Capturing Small Objects and Edges Information for Cross-Sensor and Cross-Region Land Cover Semantic Segmentation in Arid Areas

Panli Yuan, Qingzhan Zhao, Yuchen Zheng, *Member, IEEE*, Xuewen Wang, and Bin Hu

***Abstract*—In the oasis area adjacent to the desert, there is more complex land cover information with rich details, multiscales of interest objects, and blur edge information, which poses some challenges to the semantic segmentation task in remote sensing images (RSIs). In traditional semantic segmentation methods, detailed spatial information is more likely lost in feature extraction stage and the global context information is more effectively integrated into segmentation results. To overcome these land cover semantic segmentation model, FPN_PSA_DLV3+ network, is proposed in an encoder–decoder manner capturing more fine edge and small objects information in RSIs. In the encoder stage, the improved atrous spatial pyramid pooling module extracts the multiscale features, especially small-scale feature details; feature pyramid network (FPN) module realizes better integration of detailed information and semantic information; and the spatial context information at both global and local levels is enhanced by introducing polarized self-attention (PSA) module. For the decoder stage, the FPN_PSA_DLV3+ network further adds a feature fusion branch to concatenate more low-level features. We select Landsat5/7/8 satellite RSIs from the areas of north and south of Xinjiang. Then, three self-annotated time-series datasets with more small objects and fine edges information are constructed by data augmentation. The experimental results show that the proposed method improves the segmentation performance of small targets and edges, and the classification performance increases from 81.55% to 83.10% *F*1 score and from 72.65% to 74.82% mean intersection over union only using red–green–blue bands. Meanwhile, the FPN_PSA_DLV3+ network shows great generalization in cross region and cross sensor.**

***Index Terms*—Convolutional neural networks (CNNs), fine edge, remote sensing image (RSI), semantic segmentation, small objects.**

I. Introduction

The oasis–desert mosaic belt, playing the boundary role between desert and oasis ecosystems, is a representative landscape in the arid areas of northwestern China, such as Xinjiang [1]. Sustainable agriculture development and sustainability of the ecosystem affected by the desertification and urban expansion of the oasis–desert mosaic belt [2]. Therefore, we urgently need to grasp the spatial distribution and dynamic change information of land cover in an arid area, which could effectively improve economic and ecological benefits and propel the national ecological barrier construction.

With the significant development of remote sensing (RS) technology, there are vast volumes of remote sensing images (RSIs) with different resolutions available for earth observation, which are mutually constrained by the spectral resolution, spatial resolution, and richness of historical data [3]. Landsat-like satellite images have a tradeoff with the three aspects, which have a long duration, wide coverage, and repetitive capability and are beneficial to widespread applications, including land cover mapping, urban planning, land resource management, environmental monitoring, and other fields [4], [5].

Scholars have tried many approaches to RSIs classification in recent years, and there are some limitations. The traditional machine learning (ML) methods only use the texture and spectral information to classify sandy land [6] and other land cover information of oasis [7], which are prone to produce unsatisfactory classification results due to their inferior capabilities of feature extraction [8]. Object-oriented methods use the spatial and spectral characteristics of the object to classify but suffer from problems, such as insufficient generalization of classification rules and manual participation in segmentation parameters adjustment [9]. The last decade has witnessed the rapid advance of deep learning [10] in the computer vision field, especially the fundamental semantic segmentation task [11], [12]. Many improved models based on fully convolutional network (FCN) [13] have appeared one after another and successfully applied to RSIs semantic segmentation surpassing traditional methods with a large margin [14]. However, the existing convolutional neural network (CNN) based methods lose more detailed information after continuous convolution operation and are difficult to segment the small object. Meanwhile, the current segmentation methods only obtain the main smooth contour of the object instead of a sharp edge. The extraction of small objects

Manuscript received 4 March 2022; revised 30 June 2022 and 3 November 2022; accepted 23 December 2022. Date of publication 27 December 2022; date of current version 9 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 32260388, in part by the National Key Research and Development Program of China under Grant 2017YFB0504203, and in part by Xinjiang Production and Construction Corps Science and Technology Project under Grant 2017DB005. *(Corresponding authors: Qingzhan Zhao; Yuchen Zheng.)*

Panli Yuan, Qingzhan Zhao, Yuchen Zheng, and Bin Hu are with the College of Computer Science and Technology, Shihezi University, Shihezi 832003, China (e-mail: yuanpanli@stu.shzu.edu.cn; zqz_inf@shzu.edu.cn; ouczyc@outlook.com; ll@stu.shzu.edu.cn).

Xuewen Wang is with the School of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China (e-mail: demons_wxw@outlook.com).

Digital Object Identifier 10.1109/JSTARS.2022.3232409

and edge information is hindered by the similarity of spectral response, textures, sharpen, and blur edge. Some methods are low automation and features require to be manually designed [15], [16].

As an alternative, the encoder–decoder design is proposed to directly aggregate multilevel semantic features and more spatial information [17], [18]. UNet [19] and RefineNet [20] have a similar architecture with multipath skip connections to reuse multilevel feature maps. The DenseU-Net [21] extracts feature by continuous downsampling blocks and upsampling blocks to restore the spatial information of the RSIs. For those FCN-based methods, a large amount of spatial detail information is lost in the process of features transmitted from encoder to decoder and does not be restored well by decoder. Furthermore, some models enhance the ability to integrate contextual information in network structure, such as aggregating local features with decreasing dilation factor in atrous convolution [22], introducing a context encoding module to highlight the class-independent feature [23], and proposing a parallel pooling design to integrate the contextual information [24]. The atrous spatial pyramid pooling (ASPP) combining pyramid pooling module and atrous convolution [25] and applies in DeepLabV3+ [26] and Denseaspp [27]. In another way, the feature pyramid network (FPN) [28] combines strong semantic information and rich spatial information to solve the different sizes' objects segmentation problem. The attention mechanism also improves the feature extraction capability of a model by strengthening the pertinence of learning [29]. It applies learned weights to features to further extract relevant information and has better service for semantic segmentation [30], [31], [32]. However, most methods are fit for the specific case study, while they are not appropriate in other cases [33]. And the improvement of classification accuracy is limited in RSIs with complex backgrounds.

Among the above CNN-based models, global dependencies, multiscale features fusion, and attention mechanisms design bring great effect to the segmentation of small objects and fine edges. How to take advantage of the strength of these modules to establish a CNN-based network with the generalized property for refined edge and small objects segmentation is a problem worthy of exploration.

Otherwise, an effective CNN-based classification model relies heavily on the well-annotated training data as training materials. Some datasets have very high resolution but only several classes without many small objects and more detailed edge information [34], [35]. For the lack of large-scale public classification datasets, including complex land covers, such as the scattered build-up and highly fragmented landscape without clear-cut border in the arid area [36], [37], it is significant to build such dataset as data support to solve the problem of the semantic segmentation of small objects and edges, such as the oasis adjacent to the desert.

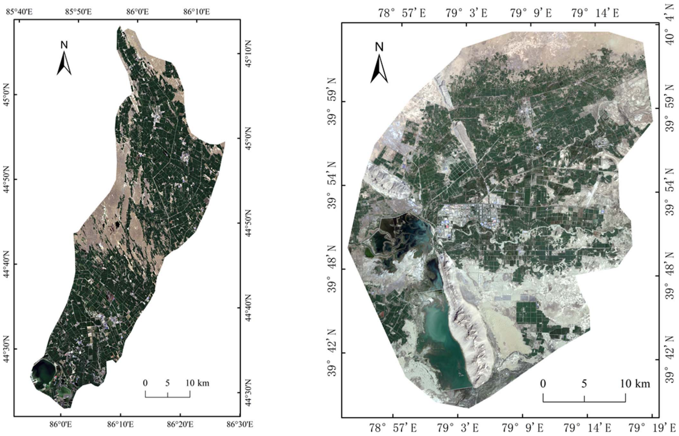
In this article, first, to monitor the desertification and urban expansion of the oasis adjacent to the desert, we select two typical arid study areas in the north and south of Xinjiang and choose long-time-series Landsat-like satellite images as the data source. Then, three self-annotation datasets, Mosuowan Landsat8 OLI Dataset (denoted as MSW_LCD), Tumushuke Landsat8 OLI Dataset (denoted as TMSK_LCD), and Tumushuke Landsat7 ETM+ and Landsat5 TM Dataset (denoted as TMSK_LELTD), are constructed. Limited by the impact of the natural environment, the edges bordering the desert are irregular and residential land is small and scattered. Therefore, there are many small objects (build-up) and complex edge information (the boundary of desert and farmland) in these three datasets, while the existing semantic segmentation datasets are more used to annotate large-scale features, and the edge information is not rich enough [38]. According to the definition of the International Society for Optical Engineering, the size of a small object is the size of the object, which is less than 0.15% of the original image [39]. Based on this definition, we will make the land cover with less than $416 \times 416 \times 0.15\%$ of pixels as the small object land cover in the subimage with the size of 416×416 pixels.

At the same time, to better solve the problem of complex land covers classification in arid areas, such as irregular edges and many scattered small objects, we propose a novel encoder–decoder semantic segmentation architecture: the FPN_PSADLV3+ network. We make targeted improvements to the model in both the encoder and decoder stages to discriminate small objects and fine parts of objects. For the encoder, the FPN_PSADLV3+ network refines multiscales semantic information via three blocks. Specifically, the FPN module makes full use of multistage feature maps from the backbone, the polarized self-attention (PSA) block [40] enhances the ability to capture semantic information, and the improved ASPP composition verify field-of-views to capture multiscale features. For the decoder, three feature fusion branches come from the encoder and the upsampling rules are improved to better fuse low-level feature information. Finally, we fine tune the model on the TMSK_LCD dataset to verify the across-region generalization of the model and on the TMSK_LELTD dataset to verify the cross-region and cross-sensor (denoted as cross-region–sensor) generalization.

The main contributions of this article are listed as follows.

1) We manually establish multisensor time-series land cover classification datasets containing many small object and edge complex land covers, namely the MSW_LCD, TMSK_LCD, and TMSK_LELTD datasets for model training, validation, and testing, which allows us to continuously monitor land surface and observe long-term regional or global changes.
2) A novel semantic segmentation network, the FPN-PSADLV3+ network, is proposed to capture small objects and edges information for semantic segmentation of RSIs. The FPN-PSADLV3+ network introduces the FPN module to fuse more detailed information, leverages the improved ASPP composition for multiscale feature maps, and imports the PSA block to capture more long-range dependency to obtain more discriminant feature representation and learn more relevant features along channel and spatial dimension improving the learning efficiency.

The rest of this article is organized as follows. Section II describes the proposed datasets in detail. Section III presents the design and architecture of the proposed FPN_PSADLV3+ network. Section IV reports the experimental results. Section V discusses the performance of the FPN_PSADLV3+ network. Finally, Section VI concludes this article.



(a) (b)

Fig. 1. Location of the study areas. (a) Mosuowan reclamation, North Xinjiang. (b) Tumushuke region, Southern Xinjiang.

II. Dataset Description and Preprocessing

We establish three datasets for land cover classification. Section II-A describes the details of two study areas. Section II-B introduces information about the source data. Section II-C presents in detail how we build the datasets and the characteristics of the datasets.

A. Study Area

There are two key points that need to decide are which geographic area to include and what types of land covers to be classified before constructing a dataset. Specifically, as the focus is on accurately the current status of land use in arid areas, we selected two interest regions of the Tumushuke region, Southern Xinjiang (39°39′N - 40°4′N, 78°53′E - 79°19′E, denoted as TMSK, see Fig. 1(a)) and Mosuowan Reclamation Area, North Xinjiang (44°23′N -45°12′N, 85 52′E - 86°19′E, denoted as MSW, see Fig. 1(b)), as the study areas, which are adjacent to the desert. And the representative types of land cover we classify are farmland, build-up, bare land, and water. Compared with the MSW region in northern Xinjiang, the TMSK region has the characteristics of decentralized and residential houses that scatter on agricultural land and urbanization development, which is relatively slow before 2013. Therefore, it is of significance to grasp the land resource utilization, urban development status, and ecological situation by mastering the distribution of land covers in the TMSK region of southern Xinjiang.

B. Data Collection and Preparation

We choose Landsat-like satellite imagery as the source data for the classification of four land covers, namely farmland, bare land, build-up, and water. What needs to be explained is that the spectral and texture characteristics of the four land covers are easier to distinguish from each other during the peak growth period (May to October of a year). The short satellite revisit period (16 days) increases the amount of candidate data. The satellite images are downloaded from the geospatial data cloud and the United States Geological Survey, and the main parameters of availability images are provided in Table I.

The radiation calibration and the atmospheric correction are the preconditions of quantitative RS study on the satellite RSI acquired by different sensors at different times. The Landsat7 ETM+ image data from 2003 to 2005 have striped data loss. With the help of the ENVI landsat_gapfill extension tool, the pixel space interpolation repair is completed. The spectral bands to build up the three datasets are the blue (0.450–0.515 μm) band, green (0.525–0.600 μm) band, and red (0.630–0.680 μm) band. The spatial dimensions are 1133 × 2874 pixels and 1237 × 1527 pixels in the TMSK and MSW region, respectively, after cutting from interest regions of the study area. Rich source data facilitate us to build the multisensor time-series satellite datasets and provide data support for subsequent work.

C. Dataset Construction

The MSW_LCD dataset is imaged by Landsat8 OLI from Mosuowan reclamation area; the TMSK_LCD is imaged by Landsat8 OLI from Tumushuke region; and the TMSK_LELTD dataset is imaged by Landsat5 TM and Landsat7 ETM+ from Tumushuke region. On May 16, 2021 and August 8, 2021, we carry out field data collection in the MSW and TMSK regions, mainly collecting the land covers corresponding to GPS information. Combining field investigations data and satellite imagery, these three datasets have complex foreground information, multiscale interest objects, and blur edges between desert and farmland. Moreover, some of these land covers may have similar spectral responses, rich texture features, and irregular geometric structures, which further increases the within-class variability and decreases separability between different land covers. Therefore, our datasets are more characteristic of typical arid areas. Here, we analyze the spectral information and texture characteristics of the images and summarize the following interpretation signs.

1) The build-up land includes towns, villages, industrial building land, etc. Build-up land is compact and massive and mostly in the form of blocks and strips in the MSW and TMSK regions after 2013, while scattering in the middle of the agricultural land TMSK before 2013. The spectral of images is mostly mixed with white, red, and blue.
2) The bare land includes gravel Gobi, barren land, bare rock, sandy area, etc. The spectral of the desert image is brown and yellow [see Fig. 2(d1), (e2), and (d2)], with complex and irregular edges at the borders of farmland and desert [see Fig. 2(f1), (e2), and (f2)].
3) The farmland includes cultivated land, shrubland, forested land, grassland, etc. The spectral and shape properties of the farmland images in Fig. 2(a1)–(c1) and (a2)–(c2) are completely diverse. Some are large continuous rectangular, while others are irregular with dark green color.
4) The water includes lakes, artificial lakes, reservoirs, wetlands, and other waterbodies. Waterbody spectral is green with different shades [see Fig. 2(j1), (j2), and (l2)]. The distribution area is relatively fixed, the boundary is more obvious, but the boundary of the small water bodies is

TABLE I
AVAILABILITY OF LANDSAT-LIKE IMAGERY IN THE MSW AND TMSK REGIONS

Satellite/ Sensor	Study Area	WRS Path/Row	Spatial Resolution	Temporal Resolution	Data range/Year	Dataset Name	Cloud cover (%)
Landsat5 TM	TMSK	147/32	30 m	16 d	1990,1992–1994, 1996–1998,2006–2011	TMSK_LELTD	0.00–14.00
Landsat7 ETM+	TMSK	147/32	30 m	16 d	1999–2005		0.00–6.00
Landsat8 OLI	TMSK	147/32	30 m	16 d	2013–2021	TMSK_LCD	0.09–6.8
	MSW	144/29			2013–2021	MSW_LCD	0.00–14.45

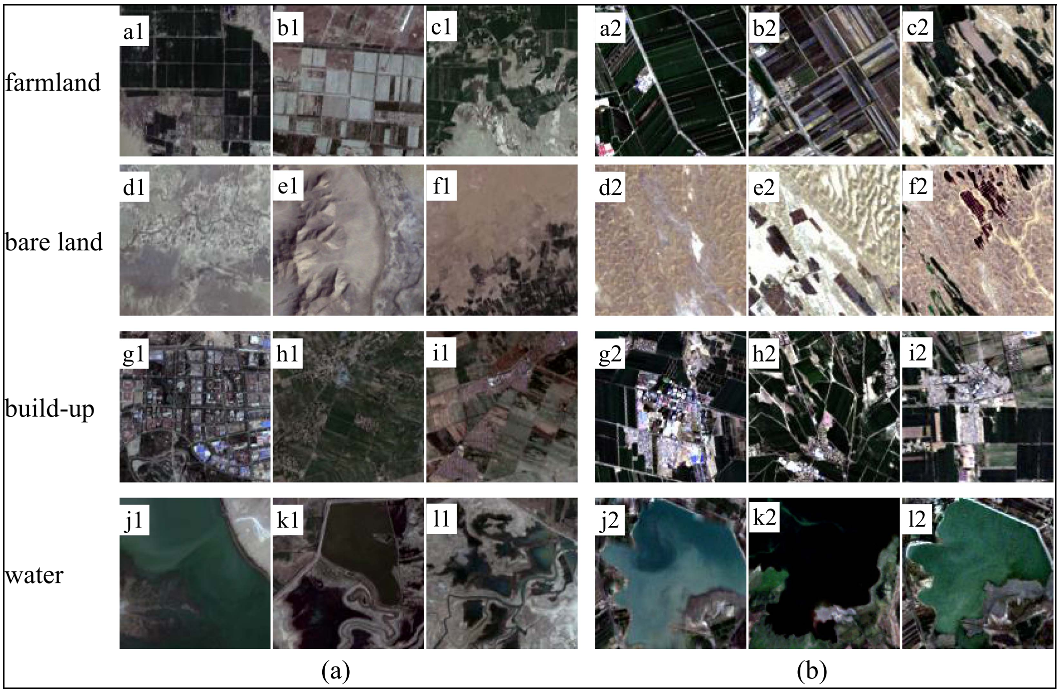


Fig. 2. Sample images from the TMSK and MSW datasets. (a) Tumushuke region. (b) Mosuowan area.

not clear with complicated shape [see Fig. 2(k1), (l1), and (k2)].

The performance of RSI classification based on deep learning relies heavily on the well-annotated large-scale training data [40]. The data input to the semantic segmentation model is cut to 416 × 416 pixels without overlapping from each other, and the blank image blocks are filtered out. Five representative land cover types are annotated: farmland, bare land, build-up, water, and background. Precise pixel-level data annotation adopts a visual interpretation method combined with high-resolution images from google earth pro and field investigations information using the LabelMe deep learning annotation tool. And then converting it into a color-coded (red, green, and blue) map with a shape of 416 × 416×3. Fig. 3 shows some annotation results.

We apply several efficient data augmentation methods to increase the number of samples, as shown in Fig. 4. The rotation and random crop resize are mimic the bird's eye view of the satellite when shooting land covers. And the Gaussian blur and random erasing are used to simulate the interference of the cloud coverage on the image quality. In all, the number of images on the MSW_LCD, TMSK_LCD, and TMSK_LELTD datasets is 4342, 2429, and 6161, respectively.

Table II presents the distribution of land covers and the proportion of small object parcels in the three datasets, and the proposed dataset has a high proportion of small targets in farmland, bare land, and build-up. The distribution ratios of water and build-up on the three datasets are similar, and bare land and farmland land are slightly different. This is related to the geographical environment of southern and northern Xinjiang. The high percentage of small objects and complex edge information of land covers will be a challenge for the exploration of land cover extraction models. The MSW_LCD, TMSK_LCD, and TMSK_LELTD datasets are available at figshare.com/articles/figure/LandCoverClassificationDataset/21445518.

III. FPN_PSA_DLV3+ NETWORK

This section presents the proposed method, FPN_PSA_DLV3+ architecture (see Section III-A), for land

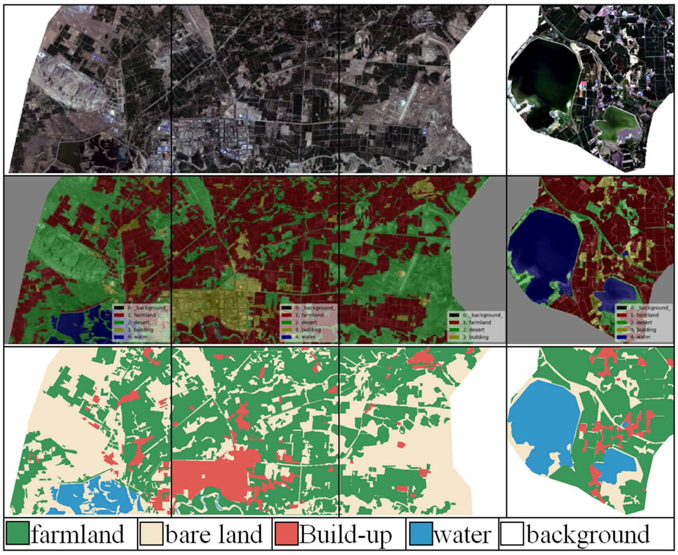


Fig. 3. Examples of images and their corresponding annotation masks. The first row is the original image, the second row is the mask of annotation and original image, and the third row is the visualization of labels.

TABLE II
AMOUNT OF DATA DISTRIBUTED PER LAND COVER ON EACH DATASET

Classes Name	Dataset Name	Proportion Count	Small Object Percentage
farmland	MSW_LCD	0.4466	0.8981
	TMSK_LCD	0.2700	0.8584
	TMSK_LELTD	0.1768	0.8518
bare land	MSW_LCD	0.1260	0.9461
	TMSK_LCD	0.4106	0.9291
	TMSK_LELTD	0.5145	0.9217
build-up	MSW_LCD	0.0298	0.7449
	TMSK_LCD	0.0330	0.9380
	TMSK_LELTD	0.0174	0.9456
water	MSW_LCD	0.0178	0.2661
	TMSK_LCD	0.0509	0.7079
	TMSK_LELTD	0.0472	0.5274

cover semantic segmentation of RSIs, which takes advantage of multistage feature maps from the backbone to integrate more detailed information with the aid of the FPN module (see Section III-B), incorporates multiscale contextual information that make the trained model robust to scale variations using the improved ASPP module (see Section III-C), and leverages PSA to improve the learning efficiency (see Section III-D).

A. Overall Model Architecture

We propose an effective encoder–decoder network, the FPN_PSA_DLV3+ network for capturing more small objects and finer edge detailed information (see Fig. 5). The FPN_PSA_DLV3+ network is based on the DeepLabv3+ [26], which is the latest improved version of the DeepLab series of networks [25], [26], [42]. We find that DeepLabv3+ may not successfully recover small objects and finer edge detailed information by applying continuous two-dimensional convolution operation of the encoder and two quadruple upsampling and pooling operations of the decoder. Some improvements are made to the DeepLabv3+ in capturing contextual information and mitigating the loss of detailed information and multiscale feature extraction.

Encoder: We take Resnet50 as the feature extractor and use the feature maps from the output of the last residual block layer of each stage to get more low-level information. Input tensor I and output tensor O of the FPN_PSA_DLV3+ network have the same dimension $C \times H \times W$. Conv1, conv2, conv3, and conv4 represent different residual blocks of Resnet50. The original size of input feature maps is denoted as f. And $\{1/2f, 1/4f, 1/8f$, and $1/16f\}$, respectively, represent the outputs of the conv1, conv2, conv3, and conv4. To preserve initial global information, the output feature maps from conv2 are sent to the decoder as low-level features instead of conv1 due to its large memory footprint. Moreover, the coarser-resolution maps from conv4 are upsampled by a factor of 2. Then, the maps are merged with maps from conv3 by elementwise addition. The generated finer-resolution maps with more detailed information, $1/8f$, are also sent to the decoder. Simultaneously, ASPP probes the incoming $1/16f$ maps from conv4 with atrous convolutions and pooling operations at multiple field-of-views to capture features at different scales. The resulting features are concatenated and pass through the parallel layout PSA module. The PSA model generates channel weighting and spatial weighting for the fed feature maps, respectively. The channel weighting is used to estimate the class-specific output scores and the spatial weighting is used to detect pixels of the same semantics. Thus, the PSA module is introduced to this position to highlight features for the above goals. After the PSA module, we apply 1 × 1 convolution operation with 256 filters to reduce the number of channels and then send this $1/16f$ high-level features to the encoder.

Decoder: Compared with the original DeepLabv3+, the FPN_PSA_DLV3+ adds a feature fusion branch to make full use of the feature maps of the backbone. Therefore, the ability of capturing semantic information is enhanced from three feature fusion branches: the first branch is from the conv2 of the backbone; the second is from the output of the FPN module; and the third branch is from the output of the ASPP5 and 1 × 1 convolution. In addition, the same semantics are highly nonlinear in nature. If the feature maps from the third branch are directly upsampled by a factor of 4, the semantic information cannot be effectively recovered, especially for the small objects and object parts. For the fine-granted segmentation, we replace bilinearly upsampling operation by a factor of 4 with two bilinearly upsampling operation by a factor of 2. And after the first upsampling operation, the second branch from FPN of the feature information is concatenated. The concatenated maps subsequently pass through the second upsampling operation by a factor of 2. After the feature fusion of the second branch and the third branch, we obtain $1/4f$ feature maps. The output feature maps concatenate the first branch feature maps from the conv2. After that, we apply a 3 × 3 convolution for features refinement followed by a simple bilinear upsampling operation by a factor of 4. Here, the upsampling operation is not replaced with two

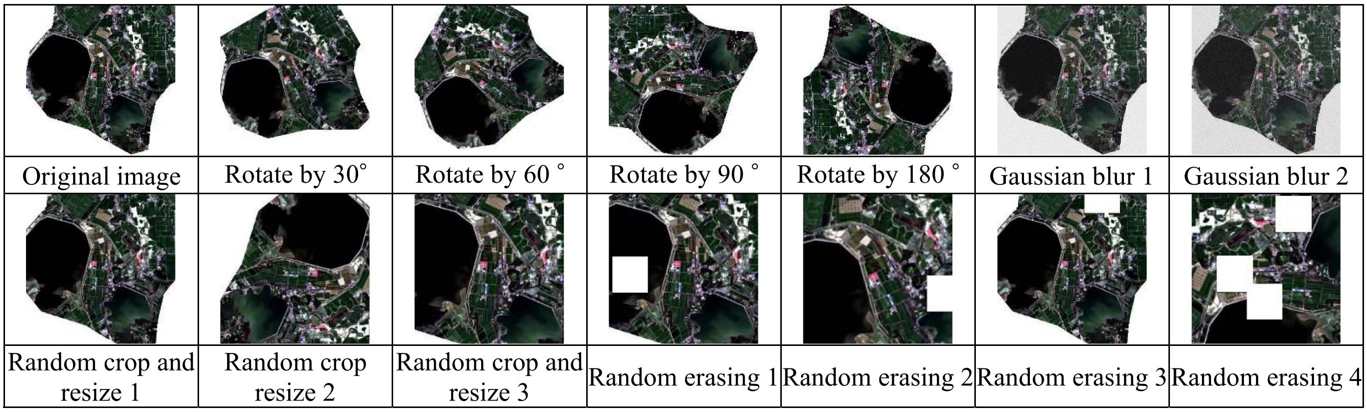

Original image
Rotate by 30°
Rotate by 60 °
Rotate by 90 °
Rotate by 180 °
Gaussian blur 1
Gaussian blur 2
Random crop and resize 1
Random crop resize 2
Random crop and resize 3
Random erasing 1
Random erasing 2
Random erasing 3
Random erasing 4

Fig. 4. Sample images after different data augmentation methods.

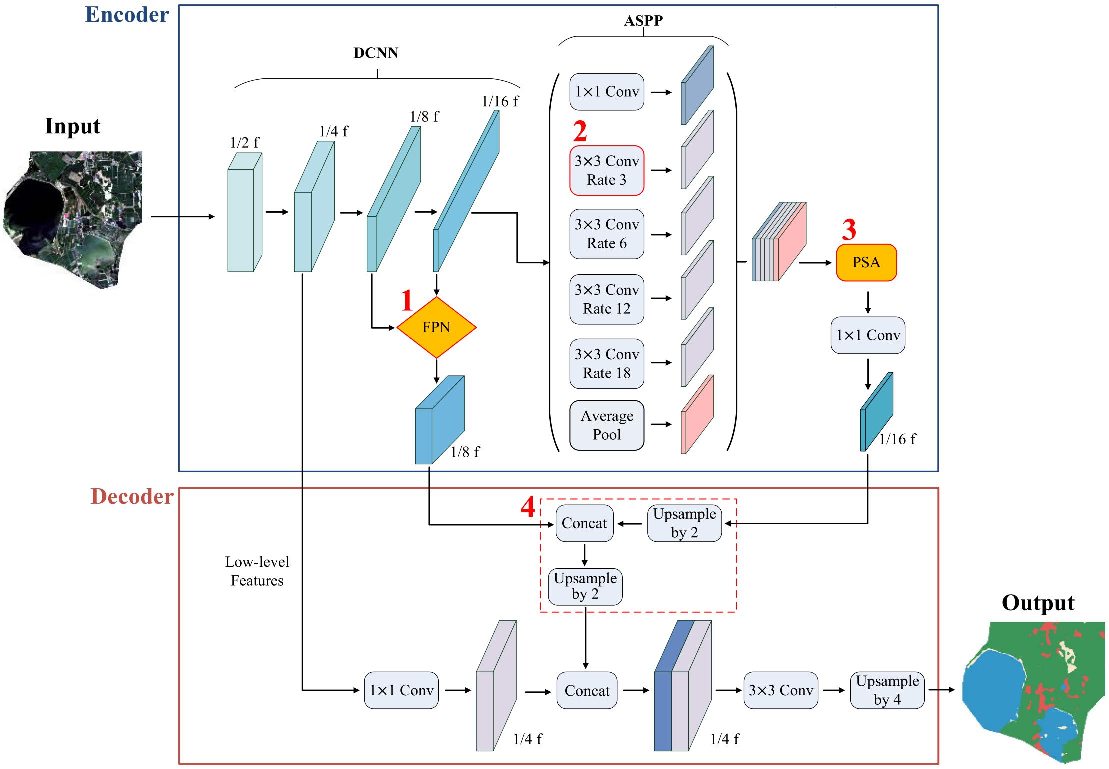


Fig. 5. Architecture of the FPN_PASDLV3+ network. PSA denotes the polarized self-attention block. FPN denotes the feature pyramid network module.

bilinear upsampling operation by a factor of 2 because there is no more detailed information that needs to be integrated into the network.

B. FPN Module

As an efficient available way to compute a multiscale feature representation, the FPN module creates a feature pyramid that alleviates large semantic gaps between low-level and high-level features. As shown in Fig. 6, the input of the FPN model comes from two parts: the finer-resolution and the coarser-resolution feature maps. And the lateral connection associates finer-resolution low-level feature maps across coarser-resolution semantic maps. First, the finer-resolution feature maps pass through a 1 × 1 convolution and add those maps with coarser-resolution feature maps after upsampling by a factor of 2. Afterward, fusion feature maps from different levels are obtained. Feature activation maps from the FPN module obtain richer semantic information and spatial information and effectively improve the performance of the whole network.

C. Improved ASPP Module

The ASPP module captures multiscale objects and context information. It is crucial to select proper atrous rates for different segmentation objects [43]. Considering that the goal of this

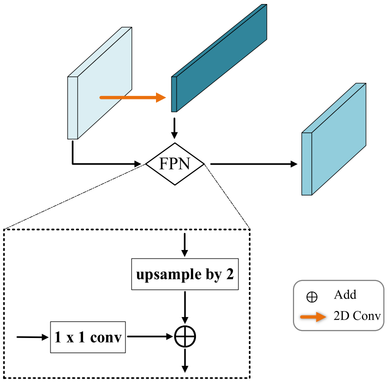


Fig. 6. Building block of FPN module.

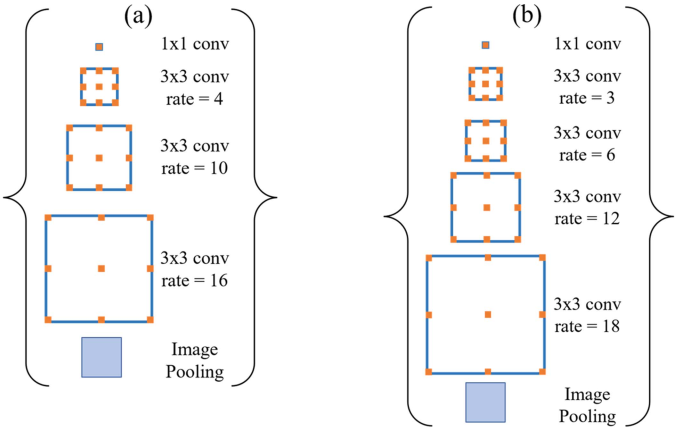


Fig. 7. Schematic diagram of the improved ASPP.

article is to capture more effective information on small objects and object parts, we improve the original ASPP by changing the size of the atrous rates or adding another atrous convolutional layer according to the characteristics of small objects, as shown in Fig. 7. On the one hand, we adjust the atrous rates of ASPP by applying the atrous rates with $\{1, 4, 10, 16\}$ (denoted as ASPP4 module). Compared with the original ASPP module, the overall atrous rates are reduced, and more detailed features are extracted without affecting the fusion of context information. On the other hand, we add a 3×3 convolution with an atrous rate of 3 with a smaller receptive field and get the ASPP5 module. ASPP5 module is a range of atrous convolutions with rates $= \{1, 3, 6, 12, 18\}$. There is a hybrid dilated convolution layers of different rates so that the convolution calculation could cover the entire feature map. To a certain degree, introducing the ASPP5 module to the proposed network aggregates multiscale information without losing resolution information and alleviates the gridding effect.

D. Polarized Self-Attention Block

PSA block is first proposed by He et al. [46], as shown in Fig. 8; PSA block is constructed for boosting long-range feature interactions, which is made up of two branches: the channel-only attention branch and the spatial-only attention branch (see Fig. 8). Specifically, the channel-only attention branch generates an attention score of $C \times 1{\times}1$ in the channel dimension, and the spatial-only attention branch generates an attention score of $1 \times H \times W$ in the spatial dimension. And then the attention scores dot product the input feature map to get the outputs O_c and O_s, respectively. The two branches greatly enhance the discriminant ability of long-range contextual feature representations in the channel and spatial dimension. There are two ways to combine the outputs of the two branches, the parallel layout PSA module [denoted as PSA (P), Fig. 8(c)] and the sequentially layout PSA module [denoted as PSA (S), Fig. 8(d)]. The PSA module preserves the highest attention resolution for both the channel of $C/2$ and spatial dimension of $[W, H]$ to

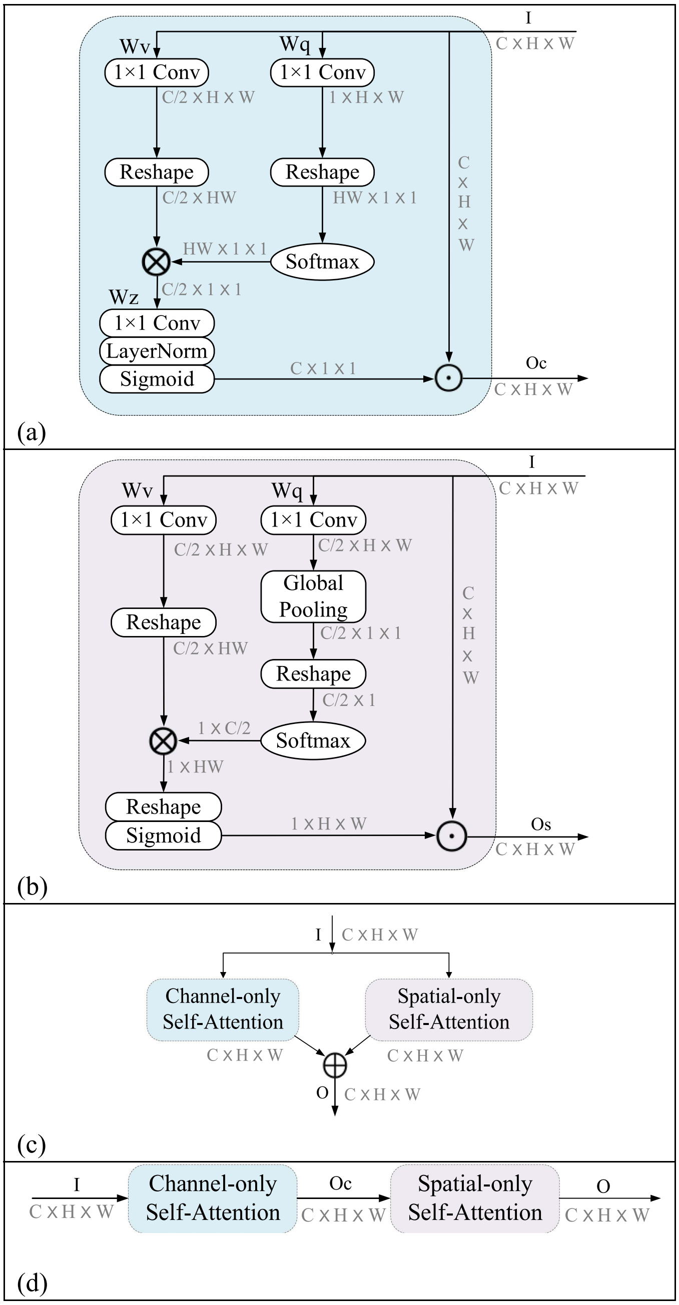


Fig. 8. Architecture of PSA block. (a) Channel-only attention block. (b) Spatial-only attention block. (c) Parallel layout PSA module. (d) Sequentially layout PSA module.

TABLE III
NUMBER OF SUBIMAGES IN TRAINING SET, VALIDATION SET, AND TEST SET

Dataset Name	training set	validation set	test set
MSW_LCD	2,288	1,524	530
TMSK_LCD	1,313	874	242
TMSK_LELTD	3,327	2,218	616

TABLE IV
PERFORMANCE OF DIFFERENT MODELS ON THE MSW_LCD DATASET

Method	OA (%)	Kappa (%)	Precision (%)	F1 (%)	MIoU (%)
PsPNet	80.17	53.86	68.64	62.29	51.57
SegNet	94.63	87.25	80.13	78.82	69.34
UNet	95.32	88.66	82.82	81.13	72.20
DeepLabv3+	**95.57**	**89.33**	**82.55**	**81.50**	**72.45**

Bold numbers are the best results.

output high-resolution semantic information. And the PSA is currently the only self-attention block that combinates nonlinear functions of SoftMax and Sigmoid to increase the dynamic range of attention and fit output distribution. Introducing the PSA module improves the learning efficiency and the effect of model classification.

IV. EXPERIMENTAL RESULTS

In this section, we first describe the experimental settings and evaluation indicators in Section IV-A. In Section IV-B, a range of popular CNN networks used in RSI classification tasks is compared in terms of classification accuracy and convergence speed. We also explore the feature extraction capabilities of different backbones on the DeepLabv3+ architecture concentrating on the classification performance gain (see Section IV-C). We then perform some ablation study where we add or replace some modules in the DeepLabv3+ architecture (see Section IV-D). Finally, we demonstrate the generalization of the FPN_PSADLV3+ architecture by focusing on cross-region and cross-sensor RS classifications (see Section IV-E).

A. *Experimental Settings and Measurement*

We train all models for 100 epochs with a batch size of 3, using Adam optimizer with a cross-entropy loss function. We set the initial learning rate to 0.001 and employ an equal interval adjustment policy with the patience of 8 and gamma of 0.5 to attenuate the learning rate gradually. One point that needs to be emphasized is that we fine tune the model with a small initial learning rate of 0.0005 on the evaluation of generalization of cross-region and cross-sensor experiments. In the process of model fine tuning, we adopt a pretrained model on the MSW_LCD dataset because a proper initialization contributes to the performance of the network and some common feature information is learned from other datasets. Our implementation is on a single NVIDIA RTX 2080Ti, Keras platform. We give the detailed training set, verification set, and test set distributions of the three datasets, as shown in Table III.

Several different metrics are used to evaluate the segmentation capability of the proposed method. The accuracy evaluation indicators include overall accuracy (OA), precision, Kappa coefficient, $F1$ score, and Mean Intersection over Union (MIoU). The recall calculates the ratio of correctly classified pixels of one land cover to the true total number of those land cover pixels in the predicted image. Kappa coefficient tests the consistency of the prediction result and the ground truth. The MIoU calculates the ratio between the intersection and the union of two sets to reflect the accuracy and completeness of the segmentation result. $F1$ score is the harmonic average of recall and precision. The calculations are listed as follows:

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{1}$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{2}$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{3}$$

$$p_e = \frac{a_1 \times b_1 + a_2 \times b_2 + +a_n \times b_n}{m \times m} \tag{4}$$

$$\mathrm{Kappa} = \frac{p_0 - p_e}{1 - p_e} \tag{5}$$

$$F1 = \frac{2}{\frac{1}{\mathrm{precision}} + \frac{1}{\mathrm{recall}}} \tag{6}$$

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

where TP, FP, TN, and FN are the numbers of true positive, false positive, true negative, and false negative pixels, respectively; p_0 is the OA; a_n represents the number of real samples of each class; b_n is the number of predicted samples of each class; n is a total of n types of land cover types; and m represents the total number of m pixels in all land covers.

B. *Comparison of Different Semantic Segmentation Methods*

We conduct experiments on the MSW_LCD dataset to evaluate the performance of four popular semantic segmentation networks. Table IV presents the specific evaluation results of different architectures on the MSW_LCD dataset. Compared with PsPNet, SegNet, and UNet, the MIoU of DeepLabv3+ is 20.88%, 3.11%, and 0.25% higher than the other three models, respectively.

There are several loss values' curves on the same validation dataset of different networks in Fig. 9. After 30 epochs, the loss curve of the DeepLabv3+ network becomes relatively flat, but the other three models need much more epochs during the training process to reach convergence. Thus, we choose the DeepLabv3+ network going through 100 epochs with the best comprehensive performance on the MSW_LCD dataset as the basic network.

We report OA, Kappa coefficient, $F1$ score, MIoU, trainable parameters, and the training time consumption for each epoch in Table V when using five different backbones, namely Xception [44], MobileNetV2 [45], Resnet50/101 [46], and

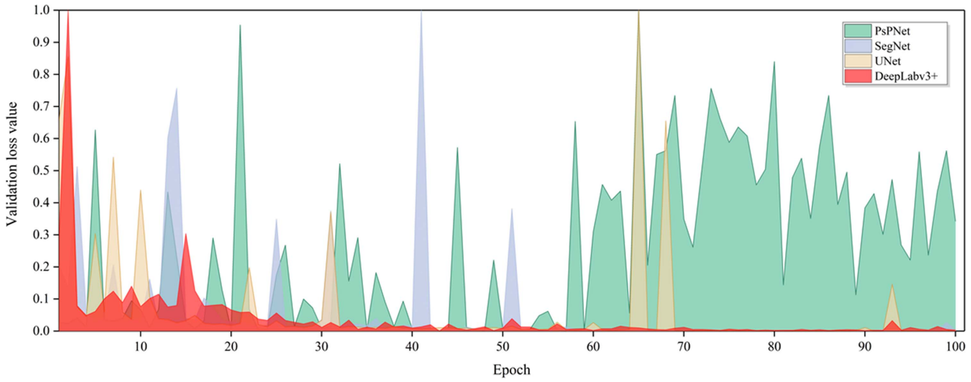


Fig. 9. Validation learning curves of four models on the MSW_LCD dataset.

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT BACKBONE NETWORKS IN THE DEEPLABV3+ ARCHITECTURE

Method	Backbone				
	Resnet50	Densenet	Resnet101	Xception	MobileNetV2
OA (%)	**95.43**	95.05	94.96	94.87	94.17
Kappa (%)	**89.01**	88.36	87.71	87.95	86.51
Precision (%)	**82.92**	79.87	82.47	80.57	83.84
Recall (%)	**80.47**	80.23	78.15	78.96	75.99
F1 (%)	**81.55**	79.84	80.10	79.55	79.46
MIoU (%)	**72.65**	70.24	70.64	70.07	69.22
Params (M)	26.80	41.25	15.30	45.82	**0.97**
Per-epoch Time (s)	233.03	450.33	355.70	538.39	**78.22**

The bolded represents the best experimental results.

TABLE VI
ABLATION EXPERIMENT ON THE MSW_LCD DATASET

FPN	Improved ASPP		PSA		F1	MIoU	Per-epoch	Params
	ASPP4	ASPP5	S	P	(%)	(%)	Time (s)	(M)
					81.55	72.65	220.55	26.80
✓					81.97	73.28	215.54	10.97
✓	✓				82.14	73.36	240.15	10.97
✓	✓		✓		82.24	73.76	227.42	13.74
✓	✓			✓	82.71	74.35	221.10	14.25
✓		✓			82.60	74.10	242.18	11.31
✓		✓	✓		82.68	74.30	233.47	15.29
✓		✓		✓	**83.10**	**74.82**	233.03	16.03

The bolded represents the best experimental results.
S: Sequential layout.
P: Parallel Layout.

Densenet121 [47] in the DeepLabv3+ model. Specifically, the DeepLabv3+ architecture with Resnet50 backbone [denoted as DeepLabv3+ (Resnet50)] outperforms other models and has a good tradeoff among accuracy, parameter amount, and time consumption.

C. Ablation Experiments

A series of ablation experiments are conducted to clarify how different modules contribute to the performance gain of the FPN module, ASPP4 composition, ASPP5 composition, PSA (S) module, and PSA (P) module on the DeepLabv3+ (Resnet50) network. The number of parameters for each model structure is also given in Table VI and reflects that the proposed model obtains optimal segmentation results at a lower computational cost. Overall, the proposed model reduces the number of parameters by about 40% compared with the DeepLabv3+ (Resnet50). Here, all results share the same hyperparameters and experimental platform.

As shown in Table VI, when introducing the FPN module and replacing bilinearly upsampled by a factor of 4 with two bilinearly upsampled by a factor of 2, the $F1$ score and MIoU are increased by 0.42% and 0.63%. Adopting ASPP4 gets only marginal performance improvement. In further

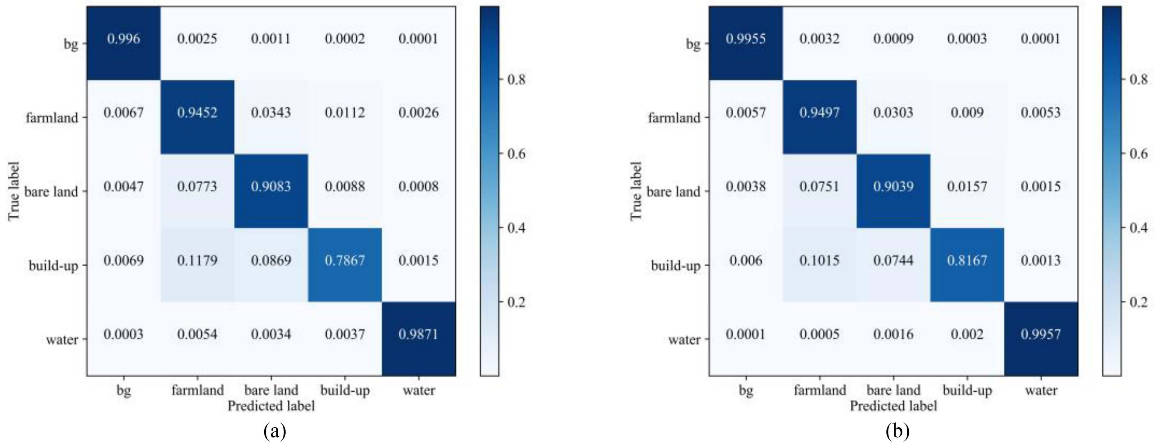


Fig. 10. Confusion matrices of two networks on the MSW_LCD dataset. (a) Confusion matrices of DeepLabv3+ (Resnet50). (b) Confusion matrices of FPN_PSA_DLV3+.

experiments, the $F1$ score and MIoU performance are further improved from 82.14% to 82.24% and from 73.76% to 73.36% when adding PSA (S). PSA (P) module boosts the network by 82.71%–82.14% $F1$ score and 74.35%–73.36% MIoU. Thereafter, we replace ASPP4 with ASPP5 module, an $F1$ score and MIoU improvement of 0.63% and 0.82% are observed. In further experimental result, it can be seen that the $F1$ score and MIoU of the introduction of the PSA (S) module are improved by 0.08% and 0.2%, and the introduction of the PSA (P) module is improved by 0.5% and 0.72% compared with the original model. DeepLabv3+ with FPN, ASPP5, and PSA (P) modules (denoted FPN_PSADLV3+) gets 84.1% $F1$ score and 74.82% MIoU. A promising result is obtained by FPN_PSADLV3+ on the MSW_LCD, which is not only better than the DeepLabv3+ with FPN module, ASPP4 composition, and PSA (S) block but also better than the DeepLabv3+ with FPN module, ASPP5 composition, and PSA (P) block.

TABLE VII
Classification Performance of Different Models on the MSW_LCD Dataset

Method	F1 score (%)			
	farmland	bare land	build-up	water
PsPNet	78.53	55.59	75.92	80.34
SegNet	88.88	67.08	70.87	84.35
Unet	90.01	69.54	74.66	**95.79**
DeepLabv3+ (Resnet50)	90.13	70.69	75.55	93.27
*_FPN	91.12	72.56	77.34	94.13
*_FPN_ASPP4	90.55	72.60	76.69	66.73
*_FPN_ASPP5	90.69	71.59	77.04	91.20
*_FPN_ASPP5_PSA (S)	90.97	72.56	76.50	92.37
*_FPN_ASPP5_PSA (P)	**91.33**	**73.26**	**77.81**	91.10

* denotes DeepLabv3+ (Resnet50).
_ denotes that adding modules on the basis of DeepLabv3+ (Resnet50) network.
S denotes the sequential layout.
P denotes the parallel layout.
Bold numbers are the best results.

D. Land Cover Segmentation

To observe the performance of the model in each land cover, we compute confusion matrices of DeepLabv3+ (Resnet50) network and FPN_PSA_DLV3+ network of five different land covers (including background class), as shown in Fig. 10, where the columns refer to predicted classes and the rows to actual classes. All five classes can be clearly distinguished and the higher the segmentation accuracy is, the darker the color will be.

Table VII reports the classification performance of each network on the four land cover types: farmland, bare land, build-up, and water. The proposed model outperforms the typical methods. It is evident that the classification accuracy of deserts with complex edges significantly improves by 2.57% to 17.67% using the FPN_PSADLV3+ network. For build-up with a high number of small objects, the proposed model improves accuracy by 2.26% to 6.94% $F1$ over the existing classification models. And classification result for farmland increases by 1.2% to 12.8% $F1$. Remarkably, compared with the DeeplabV3+ (Resnet), the FPN_PSA_DLV3+ has the highest $F1$ improvement of 2.57% for bare land, which has the highest percentage (94.04%) of small target parcels. For the build-up land cover with a sample share of only 2.98% in which the percentage of small target parcels is 74.49%, the $F1$ boost is as high as 2.26% compared with that model before the improvement. For water land cover with only 26.61% of small sample parcels, the proposed model also has no significant segmentation performance degradation. We provide segmentation visualization results of DeepLabv3+ (Resnet50) and FPN_PSA_DLV3+ network in Fig. 11. The edge details of the four land covers in the visual results greatly improve our model and the segmentation result of the network in this article is the closest to ground truth. Our model has better performance in terms of small objects land covers, such as a small area of build-up (as shown in the fourth row and the fifth row of Fig. 11) and the finer edge of land covers (as shown in the first row, second row, and third row of Fig. 11).

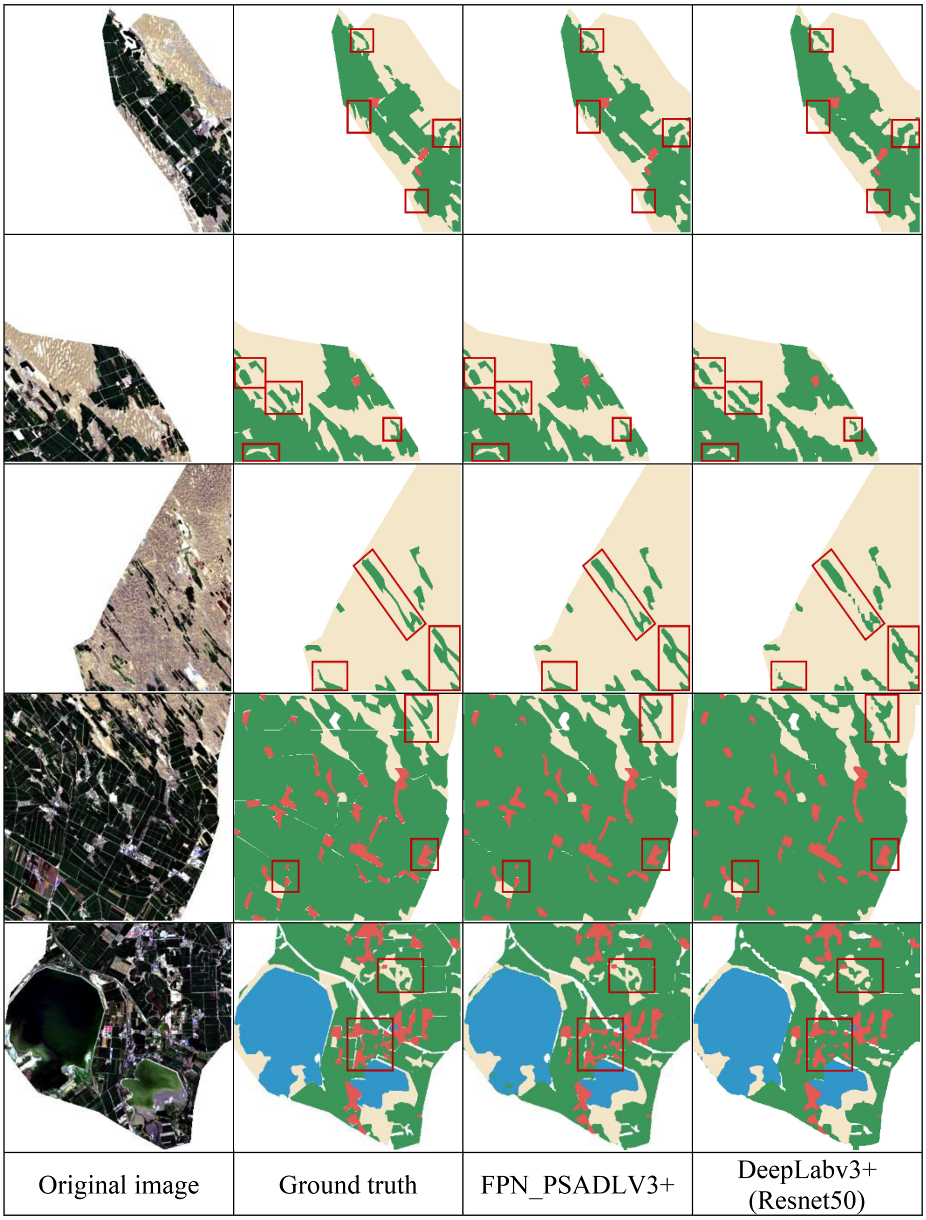


Fig. 11. Segmentation results compared with DeepLabv3+ (Resnet50) model on the MSW_LCD dataset.

V. Discussion

A. *Performance of the FPN_PSADLV3+ Model*

This article first evaluates several classical semantic segmentation networks that are often applied to RSIs classification in terms of classification accuracy, number of model parameters, training time, and model convergence speed. Table IV and Fig. 9 present that the DeepLabv3+ model not only effectively improves the classification accuracy of land covers but also has a faster convergence speed. We continue exploring the feature extraction capability by trying different backbones based on the DeepLabv3+ model. From Table V, it is evident that the DeepLabv3+ (Resnet50) performs better than the other models. Although the results generated by Densenet, Resnet101, and Xception are almost good, these models need much more average time per epoch during the training phase. And MobileNetV2 is 78.22 s per training epoch faster than ours, its kappa coefficient, recall, $F1$ score, and MIoU are worse.

Ablation experiments (see Table VI) indicate optimal segmentation of small objects and fine edges with a reasonable complexity and each module also contributes its own power: First, the FPN module incorporates more detailed information and eliminates large semantic gaps between low-level and high-level features; second, the ASPP5 module is a simple yet effective approach by adopting four dilated convolutions to extract features at different scales; The PSA (P) block preserves high-resolution semantics in attention computation at a reasonable cost and fit output distribution with a higher problem complexity by the order of output element numbers. The experiment proves that the PSA (P) in DeepLabv3+ improves the classification result marginally higher than PSA (S) with similar experimental results to the two modules in [46].

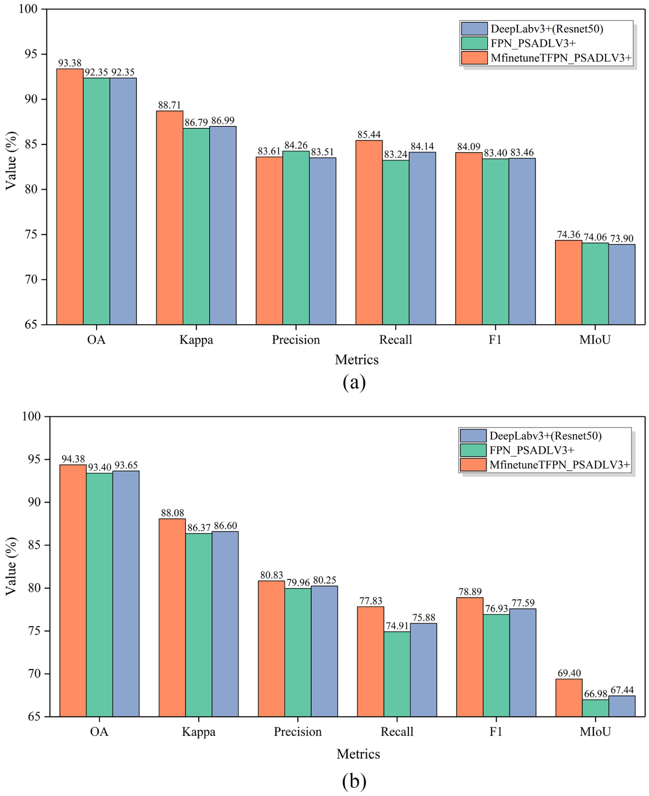


Fig. 12. Overall classification performance of three models on test dataset. The DeepLabv3+ (Resnet50) denotes the baseline. The FPN_PSADLV3+ denotes the proposed network trained from the scratch. Mfine-tuneTFPN_PSADLV3+ denotes the proposed network pretrained on MSW_LCD and then fine tuned on the TMSK _LELTD dataset. (a) Evaluation results of fine-tune subexperiment1 on the TMSK_LCD dataset. (b) Evaluation results of fine-tune subexperiment1 on the TMSK _LELTD dataset.

Although the segmentation of small samples is always a tough task for fine-grained segmentation, the FPN_PSA_DLV3+ network boosts all the other networks for small samples, build-up in Table VII. The UNet network has the best classification effect for water. There is a relatively obvious misjudgment of water in the FPN_PSA_DLV3+ network. Weeds grow in the water, which has the same spectral reflection as farmland, and there are bare land surfaces in shallow waters, which have the same spectral reflection as deserts. Therefore, water is more likely to be mistakenly divided into farmland and desert.

For the visual results in Fig. 11, the DeepLabv3+ (Resnet50) results are messy for fine details of objects and object parts and prone to be interrupted in some slender farmland and small build-up. The proposed method considers more information of small target and adds a feature extraction block with a smaller receptive field. The FPN module more outstandingly combines detailed information with high-level semantic information, which keeps more shallow information by two bilinearly upsampling with a factor of 2. Moreover, introducing PSA (P) block is expected to preserve high-resolution semantics in attention computation at a reasonable cost and fit output distribution with a higher problem complexity by the order of output element numbers. Empirically, the proposed model provides segmentation with more precise and finer details than DeepLabv3+ (Resnet50).

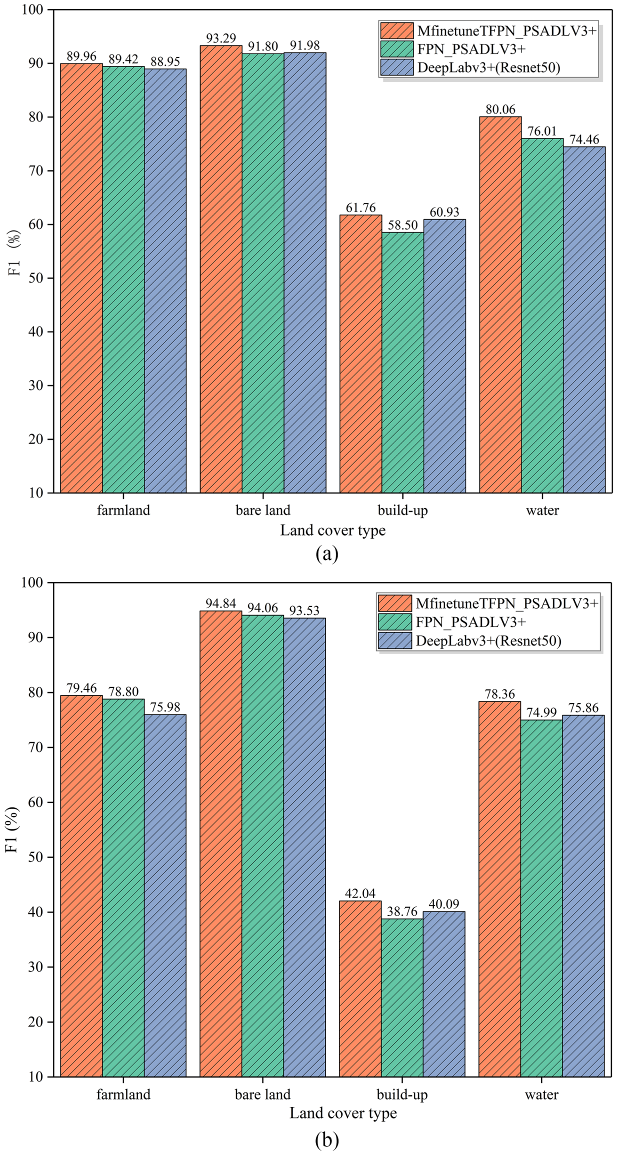


Fig. 13. Experimental results of four land covers of models with different structures on test dataset. (a) Classification performance for each land cover on the TMSK_LCD dataset. (b) Classification performance for each land cover on the TMSK_LELTD dataset.

B. Generalization Validation of the FPN_PSADLV3+ Model

A fine-tune model training skill is used to evaluate model generalizability. On the one hand, we assess how well the model generalized across two different regions of the TMSK region and MSW region with the same sensor of Landsat8 OLI sensor. Specifically, we fine tune the pretrained model from the MSW_LCD dataset with a smaller learning rate on the TMSK_LCD dataset. By fine tuning the FPN_PSADLV3+ model, a Kappa coefficient, $F1$ score, and MIoU improvement of 1.92%, 0.69%, and 0.3% are observed compared with training FPN_PSADLV3+ from scratch [see Fig. 12(a)]. On the other hand, we fine tune the model pretrained from the MSW_LCD dataset to evaluate the cross-region–sensor generalization on the TMSK_LELTD dataset. Compared with the model trained from scratch, the fine-tuned model improves the kappa coefficient, $F1$ score, and MIoU by 1.72%, 1.97%, and 2.46% [see Fig. 12(b)]. It is conclusively shown that the proposed method is successfully implemented with the different sensors and different regions.

Overall, fine-tuned FPN_PSADLV3+ model achieved the best effect whether it is on TMSK_LCD dataset or TMSK_LELTD dataset. It shows that the proposed model has strong generalization ability in different research areas, Southern Xinjiang, Northern Xinjiang, and different sensors: Landsat5, Landsat7, and Landsat8.

It is worth noting that the overall classification effect in the TMSK_LELTD dataset is slightly lower than that of the MSW_LCD dataset. One of the reasons is that there is a great gap in the build-up extraction results on the two datasets. The TMSK_LELTD dataset contains images from 1990 to 2012, the economic development of southern Xinjiang is lagging, and the build-up scales are not neatly planned. Usually, there are scattered residential houses on the cultivated land, which poses greater challenges for finer classification. The MSW_LCD dataset contains the images from 2013 to 2021. Since 2013, the urban and rural planning policies in southern Xinjiang have been gradually ameliorated, and build-up is getting denser, and the characteristics of the build-up are more prominent.

Fig. 13 shows the classification effect of the DeepLabv3+ (Resnet50), FPN_PSADLV3+ training from scratch, and the fine-tuned FPN_PSADLV3+ model on four land covers. The fine-tuned FPN_PSADLV3+ model on TMSK_LCD dataset outperforms the model trained from scratch on the segmentation of farmland, bare land, build-up, and water by 0.54%, 1.5%, 3.25%, and 4.05% $F1$, respectively [see Fig. 13(a)].

In particular, the TMSK_LCD dataset has a significant increment in the number of water samples and the percentage of small sample parcels compared with the MSW_LCD dataset, so the enhancement of water segmentation after fine tuning on TMSK_LCD dataset reaches up to 4.05%. In the MSK_LCD dataset with the occupancy of small target parcels up to 93.8% of build-up, the segmentation performance promotion of fine-tuned FPN_PSADLV3+ model for build-up is 3.26%. And the fine-tuned FPN_PSADLV3+ model on the TMSK_LELTD dataset outperforms FPN_PSADLV3+ model trained from scratch on the segmentation of farmland, bare land, build-up, and water by 0.66%, 0.79%, 3.27%, and 3.37% $F1$, respectively [see Fig. 13(b)]. The distribution about small object parcels on TMSK_LELTD dataset and the performance segmentation improvement after fine tuning are similar to those on TMSK_LCD dataset. The above validation process has a better estimation of the generalization performance of the model. The fine-tuning model effectively avoids the problems of insufficient parameter optimization and significantly improves the accuracy of each land cover, especially for water and build-up, which are prone to misclassification.

The differences between Landsat sensors and others are not only in spatial resolution and band settings but also may be in the corresponding spatial response functions and spectral response functions. However, the source data we use are all from the Landsat-like satellite imagery and the cross-sensor generalization performance we verified is to explore the generalization among Landsat8 OLI, Landsat7 ETM+, and Landsat5 TM. We further attempted to use the trained FPN_PSADLV3+ network for semantic segmentation of Sentinel-2 images. Specially, we use the pretrained model from the MSW_LCD dataset to fine tune the FPN_PSADLV3+ model using Sentinel-2 satellite data from Yuli Comt, Bayinguoleng Mongolian Autonomous Prefecture, Xinjiang Uygur Autonomous Region, China. By fine tuning the FPN_PSADLV3+ model, the model achieved 92.25% Kappa coefficient, 80.09% $F1$ score, and 71.66% MIoU. The proposed network can effectively extract small objects and edges almost close to the ground truth, as shown in Fig. 14. This further illustrates the cross-region and cross-sensor generalization potential of the FPN_PSADLV3+ model.

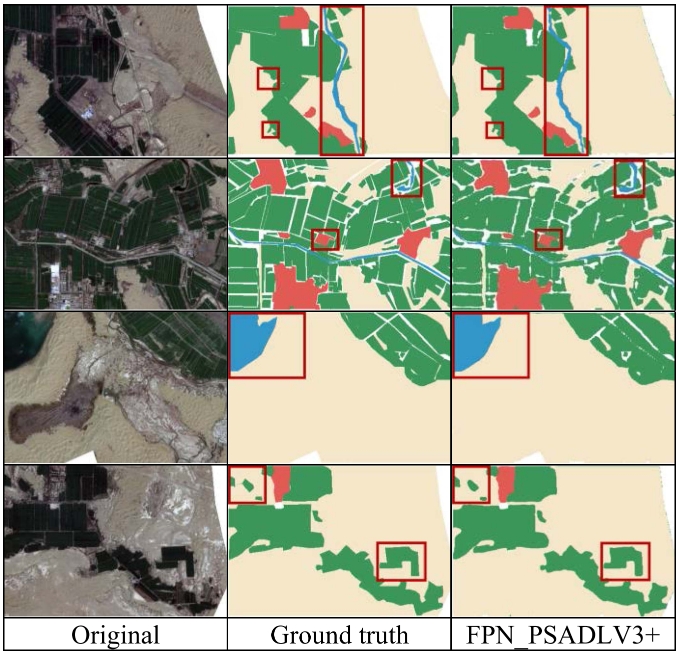


Fig. 14. Land cover segmentation results of FPN_PSADLV3+ network on Sentinel-2A images.

In the future, quantitative tests are still needed to be done on more types of datasets to further confirm the generalization performance of the network. We will exploratively extend the current approach by integrating CNN-based models and transformer-based models to further improve the segmentation performance of fine edge and small objects in RSIs. Moreover, real-time semantic segmentation with context aggregation architectures for RSIs will be explored for multiscenario applications.

VI. Conclusion

In this article, to meet the challenge of small objects and fuzzy edges of land cover in arid areas adjacent to the desert, FPN_PSADLV3+ architecture with the best comprehensive performance is introduced. First, a benchmark network, DeepLabV3+ (Resnet50), with a moderate number of model parameters and fast convergence, is selected to which the FPN, ASPP5, and PSA modules are added for improving performance of the model in mitigating loss of detail information, multiscale feature extraction, and capturing contextual information. And ablation experiments are carried out to prove the effectiveness of each module. The proposed semantic segmentation model overall achieves a 1.55% $F1$ score improvement and 2.17% MIoU improvement on the MSW_LCD dataset with a reasonable computational complexity to DeepLabv3+. The small objects, build-up, get a 2.26% $F1$ score improvement, and farmland and bare land with blur boundary get 1.2% and 2.57% improvement. Moreover, the FPN_PSADLV3+ model has great generalization in cross region and cross sensor. The proposed model has a tradeoff between efficiency and accuracy on the segmentation of small objects and fine edges in RSIs and provides good solution for land cover segmentation and, thus, quantitative for built-up expansion and desertification monitoring in complex arid scenarios.

References

[1] X. Yang, L. Ci, and X. Zhang, "Dryland characteristics and its optimized eco-productive paradigms for sustainable development in China," *Natural Resour. Forum*, vol. 32, no. 3, pp. 215–227, 2008.

[2] C. Liu, F. Zhang, V. C. Johnson, P. Duan, and H.-T. Kung, "Spatio-temporal variation of oasis landscape pattern in arid area: Human or natural driving," *Ecol. Indicators*, vol. 125, Jun. 2021, Art. no. 107495.

[3] C. Elachi and J. J. van Zyl, *Introduction to the Physics and Techniques of Remote Sensing*. Hoboken, NJ, USA: Wiley, 2021.

[4] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.

[5] J. Li, Y. Pei, S. Zhao, R. Xiao, X. Sang, and C. Zhang, "A review of remote sensing for environmental monitoring in China," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1130.

[6] X. Wang, Z. Li, and Z. Gao, "Monitoring sandified land changes using multi-temporal landsat TM/ETM+ data in Dengkou County of Inner Mongolia, China," in *Proc. 4th Int. Congr. Image Signal Process.*, 2011, pp. 1646–1651.

[7] Q. Yu et al., "The optimization of urban ecological infrastructure network based on the changes of county landscape patterns: A typical case study of ecological fragile zone located at Deng Kou (Inner Mongolia)," *J. Cleaner Prod.*, vol. 163, pp. S54–S67, 2017.

[8] S. Talukdar et al., "Land-use land-cover classification by machine learning classifiers for satellite observations—A review," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1135.

[9] C. Luo et al., "Using time series Sentinel-1 Images for object-oriented crop classification in google earth engine," *Remote Sens.*, vol. 13, 2021, Art. no. 561.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[11] G. Alberto et al., "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*.

[12] M. Orsic and S. Šegvić, "Efficient semantic segmentation with pyramidal fusion," *Pattern Recognit.*, vol. 110, 2021, Art. no. 107611.

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

[14] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 447–456.

[15] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 680–688.

[16] S. Du, Shihong Du, B. Liu, and X. Zhang, "Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images," *Int. J. Digit. Earth*, vol. 14, no. 3, pp. 357–378, 2021.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 52, pp. 436–444, 2015.

[18] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899–6910, Dec. 2018.

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

[20] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5168–5177.

[21] R. Dong, X. Pan, and F. Li, "DenseU-net-based semantic segmentation of small objects in urban remote sensing images," *IEEE Access*, vol. 7, pp. 65347–65356, 2019.

[22] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1442–1450.

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

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

[25] L. Chen et al., "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[26] 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. 833–851.

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

[28] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[29] 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.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

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

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

[33] M. Langkvist, A. Kiselev, M. Alirezaie, and A. Loutfi, "Classification and segmentation of satellite orthoimagery using convolutional neural networks," *Remote Sens.*, vol. 8, 2016, Art. no. 329.

[34] F. Rottensteiner et al., "The ISPRS benchmark on urban object classification and 3D building reconstruction," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 1–3, pp. 293–298, 2012.

[35] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 1–9.

[36] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, 2018.

[37] L. Weng, M. Qian, M. Xia, Y. Xu, and C. Li, "Land use/land cover recognition in arid zone using a multi-dimensional multi-grained residual forest," *Comput. Geosci.*, vol. 144, 2020, Art. no. 104557.

[38] X.-Y. Tong, Q. Lu, G.-S. Xia, and L. Zhang, "Large-scale land cover classification in Gaofen-2 satellite imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 3599–3602.

[39] W. Zhang, M. Cong, and L. Wang, "Algorithms for optical weak small targets detection and tracking: Review," in *Proc. Int. Conf. Neural Netw. Signal Process.*, 2003, pp. 643–647.

[40] H. Liu et al., "Polarized self-attention: Towards high-quality pixel-wise regression," 2021, *arXiv:2107.00782*.

[41] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. Int. Conf. Artif. Neural Netw.*, 2018, pp. 270–279.

[42] 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.

[43] P. Wang et al., "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1451–1460.

[44] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807.

[45] S. Mark et al., "Mobilenet V2: Invented residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[46] 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.

[47] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

Panli Yuan received the B.E. degree in computer science and technology, in 2020 from the Department of Computer Science and Technology, Shihezi University, Shihezi, China, where she is currently working toward the M.E. degree in electronic information with the Department of Computer Science and Technology.

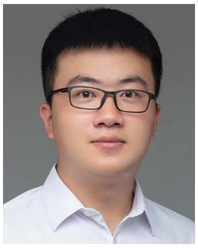
Her research interests include intelligent understanding of remote sensing images and change detection.

Qingzhan Zhao received the B.E. degree in vehicle engineering from Shihezi University, Shihezi, China, in 1995, the M.E. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 1998, the M.E. degree in industrial economics from Shihezi University, Shihezi, China, in 2007, and the master's degree in admission in 2012.

He is currently working as a Professor with the Department of Computer Science and Technology, Shihezi University, Shihezi, China. His research interests include remote sensing image processing and pattern recognition.

Yuchen Zheng (Member, IEEE) received the B.E. and M.E. degrees in computer science and technology from the Department of Computer Science and Technology, Ocean University of China, Qingdao, China, in 2014 and 2017, respectively, and the Ph.D. degree in information intelligence engineering from the Department of Advanced Information Technology, Kyushu University, Fukuoka, Japan, in 2020.

He is currently an Associate Professor with the Department of Computer Science and Technology, Shihezi University, Shihezi, China. His research interests include biometrics and security, pattern recognition, and neural networks.



Xuewen Wang received the B.E. degree in in computer science and technology and M.E. degree in agriculture engineering from the Department of Computer Science and Technology, Shihezi University, Shihezi, China, in 2018 and 2020, respectively. He is currently working toward the Ph.D. degree in earth exploration and information technology with the School of Geophysics and Geomatics, China University of Geosciences, Wuhan, China.

His research interests include remote sensing and artificial intelligence.

Bin Hu received the B.E. degree in computer science and technology from Yangtzeu University, Jingzhou, China, in 2019. He is currently working toward the M.E. degree in electronic information with the Department of Computer Science and Technology, Shihezi University, Shihezi, China.

His research interests include deep learning and data analysis.