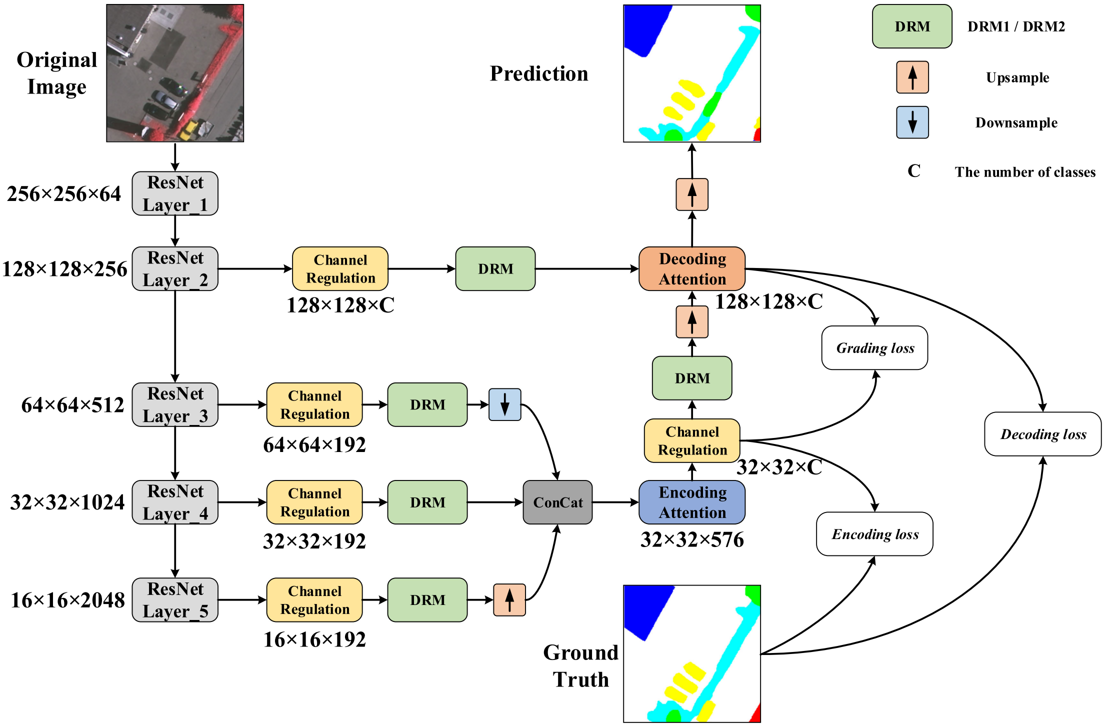


Fig. 1. General framework of the proposed Detail Injection Net (DI-Net).

the receptive field (RF) by gradually stacking convolutional structures, which facilitated feature nonlinear mapping to assist in land cover classification. In addition, the skip connections were also employed to assist the feature fusion. For the high resolution UAVs aerial imagery semantic labeling, Pan *et al.* [31] proposed a dense pyramid network, in which a pyramid pooling module combined with two convolutional layers was designed to improve the feature representation ability. In general, these land cover classification methods mostly focused on exploring efficient encoding-to-decoding frameworks for accurate land cover classification prediction. However, the fine-scale detail information is gradually diminished by convolutional structure in the encoding-to-decoding process. Previous works mostly recovered details by fusing multistage features with direct shortcut path appended with addition or concatenation, which was inadequate due to the lack of weight assignment and adaptability. In order to further improve the refined land cover classification performance, a series of attention mechanism based methods have been proposed. Luo *et al.* [32] proposed the deep FCN with channel attention mechanism to realize semantic segmentation from high resolution aerial images. Then, Yuan *et al.* [33] also incorporated a wide-range attention unit for feature selection into a densely connected U-Net (WRAU-Net) to achieve the road segmentation from remote sensing images. In our previous work of full receptive field network (FRF-Net) [34], feature-wise attention mechanisms were proposed to design a novel encoding-to-decoding process, which captured the globally consistent information efficiently. These abovementioned methods utilized attention mechanisms in the channel or feature domain to select more significant features or enhance their effectiveness, while the detail features are easy to be ignored. Nevertheless, the accurate projection of image details (i.e., boundary and complex interior texture pixels) is important for improving the legibility of land cover mapping results.

Therefore, for implementing the refined land cover classification from VHR optical remote sensing images, the detail injection network (DI-Net) is proposed as shown in Fig. 1. First, two types of decoupling refinement modules (DRMs) are proposed to strengthen the feature representation ability before encoding-to-decoding process. Second, we redesign a new encoding-to-decoding pattern based on the attention mechanism combining with channel residual structure to model details in the feature domain. Third, the proposed specific stage grading (SG) loss assists DI-Net in rectifying the feature description of detailed regions (e.g., boundaries and complex interior textures), where misclassifications frequently occur, by exploring structural relation constraint between encoder and decoder. Finally, extensive experiments are carried out on International Society for Photogrammetry and Remote Sensing (ISPRS) and Gaofen Image Dataset (GID) datasets. The results show that the proposed DI-Net provides better performances from VHR optical remote sensing images, especially by giving more accurate pixel-level prediction of boundaries and more consistent interior textures than the state-of-the-art methods. In general, the contributions of this article could be summarized as three aspects:

1) A novel detail injection CNN-based network is proposed to improve the VHR optical remote sensing land cover classification performance, especially in producing accurate pixel-level boundaries and complex interior texture predictions.

2) A powerful DRM refinement module and encoding-to-decoding pattern based on attention mechanism embedding with channel residual structure are proposed to generate fine-scale feature description for refined land cover classification.
3) A novel SG loss is proposed to explore the structural constraint between the encoding and decoding stages, which is leveraged to retrieve detailed features.

The rest of this article is organized as follows: In Section II, the related studies in refined land cover classification are presented. In Section III, we elaborate on the proposed DI-Net in four subsections. Next, extensive experiments and discussions are presented in Section IV. Finally, the conclusion is given in Section V.

II. Related Work

Recently, CNN-based semantic segmentation methods have been widely applied in refined land cover classification from VHR optical remote sensing images. Powerful refinement modules, efficient encoding-to-decoding frameworks, and suitable loss functions are proposed as the components necessary to produce finer results. We then examine these aspects to introduce several related works.

Refinement Module: For achieving the refined land cover classification results from VHR optical remote sensing images, an intractable problem is that the detail information would be severely corrupted by gradually convolutional structure. Therefore, to rectify the feature maps and to extract detail information, several research projects began to emphasize the design of the refinement module. For refining fine-scale details, global convolution network (GCN) [15] implements a boundary refinement module with small convolutional kernels. Then, the shuffle module [35] based on the principle of dimension transposing overcomes the side effects brought by group convolutions and achieves channel-wise information flows. Next, efficient spatial pyramid (ESP) [36] refinement module combined with point-wise convolution is proposed for preserving large RF with fewer parameters and memory footprints for the detail feature description. In addition, the refinement module of FRF-Net [34] is designed for unifying channel numbers of the different feature layers and adjusting the feature maps slightly to improve classification performance. Thus, similar to depth-wise separable convolution, a 1×1 convolution is used to extract the channel-wise dependency, while 3×3 convolutions focused on the spatial-wise dependency extraction. Related to abovementioned works, smaller convolution kernels are adopted to learn the fine-scale features. Therefore, in the proposed DI-Net, we also follow this view point and employ several small convolution kernels to build the powerful refinement module for refining fine-scale features.

Encoding-to-Decoding Framework: The encoding-to-decoding framework has been extensively applied in many semantic segmentation methods. In general, an encoding-to-decoding pattern must contain two parts: one encoder module capturing the higher-level semantic information, and one decoder module recovering detailed spatial information. MUnet [30] utilizes the encoding-to-decoding process with skip connections to implement land cover mapping. RiFCN [28] adopts a recurrently connected encoding-to-decoding framework to produce accurate land cover classification results. DeepLab v3+ [14] utilizes an encoding-to-decoding structure to encode abundant contextual information, and a simple yet effective decoder is adopted to recover the detailed information. In FRF-Net [34], an encoding-to-decoding process is designed based on two types of attention mechanism. Self-attention is set up to build the encoder for capturing long-range dependence. In addition, a fusion attention decoder is designed to efficiently fuse low-level feature with high-level feature. In this article, a novel encoding-to-decoding framework is designed for reserving detail information, with the aim of achieving powerful boundary and complex interior texture descriptions for VHR optical remote sensing land covers.

Loss Function: Related to the semantic segmentation task, cross-entropy loss is widely used in most research. However, simple cross entropy loss could barely meet the requirements of refined semantic segmentation. For encoding-to-decoding structures, there is more auxiliary information that could be utilized for predicting accurate detailed projections. Much research focuses on exploring dynamic loss function for facilitating network to learn the sharper details. Zhang *et al.* [38] propose the semantic encoding loss (SE-loss) function that predicts the presence of the object classes in terms of the encoded semantics. By this means, SE-loss gives equal contributions for multiscale objects, which make up for the defects of the per-pixel loss. Bilateral segmentation network [39] utilizes a multistage cross entropy loss to supervise the training process, in which the principal loss function is applied to supervise the output of the entire network, and there are two auxiliary losses utilized for optimizing the immature feature maps. In general, these auxiliary losses significantly accelerate the training speed. In contrast, our method aims at retrieving detailed information from the encoding-to-decoding process for generating finer land cover maps.

Focusing on these three problems, we designed DI-Net and performed experiments on the ISPRS Vaihingen two-dimensional semantic labeling contest and the GID datasets. The ISPRS is a dataset composed of VHR UAVs aerial images. The land covers in the ISPRS dataset are labeled with six classes: *Impervious Surfaces (IS)*, *Building (B)*, *Low Vegetation (LV)*, *Tree (T)*, *Car (C),* and *Clutter (CL)*. The GID is a large-scale land-use dataset, which has complex interior textures with large intraclass differences and similar interclass diversities of the six classes: *Built-up (B)*, *Farmland (Far)*, *Forest (F)*, *Meadow (M)*, *Waters (W),* and *Unknown (Un)*.

III. Proposed Methodology

This section elaborates on the principles of the proposed DI-Net, as shown in Fig. 1. First, ResNet-101 is applied for feature extraction. Then, the DRM is designed, in which the multiscale representation and channel shuffle operations are both applied to generate powerful feature description. Next, the defect of feature inactivation phenomenon in the traditional attention mechanism impeding the accurate mapping of details is analyzed. Then, novel attention modules are proposed to resolve detail misclassification problem by alleviating the feature inactivation phenomenon with channel residual structure. Finally, for further optimizing DI-Net convergence and retrieving detail feature, the SG loss is proposed based on the structural relation constraint between the encoder and decoder to effectively refine the land cover classification predictions on complex interior textures and boundary areas. Consistent with the sequence of the workflow, we describe the proposed DI-Net individually, including the DRM in Section I-A, the encoding attention module in Section I-B, the decoding attention module in Section I-C, and the SG loss in Section I-D.

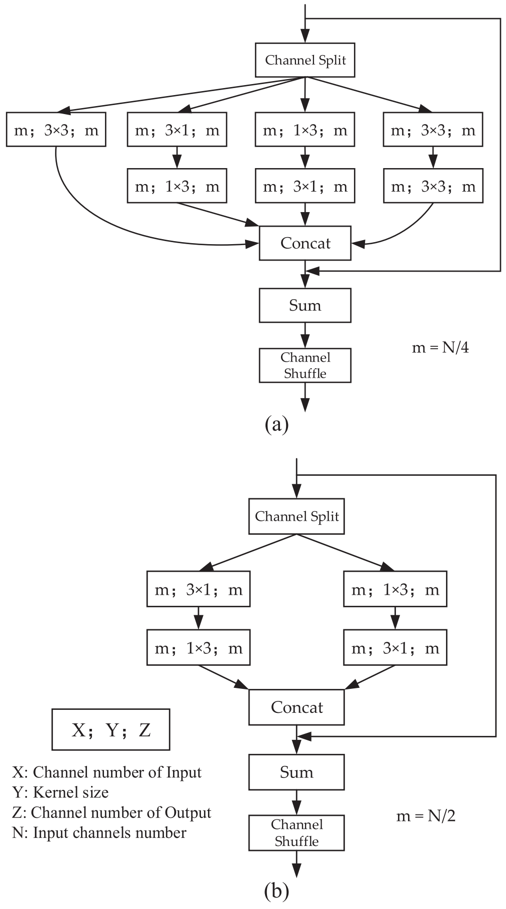


Fig. 2. (a) DRM1. (b) DRM2.

A. Decoupling Refinement Module

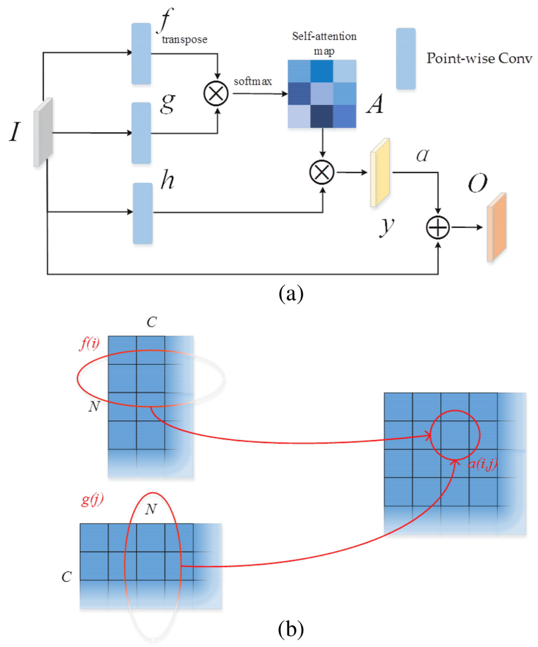
As illustrated in Fig. 1, before the encoding-to-decoding process, the DRM in DI-Net is proposed for refining different stage features, which are extracted from the ResNet-101 backbone. The channel regulation modules compressing the channels are adopted to reduce the computational load and memory consumption of the following refinement module and encoding-to-decoding process. However, method [35] proved that due to the complexity constraints in different channels, the channel compression by expensive point-wise convolutions would restrict the accuracy of land cover classification. Therefore, before the DRM refinement module, we take the complexity constraint into account and utilize grouped point-wise convolutions to construct the channel regulation modules.

After the channel regulation module applied for different scale features, the DRM is utilized to refine detail information. As shown in Fig. 2(a) and (b), the DRM includes two different forms based on a uniform idea. Specially, DRM is based on channel splitting and shuffle operation. The input feature is split according to the channel, and then split features are calculated in different convolutional paths, respectively. Thus, the channel splitting operation could be considered as one kind of grouping operation to keep different eigenmode in each path. Then, we also consider adopting several small convolutional kernels for learning the fine-scale information in feature maps. However, solely adopting unique convolutional kernel would limit the feature representation ability. Increasing cardinality (i.e., the size of the set of paths) with multiscale kernels is more effective than stacking layers blindly. Therefore, we utilize several $1 \times 3, 3 \times 1$, and 3×3 convolution kernels to achieve the multiscale feature description after input split by N channels into N/m branches as shown in Fig. 2(a) and (b). Then, each branch has an equal number of channels m. The convolution kernels of various sizes can capture different scale context information benefiting feature representation in DI-Net. After the multiscale feature extraction by parallel convolutions, the results of all branches are concatenated to keep the channels number invariable with the input. The result is summarized with the input to optimize the training process. Finally, a channel shuffle operation based on dimension transposing is used to ensure that the information between all branches can be communicated and combined properly. In general, the designed DRM can enhance the fine-scale feature extraction capability, which is demonstrated in experiments and analysis Section IV-B (1). Then, following the architecture of DI-Net in Fig. 1, after the DRMs, the encoding-to-decoding process based on two types of encoding and decoding attention modules are introduced as follows.

B. Encoding Attention Module

In general, a better land-cover classification result must have two strong capabilities. First, the inside of the object must be predicted correctly and with great consistency. Second, the results must present more accurate boundary predictions between multiple land covers. We hold the opinion that encoder can facilitate the complex interior texture prediction. The RF is enlarged by gradually convolutions or large convolution kernels in encoder, which could extract global context semantic to facilitate the interior classification [15].

For the encoding process, a great amount of research [22]–[34] focuses on exploring RF and capturing long-range dependency to improve pixel-level interpretation ability. The self-attention mechanism [37] is adopted as an efficient way to capture the long-range dependency in the feature domain as shown in Fig. 3(a). As shown in Fig. 3(b), the self-attention module is built on the principle of intracalculation to calculate the statistics correlation of the features. However, the ordinary self-attention benefits in global consistency at the expense that



(a)

(b)

Fig. 3. Traditional self-attention mechanism. (a) Global view of the whole process. (b) Calculation process to obtain the attention map A.

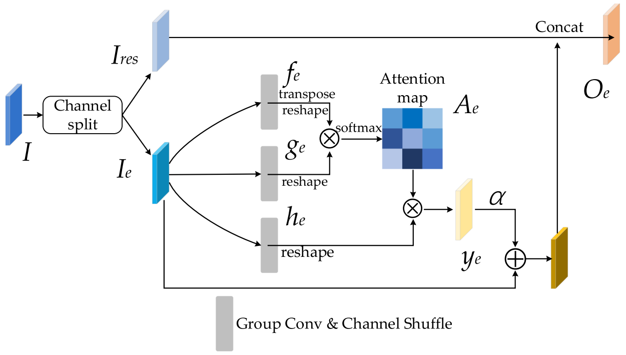


Fig. 4. Encoding attention module in DI-Net. ⊗ denotes matrix multiplication. SoftMax operation is performed on each row. And ⊕ denotes element-wise summation process.

the weak correlation features, especially for the features of the details (i.e., boundary and complex interior texture), suffer from inactivation and nondiscrimination. The details of images always be projected with weak response in the deep layer, while ordinary attention mechanism weakens detail response further.

For avoiding the feature inactivation and generating discriminative feature description, we employed the self-attention mechanism combined with channel residual structure to build the encoding attention module in DI-Net as shown in Fig. 4. Here, one part of channel group I_{res} is reserved from the self-attention calculation for keeping the original feature information. Then, 1×1 group convolution combined with the channel shuffle operation is embedded with self-attention mechanism to capture the available long-term dependency and alleviate the feature inactivation phenomenon of the detail features in the I_e channel group. Therefore, in Fig. 4, the input $I \in R^{C_I \times H \times W}$ is split into two parts of I_e and I_{res}, which contain β and $C_I - \beta$ channels separately. I_{res} would be directly concatenated last, and the feature map $I_e \in R^{\beta \times H \times W}$ is fed into three 1×1 group convolutions appended with channel shuffle operation, respectively, to generate the three feature maps f_e, g_e, and h_e which have the same channel number β. Then, they are both reshaped into $R^{\beta \times N}$, where $N = H \times W$. Next, the attention map A_e is calculated by a matrix multiplication between f_e and g_e, and the attention map $A_e = R^{N \times N}$ is obtained after the SoftMax layer. Therefore, element $a_{j,i}$ in A_e could be expressed as (1)

$$a_{j,i} = \frac{e^{s_{ij}}}{\sum_{i=1}^{N} e^{s_{ij}}},$$

$$\text{where } s_{i,j} = (f_e^T \otimes g_e)_{i,j} = \sum_{n=1}^{\beta} (f_e)_{i,n}^T \otimes (g_e)_{n,j}. \quad (1)$$

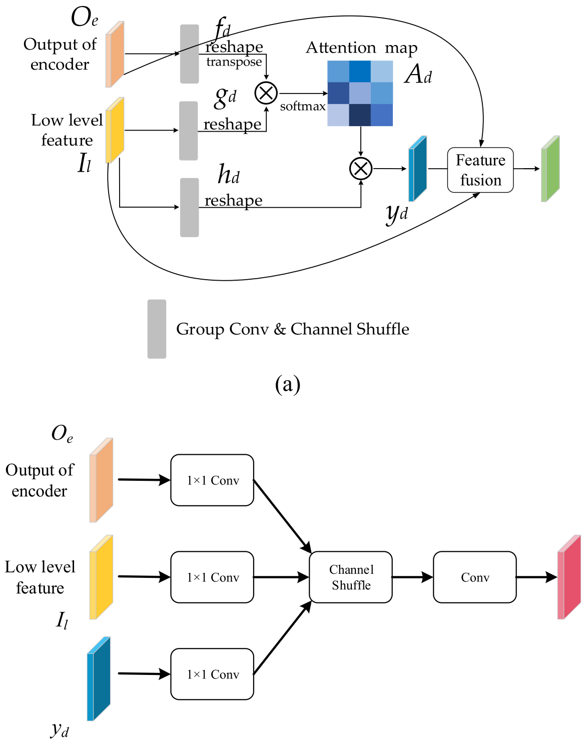
After the attention map calculation, a matrix multiplication is taken between feature map h_e and attention map A_e to produce $y_e \in R^{\beta \times H \times W}$. Then, the result y_e is weighted by a learnable parameter α and added back to input I_e. At the end, I_{res} is concatenated with the result of αy_e added with I_e to produce O_e by (2) to generate the encoding feature in DI-Net.

$$O_e = \text{concat}(I_{\text{res}}, \alpha y_e + I_e), \ \alpha \text{ is initialized as } 0. \quad (2)$$

Following (2) as shown in Fig. 4, the encoding attention module could produce the powerful feature description for semantic generalization, and it further facilitates the pixel-level feature description of details.

C. Decoding Attention Module

As mentioned at the beginning of Section III-B, to achieve refined land cover classification, the decoding attention module recovers fine-scale detail information into high-level feature from encoding process. To recover fine-scale details reasonably, an effective way is decoding the high-level feature with the participation of the low-level feature. Based on this ideal, we proposed the decoding attention module as shown in Fig. 5. In Fig. 5(a), output O_e of the encoder and the low-level feature I_l are both fed into a group point-wise convolution layer appended with the shuffle operation, respectively. Then, in decoding attention module, the calculations of attention map A_d and multiplication result y_d are identical to the encoding attention module. Then, an efficient feature fusion strategy is proposed to generate the final predicted features from multiple features (e.g., O_e, I_l, and y_d). First, three 1×1 convolutions are applied for I_l, O_e, and y_d feature maps to achieve the channel compression, respectively, as shown in Fig. 5(b). In relation to O_e, I_l, and y_d feature fusion, we expect that they all make contributions to the final pixel-level refined land cover classification prediction. Therefore, the proportions of the channel numbers are fixed into 2:1:1 for O_e, I_l, and y_d feature maps. Then, the compressed feature maps of I_l, O_e, and y_d are concatenated and appended with convolution block for refined land cover classification. In general, following the proposed decoding attention module, DI-Net could generate more refined land cover classification results.



(a)

(b)

Fig. 5. Decoder attention module for recovering the fine-scale features.

D. Stage Grading Loss

Motivated by further leveraging the finer details to produce more refined land-cover classification results, SG loss is proposed for supervising the training phase. The multistage features are graded and constructed with adaptive weights to achieve detail information retrieval. Here, the proposed SG loss could be roughly divided into two parts as shown in (3)

$$\text{loss} = \sum_{\text{s=encoder}}^{\text{decoder}} \lambda_s L\left\{P^{(s)}, T\right\} + L\left\{p_T^{(\text{encoder})}, p_T^{(\text{decoder})}\right\}. \tag{3}$$

In (3), the first term is composed of the cross entropy loss for both the encoder and decoder. $P^{(s)}$ means the prediction result of stage s which indicates the encoding or decoding stage as "0" or "1," and T means the target label. Then, we consider that the output $P^{(1)}$ of the decoder is actually the decision-making feature layer, which directly decides the prediction accuracy of DI-Net. Thus, the cross-entropy loss defined between the output $P^{(1)}$ of the decoder and the target T is supposed to be the principal loss. Moreover, the output feature of the encoder is considered as an important part for optimizing training process and producing better land-cover classification performance. Therefore, we also employed the output $P^{(0)}$ of the encoder and target T as an auxiliary loss to supervise the output of the encoder for easily optimizing the model performance. Here, the parameter λ_s is used to balance the weights of the decoder and encoder losses. In (3), the second term is defined as a grading loss, which can be formed as (4)

$$L\left\{p_T^{(\text{encoder})}, p_T^{(\text{decoder})}\right\} = \left\|\max\left\{p_T^{(\text{encoder})} - p_T^{(\text{decoder})} + \text{margin}, 0\right\}\right\|_2. \tag{4}$$

In (4), p_T means the prediction probability map on the correct target label T, and the $\|(\cdot)\|_2$ means the L_2 norm of the given matrix. Actually, the encoder and decoder sometimes produce different probability distribution, which means they make the opposite decisions for a certain pixel. The main reason for this phenomenon is that the decoder recovers detail features with the appearance information, which introduces extra detail information. To be specific, appearance information of details (i.e., boundary and complex interior texture) is introduced from shallow layer to the encoder output by the proposed adaptive decoding process, which causes the decision divergence with encoder. The feature fusion process in decoder is essential for recovering the more accurate detailed prediction. Therefore, the grading loss of (4) is designed for utilizing the structure-level relation of decoder and encoder to rectify the feature representation in training phase. We consider the decoder more valuable for reconstructed discriminative details into high-level features from encoding process, and it must be endowed with more weight than the encoder to predict land cover classification with an accurate detail projection. Consequently, in grading loss, there is an expected relationship $p_T^{(\text{decoder})} > p_T^{(\text{encoder})} + \text{margin}$ set up to retrieve detailed information in the training phase. In addition, related to the setup relation of encoder and decoder, λ_s of decoder in (3) should be relatively larger, because it can ensure that the training process puts more emphasis on decoder optimization for refining the details prediction.

IV. Experiments and Analysis

To show the performance of the proposed DI-Net, extensive experiments are performed on published ISPRS and GID datasets and are compared with several state-of-the-art algorithms, i.e., DeepLab v3+ [14], GCN [15], PSPNet [17], U-Net [18], and FRF-Net [34]. The experimental results demonstrate that the proposed DI-Net can achieve the best performance on pixel-wise land cover classification. As follows, we first introduce datasets and implementation details, and then we perform a series of ablation experiments and comparison analyses.

A. Datasets and Implement Details

The GID dataset is composed of images with spatial resolution of 4 m multispectral. For GID dataset, we just adopt the RGB channels and regulate the images with a down sampling factor of 4 and clip with a size of 360 × 340. Then, 3750 images obtained with 16 m resolution are split into 2000 training and 1750 validation. For the ISPRS dataset, there are 33 tiles available for training. Among them, the 16 available tiles are divided into 11 training tiles and 5 validation tiles in our experiments. Next, the rest of this dataset is employed as a testing set for evaluating the performances of the algorithms. In addition, only three channels of RGB, i.e., the TOP channels, are used in

TABLE I
COMPARISON OF REFINEMENT MODULES WITH DIFFERENT BACKBONES

Backbone	Refinement Module	*mIoU*(%)
Res101	Base	66.21
	MobileNet	67.49
	ShuffleNet	67.72
	GCN	67.64
	ASPP	67.34
	ESP	67.18
	FRF	67.83
	DRM1	68.37
	DRM2	68.00
Res50	Base	66.71
	ASPP	67.20
	ESP	67.60
	DRM1	68.07
	DRM2	67.66
Res152	Base	67.53
	ASPP	67.83
	ESP	68.18
	DRM1	68.32
	DRM2	68.20

TABLE II
EXTENT ANALYZATION THAT THE SPLIT PROPORTION EFFECTS THE SCORE

Channel Split proportion ω	*mIoU*(%)	*PA*(%)
0	67.45	86.67
1/8	67.78	86.46
2/8	68.00	86.64
3/8	67.65	86.45
4/8	67.80	86.67
5/8	67.56	86.59
6/8	66.99	86.45
7/8	67.19	86.36
1	67.10	86.56

all of our experiments, while the DSMs are abandoned. The ground sampling distance of ISPRS is 9 cm. Next, all comparison methods are implemented on PyTorch 1.0, and experiments are performed on an NVIDIA TITAN Xp GPU. We employ a polynomial learning rate scheduler. And we use a standard stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. Some data augmentations are applied to avoid overfitting. In addition, we empirically set λ_s of the encoder and decoder in (3) to 0.5 and 1.5, respectively. The margin in (4) is set to 0.05. The number of groups of the group convolution used in the channel regulation module is 4. To compare DI-Net with the state-of-the-art methods, we report the Pixel Accuracy (*PA*) [34] and mean of Intersection over Union (*mIoU*) [34] averaged over all classes.

B. Ablation Experiments

In this section, for the proposed DRM, encoding-to-decoding attention modules and SG loss discussions, several ablation experiments are carried out on ISPRS and GID datasets to demonstrate their effect on refined land cover classification in DI-Net.

1) Decoupling Refinement Module: In Section III-A, the proposed DRM contains two variant forms, DRM1 and DRM2, for detail feature extraction. There are several refinement modules of GCN boundary refinement module [15], ShuffleNet refinement module [35], MobileNet refinement module [40], Atrous Spatial Pyramid Pooling [14], ESP [36], refinement modules in FRF-Net [34], and two types of DRM tested, respectively. For better comparisons, we set the baseline without any special refinement module in DI-Net, and it just uses point-wise convolutions to regulate the channel number. Then, as the results shown in Table I, based on the ResNet-101 backbone, the performance of DI-Net with DRM1 surpasses the baseline by a significant margin of 2.16%. The models with DRM1 and DRM2 achieve higher scores than the other refinement modules as well. In addition, related to backbone ResNet-50 and ResNet-152, all results show our proposed DRM has the better performance. Moreover, the performance with DRM1 is evidently higher than that with DRM2. This is because DRM1 provides more eigenmodes, and a larger cardinality (the size of the set of paths) with multiscale kernel, which is effective in improving model performance. DRM leverages the rectification of different eigenmodes in each path. The input feature of each path is obtained by a channel splitting operation, through which the number of channels is reduced. Compared with traditional compression strategy by 1×1 convolution, the calculation cost of the channel splitting operation is lower. Compared with DRM2, DRM1 expands the eigenmodes instead of increasing convolution depth. Due to the splitting-convolution-shuffle pattern, the parameters and calculation loads of DRM1 not increase largely. However, the DRM1 with more eigenmodes achieves higher accuracy proving that the increase on eigenmodes is efficient. Therefore, these comparison results prove that the proposed DRM could effectively attain refined feature tuning before encoding-to-decoding process with less computation cost.

2) Encoding-to-Decoding Attention Module: As the description in Section III-B, we utilize the attention mechanism combined with the channel residual structure to perform encoding attention module in DI-Net. The input channels are split into two parts which include I_{res} and I_e, respectively. Then, the channel proportion of I_{res} and I_e can be defined as ω, where$\omega = C_{I_{\text{res}}}/(C_{I_e}+C_{I_{\text{res}}})$. The $C_{I_{\text{res}}}$ and C_{I_e} refer to the number of channels of I_{res} and I_e, respectively. Next, we discuss the split proportion ω and how it is affecting the refined land-cover classification, and testing results evaluated by *mIoU* and *PA* scores are shown in Table II. In Table II, when ω is equal to "0," the encoding attention module degrades to an original self-attention module. However, if ω is equal to "1," it means that there is no encoding attention structure in the network and the encoder is simply constructed by the DRM. Related to these two results in first and last lines of Table II, the *mIoU* of "$\omega = 0$" is evidently higher than that of "$\omega = 1$," which indicates that the encoder with encoding attention is more efficient than encoder without encoding attention. Thus, we can see that the attention mechanism can capture the long-range dependence to achieve feature generalization in the encoding process. Next, when we endow a certain proportion of ω for I_{res} and I_e, the results show

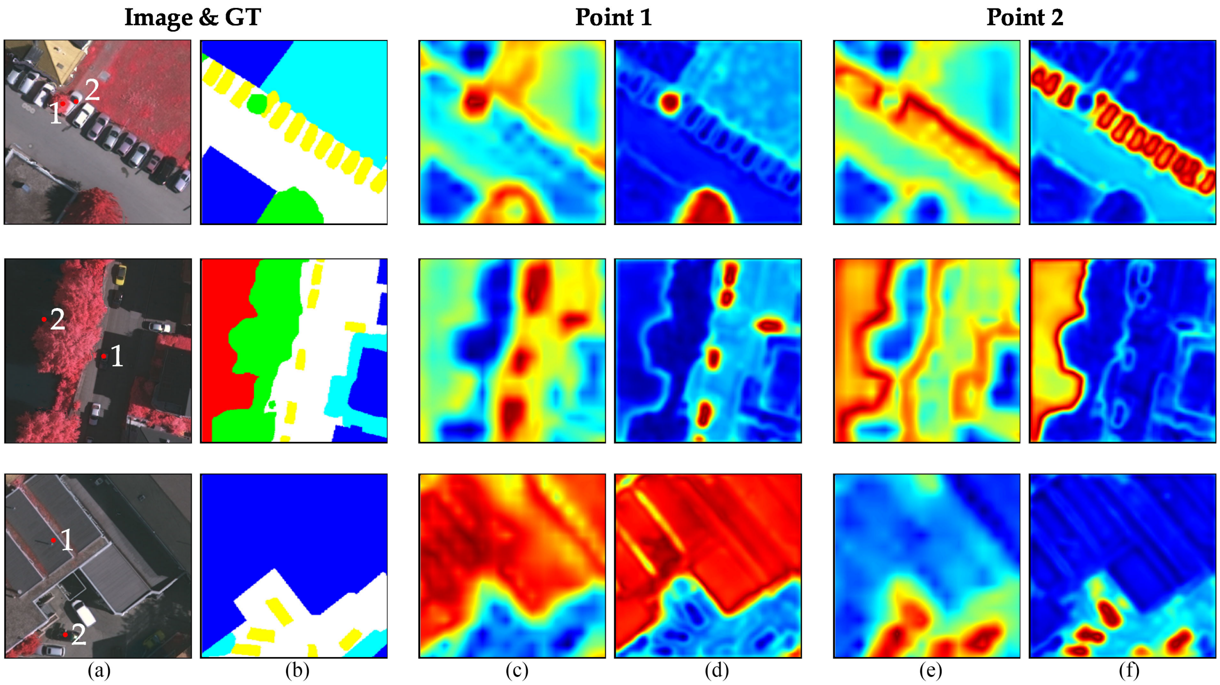


Fig. 6. Feature similarities maps to a pointed pixel, in which hotter color represents more similarities. (a) Input original images. (b) Ground-truths. (c) Results with point 1 after the attention modules are removed. (d) Results with point 1 with attention modules. (e) Results with point 2 after the attention modules are removed. (f) Results with point 2 with attention modules.

TABLE III
STUDY ON ATTENTION MODULES IN ENCODING-TO-DECODING FRAMEWORK

Backbone	Encoding Attention	Decoding Attention	*mIoU*(%)
Res50	×	×	62.36
	√	×	62.71
	×	√	67.59
	√	√	68.07
Res101	×	×	63.80
	√	×	64.23
	×	√	67.80
	√	√	68.37

The √ and × means the presence or absence of attention module, respectively.

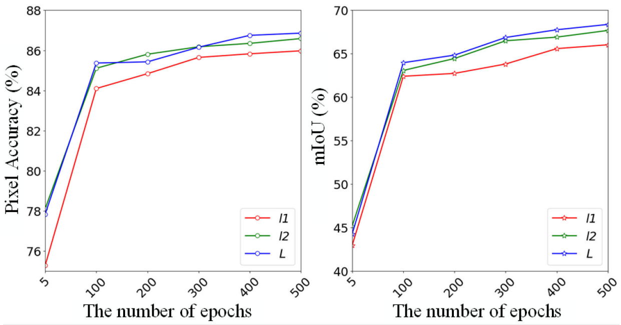
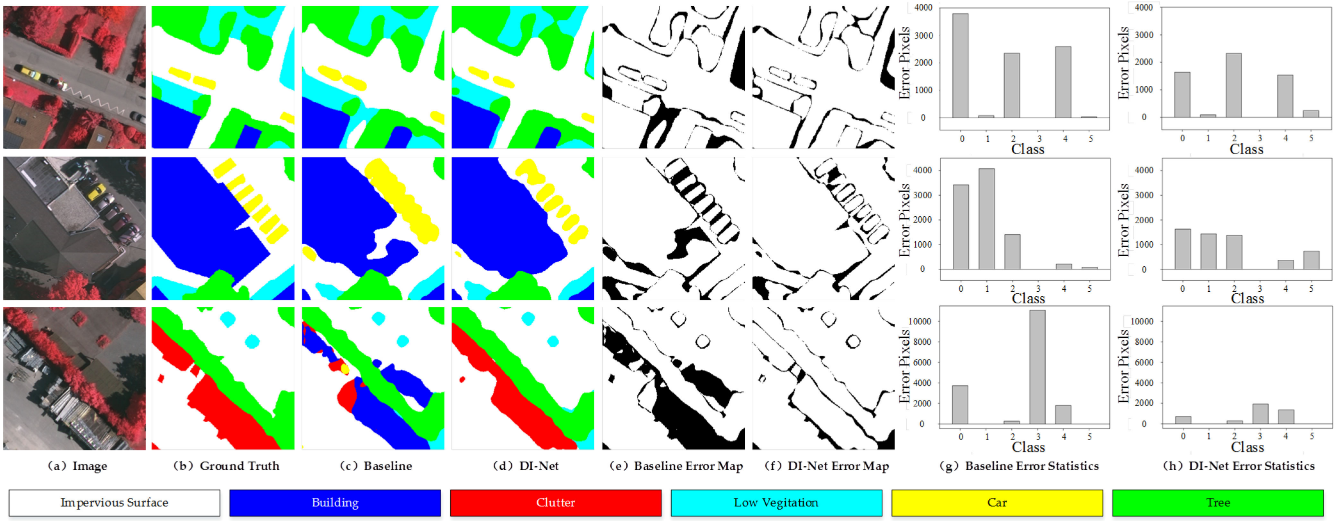


Fig. 7. Extents of general PA and mIoU scores that effected by losses l_1, l_2, and L.

that the reserved channels can facilitate the feature description to resolve the detail feature inactivation in encoding process. Then, related to the channel residual structure demonstration and striking a balance between *mIoU* and *PA*, we observe that the encoding attention module achieves the best performance with the split proportion of $\omega = 0.25$ empirically.

Next, for further discussing the attention structure effect in proposed DI-Net, several experiments are performed on the ISPRS dataset shown in Table III to demonstrate the effectiveness of the encoding and decoding attention modules. From Table III, compared with the baselines (e.g., no attention structure) of Res50 and Res101, complete DI-Net equipped with attention modules surpasses 4.57% on Res-101 and 5.71% on Res-50. Note that the decoding attention module could improve performance much more than the encoding attention module, thus we can find that performance gains of the refined land cover classification from DI-Net primarily come from the decoding process of the feature fusion. In addition, there are the several visualized evaluation results of the attention mechanism as illustrated in Fig. 6. In Fig. 6, the final predicted feature layer is presented for evaluating the description ability of detail features on the condition of with or without the attention structure in encoding-to-decoding process. The Euclidean distance is employed for measuring pointed pixels comparing with whole feature map pixels. Each line in Fig. 6 shows the examples of two selected points (i.e., points 1 and 2) in three UAVs remote sensing scenes. The feature similarities maps are individually calculated with chosen pixels. In Fig. 6(c)–(f), the red color indicates that the feature is similar to the selected points and blue color represents the nonsimilar. In the second line, the two pointed pixels (i.e., points 1 and 2) are separately chosen from the inside of “car” and the boundary of “clutter.” If the attention



(a) Image (b) Ground Truth (c) Baseline (d) DI-Net (e) Baseline Error Map (f) DI-Net Error Map (g) Baseline Error Statistics (h) DI-Net Error Statistics

Fig. 8. Due to the ability to recover sharp feature, DI-Net can correctly classify confusing boundaries and can enhance the legibility of land cover maps.

modules are abandoned, the feature similarities would be dispersive and disordered, in which case there is not enough context information to support correct classification. Consequently, the segmentation results in this case are easily to be misclassified or out of shape. The first line in Fig. 6 shows the performance that the attention modules deal with the intensive appeared small objects. The point 2 is pointed at the "car" that is partly below the tree. The response similarities show that the model with attention modules could still segment the "car" successfully, in situations where the "cars" are densely arranged. The third line in Fig. 6 denotes the attention modules' ability of maintaining the inner consistency in case of a large-area object. We hold the opinion in Section III-B that sufficient contextual information is critical for land cover classification, especially for easily confused categories and large-scale objects. As shown in Fig. 6, in the ISPRS dataset, some land cover categories have similar appearances (such as "Tree" and "Low-Vegetation") and have large gaps in scale (such as "Building"), for which extracting adequate context information is more significant. On the contrary, limited RF leads to indistinguishable feature prediction as shown in Fig. 6(c) and (e), which results in pixel misclassifications. In our work, we introduce attention mechanisms to resolve this challenge. Similarities maps show that the existence of the attention modules could significantly model context information and predict more accurate results in our encoding-to-decoding framework.

3) Stage Grading Loss: In Section III-D, the SG loss is designed for optimizing the training process and generating accurate detail retrieval maps. Here, we set up several experiments to test the effect of the SG loss function. On the basis of decoder entropy loss, we tried to expound the effects of the constraint on the encoder output and the grading loss between the encoder and decoder. There are three curves in Fig. 7 separately referring to the training process with: only entropy loss of decoder l_1, the sum of entropy loss of decoder and encoder l_2 and the proposed SG loss function L. The tested losses l_1, l_2, and L could be expressed by formulas as shown in (5)–(7). Here, l_1 could be considered as the baseline.

$$l_1 = L\left\{P^{(\text{decoder})}, T\right\} \tag{5}$$

$$l_2 = \sum_{s=\text{encoder}}^{\text{decoder}} \lambda_s L\left\{P^{(s)}, T\right\} \tag{6}$$

$$L = \sum_{s=\text{encoder}}^{\text{decoder}} \lambda_s L\left\{P^{(s)}, T\right\} + L\left\{p_T^{(\text{encoder})}, p_T^{(\text{decoder})}\right\}. \tag{7}$$

As shown in Fig. 7, the curve of the L loss achieves the best performance of *PA* and *mIoU* compared with l_1 and l_2. Here, the entropy loss defined on the encoder supervises the optimization of the encoding process and accelerates the convergence of the whole network. Then, the grading loss is designed for taking the prediction from encoder as reference to enforce the decoder generating more confident and stable prediction based on the set-up relation. Furthermore, in grading loss term, a lot of details are modeled in final predicted feature layers. Therefore, the proposed SG loss can effectively improve the quality of the refined land cover classification results.

In addition, for demostrating the detail projection ability of DI-Net, our previously work FRF-Net [34] which is a powerful land cover classification network is employed as a baseline. Fig. 8 illustrates three examples that have the complex interior texture and the boundary information. Then, we set FRF-Net [34] as a baseline, and the error maps are set up according to the baseline and DI-Net in Fig. 8(e) and (f), respectively. In the error maps, the black areas denote the misclassified pixels, and we can see that the misclassifications of boundaries or complex interior textures frequently occur in the baseline module. Fuzzy boundary predictions constrict the legibility and usability of the results, especially for small objects (e.g., "car") and sporadic objects (e.g., "tree" and "clutter"). In Fig. 8(g) and (h), there are histograms to count the error pixels of each class in ISPRS for the selected examples. These histograms show that the proposed

TABLE IV
PER-CLASS COMPARISON RESULTS OF DIFFERENT METHODS ON ISPRS AND GID DATASETS

ISPRS dataset

Methods	IS (%)	CL (%)	T (%)	B (%)	C (%)	LV (%)	*PA* (%)	*mIoU* (%)
DeepLab V3+ [14]	**89.77**	43.66	***88.27***	91.10	61.01	76.31	86.14	65.86
GCN [15]	87.08	**49.14**	82.81	89.27	57.41	**80.44**	84.52	63.02
PSPNet [17]	89.22	38.50	86.30	90.99	61.20	78.04	85.81	65.07
U-Net [18]	84.83	7.22	82.53	82.15	14.84	63.86	77.58	46.04
FRF-Net [34]	88.19	38.32	87.43	***92.44***	***71.24***	76.88	86.04	65.74
DI+DRM2	88.41	46.83	**88.70**	**92.52**	69.94	77.60	***86.64***	***68.00***
DI+DRM1	***89.52***	***48.20***	87.39	91.66	**72.48**	***79.51***	**86.86**	**68.37**

The abbreviation in ISPRS: IS-*Impervious Surfaces*; B-*Building*; LV-*Low Vegetation*; T-*Tree*; C-*Car*; CL-*Clutter*.

GID dataset

Method	Un (%)	B (%)	M (%)	W (%)	Far (%)	F (%)	*PA* (%)	*mIoU* (%)
Deeplab V3+ [14]	71.74	***81.95***	84.60	85.33	74.20	**72.74**	75.33	62.85
GCN [15]	**77.74**	74.65	76.84	83.29	69.97	66.59	75.34	62.46
PSPNet [17]	***76.83***	78.45	**85.82**	**86.13**	70.27	64.09	75.85	63.48
U-Net [18]	72.37	69.86	79.12	79.48	72.11	49.59	72.07	57.01
FRF-Net [34]	76.22	81.00	***85.56***	84.88	71.80	65.96	76.24	64.17
DI+DRM2	76.50	78.55	78.42	84.28	***74.31***	***68.78***	***76.69***	***64.33***
DI+DRM1	75.88	**83.22**	81.63	***85.46***	**74.66**	68.42	**77.04**	**64.38**

The abbreviation in GID: B-*Built-up*; Far-*Farmland*; F-*Forest*; M-*Meadow*; W-*Waters*; Un-*Unknown*.
Bold numbers represent the best score for a class, italic numbers the second-best score.

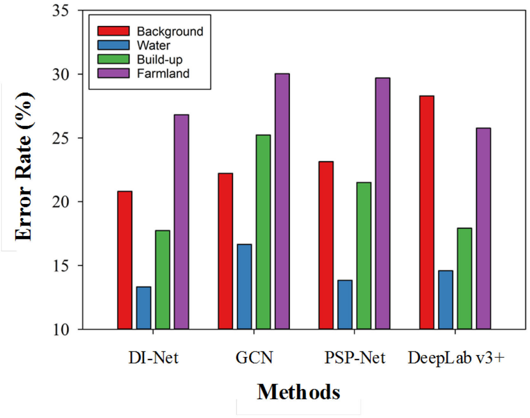


Fig. 9. Pixel misclassification rate of difficult context areas in GID.

DI-Net has a lower pixel error rate than the baseline and a more accurate performance for boundary pixels, which greatly improves the legibility of the produced land cover mapping. In addition, the land covers of water, background, built-up, and farmland in the GID dataset have more complex interior textures, which are easy to be misclassified. Then, we also employ the performances of complex objects (i.e., water, background, built-up, and farmland) in the GID dataset to evaluate the detail projecting ability of DI-Net as well. The results are shown in Fig. 9, and we can see that DI-Net also has a lower pixel error rate comparing with GCN [15], PSP-Net [17], and DeepLab v3+ [14], respectively, for the water, background, built-up, and farmland land cover classes.

C. Comparisons Analysis

In this section, we focus on evaluating and comparing the DI-Net with several state-of-the-art approaches. A series of comparison experiments are taken on GID and ISPRS datasets. As shown in Table IV, the proposed DI-Net achieves a better overall performance than other state-of-the-art methods, whether in ISPRS or GID datasets. Here, the GID dataset has more complex land-cover distributions which have a complex interior texture and complicated boundaries. Therefore, it presents a great challenge to refined land-cover classification. Then, the ISPRS dataset includes VHR UAVs images which contain several difficult classified land covers such as "car," "low vegetable," and "clutter." From Table IV, we can see that the DI-Net can obtain the best performance for "build-up," "water," and "farmland" on GID, and these land covers all have a varied interior texture and irregular edges. Otherwise, related to the ISPRS dataset, DI-Net achieves the best performance for "car" and "low vegetable" and reached the second-best performance for "clutter." Furthermore, when we employed the DRM1 as refinement module, DI-Net can produce 77.04% *PA* and 64.38% *mIoU* on GID and 86.86% *PA* and 68.37% *mIoU* on ISPRS. These indexes significantly demonstrate that the proposed DI-Net has powerful refined land cover classification abilities. To further illustrate the DI-Net performance, several visualization results are shown in Figs. 10 and 11. From Fig. 10, we can see that DI-Net can provide refined land-cover classification results on

TABLE V
COMPUTATIONAL CONSUMPTIONS OF DIFFERENT METHODS

Methods	Time(ms)	Frame(fps)	FLOPs(G)	Params(M)
DeepLabV3+	36.76	27.20	22.17	59.34
GCN	7.35	136.05	15.50	58.94
PSPNet	147.88	6.76	63.81	65.58
U-Net	6.94	144.09	15.23	31.04
FRF	17.97	55.65	19.98	48.33
DI+DRM2	19.27	51.89	21.90	45.81
DI+DRM1	19.31	51.79	21.53	45.08

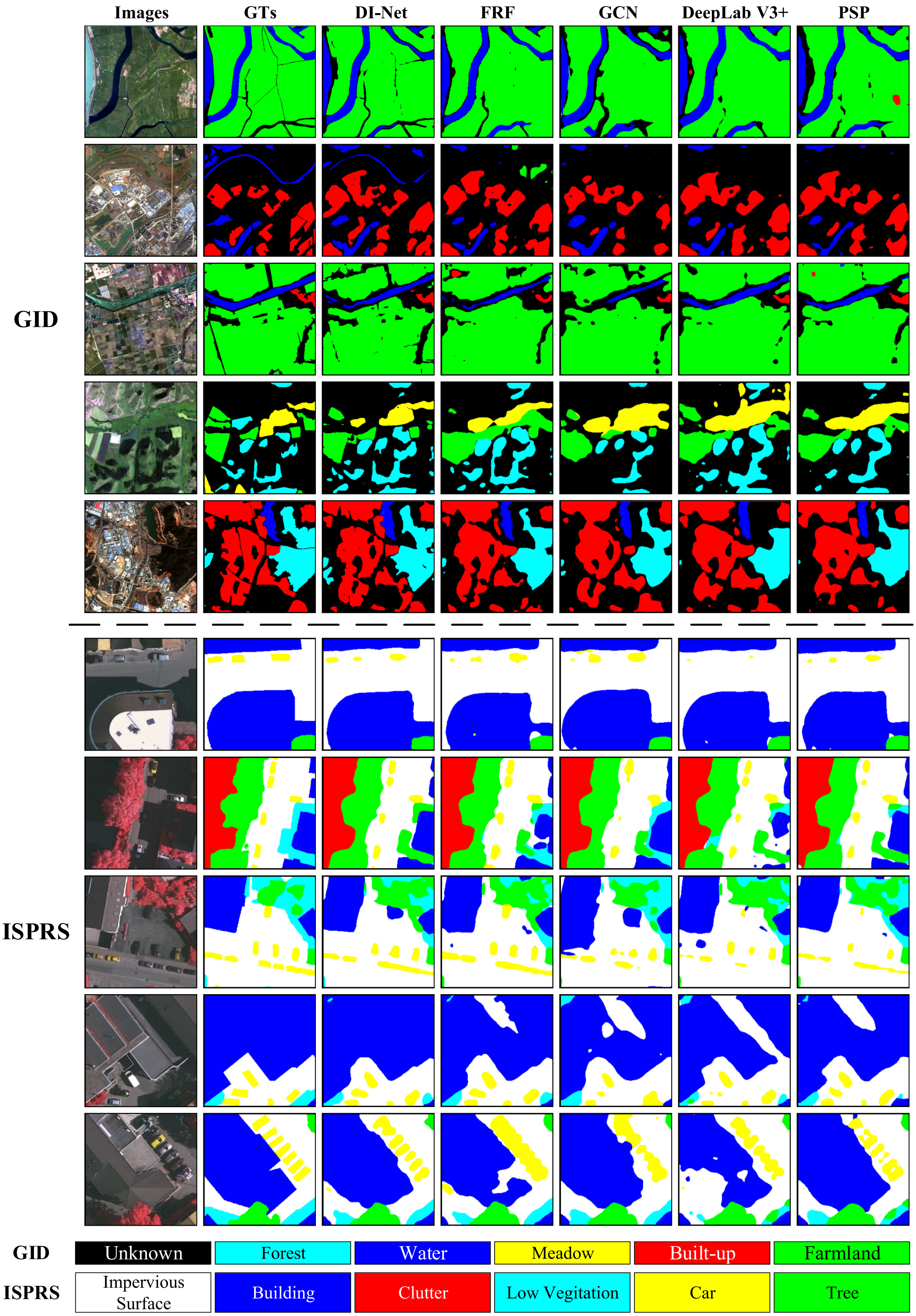


Fig. 10. Visualization results of refined land cover classification from GID and ISPRS.

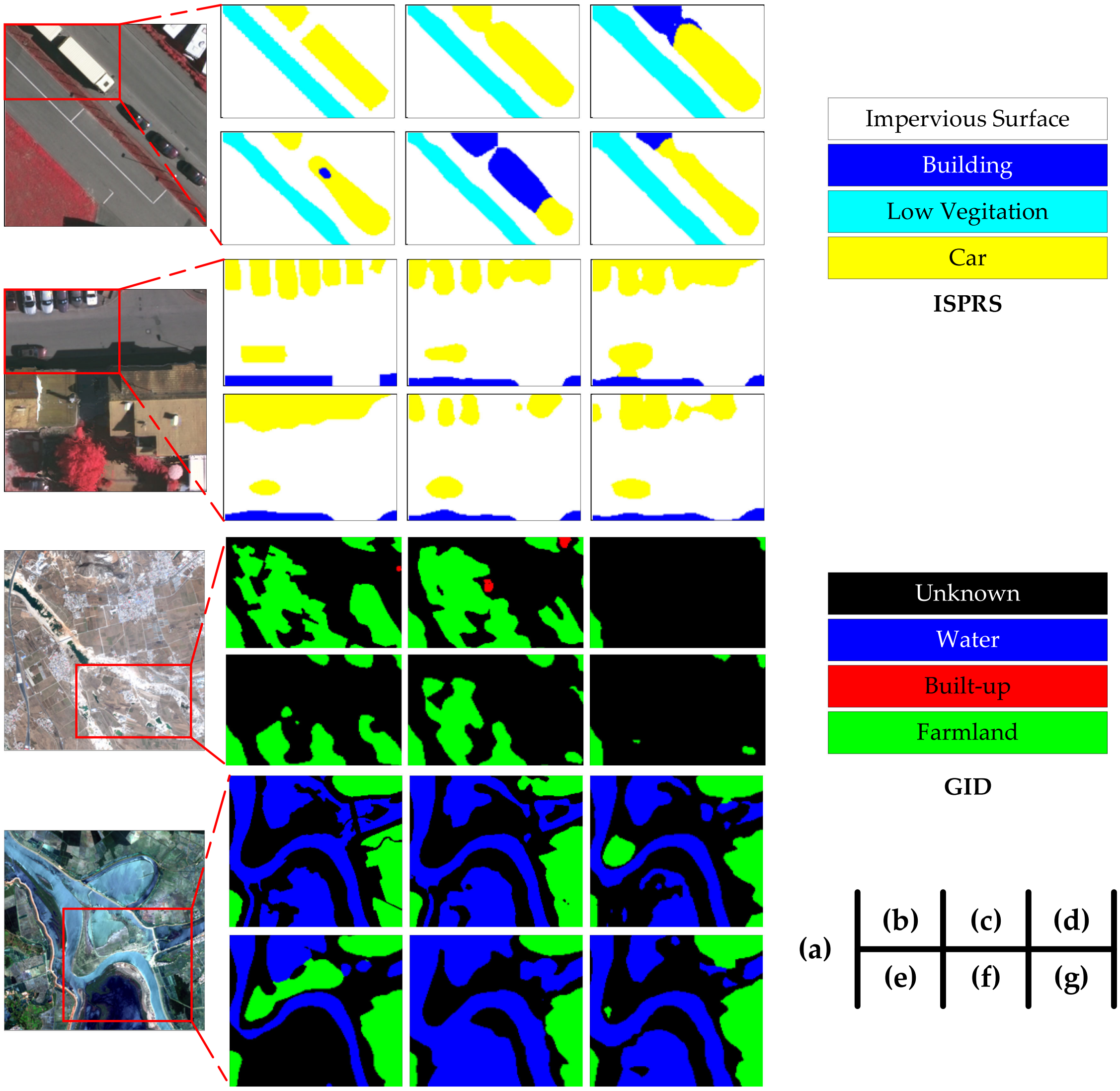


Fig. 11. Local subtle comparison on ISPRS and GID dataset. (a) Image. (b) GroundTruths. (c) DI-Net. (d) FRF-Net. (e) GCN. (f) DeepLabv3+. (g) PSPNet.

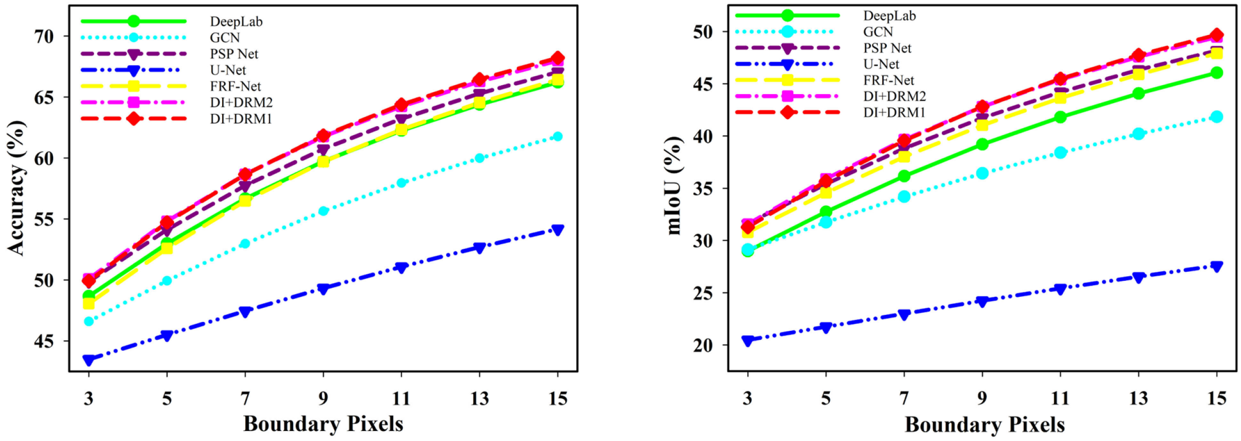


Fig. 12. Comparison on performance of boundaries classification.

GID and ISPRS, which are more similar to the ground truth than the other state-of-the-art methods. Next, the local detailed results of refined land-cover classification are shown in Fig. 11. We also come to the conclusion that the proposed DI-Net produces more accurate pixel-level predictions. For quantifying the boundary pixel predicted accuracy further, we defined the boundary area as several pixels to test DI-Net and the comparisons boundary prediction performance. In Fig. 12, the horizontal axis represents the defined number of boundary pixels, and the vertical axis represents the evaluation indexes of *accuracy* and *mIoU*. As shown in Fig. 12, DI-Net produces a better boundary prediction performance than the others with the boundary areas defined as any number of pixels. Finally, the computational complexities and the times are also evaluated, and the results are shown in Table V. Here, Res101 is employed as the backbone for DI-Net and comparisons. Based on time, frame, float-point operations (FLOPs) and the number of parameters (Params) evaluations, DI-Net has a lower computational consumption, which reduces Params and FLOPs further and achieves a better performance without a large increase in time costs as compared with the state-of-the-art methods.

V. Conclusion

For exploring the more refined land cover classification results with exquisite details (i.e., confusing boundaries and complex interior textures) from VHR optical UAVs remote sensing images, we propose a novel detail injection CNN (DI-Net). In our work, the DRM is proposed for achieving fine-scale feature extraction with multiscale representation. Then, a novel encoding attention module is proposed to address the detail misclassification caused by the feature inactivation phenomenon and efficiently capture long-range dependence for semantic generalization. Deep feature and shallow feature are adaptively fused by the proposed decoding attention module to reasonably recover detail information. Moreover, based on the designed DI-Net structure, the specific SG loss is proposed to further construct the relation constraints between encoder and decoder for the retrieval of detail information in final predicted fusion feature layers. Finally, extensive experiments are carried out on GID and ISPRS datasets, respectively, and the results show that the proposed DI-Net can achieve more refined land cover classification results than the state-of-the-art methods, especially for fine-grained details.

References

[1] C. Gómez *et al.*, “Optical remotely sensed time series data for land cover classification,” *ISPRS J. Photogramm. Remote Sens.*, vol. 115, pp. 55–72, 2016.

[2] G. Cui *et al.*, “Refining land cover classification maps based on dual-adaptive majority voting strategy for very high resolution remote sensing images,” *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1238.

[3] W. L. Stefanov *et al.*, “Monitoring urban land cover change: An expert system approach to land cover classification of semiarid to arid urban centers,” *Remote Sens. Environ.*, vol. 77, no. 2, pp. 173–185, 2001.

[4] C. J. Tucker *et al.*, “African land-cover classification using satellite data,” *Science*, vol. 227, no. 4685, pp. 369–375, 1985.

[5] K. E. Joyce *et al.*, “A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters,” *Prog. Phys. Geography*, vol. 33, no. 2, pp. 183–207, 2009.

[6] Z. Xie *et al.*, “Classification of land cover, forest, and tree species classes with ZIYuan-3 multispectral and stereo data,” *Remote Sens.*, vol. 11, no. 2, 2019, Art. no. 164.

[7] N. Joshi *et al.*, “A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring,” *Remote Sens.*, vol. 8, no. 1, 2016, Art. no. 70.

[8] W. Chen *et al.*, “A review of fine-scale land use and land cover classification in open-pit mining areas by remote sensing techniques,” *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 15.

[9] J. Jung, E. Pasolli, S. Prasad, J. C. Tilton, and M. M. Crawford, “A framework for land cover classification using discrete return LiDAR data: Adopting pseudo-waveform and hierarchical segmentation,” *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 7, no. 2, pp. 491–502, Feb. 2014.

[10] A. Krizhevsky *et al.*, “ImageNet classification with deep convolutional neural networks,” *Adv. Neural Inf. Process. Syst.*, vol. 60, pp. 1097–1105, 2012.

[11] K. Simonyan *et al.*, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Representation*, 2015.

[12] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

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

[14] L. C. Chen *et al.*, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[15] P. Chao, X. Zhang, G. Yu, G. Luo, and J. Sun, “Large kernel matters improve semantic segmentation by global convolutional network,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4353–4361.

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

[17] H. S. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[18] O. Ronneberger *et al.*, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[19] V. Badrinarayanan, A. Hanada, and R. Cipolla, “SegNet: A deep convolution encoder-decoder architecture for robust semnatic pixel-wise labelling,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2015, pp. 1–14.

[20] Y. Wang *et al.*, “A land-cover classification method of high-resolution remote sensing imagery based on convolution neural network,” in *Proc. Earth Observ. Syst. XXIII Int. Soc. Opt. Photon.*, 2018, Art. no. 107641Y.

[21] Y. Hu *et al.*, “A deep convolution neural network method for land cover mapping: A case study of Qinhuangdao, China,” *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 2053.

[22] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, “Deep learning classification of land cover and crop types using remote sensing data,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[23] L. Ma *et al.*, “Deep learning in remote sensing applications: A meta-analysis and review,” *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, 2019.

[24] E. Kroupi *et al.*, “Deep convolutional neural networks for land-cover classification with Sentinel-2 images,” *J. Appl. Remote Sens.*, vol. 13, no. 2, 2019, Art. no. 024525.

[25] L. Wan *et al.*, “A small-patched convolutional neural network for Mangrove mapping at species level using high-resolution remote-sensing image,” *Ann. GIS*, vol. 25, no. 1, pp. 45–55, 2019.

[26] Y. Liu *et al.*, “Efficient patch-wise semantic segmentation for large-scale remote sensing images,” *Sensors*, vol. 18, no. 10, 2018, Art. no. 3232.

[27] H. Lin, Z. Shi, and Z. Zou, “Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network,” *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 480.

[28] L. Mou and X. Zhu, “RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images,” 2018, *arXiv:1805.02091*.

[29] M. Wurm *et al.*, “Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks,” *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 59–69, 2019.

[30] L. Garg *et al.*, “Land use land cover classification from satellite imagery using mUnet: A modified Unet architecture,” in *Proc. 14th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2019, pp. 359–365.

[31] X. Pan *et al.*, "High-resolution aerial imagery semantic labeling with dense pyramid network," *Sensors*, vol. 18, no. 11, 2018, Art. no. 3774.

[32] H. Luo, C. Chen, L. Fang, X. Zhu, and L. Lu, "High-resolution aerial images semantic segmentation using deep fully convolutional network with channel attention mechanism," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3492–3507, Sep. 2019.

[33] M. Yuan *et al.*, "Using the wide-range attention U-Net for road segmentation," *Remote Sens. Lett.*, vol. 10, no. 5, pp. 506–515, 2019.

[34] Q. Sang, Y. Zhuang, S. Dong, G. Wang, and H. Chen, "FRF-Net: Land cover classification from large-scale VHR optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1057–1061, Jun. 2020.

[35] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[36] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9190–9200.

[37] H. Zhang *et al.*, "Self-attention generative adversarial networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

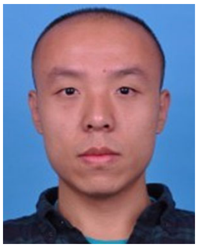
[38] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7151–7160.

[39] C. Yu *et al.*, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 334–349.

[40] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

Qianbo Sang was born in Shandong, China, in 1996. He received the B.S. degree in information engineering from Beijing Institute of Technology, Beijing, China, in 2018. He is currently working toward the M.S. degree in information and communication engineering from Beijing Institute of Technology, Beijing, China.

His research interests include remote sensing land cover classification, super resolution, and domain adaptation.



Yin Zhuang (Member, IEEE) was born in Henan, China, in 1990. He received the B.S. degree from the University of Sussex, Brighton, U.K., in 2013, and the Ph.D. degree from Beijing Institute of Technology, Beijing, China, in 2018, both in signal and information processing.

From 2018 to 2020, he was a Postdoctoral with the School of Electronic Engineering and Computer Sciences, Peking University, Beijing, China. His research interests include remote sensing target detection and recognition.

Shan Dong was born in Baoding, China, in 1992. She received the B.S. degree in electronic and information engineering from Dalian University of Technology, Dalian, China, in 2014, and the M.S. degree in electronic and communication engineering from Beijing Institute of Technology, Beijing, China, in 2016.

Her research interests include remote sensing land cover classification and recognition.

Guanqun Wang (Student Member, IEEE) was born in Heilongjiang, China, in 1995. He received the B.S. degree in information and communication engineering from Beijing Institute of Technology, Beijing, China, in 2017. He is currently working toward the Ph.D. degree in signal and information processing at Beijing Institute of Technology.

His research interests include object detection and recognition in remote sensing images.

He Chen was born in Shenyang, China, in 1970. She received the Ph.D. degree in electronic engineering from Harbin Institute of Technology, Harbin, China, in 1998.

She is currently a Professor and Dean with School of Information, Beijing Institute of Technology, Beijing, China. Her research interests include system-on-chip design, remote sensing data intelligent processing, VLSI architectures for real-time image, and signal processing.

Lianlin Li (Senior Member, IEEE) was born in Shan'xi, China, in 1980. He received the Ph.D. degree in electrical engineering from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2006.

Since July 2012, he has been with the School of Electronic Engineering and Computer Sciences, Peking University, Beijing, China, as a Hundred Talented Program Professor. He has authored over 80 peer-reviewed journal papers in the *Nature Communications, Advanced Science*, and a series of the IEEE.

Dr. Li serves as a Guest Editor of the special issue on radar imaging and detection for concealed target in *Journal of Radar* (Chinese). He serves as a Regular Reviewer for the multiple journals such as IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, *Inverse Problem, Chinese Physics Letters*, and so on. He was the recipient of the Excellent Doctoral Dissertation Award from the Chinese Academy of Sciences, in 2008 and the National Excellent Doctoral Dissertation Nomination Award, in 2009, and the URSI Young Scientist Award, in 2011 and 2014. He will serve as a Technical Program Committee Member and a Session Chair for URSI Atlantic Radio Science Conference, in 2015. He has been a Senior Member of the Institute of Electronics (China), since 2014.