Light-Weight Semantic Segmentation Network for UAV Remote Sensing Images

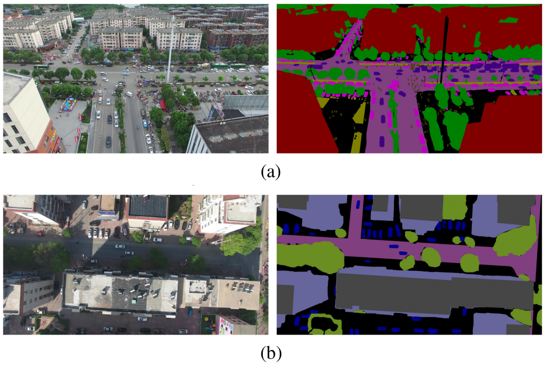
Siyu Liu, Jian Cheng, Leikun Liang, *Graduate Student Member, IEEE*, Haiwei Bai, and Wanli Dang

***Abstract*—Semantic segmentation for unmanned aerial vehicle (UAV) remote sensing images has become one of the research focuses in the field of remote sensing at present, which could accurately analyze the ground objects and their relationships. However, conventional semantic segmentation methods based on deep learning require large-scale models that are not suitable for resource-constrained UAV remote sensing tasks. Therefore, it is important to construct a light-weight semantic segmentation method for UAV remote sensing images. With this motivation, we propose a light-weight neural network model with fewer parameters to solve the problem of semantic segmentation of UAV remote sensing images. The network adopts an encoder–decoder architecture. In the encoder, we build a light-weight convolutional neural network model with fewer channels of each layer to reduce the number of model parameters. Then, feature maps of different scales from the encoder are concatenated together after resizing to carry out the multiscale fusion. Moreover, we employ two attention modules to capture the global semantic information from the context and the correlation among channels in UAV remote sensing images. In the decoder part, the model obtains predictions of each pixel through the softmax function. We conducted experiments on the ISPRS Vaihingen dataset, UAVid dataset, and UDD6 dataset to verify the effectiveness of the light-weight network. Our method obtains quality semantic segmentation results evaluated on UAV remote sensing datasets with only 9 M parameters the model owns, which is competitive among popular methods with the same level of parameters.**

***Index Terms*—Attention mechanism, light-weight network, remote sensing, semantic segmentation, unmanned aerial vehicle images.**

I. Introduction

THE application of remote sensing images captured by unmanned aerial vehicles (UAVs) is of great significance. With the help of photography devices installed on the UAVs, low-altitude high-resolution aerial images can be collected more conveniently and economically. Compared with the satellite-borne remote sensing platform, the UAV-borne photographic platform has a lower flight height and could fly close to the ground for improving the resolution of objects. These characters enable UAV remote sensing images to distinguish detailed objects, such as nonmotor vehicles, pedestrians, and infrastructures.


(a)

(b)

Fig. 1. Two typical UAV remote sensing images of street scene and their semantic annotations from the UAVid dataset [1] and UDD6 dataset [2]. The view of UAV remote sensing photography is divided into two types. (a) Oblique projection. (b) Orthographic projection.

In terms of flexibility, the UAVs support a controllable flight path to get multitemporal data without the restriction of revisit interval of the platform. It is especially important in tasks that require a rapid response, such as natural disaster damage assessment [3]–[5], automatic warning system [6], and traffic analysis [7]–[9]. In addition, as shown in Fig. 1, the UAV photography can also perform oblique projection imaging to acquire multiview remote sensing images. It can provide a lot of information for structural damage mapping of buildings [10]–[12].

In many kinds of image information acquisition tasks for UAV remote sensing images, semantic segmentation is one of the emerging and challenging research focus. The target of semantic segmentation is to predict pixel-level labels according to the semantic information represented by pixels of the image. In recent years, convolutional neural networks (CNNs) have been developed considerably and proved to have significant effects in semantic segmentation tasks [13]–[15]. In remote sensing, semantic segmentation approaches are used to effectively determine the type of land, analyze the ground objects, and extract the roads.

Conventional UAV image analysis system only uses the UAV platform for image acquisition, while the images are uploaded

Manuscript received April 30, 2021; revised June 27, 2021 and July 25, 2021; accepted August 4, 2021. Date of publication August 12, 2021; date of current version September 1, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 62071104, in part by the Sichuan Science and Technology Program under Grant 2020YFG0085 and Grant 2021YFG0328, and in part by the Intelligent Terminal Key Laboratory of Sichuan under Grant SCITLAB-0017. *(Corresponding author: Jian Cheng.)*

Siyu Liu, Jian Cheng, Leikun Liang, and Haiwei Bai are with the University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: syliu@std.uestc.edu.cn; chengjian@uestc.edu.cn; liang_leikun@163.com; hwbaymax@std.uestc.edu.cn).

Wanli Dang is with the University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Second Research Institute of Civil Aviation Administration of China, Chengdu 610041, China (e-mail: dangwanli@caacsri.com).

Digital Object Identifier 10.1109/JSTARS.2021.3104382

to large-scale computing equipment for image processing. However, considering the agility and real-time analysis requirements of UAV remote sensing tasks, a more flexible way is to compute by devices on UAV individually. In this way, UAV equipments can only upload calculation results instead of original images, or directly make decisions based on these results. Nevertheless, many popular semantic segmentation approaches designed for satellite-borne remote sensing images are not suitable for tasks on UAVs in general. This is because of the specificity of UAV remote sensing images. One is that the UAV remote sensing images have high resolution and large amount of data, but the memory and storage capacity of the mobile device are insufficient. Popular semantic segmentation approaches try to apply deeper networks with more layers and channels, which strive to get richer information for better performance, such as ResNet [16]. However, those methods based on large-scale neural networks have inherent disadvantages, that is excessive parameter amount. Due to the limitations of UAV platforms, it is difficult to deploy those large-scale models on mobile devices with low memory. In addition, models with numerous parameters are more difficult to train, which contributes to difficulties when training if annotated images are insufficient.

Another imperfection of popular methods is that the purpose of UAV remote sensing differs from that of satellite-borne. The mission of satellite-borne remote sensing focuses on ground object detection and land statistic in a larger area, while the observation range of a UAV image is only about a street, a crossroad, or a block, depending on the height and view of UAV platform. Therefore, the goal of a large number of low-altitude UAV remote sensing tasks is to analyze the objects in the street view more accurately through high-resolution imaging. There is richer contextual information that needs to mine in these scenes captured up close by UAVs.

In other words, the current popular methods cannot perform the semantic segmentation task of UAV images well. To overcome shortcomings mentioned, this article proposes a light-weight semantic segmentation network under attention mechanism for UAV remote sensing images. The network's encoder is simplified with a few parameters in consideration of the limitation for UAV application, while maintaining the capability of feature representation. At the end of the encoder, we apply a multiscale feature fusion method to merge the information of multiple scales obtained from different stages of the encoder. For capturing richer relationships, the attention modules [17] are introduced to our method, including the channel attention module and the spatial attention module. The former acquires the correlation among channels of feature maps for effective representations, while the latter provides global contextual information among pixels. These two parts of features are fused to generate the third part of the feature map, which together serve as the representation of the input. To summarize, the contributions of this article are listed as follows.

- We propose an efficient light-weight network for semantic segmentation of UAV remote sensing images with fewer parameters.
- Attention mechanism is introduced in our method to capture global relationships among pixels and correlation among feature representations. Feature fusion methods are used to fuse features containing different information.
- Experimental results conducted on UAV-borne remote sensing datasets UAVid and UDD6 and satellite-borne dataset ISPRS Vaihingen indicated the effectiveness of our approach in comparison with popular methods.

In our previous work [18], we designed a light-weight semantic segmentation network and verified its performance on a satellite-borne remote sensing dataset. In this article, we added features from shallow layers of neural network in the multiscale feature fusion. Then, we test and discuss the effectiveness of attention mechanism and feature fusion strategy. Finally, we apply this model to the UAV remote sensing semantic segmentation task, and conduct more experiments and analysis on large-scale UAV remote sensing datasets.

The rest of this article is organized as follows. Section II introduces the development and related works of remote sensing of UAVs, semantic segmentation, and attention mechanism. Section III first describes the structure of the proposed network and subsequently introduces the employ-of-attention mechanism. Section IV shows the implementation details and experimental results on the UAVid [1] and UDD6 [2] dataset, and shows the comparison with other approaches. Finally, Section V summarizes the work of this article.

II. Related Works

In this section, the research related to our approaches will be introduced in detail, including the work of UAV remote sensing, semantic segmentation, and attention mechanism.

A. *UAV Remote Sensing*

Researches on UAV remote sensing have been more abundant by reason of the flexibility and mobility of UAV photography platform. These studies cover many fundamental tasks in the field of computer vision and remote sensing. Many researches of UAV remote sensing process their focuses on matching the characteristics of UAV remote sensing images, such as small object detection and reidentification. In the object detection and tracking task, aiming at the problem of detecting small-sized objects in UAV remote sensing images, Liu *et al.* [19] proposed a detection network for small objects based on YOLOv3 [20]. On individual tree crown detection and delineation task, Huang *et al.* [21] applied the bias field estimation to reduce the spectral heterogeneity in the UAV images. Du *et al.* [22] developed a UAV benchmark of object detection and tracking focusing on complex scenarios, including 80-k annotated frames up to 14 kinds of attributes. Another dataset called UAV-VeID [23] is conducted for vehicle reidentification, which contains 4601 vehicles in multiple images taken from different viewpoints. Remote sensing based on UAV imagery is currently under development. Many computer vision tasks and popular methods are gradually applyied to UAVs.

B. *Semantic Segmentation*

Semantic segmentation is a fundamental task in computer vision field. The goal of this task is to predict the label of each pixel in images or videos. At present, feature extraction approaches based on CNN have been proved effective in many tasks, and semantic segmentation methods based on CNN have gradually become the mainstream.

In earlier CNN-based pixel-wise analysis studies, FCN [13] improves the feature extractor for semantic segmentation, and concatenates the feature maps of different layers to fuse the multiscale features. U-Net [24] follows the improved idea of FCN. It stitched the output information of the corresponding layer in the process of continuous upsampling decoding. This method has an outstanding performance in medical image analysis. PSPNet [25] proposes a pyramid semantic segmentation network to embed scenery context features in scene images, improving the ability to get global information. DeepLab [15] proposes an atrous spatial pyramid pooling (ASPP) module to obtain multiscale information meanwhile maintaining high-resolution feature maps.

C. Light-Weight Networks

The lightweight of CNNs is of great significance for accelerating the calculation speed and reducing the difficulty of deployment. Generally speaking, there are two ways to achieve lightweight. One is to adopt methods such as pruning and model compression after the network is designed and trained. This method is widely used in engineering. Another method is to design a more efficient neural network calculation method, such as using small-size convolution kernels, or reducing the number of feature channels.

In this category of lightweight, MobileNet [26] proposed a depthwise separable convolution method, using depthwise convolution and pointwise convolution operations instead of regular convolution, reducing the amount of calculation of convolution operations. ShuffleNet [27] proposed pointwise group convolution and channel shuffle to reduce computation cost. EfficientNet [28] used a neural architecture search (NAS) method to explore the trade-offs between neural network width, depth, and resolution.

In semantic segmentation tasks, DFANet [29] is based on a lightweight backbone architecture for deep feature aggregation, and has achieved extremely high efficiency. In order to avoid the loss of information caused by model lightweighting and acceleration, BiSeNet [30] uses spatial path and context path to extract information separately, and then integrates them to achieve a good balance between speed and effect.

D. Attention Mechanism

The term *attention* comes from the selective attention mechanism of human vision, in which humans efficiently allocate limited resources of attention when watching. In recent years, it has been widely used in deep learning to the selection of information that is more critical to the tasks. According to its different function in the neural network, it can be roughly divided into spatial attention, channel attention, and so on.

The purpose of spatial attention is to enlarge the receptive field of neural network and collect the global context information in the feature maps. Nonlocal [31] introduces nonlocal module in the high-level semantic features, which play the role of increasing the receptive field. This method solves the problem that the convolution operation cannot obtain the global correlation and brings richer information to the subsequent layers. CCNet [32] proposes a criss-cross module for simplifying nonlocal operation, which has higher computational efficiency. In the research on natural language processing (NLP), Transformer [33] relies on self-attention to compute representations of its input. In Microsoft Research Asia's empirical research [34], spatial attention is divided into four different types of factors according to the determination of weight for key and query. It also brings deformable convolution and dynamic convolution into the category of spatial attention.

Modeling the relationship among representations is also one function of the attention mechanism. SENet [35] starts with the relationship between feature representations and models the interdependence among channels explicitly. Specifically, the importance of each feature channel is automatically gained through learning. Then, useful features are enhanced and features that are not useful for the current task are suppressed according to this importance. DANet [17] designs a channel attention module to capture the relation among channels. It could focus on the contribution of objects in the feature map, reducing the influence of background.

III. Proposed Method

In this section, the introduction of our light-weight CNN is given. This section first introduces the structure and workflow of the proposed network, including the encoder part and the decoder part. Then, applications of the channel attention module and the spatial attention module are illustrated.

A. Network Architecture

Our light-weight semantic segmentation network is shown in Fig. 2. In the encoder part, the network takes the EfficientNet-b1 [28] as a backbone, which uses an efficient feature extraction structure with several stages designed by the NAS approach, and has achieved excellent performance on classification tasks with a few parameters.

The backbone network contains several convolution blocks, and each block is composed of several convolution layers, rectified linear unit (ReLU) layers, pooling layers, and other basic components of the CNN. The ith feature extraction function of each convolution block can be expressed as $F_i(X_i, w_i)$, where $F_i(\cdot,\cdot)$ is the predicting function of a convolution block, and X_i are the input images or features of this block, depending on the location of this block. w_i represent the weight parameters. The feature map D_n generated after the nth convolution block cascade can be expressed as

$$D_n = \odot_{i=1}^{n} F_i\left(X_i, w_i\right) \tag{1}$$

where $\odot$ expresses the cascade of blocks. Feature maps with different resolutions from encoder contain multiscale and multisemantic information. In order to collect these multiscale information comprehensively, our method extracts fixed multiscale feature maps, unifies the size through upsampling to 64×64,

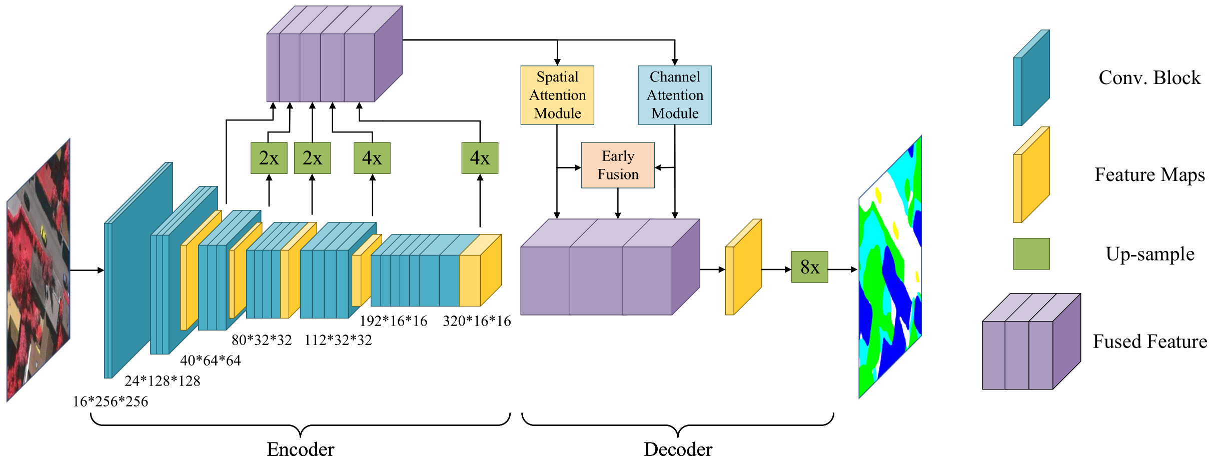


Fig. 2. The light-weight attention network we proposed. The encoder uses an efficient feature extraction structure, and the decoder aggregates feature maps to gain multiscale information. The attention mechanism is introduced in order to obtain contextual relationship.

TABLE I
OUR BACKBONE BASED ON EFFICIENTNET-B1 [28]. EACH ROW REPRESENTS AN OPERATOR WITH ITS INPUT RESOLUTION, OUTPUT CHANNELS, AND LAYERS OF CONVOLUTIONS IN IT

Stage	Operator	Input Resolution	Output Channels	Layers
1	Input Conv.	512*512	32	1
2	Conv. Block	256*256	16	2
3	Conv. Block	256*256	24	3
4	Conv. Block	128*128	40	3
5	Conv. Block	64*64	80	4
6	Conv. Block	32*32	112	4
7	Conv. Block	16*16	192	5
8	Conv. Block	16*16	320	1

and finally merges them as the input of the decoder. This operation can be expressed as

$$Concat(D'_{n_1}, D'_{n_2}, \ldots), n_i \in S_{\text{feature}} \tag{2}$$

where S_{feature} is the set of feature maps that need to be extracted, and each D'_i is generated by D_i through upsampling in order to unify the same feature map size. The detailed network structure is shown in Table I. The input convolutional layer contains normalization, zero padding, convolution, batch-normalization, and activation layers, and each of the remaining blocks contains convolution, batch-normalization, and activation layers. When reducing the size of the feature map, global average pooling operator is used.

These feature maps after uniform are concatenate function through $Concat(\cdot)$. Compared with the previous work, this method introduces the shallower feature maps sized 64×64, which can increase the diversity of multiscale feature fusion.

In the decoder part, we use 1×1 convolution to reduce the number of channels and perform feature fusion, which can significantly reduce the model parameters and calculations.

Subsequently, these feature maps are processed by attention modules mentioned in Section III-B. Through the output convolutional layer, upsampling, and softmax processing, we could obtain the categories of pixels in remote sensing images.

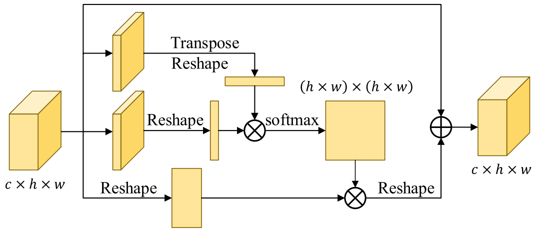


Fig. 3. Structure of the spatial attention module.

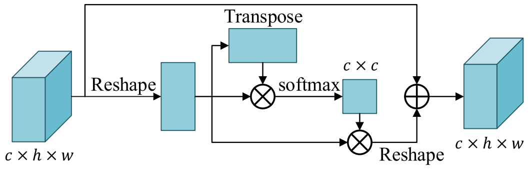


Fig. 4. Structure of the channel attention module.

B. Attention Mechanism and Feature Fusion

General CNNs only have a local receptive field. Therefore, approaches based on CNNs lack the effective representation of global information. The attention mechanism can effectively capture context information, so we introduce an attention mechanism in our network.

In order to capture the spatial global context information, we utilize the spatial attention module. The structure of this module is shown in Fig. 3. Through the operation of convolution and matrix multiplication, the module gains the spatial attention matrix of size $(h \times w) \times (h \times w)$, where h and w are the height and width of feature maps.

In addition, we use a channel attention module to mine the relationships among channels. The structure of this module is shown in Fig. 4. This channel attention module captures the information connection among channels by calculating a matrix sized $c \times c$, where c is the number of channels.

After the two attention modules output the features independently, our method first performs the joint modeling of the attention mechanism by combining two sets of feature maps from two modules, and generating a third set of features. This step is called early fusion. The purpose of it is to combine representations of multiple attention modules. Finally, three sets of feature maps are integrated together and fused through 1×1 convolution layers. See the analysis in Section IV-D later on the ablation study of different attention modules and attention early fusion.

IV. Experiment

To verify the performance of our method in the semantic segmentation task, we set a series of experiments to validate the performance of our method on three datasets, including ISPRS Vaihingen dataset, UAVid dataset, and UDD6 dataset. After that, the analysis of the results, the ablation study of modules, and the comparison with other popular approaches are discussed.

A. Datasets

Validation on multiple large-scale datasets is an important way to verify the generalization performance of approaches. In this article, we first conducted semantic segmentation experiments on the ISPRS Vaihingen 2D Semantic Labeling dataset[1] to validate the performance of our model, and compare it with other popular methods. Then, we carry out a two sets of experiments on challenging UAV high-resolution remote sensing semantic labeling datasets named UAVid [1] and UDD6 dataset [2].

ISPRS Vaihingen dataset is a popular satellite-borne remote sensing dataset, which contains 33 patches of different size. Each of them corresponds to the near-infrared (IR), red (R), and green (G) bands delivered by the camera. The pixels in the picture are divided into six categories, namely *Impervious surfaces*, *Buildings*, *Low vegetation*, *Trees*, *Cars*, and *Clutters*. To conduct the comparison with other approaches, we use the same training set and validation set as in Ref. [43]. The training set includes 11 patches of images (1, 3, 5, 7, 13, 17, 21, 23, 26, 32, 37), and the validation set includes five patches (11, 15, 28, 30, 40).

UAVid dataset[2] contains 42 UAV-borne remote sensing image sequences of the street scene in total. Each sequence has ten images of 4K resolution, and corresponds to the red (R), green (G), and blue (B) bands. Each pixel in images is labeled as one of eight categories, including *Building*, *Road*, *Tree*, *Low vegetation*, *Static car*, *Moving car*, *Human*, and *Background clutter*. Unlike most popular remote sensing datasets, the UAVid dataset has a category label of *Human*. Because of the high resolution of UAV-borne aerial images, the *Human* class can be detected, while the classification of this category is more challenging.

UDD-6[3] dataset is another UAV image dataset collected at multiple cities. The full name of it is Urban Drone Dataset. It contains 141 UAV-borne remote sensing images with a resolution of 4000×3000 or 4096×2160 pixels, which is divided into a training set of 106 pictures and a validation set of 35 pictures. It also corresponds to the red (R), green (G), and blue (B) bands. The pixels of the pictures in the dataset are divided into six categories, namely *Facade*, *Road*, *Vegetation*, *Vehicle*, *Roof*, and *Other*. It is emphasized that the latest version of the UDD6 dataset subdivides the label of buildings into the *Roof* and *Facade*, which depends on the oblique photographic characteristics of UAV remote sensing images.

B. Implementation Details

The light-weight network and attention module are constructed in the form mentioned in Sections III-A and III-B. The encoder part is initialized with parameters pretrained on ImageNet, while the decoder part is trained from scratch. Our feature early fusion method is implemented by channel-by-channel addition. As the high resolution of each patch, we split each patch into 512×512 with 25% overlaps, avoiding destroying the ground objects when splitting images.

To evaluate the effect of the attention mechanism, we take the light-weight network (LWN) with no attention modules as the baseline approach and compare it with the network after using attention modules. In addition, several popular approaches have been tested on these datasets as a control group.

The experiment is on the environment of Pytorch-1.3. In the training course, we choose stochastic gradient descent (SGD) as the optimizer and set $momentum = 0.9$. Considering the situation of different datasets, we designed different training strategies for them. On different datasets, suitable training epochs and initial learning rates are different. For the training processes in Vaihingen dataset, GTX 2080Ti is used for 500 training epochs under the condition of $batch\_size = 8$. The initial learning rate $lr_{init} = 0.002$. As the training goes on, the learning rate changes according to the following formula:

$$lr = lr_{init} \times \left(1 - \frac{current\_iterations}{max\_iterations}\right)^{0.9}. \tag{3}$$

In the UAVid dataset, as a result of the increase in the amount of training data, the training only lasts 100 epochs, and the initial learning rate $lr_{init} = 0.004$. And in UDD6, lr_{init} is set to 0.01 empirically.

In this article, the semantic segmentation of remote sensing images is considered to be a pixel-wise dense prediction task. We apply the 2D cross-entropy loss to measure the difference between the predicted result p and the ground truth y, which is defined as

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{H\times W} y_{i,n} \log p_{i,n} \tag{4}$$

where N represents the number of classes.

C. Evaluation Metric

Compare the test result of the model with the real label to get the confusion matrix $P = \{p_{ij}\} \in \mathbb{N}^{k\times k}$ of the result, where p_{ij} represents the number of pixels for which belong to the ith category and is classified into the jth category, and k is the number of categories. To be specific, the diagonal elements p_{ii} represent the number of pixels for which the prediction is equal to the ground truth. In order to evaluate the effect of the model,

[1] [Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/

[2] [Online]. Available: https://uavid.nl/

[3] [Online]. Available: https://github.com/MarcWong/UDD

TABLE II
ABLATION STUDY ON ISPRS VAIHINGEN AND UAVID DATASET

Methods	SA	CA	EF	OA on Vaihingen	mIoU on UAVid
LWN	-	-	-	88.27	67.82
LWN-SA	✓	-	-	88.37	68.40
LWN-CA	-	✓	-	88.45	68.02
LWN-A	✓	✓	-	88.72	69.00
LWN-A-F	✓	✓	✓	**88.85**	**69.02**

we use three different indicators to measure the effectiveness of segmentation results of the model, namely F_1 score, OA, and mIoU.

We calculate F_1 score with the following formula:

$$F_1\text{score} = 2 \times \frac{P \times R}{P + R} \tag{5}$$

where

$$P = \frac{1}{k}\Sigma_{i=1}^{k}\frac{p_{ii}}{\Sigma_{j=1}^{k}p_{ji}}, R = \frac{1}{k}\Sigma_{i=1}^{k}\frac{p_{ii}}{\Sigma_{j=1}^{k}p_{ij}}. \tag{6}$$

It can be analyzed from the formula that F_1 score can balance the precision rate and the recall rate.

The overall accuracy (OA) can be calculated with the following formula:

$$\text{OA} = \frac{\Sigma_{i=1}^{k}p_{ii}}{\Sigma_{i=1}^{k}\Sigma_{j=1}^{k}p_{ij}}. \tag{7}$$

The overall accuracy represents the proportion of the number of correctly classified pixels to the total number of pixels.

The mean Intersection over Union (mIoU) can be calculated with the following formula:

$$\text{mIoU} = \frac{1}{k}\Sigma_{i=1}^{k}\frac{p_{ii}}{\Sigma_{j=1}^{k}p_{ij} + \Sigma_{j=1}^{k}p_{ji} - p_{ii}}. \tag{8}$$

Compared with OA, mIoU is more sensitive to the analytical results of objects in a small proportion.

The official benchmark of each dataset chooses different measurement indicators: Vaihingen uses F_1 score and OA, where F1 scores only calculate the classification results of the first five categories except $Clutters$. The two UAV remote sensing datasets choose mIoU and OA.

D. Ablation Study

Our method uses multiple modules with different functions. In order to verify the effectiveness of them independently introduced in Section III-B, an ablation study was carried out in this article, and the results are in Table II. These experiments use the same backbone LWN with no attention modules. SA in the table represents whether to use the spatial attention module, and CA represents the channel attention module, and EF represents the early fusion method. The ablation study was carried out on the Vaihingen dataset and UAVid to verify the module's ability on the satellite-borne and UAV-borne remote sensing datasets.

Through the comparison in Table II, we find that each attention module can improve the performance of the network when works independently. Compared with the baseline method LWN, the LWN-SA using the spatial attention module only improved OA by 0.1% on Vaihingen dataset, and the mIoU on UAVid was improved by 0.58%. For LWN-CA using the spatial attention module, these figures are 0.18% and 0.2%, respectively. The data shows that spatial attention has a higher performance improvement on the UAV remote sensing dataset. This is because UAV photography has a stronger ability to distinguish objects, which makes spatial correlation information more necessary.

When two sets of modules work together, the performance of the network is improved more obviously. LWN-A, which does not use the feature early fusion method, improves OA by 0.45% on Vaihingen and mIoU by 1.18% on UAVid. For the LWN-A-F that uses early fusion method, these two results are increased to 0.58% and 1.20%, respectively.

Ablation study has shown that the effect of a single module is a bit limited, and the modules work together to bring better results. It should be noted that the effect of the feature early fusion method is effective but partially influenced by the dataset.

E. Results and Analysis

In this subsection, we will show the experimental results, including data, pictures, and analysis based on the results. We conducted the experiment under conditions in Section IV-B. When testing on those three datasets, we calculated F_1 score, mIoU, OA, and the amount of parameters according to Section IV-C.

1) Result on Vaihingen Dataset: Table III shows the test results of LWN, LWN-A-F, and some popular approaches on the ISPRS Vaihingen Dataset, including RF+dCRF [36], FCN [37], FCN-dCRF [38], SCNN [39], Dilated FCN [38], PSPNet(VGG16) [25], RotEqNet [40], U-Net [24], SegNet [14], ERFNet [41], and BiSeNetV2 [42]. A part of experimental results in that table are referenced from the paper [43], with * on them. Models marked with ‡ take $VGG16$ as backbone with batch-normalization. In the following table, models marked with ‡ have the same settings as above.

Results show that our LWN-A-F model achieved the best mean F_1 in five categories compared to other approaches, and got the best 88.85% overall accuracy, which is 0.58% higher than that of LWN. Compared with U-Net whose model parameters are close, LWN and LWN-A-F have achieved 1.04% and 1.87% improvement on mean F_1, respectively. Especially in the Car category, LWN-A-F with attention mechanism achieved an F1 score of 84.43%, which exceeded the LWN result by 1.68%. These results show that the attention mechanism can effectively extract the global scene context information and improve the performance of the network in the HRRS images semantic segmentation task.

The segmentation masks in the ISPRS Vaihingen dataset are presented in Fig. 5. Significantly, comparing with the results of other methods, our segmentation masks show fewer holes in the building and sharper edges. It shows that our method can better guarantee the integrity of the segmentation results.

2) Result on UAVid Dataset: As a relatively novel dataset, there are a few results tested on UAVid benchmark. As shown in Table IV, this article selects FCN [37], SegNet [14], DeepLab-V3 [45], DeepLab-V3+[15], U-Net [24], MSD [1], ERFNet [41],

TABLE III
EXPERIMENTAL RESULTS ON ISPRS VAIHINGEN VALIDATION DATASET COMPARED WITH SOME APPROACHES

Methods	class F_1%					meam F_1%	mIoU%	OA%	Params.
	Imp. surf.	Build.	Low veg.	Tree	Car				
RF+dCRF [36]*	86.90	92.00	78.30	86.90	29.00	74.60	-	85.90	-
FCN-8s [37]‡	89.45	93.25	78.64	87.79	81.55	86.14	76.12	87.63	19M
FCN-dCRF [38]*	88.80	92.99	76.58	86.78	71.75	83.38	72.28	86.65	-
SCNN [39]*	88.21	91.80	77.17	87.23	78.60	84.40	73.73	86.43	-
Dilated FCN [38]*	90.19	94.49	77.69	87.24	76.77	85.28	-	87.70	-
PSPNet(VGG16) [25]*	89.92	94.36	78.19	87.12	72.97	84.51	73.97	87.62	-
RotEqNet [40]*	89.50	94.80	77.50	86.50	72.60	84.18	-	87.50	-
U-Net [24]	90.15	93.72	78.46	87.85	78.59	85.75	75.66	87.93	13M
SegNet [14]‡	90.56	94.21	78.80	87.61	80.33	86.30	76.42	88.12	44M
ERFNet [41]	89.49	93.21	78.53	87.74	77.53	85.30	74.87	87.54	2M
BiSeNetV2 [42]	87.53	91.10	77.30	87.24	73.87	83.41	72.02	86.11	12M
LWN	90.59	94.41	78.21	87.98	82.75	86.79	77.11	88.27	9M
LWN-A-F	**91.00**	**94.85**	**79.24**	**88.60**	**84.43**	**87.62**	**78.38**	**88.85**	15M

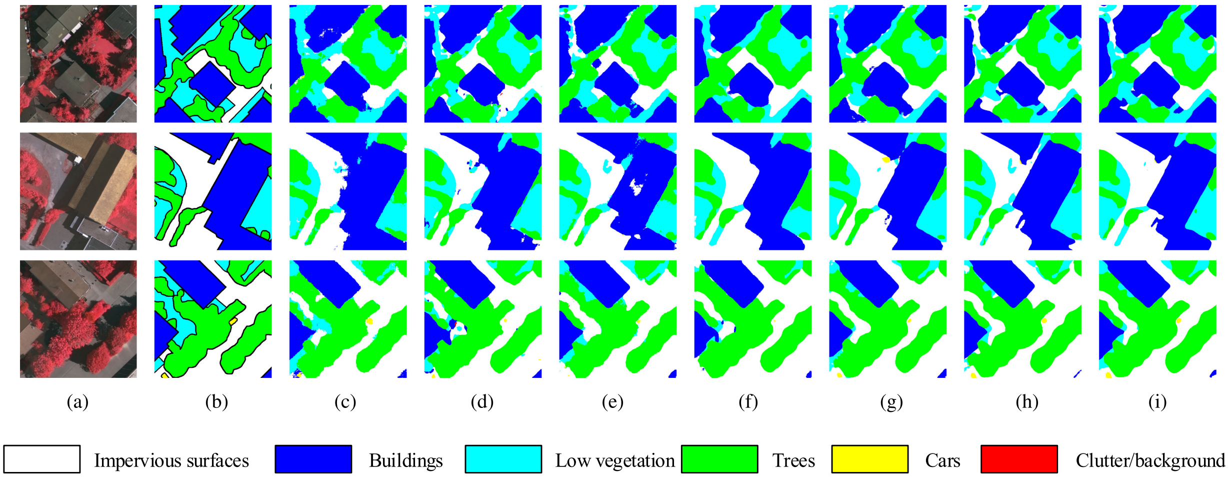


Fig. 5. Semantic segmentation results of three mini patches (512×512) in the ISPRS Vaihingen dataset with FCN, U-Net, SegNet, BiSeNetV2, ERFNet, LWN, and LWN-A-F, respectively. Different colors represent different labels. (a) Image (b) GT (c) FCN (d) U-Net (e) SegNet (f) BiSeNetV2 (g) ERFNet (h) LWN (i) LWN-A-F.

TABLE IV
EXPERIMENTAL RESULTS ON UAVID TEST DATASET COMPARED WITH SOME APPROACHES

Methods	class IoU%								mIoU%	OA%	Params.
	Clutter	Building	Road	Tree	Low veg.	Mov. car	Sta. car	Human			
Dilation Net [44]*	45.40	80.70	65.10	73.80	45.50	53.60	24.50	0.00	48.60	-	-
FCN-8s [37]‡	63.91	84.72	76.51	78.32	61.88	65.87	45.54	22.26	62.38	84.55	19M
SegNet [14]‡	65.62	85.89	79.23	78.78	63.73	68.94	52.10	19.29	64.20	85.54	44M
U-Net [24]	61.80	82.94	75.15	77.27	62.03	59.59	29.98	18.62	58.42	83.43	13M
MSD [1]	57.00	79.80	74.00	74.50	55.90	62.90	32.10	19.70	57.00	-	-
ERFNet [41]	64.50	85.58	77.34	77.87	62.21	60.64	46.13	0.00	59.28	84.69	2M
BiSeNetV2 [42]	61.18	81.62	77.11	75.97	61.30	66.36	38.51	15.40	59.68	83.10	12M
DeepLab-V3 [45]	69.70	88.50	82.10	80.23	65.76	71.75	61.43	21.37	67.60	87.27	59M
DeepLab-V3+ [15]	68.86	87.62	**82.22**	79.76	65.88	69.86	55.39	**26.07**	66.96	86.94	59M
LWN	69.11	88.43	80.55	80.50	65.40	71.12	66.10	21.32	67.82	87.13	9M
LWN-A-F	**70.45**	**88.90**	81.53	**80.89**	**66.72**	**72.68**	**67.68**	23.33	**69.02**	**87.66**	15M

BiSeNetV2 [42], and Dilation Net [44] for comparative experiments on UAVid test set. The result of Dilation Net is cited from MSD [1] marked with *. We analyze the performance as online evaluation. The official benchmark provides three measurement indicators: class IoU, mIoU, and OA on UAVid test set.

On the test set of UAVid, our models achieved quality segmentation results. LWN-A-F got the highest mIoU in six out of eight categories. Compared with the LWN, the mIoU and OA of LWN-A-F increase by 1.20% and 0.53%, respectively, which also exceeded the results of other methods. In the special

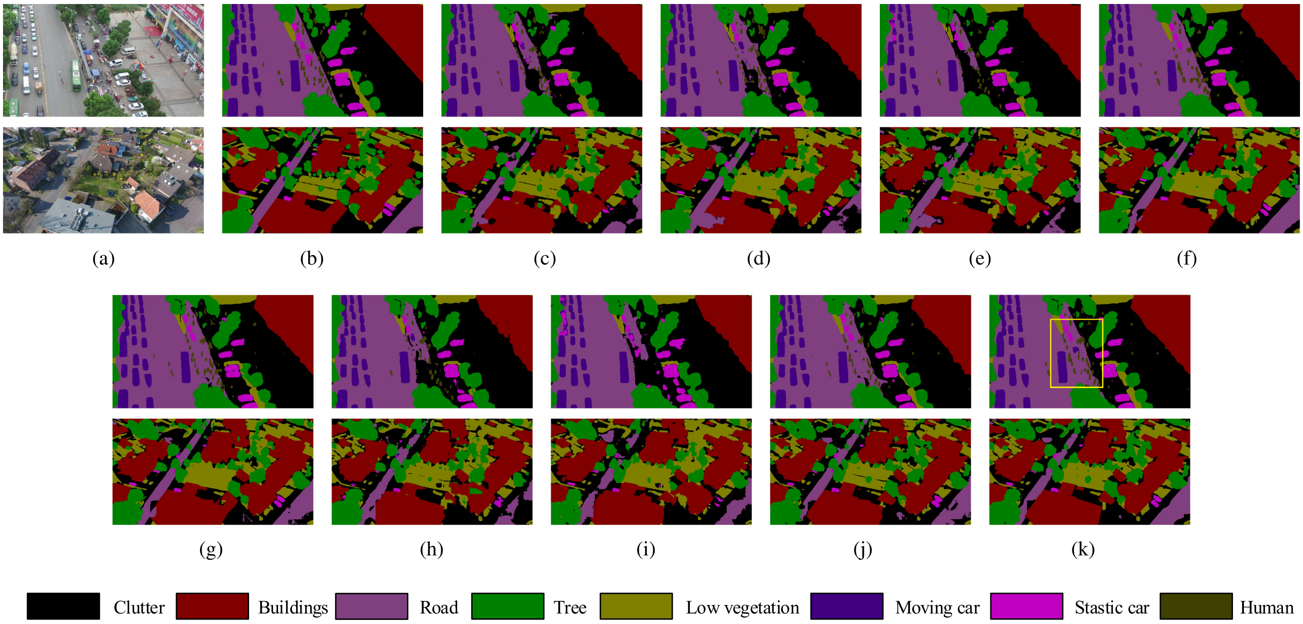


Fig. 6. Semantic segmentation results of 2 patches sized 1920 × 1080 and 4096 × 2160 on the UAVid dataset with FCN, U-Net, SegNet, DeepLab-V3, DeepLab-V3+, BiSeNetV2, ERFNet, LWN, and LWN-A-F, respectively. Different colors represent different labels. (a) Image (b) Ground Truth (c) FCN (d) Segnet (e) U-Net (f) DeepLab-V3 (g) DeepLab-V3+ (h) BiSeNetV2 (i) ERFNet (j) LWN (k) LWN-A-F.

TABLE V
EXPERIMENTAL RESULTS ON UDD6 VALIDATION DATASET COMPARED WITH SOME APPROACHES

Methods	class IoU%						mIoU%	mean F_1%	OA%	Params.
	Other	Facade	Road	Veg.	Vehicle	Roof				
FCN-8s [37]‡	88.38	67.70	66.12	66.22	83.53	58.65	71.77	83.14	85.55	19M
SegNet [14]‡	89.00	71.88	68.72	67.62	86.53	60.70	74.07	84.71	87.08	44M
U-Net [24]	89.34	67.31	65.18	62.56	82.54	58.08	70.84	82.44	85.27	13M
GCN [46]*	-	-	-	-	-	-	69.44	-	85.20	-
ENet [47]*	-	-	-	-	-	-	72.58	-	86.54	-
ERFNet [41]	89.09	72.28	67.40	65.37	86.68	59.24	73.35	84.17	86.91	2M
BiSeNetV2 [42]	89.78	67.99	66.84	58.74	84.44	59.51	71.22	82.65	86.10	12M
DeepLab-V3 [45]	89.82	75.65	**71.68**	71.56	88.42	**64.73**	76.98	86.69	88.74	59M
DeepLab-V3+ [15]	89.71	75.36	71.40	71.54	88.79	64.40	76.87	86.61	88.66	59M
LWN	**90.07**	75.80	70.17	70.86	89.76	64.03	76.78	86.19	88.75	9M
LWN-A-F	90.04	**76.51**	69.98	**71.88**	**90.19**	64.52	**77.19**	**86.79**	**88.93**	15M

category *Human* of the dataset, LWN-A-F with attention also achieved excellent results, and its IoU exceeded LWN without attention mechanism by 2.01%. This result also exceeds those of other light-weight models. The effect diagram of this group of experiments is shown in Fig. 6. Both our LWN and LWN-A-F methods achieved quality segmentation results. Observing the part marked by the yellow box in Fig. 6(k) where pedestrians and the road intersect in the first group of pictures, it can be found that the method with attention modules has a strong resolving ability of capturing relationships among pixels and features, so it can ensure the segmentation integrity of the road surface.

Taking the number of parameters into account, our LWN (9 M parameters) and LWN-A-F (15 M parameters) have much fewer parameters in comparison with other large-scale networks like SegNet (44 M parameters) and DeepLab-V3+ (59 M parameters). Smaller models are conducive to deployment on edge computing devices with low memory constraints, which can be applied to UAVs easily.

3) Result on UDD6 Dataset: As shown in Table V, this article selects FCN [37], SegNet [14], U-Net [24], DeepLab-V3 [45], DeepLab-V3+[15], GCN [46], ENet [47], ERFNet [41], and BiSeNetV2 [42] for comparative experiments on UDD6 validation set. The result of GCN and ENet is cited from UDD6 benchmark,[4] and they are marked with * in the table. We analyze the class IoU, mIoU, mean F_1, and OA on it.

Similarly, the model proposed in this article has obtained high-quality results. Compared with the baseline method LWN, LWN-A-F achieved 0.41% and 0.60% improvement on mIoU and meam F_1, which were higher than other comparison methods. The results on UDD6 are shown in Fig. 7. On the special category *Facade* of UAV oblique projection, our method achieves a high effect of 76.51%. Relying on the attention module to capture the associated information, the *Facade* part of the building can be understand completely as marked by the yellow box in Fig. 7(k).

[4] [Online]. Available: https://github.com/MarcWong/UDD

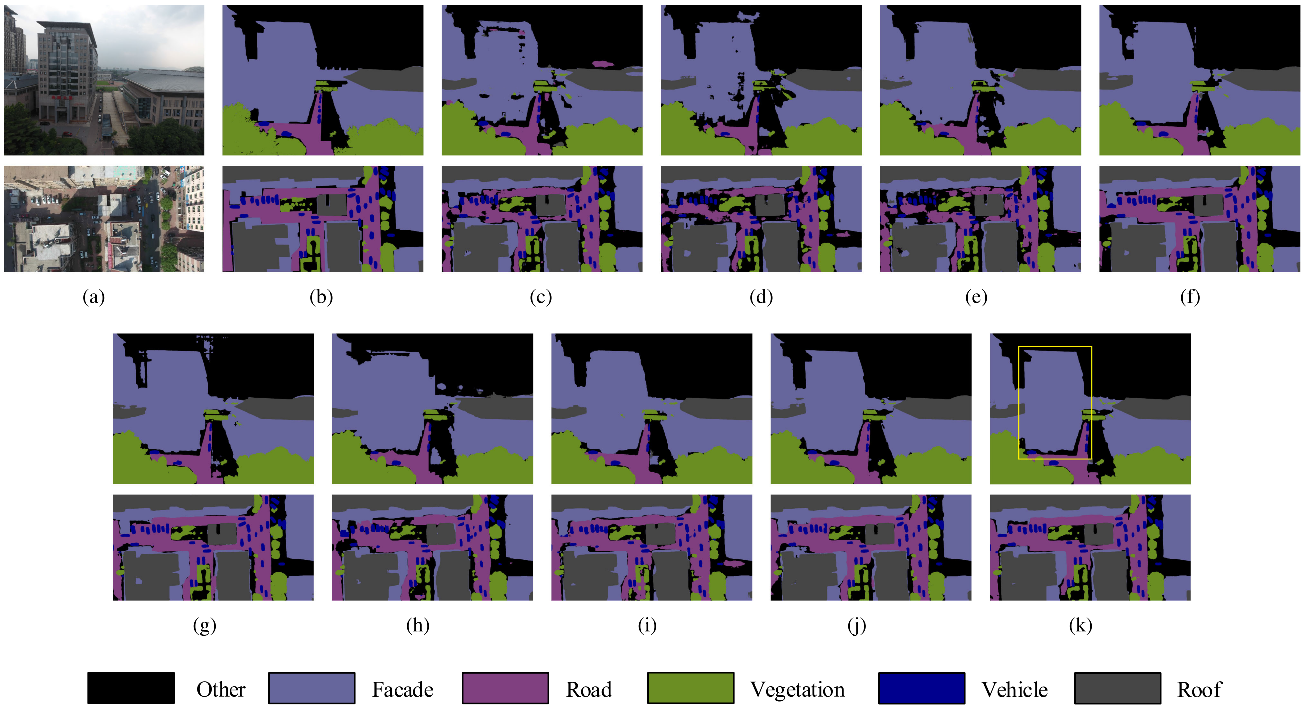


Fig. 7. Semantic segmentation results of two patches sized 4000 × 3000 and 4096 × 2160 on UAVid dataset with FCN, U-Net, SegNet, DeepLab-V3, DeepLab-V3+, BiSeNetV2, ERFNet, LWN, and LWN-A-F, respectively. Different colors represent different labels. (a) Image (b) Ground Truth (c) FCN (d) Segnet (e) U-Net (f) DeepLab-V3 (g) DeepLab-V3+ (h) BiSeNetV2 (i) ERFNet (j) LWN (k) LWN-A-F.

This provides a good foundation for building structural damage detection.

V. Conclusion

In this article, we proposed a light-weight semantic segmentation network with attention modules for UAV remote sensing images. The results of experiments demonstrate that our light-weight network has a high-quality semantic segmentation effect on UAV remote sensing datasets with fewer parameters. In addition, the application of the attention mechanism enhances the performance of our method, and improves the completeness and accuracy of segmentation. In the future work, we would continue to carry out the light-weight design of the model and find a more suitable attention module and feature fusion mechanism.

References

[1] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 165, pp. 108–119, 2020.

[2] Y. Chen, Y. Wang, P. Lu, Y. Chen, and G. Wang, "Large-scale structure from motion with semantic constraints of aerial images," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, 2018, pp. 347–359.

[3] M. Rahnemoonfar, T. Chowdhury, R. Murphy, and O. Fernandes, "Comprehensive semantic segmentation on high resolution UAV imagery for natural disaster damage assessment," in *Proc. IEEE Int. Conf. Big Data*, 2020, pp. 3904–3913.

[4] M. Erdelj and E. Natalizio, "UAV-assisted disaster management: Applications and open issues," in *Proc. Int. Conf. Comput. Netw. Commun.*, 2016, pp. 1–5.

[5] L. Hashemi-Beni and A. A. Gebrehiwot, "Flood extent mapping: An integrated method using deep learning and region growing using UAV optical data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2127–2135, 2021.

[6] N. Tijtgat, W. V. Ranst, B. Volckaert, T. Goedeme, and F. D. Turck, "Embedded real-time object detection for a UAV warning system," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2017, pp. 2110–2118.

[7] M. Khan, W. Ectors, T. Bellemans, D. Janssens, and G. Wets, "Unmanned aerial vehicle-based traffic analysis: A case study for shockwave identification and flow parameters estimation at signalized intersections," *Remote Sens.*, vol. 10, no. 3, Mar. 2018, Art. no. 458.

[8] R. Ke, Z. Li, J. Tang, Z. Pan, and Y. Wang, "Real-time traffic flow parameter estimation from UAV video based on ensemble classifier and optical flow," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 54–64, Jan. 2019.

[9] J. Zhu *et al.*, "Urban traffic density estimation based on ultrahigh-resolution UAV video and deep neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4968–4981, Dec. 2018.

[10] J. F. Galarreta, N. Kerle, and M. Gerke, "UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning," *Natural Hazards Earth Syst. Sci.*, vol. 15, no. 6, pp. 1087–1101, 2015.

[11] N. Kerle, F. Nex, M. Gerke, D. Duarte, and A. Vetrivel, "UAV-based structural damage mapping: A review," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 1, 2020, Art. no. 14.

[12] F. Nex, D. Duarte, A. Steenbeek, and N. Kerle, "Towards real-time building damage mapping with low-cost UAV solutions," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 287.

[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] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[15] 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.

[16] 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.

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

[18] S. Liu, C. He, H. Bai, Y. Zhang, and J. Cheng, "Light-weight attention semantic segmentation network for high-resolution remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2595–2598.

[19] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "UAV-YOLO: Small object detection on unmanned aerial vehicle perspective," *Sensors*, vol. 20, no. 8, 2020, Art. no. 2238.

[20] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[21] H. Huang, X. Li, and C. Chen, "Individual tree crown detection and delineation from very-high-resolution UAV images based on bias field and marker-controlled watershed segmentation algorithms," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2253–2262, Jul. 2018.

[22] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 375–391.

[23] S. Teng, S. Zhang, Q. Huang, and N. Sebe, "Viewpoint and scale consistency reinforcement for UAV vehicle re-identification," *Int. J. Comput. Vis.*, vol. 129, no. 3, pp. 719–735, 2021.

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

[25] 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.

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

[27] 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.

[28] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[29] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9522–9531.

[30] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325–341.

[31] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

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

[33] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[34] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6687–6696.

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

[36] N. T. Quang, N. T. Thuy, D. V. Sang, and H. T. T. Binh, "An efficient framework for pixel-wise building segmentation from aerial images," in *Proc. Int. Symp. Inf. Commun. Technol.*, 2015, pp. 282–287.

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

[38] L. 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.

[39] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial CNN for traffic scene understanding," in *Proc. AAAI Conf. Artif. Intell*, 2018, pp. 7276–7283, [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/12301.

[40] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 96–107, 2018.

[41] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.

[42] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation," 2020, *arXiv:2004.02147*.

[43] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12408–12417.

[44] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Representations*, 2016, [Online]. Available: https://dblp.uni-trier.de/rec/journals/corr/YuK15.html?view=bibtex.

[45] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[46] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4353–4361.

[47] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

Siyu Liu received the B.Eng. degree in electronic information engineering in 2019 from the University of Electronic Science and Technology of China, Chengdu, China, where he is currently working toward the Ph.D. degree in information and communication engineering with the School of Information and Communication Engineering.

His research interests include human action recognition in computer vision and pattern recognition, and remote sensing image processing.

Jian Cheng received the Ph.D. degree in pattern recognition and intelligent system from Shanghai Jiao Tong University, Shanghai, China, in 2006.

From 2006 to 2007, he was an Assistant Researcher with the Chengdu Information Technology of Chinese Academy of Sciences Co., Ltd. He is currently a Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include machine learning, computer vision, remote sensing image analysis, multimodal image classification, video surveillance and scene understanding, and human behavior analysis.

Leikun Liang (Graduate Student Member, IEEE) received the B.Eng. degree in electronic information engineering in 2019 from the University of Electronic Science and Technology of China, Chengdu, China, where he is currently working toward the master's degree in information and communication engineering.

His research interests include reidentification in computer vision and pattern recognition, and remote sensing image processing.

Haiwei Bai received the bachelor's degree in electronic information engineering from the University of Electronic Science and Technology of China, Chengdu, China, where he is currently working toward the Ph.D. degree in information and communication engineering with the School of Information and Communication Engineering.

His research interests include computer vision, image processing, deep learning, and the method and application of semantic segmentation of high-resolution remote sensing images.

Wanli Dang is currently working toward the Ph.D. degree in electronics and information technology with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China.

She is currently an Engineer with The Second Research Institute, Civil Aviation Administration of China, Beijing, China. Her research interests include human action recognition for airport, pedestrian detection, and tracking.