Road and Car Extraction Using UAV Images via Efficient Dual Contextual Parsing Network

Yueming Sun, Zhenfeng Shao, Gui Cheng, Xiao Huang, and Zhongyuan Wang, *Member, IEEE*

***Abstract*— The rapid development and commercialization of unmanned aerial vehicle (UAV) technology has made it possible to conduct urban traffic information extraction using UAV images. However, the large variations of targets in urban environments, complex foregrounds and backgrounds in cities, and severe tree and shadow occlusions pose great challenges in car and road extraction using UAV images. In this study, we propose a lightweight, efficient dual contextual parsing network (EDCPNet) to address the above issues. The proposed efficient dual contextual parsing (EDCP) module in EDCPNet is mainly composed of spatial contextual parsing (SCP) and channel contextual parsing (CCP), which can effectively acquire rich contextual features in both spatial and channel dimensions, adaptively recalibrate the attention weights, perceive the salient features of targets in images, and suppress the importance of irrelevant elements. It, thus, leads to improved performance and adaptability that facilitate the practical applications of large-scale urban traffic monitoring in complex urban scenes. We conduct experiments on two benchmark datasets [UAV image dataset (UAVid) and urban drone dataset (UDD)] by comparing the proposed EDCPNet with six other competing methods, i.e., U-Net, PSPNet, Deelabv3+, SegNet, ESNet, and ERFNet, and validate the effectiveness of the proposed EDCP module via extensive ablation studies. The results suggest that the proposed network outperforms all competing methods in car and road extraction from UAV images with a balanced computational cost. Its great performance and low computational demand (with only 2.37M model parameters) facilitate its deployment on edge computing devices with memory constraints.**

***Index Terms*— Attention mechanism, car extraction, lightweight network, road extraction, unmanned aerial vehicle (UAV) images.**

Manuscript received 2 April 2022; revised 6 September 2022; accepted 6 October 2022. Date of publication 14 October 2022; date of current version 27 October 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42090012, in part by the Guangxi Science and Technology Program (Guangxi Key Research and Development Plan) under Grant GuiKe 2021AB30019, in part by the 03 Special Research and 5G Project of Jiangxi Province in China under Grant 20212ABC03A09, in part by the Zhuhai Industry University Research Cooperation Project of China under Grant ZH22017001210098PWC, in part by the Sichuan Science and Technology Program under Grant 2022YFN0031, in part by the Hubei Key Research and Development Plan under Grant 2022BAA048, and in part by the Zhizhuo Research Fund on Spatial-Temporal Artificial Intelligence under Grant ZZJJ202202. *(Corresponding author: Zhenfeng Shao.)*

Yueming Sun, Zhenfeng Shao, and Gui Cheng are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: sunyueming@whu.edu.cn; shaozhenfeng@whu.edu.cn; chenggui@whu.edu.cn).

Xiao Huang is with the Department of Geosciences, University of Arkansas, Fayetteville, AR 72701 USA (e-mail: xh010@uark.edu).

Zhongyuan Wang is with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: wzy_hope@163.com).

Digital Object Identifier 10.1109/TGRS.2022.3214246

I. Introduction

THE rapid development of unmanned aerial vehicle (UAV) technology allows consumers from various fields to receive high spatial resolution aerial images in a more easy and economically friendly manner [1]. UAV remote sensing images are being gradually applied in various fields, such as land and resource surveys [2], urban construction and planning [3], agricultural production [4], natural disasters [5], and crisis management [6]. Even today, the number of flights of different types of UAVs is still increasing, with a developing trajectory toward high definition, miniaturization, and high endurance time. UAV remote sensing plays an increasing role in urban management as data support. Among many UAV-based tasks, traffic monitoring is considered one of the main challenges that remain unsolved [7], [8], [9].

In the task of traffic monitoring, in addition to the extraction of the car, retrieving road locations in images is equally important for the collection of traffic information and road surface information. As important research objects in traffic engineering science, road targets and car targets have been investigated in traffic management, traffic prediction, traffic planning, traffic control and guidance, and traffic command, to list a few [10]. Traditional road and car extraction surveys are mostly based on static surveillance cameras or satellite-based remote sensing platforms; however, static surveillance cameras have a limited field of view and are prone to blind spots, and satellite-based remote sensing platforms cannot meet the traffic monitoring requirements for real-time information [11]. In comparison, the UAV platform is more flexible and mobile, able to obtain better details of road surfaces and cars, support controlled flight paths, and capture multiview images (broader detection scope). They are not limited by platform revisit intervals and are not limited by road congestion, which is especially important in traffic regulation tasks that require rapid response, providing an ideal platform for car and road extraction.

With the current popularity of artificial intelligence and deep learning, the trend of using convolutional neural networks (CNNs) to extract information from images has been increasingly evident since the birth of AlexNet [12] in 2012. Among them, Long et al. [13] proposed a fully convolutional network (FCN) that consists of an encoder and a decoder, where the encoder is similar to a traditional CNN extracting deep abstract features and the decoder recovers these features into a prediction map of the same size as the input image. FCN uses convolutional layers instead of fully connected layers to achieve end-to-end semantic segmentation, overcoming the inefficiency of sliding window-based CNN methods that can extract semantic text information while preserving the target spatial information [14]. Building upon the basic FCN model, U-Net [15], DeconvNet [16], and SegNet [17] adopt an encoder–decoder structure to fuse the deep features in a top-to-low layer manner and improve the segmentation accuracy. Most of the deep learning algorithms used for road extraction are based on CNN and FCN [18]. Zhang et al. [19] improved the U-net by combining residual learning to improve the generalization ability of the network model while reducing the number of parameters by multilayer jump connections. Wan et al. [20] used a shallow encoder–decoder framework with densely connected blocks to build DA-Road Net, which reduces information loss and integrates road features in spatial and channel dimensions to improve road extraction continuity. DA-CapsUNet [21] is a dual-attention capsule U-Net with feature attention modules to extract and fuse multiscale contextual information to enhance road extraction accuracy. In summary, deep learning can extract deep abstract features hidden in images and has a better performance on complex scenes. There are also many studies on UAV image road extraction. Kestur et al. [22] proposed a U-shaped FCN (UFCN) for road extraction and demonstrated its great potential for UAV image road extraction, but the dataset used was small and the network failed to be well trained. Varia et al. [23] applied the fully convolutional network FCN-32 and the generative adversarial network (GAN) for UAV road extraction. However, it is prone to misclassifying nonroad regions as road regions in certain regions of high complexity, reducing the completeness and accuracy of the final output. Senthilnath et al. [24] used deep transfer learning with ensemble classifier (TEC) (deep migratory learning with integrated classifiers) for road extraction from UAV images, using conditional GANs, recurrent GANs, and FCNs to pretrain and then integrate the classification; however, the integration of multiple deep learning networks caused a decrease in computational efficiency.

Existing car extraction methods can be roughly classified into two categories: car detection and car semantic segmentation. The purpose of car detection is to detect all cars and locate them in the image, usually in the form of bounding boxes with confidence scores [25], [26], [27]. In contrast, semantic segmentation of the car can determine the car range by classifying each pixel in addition to localization. Compared with car detection, it can give more accurate pixel extraction results and can better determine the car's conditions. In particular, using only bounding box detection is not a good solution in the case of high car traffic [28]. A relatively small amount of work addresses this problem. In [29], a deep learning-based "segment-before-detect" method is proposed for semantic segmentation and subsequent classification of cars in high-resolution remote sensing images. Masouleh and Shah-Hosseini [30] combined Gaussian–Bernoulli restricted Boltzmann machine (GB-RBM) and CNN to improve the performance of a deep learning model for segmenting ground cars from UAV-based thermal infrared images. Masouleh and Shah-Hosseini [31] combined a deep CNN and an RBM to construct RBMDeepNet to achieve semantic segmentation of cars from high-resolution airborne images; however, it is prone to treating the redundant shadow pixels as the car.

Several research works addressed the problem of simultaneous extraction of road and car. DLT-Net [32] proposed a unified neural network to simultaneously detect drivable areas and cars, constructing context tensors between subtask decoders to share specified weights between tasks. Bianco et al. [33] jointly detected car and lane based on the ERFNet semantic segmentation network, using weak and strong labels to combine different annotation sources in order to generate a new dataset containing all categories of interest. However, these methods are limited to images from in-car cameras with a simpler image background, a single view, and no access to global information about the traffic [34].

In addition, the large variations of targets in urban environments, complex foregrounds and backgrounds in cities, and severe tree and shadow occlusions pose great challenges in car and road extraction on UAV images [35]. Due to the tilted camera view of UAVs, large variations of objects at different distances or in different categories in UAV remote sensing images can occur. In addition, road objects can be so large that they may exceed the neural network perceptual field, while small car objects are difficult to identify. In urban scenes, the backgrounds and foregrounds in UAV remote sensing images can be complex, while other objects are on the ground. What is more, the occlusion of trees and shadows can lead to the loss or misrepresentation of certain road or car features, resulting in incorrect identification and classification.

The attention mechanism aims to direct the attention of deep networks to more relevant features of the corresponding tasks. The attention mechanism acquires weight parameters and the action of the weight parameters on the original feature map, thus enhancing the representation of the more informative features and suppressing the representation of the less informative features. SENet [36] calibrates the importance of a channel using channelwise global information. However, this direct approach by constructing a global representation fails to consider the dependencies between different feature values within the same channel. NLNet [37] obtains an attentional relationship between the feature value of each position and the feature values of all other positions using a self-attentive mechanism. This approach, however, demands massive computational resources. In addition, Woo et al. [38] proposed CBAM that pools the middle layer feature tensor of the network along the channel and spatial dimensions, respectively, and further uses CNN to learn its attention weights. However, the limited perceptual field of CNNs makes it difficult to infer the importance of local features using global information.

To mitigate the aforementioned issues, we propose an efficient dual contextual parsing network (EDCPNet) for car and road extraction from UAV remote sensing images. The main contributions of this article are given as follows.

1) We propose a UAV-based lightweight car and road extraction network, named EDCPNet, to acquire rich contextual features and perceive local salient characteristics of salient targets, thus enhancing the model's ability

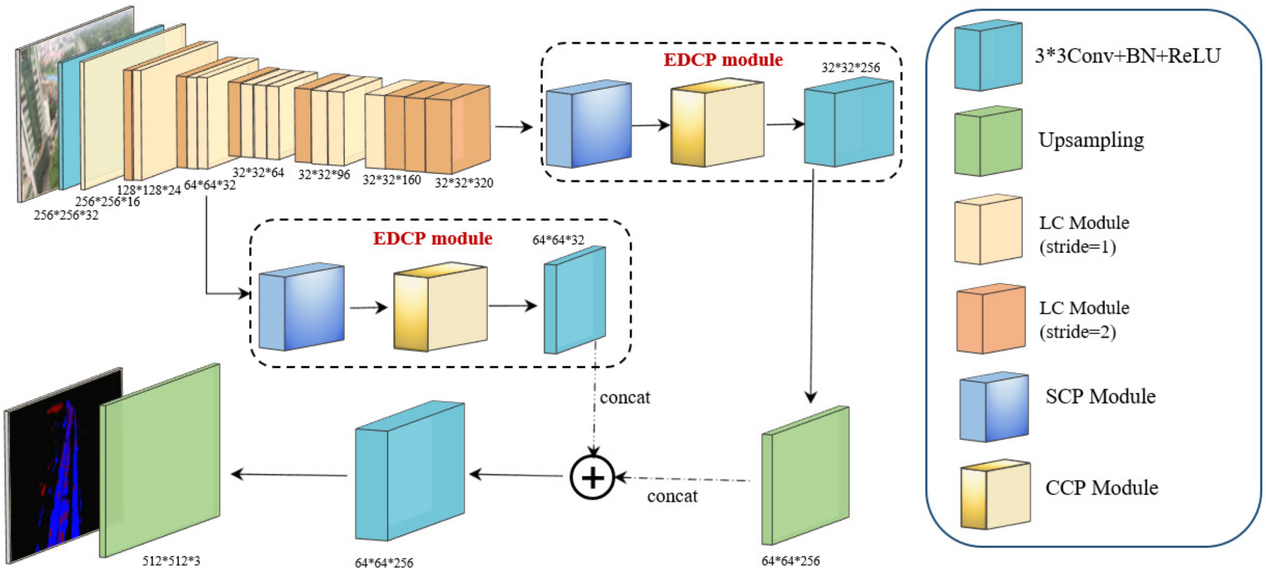


Fig. 1. Overall structure of the proposed EDCPNet.

to resolve target scale variations and perform road and car detection in complex urban scenes.

2) We propose an efficient dual contextual parsing (EDCP) module mainly composed of spatial contextual parsing (SCP) and channel contextual parsing (CCP), which is capable of acquiring contextual information and global information about the target in both spatial and channel dimensions to adapt to the target context and learn richer feature representations. The proposed EDCP module can adaptively recalibrate attention weights to perceive salient features of salient targets in images and suppress the importance of irrelevant elements by parsing the target context.

3) We conduct extensive experiments on the UAV remote sensing semantic segmentation datasets, UAV image dataset (UAVid) and urban drone dataset (UDD), and ablation studies to validate the great performance of the proposed method in both qualitative and quantitative results.

The rest of this article is organized as follows. Section II describes the proposed EDCPNet in detail. Sections III and IV describe the contrast and ablation experimental procedure in detail and present the experimental results, respectively. Section V concludes this study.

II. Proposed Framework

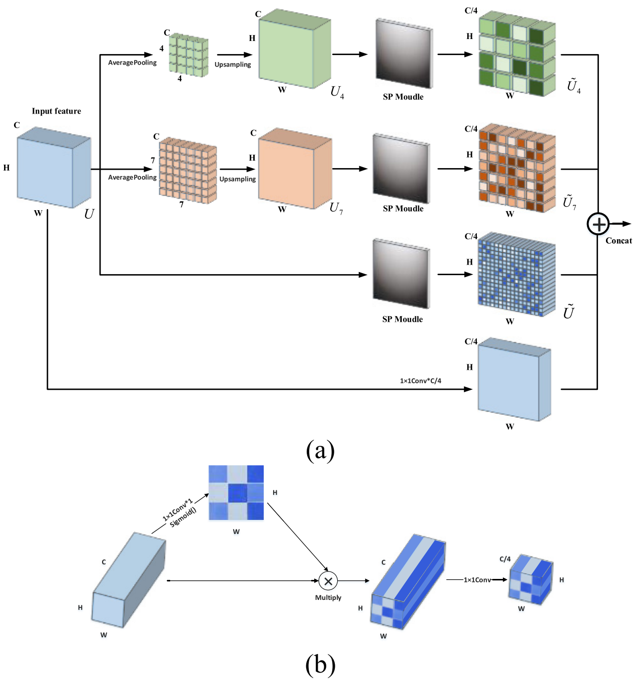
In this section, we describe our proposed EDCPNet in detail. We first describe the basic structure of the EDCPNet and then present the key components that include the SCP module, the CCP module, the EDCP module, and the lightweight convolutional (LC) module.

The network structure proposed in this article is shown in Fig. 1. To better adapt to the platform of UAV applications, in the feature extraction part, we build an LC module for feature extraction by adopting the inverted residuals and linear bottleneck structures from the Mobilenetv2 [36], aiming to reduce the number of parameters in the network while maintaining the extraction accuracy. The 64 × 64 × 32 feature maps with eightfold downsampling and the 32 × 32 × 320 feature maps with 16-fold downsampling are fed into our EDCP module, which includes the SCP module and then the CCP module, to integrate and parse the global information and contextual information of the target. The proposed EDCP module adaptively recalibrates attention weights to parse the target context and learn richer feature representations. The module leads to the perception of the salient features of salient targets in images, suppressing the importance of irrelevant elements and establishing cross-dimensional interactions. Furthermore, the 32 × 32 × 256 feature map processed by the EDCP module is performed by a twofold upsampling and concatenated to the 64 × 64 × 32 shallow features processed by the EDCP module with the same spatial resolution. The fusion of deep and shallow features is a fusion of detailed features (shallow features) and global features (deep features) of the target, after which the concatenated features are refined with a 3 × 3 convolution, followed by a bilinear upsampling with a multiplicity of 8, so as to generate the final prediction map.

A. Spatial Contextual Parsing Module

In car and road extraction from UAV images, there exist huge scale differences between car and road targets; therefore, attention needs to be focused on various target objects in different scale contexts.

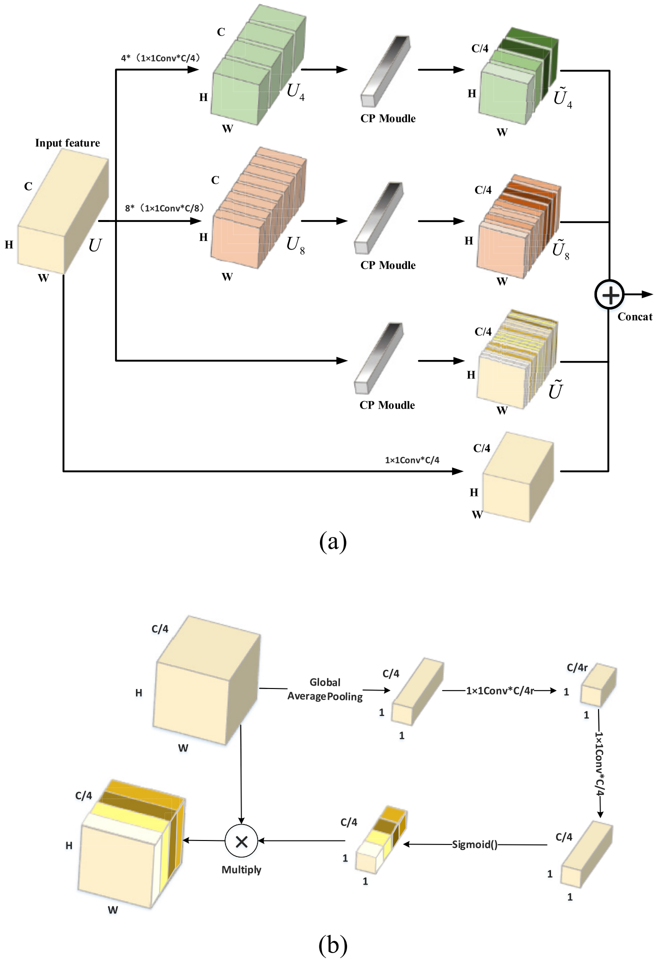
Inspired by the squeeze-excitation (SE) mechanism [36], we aim to design a new network with pixel-level attention, which can extract accurate features of cars and roads. In consideration of that, spatial pyramid structures can extract feature information at different scales and increase the pixel-level field of perception but lack context-first attention to features at different scales, while the simple use of the SE mechanism cannot effectively extract multiscale features and lacks the



(a)

(b)

Fig. 2. Structure of the SCP module. (a) Structure of the SCP module. (b) Structure of the SP module in SCP.



(a)

(b)

Fig. 3. Structure of the CCP module. (a) Structure of the CCP module. (b) Structure of the CP module in CCP.

difference in attention to target objects at different scales. We combine the attention mechanism and spatial pyramid to extract precise dense features for pixel labeling. Different from most existing works, we design the SCP module that utilizes the SE mechanism at multiple spatial pooling scales to focus on global, car, and road contextual pixel information at different scales. The SCP module performs SCP operations, as shown in Fig. 2(a). The SCP module consists of a spatial context feature extraction unit and a spatial parsing (SP) unit, as shown in Fig. 2(b).

$U = [u^{1,1}, u^{1,2}, u^{i,j}, \ldots, u^{H,W}]$ is the input feature, where $u^{i,j} \in \mathrm{R}^{1\times 1\times C}$ represents the channel features at each spatial location, $i \in \{1, 2, \ldots, H\}$, $j \in \{1, 2, \ldots, W\}$, and H, W, and C refer to the height, the width, and the number of input channels, respectively. In the spatial context feature extraction unit, we use different sizes of average pooling kernels and step sizes to obtain different subregion spatial grids to further reduce the loss of contextual information between different subregions and preserve information of different scales. Considering the scale of road and car, the size of the spatial grid is set to 3 × 3 and 6 × 6, respectively, for road and car. In addition, to obtain a compact and discriminative descriptor, the size of the average pooling window is set to $[2 \times W/(N+1)]$ and $[2 \times H/(N+1)]$ in steps of $[W/(N+1)]$ and $[H/(N+1)]$ to have approximately 50% overlap on each side, and the output feature map size is $[(N+1) \times (N+1) \times C]$. To maintain the spatial information of the features, we upsample the feature maps of different scales to the size of the original feature maps to obtain the set of spatial feature maps U_{PS}

$$U_{\mathrm{PS}} = [U, U_4, U_7]. \tag{1}$$

In the SP unit, each spatial feature map in the set is first convolved by a standard 1 × 1 convolution layer to produce a 2-D spatial attention map q, and $q^{i,j}$ represents a linear combination of all channels C at a spatial location $(i,\ j)$ in that map. The obtained q is activated by a sigmoid function σ, and the obtained spatial weight values are used to recalibrate the feature maps, attaching more importance to the relevant spatial locations while ignoring the irrelevant ones

$$\hat{U} = \left[\sigma\left(q^{1,1}\right)u^{1,1}, \ldots, \sigma\left(q^{i,j}\right)u^{i,j}, \ldots, \sigma\left(q^{H,W}\right)u^{H,W}\right]. \tag{2}$$

Then, the feature channel dimension is reduced to 1/4 of the original one by a 1 × 1 convolution layer, i.e., $\tilde{U} = W_S * \hat{U}$, where $W_S \in \mathrm{R}^{1\times 1\times (C/4)}$ to obtain the SCP feature map set $\tilde{U}_{\mathrm{PSA}}$

$$\tilde{U}_{\mathrm{PSA}} = \left[\tilde{U}, \tilde{U}_4, \tilde{U}_7\right]. \tag{3}$$

Finally, the channel dimension of the original feature map U is reduced to 1/4 of the original one and aggregated with $\tilde{U}_{\mathrm{PSA}}$ into the final output feature O, thus accelerating the model learning and alleviating the gradient disappearance.

B. Channel Contextual Parsing Module

In CNN, information differs in each channel. The proposed CCP module performs CCP operations, as shown in Fig. 3(a). Similarly, the CCP module includes a channel context feature extraction unit and a channel parsing (CP) unit, as shown in Fig. 3(b). In the channel context feature extraction unit, the input feature map U is divided into K parts in the channel dimension $U = [u_1, u_i, \ldots, u_K]$, and the number of channels in each part is $C' = C/K$, i.e., the ith feature map $u_i \in \mathrm{R}^{H\times W\times C'}$ where $i \in \{1, 2, \ldots, K\}$, and K needs to be divided by C. We further carry out channel information extraction in different dimensions and obtain the channel feature map set U_{PC}

$$U_{\mathrm{PC}} = [U, U_4, U_8] \tag{4}$$

where $U = W_c * U$ and $W_c \in \mathrm{R}^{1\times 1\times C'}$.

In the CP unit, weights are added to each channel to highlight the channelwise correlation. First, global average pooling is used to encode the entire features on the channel as a global feature to produce a vector $S \in \mathrm{R}^{1\times 1\times C}$

$$s_z = \frac{1}{H\times W}\sum_{i}^{H}\sum_{j}^{W} u_z(i, j). \tag{5}$$

To combine the linear information between channels and reduce the model complexity, a bottleneck structure that contains two fully connected layers is used, where the first FC layer is downscaled and then activated using the ReLU function, and the final FC layer restores the original dimensionality. The channel weight values are then obtained through the sigmoid layer, which falls between [0, 1]

$$\hat{S} = \sigma(W_2(\varsigma(W_1 S))) \tag{6}$$

where ς represents the ReLU activation function, and $W_1 \in \mathrm{R}^{C\times(C/r)}$ and $W_2 \in \mathrm{R}^{(C/r)\times C}$ represent the fully connected layers. σ represents the sigmoid activation function, and the obtained channel weight values recalibrate the feature map

$$\hat{U} = [\hat{s}_1 u_1, \hat{s}_i u_i, \ldots, \hat{s}_c u_c]. \tag{7}$$

Same as the SCP module, the feature dimension is reduced to 1/4 of the original one by a 1 × 1 convolution layer to obtain a multiscale set of spatial attention feature maps $\tilde{U}_{PCA}$

$$\tilde{U}_{PCA} = \left[\tilde{U}, \tilde{U}_4, \tilde{U}_8\right]. \tag{8}$$

Finally, the channel dimension of the original feature map U is reduced to 1/4 of the original and aggregated with $\tilde{U}_{PCA}$, forming the final output feature O.

C. Efficient Dual Contextual Parsing Module

We combine two modules, i.e., SCP and CCP described above, to simultaneously obtain contextual features in both spatial and channel dimensions and recalibrate the input features, which encourages the network to further parse the scenes of the target context and establish the association of feature maps spatially and channelwisely. After extensive experiments, we found that the EDCP module where the SCP modules are tandem first and channel integration is followed by channel integration in spatial priority order (see Fig. 4) yields the best results.

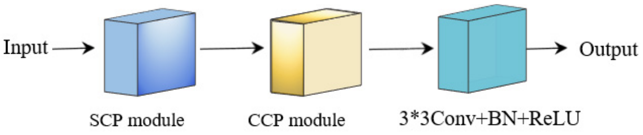
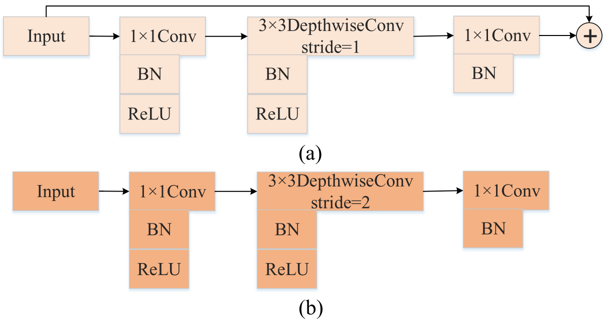


Fig. 4. Structure of the EDCP module.



(a)

(b)

Fig. 5. Structure of the LC module. (a) Stride = 1 block. (b) Stride = 2 block.

D. Lightweight Convolutional Module

Mobilenetv1 [39] is a lightweight CNN network proposed by Google, designed to separate the standard convolutional process into depthwise convolution (DC) and point convolution (PC). DC assigns a separate convolutional kernel to each channel of the input feature map for convolution, and PC convolution uses 1 × 1 convolution to perform standard convolutional operations on the results of DC. Depthwise separable convolution (DSC) reduces the total computation to 1/4 of the standard convolution by decomposition. At the same time, almost all computations are concentrated in the 1 × 1 convolution operation, which does not require reordering in memory, leading to accelerated computation. Mobilenetv2 [40] introduces inverse residual and linear bottleneck structures compared to Mobilenetv1. The inverted residual structure replaces the 3 × 3 standard convolution with 3 × 3 DSC, which significantly reduces the computational demand so that a layer of 1 × 1 PC can be added before the DSC to boost the number of channels, thus leading to improved performance. After the 3 × 3 DSC, the bottleneck design is followed by the 1 × 1 PC to reduce the dimensionality and summarize the input. The resulting LC module that we constructed is shown in Fig. 5, with the LC module with stride = 1 adding residual connections to the LC module with stride = 2.

E. Loss Functions

The commonly used loss function for classification problems is the cross-entropy loss

$$L_{\mathrm{ce}} = -\sum_{i=1}^{N} p_i \log(q_i) \tag{9}$$

where p_i is the predicted probability value of the ith pixel of the image, q_i is the true value of the ith pixel of the image, and N is the total number of pixels. In the road and car extraction tasks of UAV remote sensing images, the total number of pixels differs between the extraction target and the background. Thus, we introduce dice loss to increase the intersection ratio of predicted and true results during the training process, which is expected to mitigate data imbalance issues and improve the extraction accuracy. The dice loss (L_{dice}) is computed as

$$L_{\text{dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i q_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} q_i^2}. \tag{10}$$

The overall loss function L_{total} is a combination of dice loss and cross-entropy loss: $L_{\text{total}} = L_{\text{ce}} + L_{\text{dice}}$.

III. Experimental Results and Analysis

To evaluate the performance of the proposed EDCPNet in road and car extraction tasks from the UAV remote sensing images, we conducted experiments on two UAV image semantic segmentation datasets, i.e., UAVid and UDD. The details regarding these two datasets and the experimental settings are described in Sections III-A and III-B, respectively. Section III-C presents the evaluation criteria used in this article. Section III-D presents the comparison results on UAVid and UDD.

A. Datasets

The UAVid [41] is a semantic segmentation dataset that contains high-resolution UAV remote sensing images in street scenes. It contains 300 UAV remote sensing images of 4096 × 2160 or 3840 × 2160 pixels. Unlike other popular remote sensing datasets, the UAVid contains images captured from a tilted photographic view, which brings a large-scale variance and is more challenging in urban street scenes with complex foreground–background objects. Due to the large size of the training set images, we crop 10 000 random nonoverlapping 512 × 512 small patches from the UAVid, whose 8000 are used as the training set, 1000 as the validation set, and 1000 as the test set. For our task requirements, each pixel in the image is labeled as one of three categories: road, car, and background.

The UDD [42] is another semantic segmentation dataset that contains high-resolution UAV images collected in multiple cities. It contains a total of 141 UAV remote sensing images with resolutions of 4000 × 3000 or 4096 × 2160 pixels. Unlike the UAVid dataset, view angles in UDD vary, and its images are mainly targeted at urban scenes in living areas, e.g., residential areas and roads inside the campus. Similarly, we randomly crop 10 000 patches with 512 × 512 in size (without overlapping), whose 8000 are used as the training set, 1000 as the validation set, and 1000 as the test set. Similarly, each pixel in the image is labeled as either road, car, or background.

B. Experimental Platform and Parameter Configuration

In the following, we list the hardware configuration of the algorithm experiment platform.

1) CPU: Intel[1] Xeon[1] E5-2687W v4 cores at 3.00 GHz; base frequency: 3.00 GHz; and memory: 16 GB; GPU: NVIDIA GRID RTX8000-8Q GPU; memory: 12 GB. Batch size = 16, initial learning rate = 1e-4, and epoch = 100. The model performs gradient descent using the Adam optimizer, and the learning rate is reduced to half when the loss value of the validation set does not drop three times.

C. Evaluation Indicators

The evaluation criteria used in this article are F1 score, category mean pixel accuracy (MPA), and Mean Intersection over Union (MIoU) ratio.

We calculate the F1 score with the following formula:

$$\mathrm{F1} = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}. \tag{11}$$

The MPA can be computed as

$$\mathrm{MPA} = \frac{1}{k}\sum_{i=0}^{k} \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{12}$$

The MIoU can be computed as

$$\mathrm{MIoU} = \frac{1}{k}\sum_{i=0}^{k} \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{13}$$

where the precision rate $\mathrm{P} = (\mathrm{TP}/\mathrm{TP} + \mathrm{FP})$; recall rate $\mathrm{R} = (\mathrm{TP}/\mathrm{TP} + \mathrm{FN})$; k is the number of categories; and TP, FP, and FN, respectively, represent the number of true positives, false positives, and false negatives.

D. Experimental Results and Analysis

1) Experiments on UAVid: We conduct road and car extraction experiments on the UAVid dataset. The purpose of these experiments is to evaluate the performance of the proposed EDCPNet in complex urban scenarios (common scenarios for traffic monitoring tasks) and verify the superiority of EDCPNet compared to other competing algorithms. Selected detection results obtained by EDCPNet with other competing algorithms, i.e., U-Net [15], SegNet [17], PSPNet [43], DeepLabv3+ [44], ESNet [45], and ERFNet [46], are presented in Fig. 6, while quantitative comparisons with competing algorithms are shown in Table I.

In general, our EDCPNet method achieves high-quality segmentation results on the test set of UAVid, presenting a notable advantage over the other six models. As can be seen from Table I, compared with other models, the proposed EDCPNet achieves the highest accuracy in the road and car classes, with MIoU and MPA reaching 86.97% and 94.21%, respectively. The MIoU of the proposed model is 2.02% and 1.38% higher than U-Net and DeepLabv3+, 6.09% and 6.16% higher than PSPNet and SegNet, and 2.37% and 3.12% higher than the lightweight models ESNet and ERFNet, respectively. We observe that EDCPNet presents higher classification accuracy for roads, evidenced by its capability in accurately segmenting the general road contours. In addition,

[1] Registered trademark.

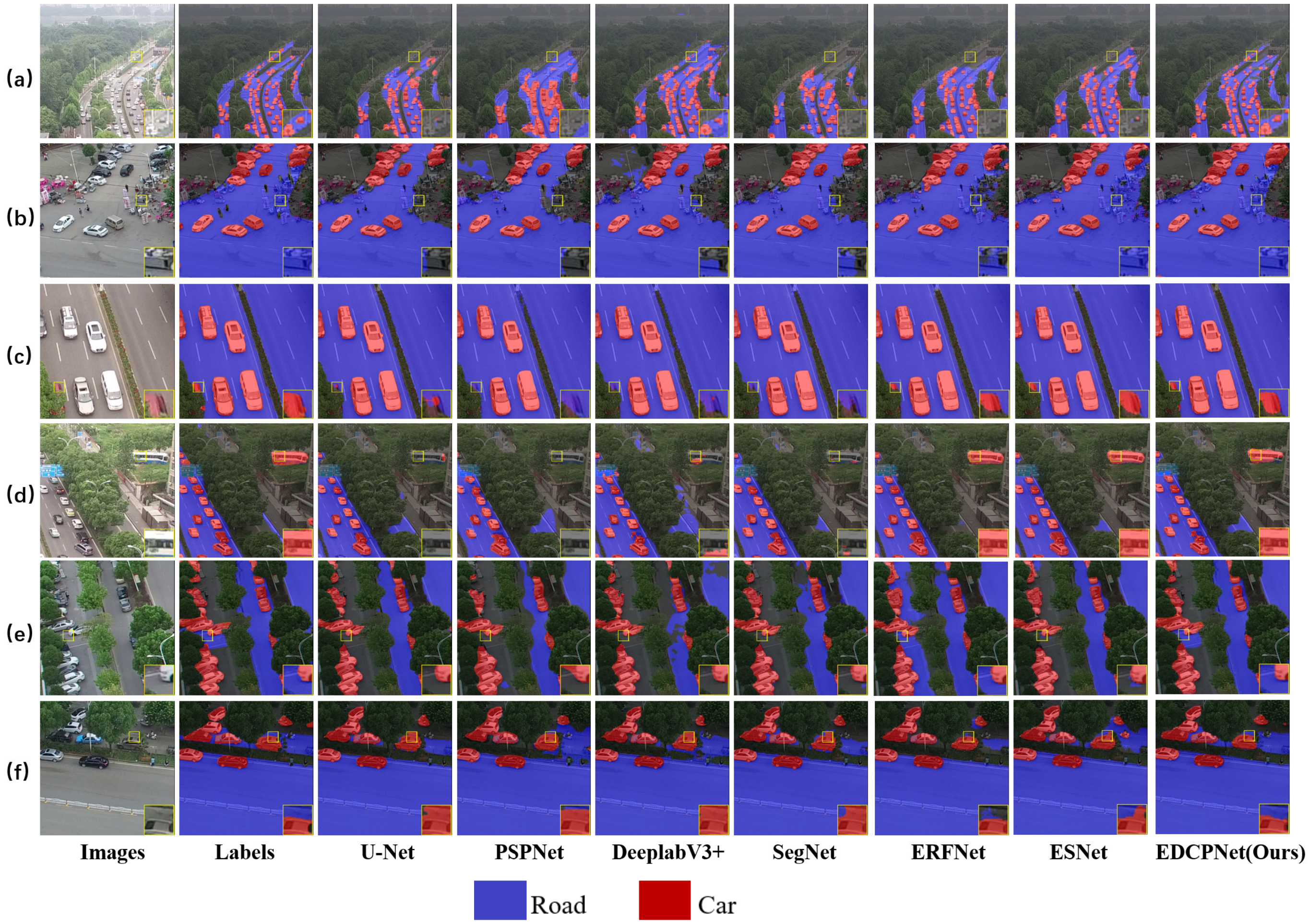


Fig. 6. Semantic segmentation results on the UAVid dataset with U-Net, PSPNet, DeepLabv3+, SegNet, ERFNet, ESNet, and the proposed EDCPNet.

TABLE I

QUANTITATIVE EVALUATION RESULTS ON UAVID DATASET COMPARED WITH OTHER COMPETING METHODS

Methods	IoU (%)		F1 (%)		MIoU (%)	MPA (%)	Parameters
	road	car	road	car			
U-Net [15]	82.68	74.31	90.08	85.06	84.95	90.35	24.79M
PSPNet [43]	78.26	66.93	87.8	80.19	80.88	90.07	46.66M
DeepLabv3+ [44]	81.99	77.08	90.1	87.06	85.59	93.23	41.15M
SegNet [17]	78.93	66.37	88.22	79.78	80.81	87.58	11.44M
ESNet [45]	80.68	75.55	89.31	86.07	84.61	92.64	**1.66M**
ERFNet [46]	80.33	73.72	89.09	84.87	83.85	91.73	2.07M
EDCPNet (Ours)	**82.72**	**80.34**	**90.54**	**89.1**	**86.97**	**94.21**	2.37M

it significantly improves the classification accuracy in the car category compared with other methods. The IoU and F1 of EDCPNet in the car category are, respectively, 3.26% and 2.04% higher than DeepLabV3+, notably outperforming the other algorithms.

In urban contexts, road features (e.g., lane lines, pedestrian crossings, and green belts) can be very complex. In addition, the tilted camera views of UAVs lead to the large-scale disparity of the extracted objects. Those challenges are usually coupled with noisy backgrounds and occlusions by trees and buildings, responsible for the poor road and car extraction performance from traditional models. However, our model can effectively alleviate these problems, as shown in Fig. 6. From Fig. 6(a), we can see that PSPNet and SegNet present poor extraction results for dense traffic with aggregated car objects, while ESNet, ERFNet, and U-Net fail to identify the most distant road patches due to their smaller scale. In comparison, our method is able to maintain the continuity of the extracted road under the large-scale disparity. Such an advantage is important in situations where large-scale traffic

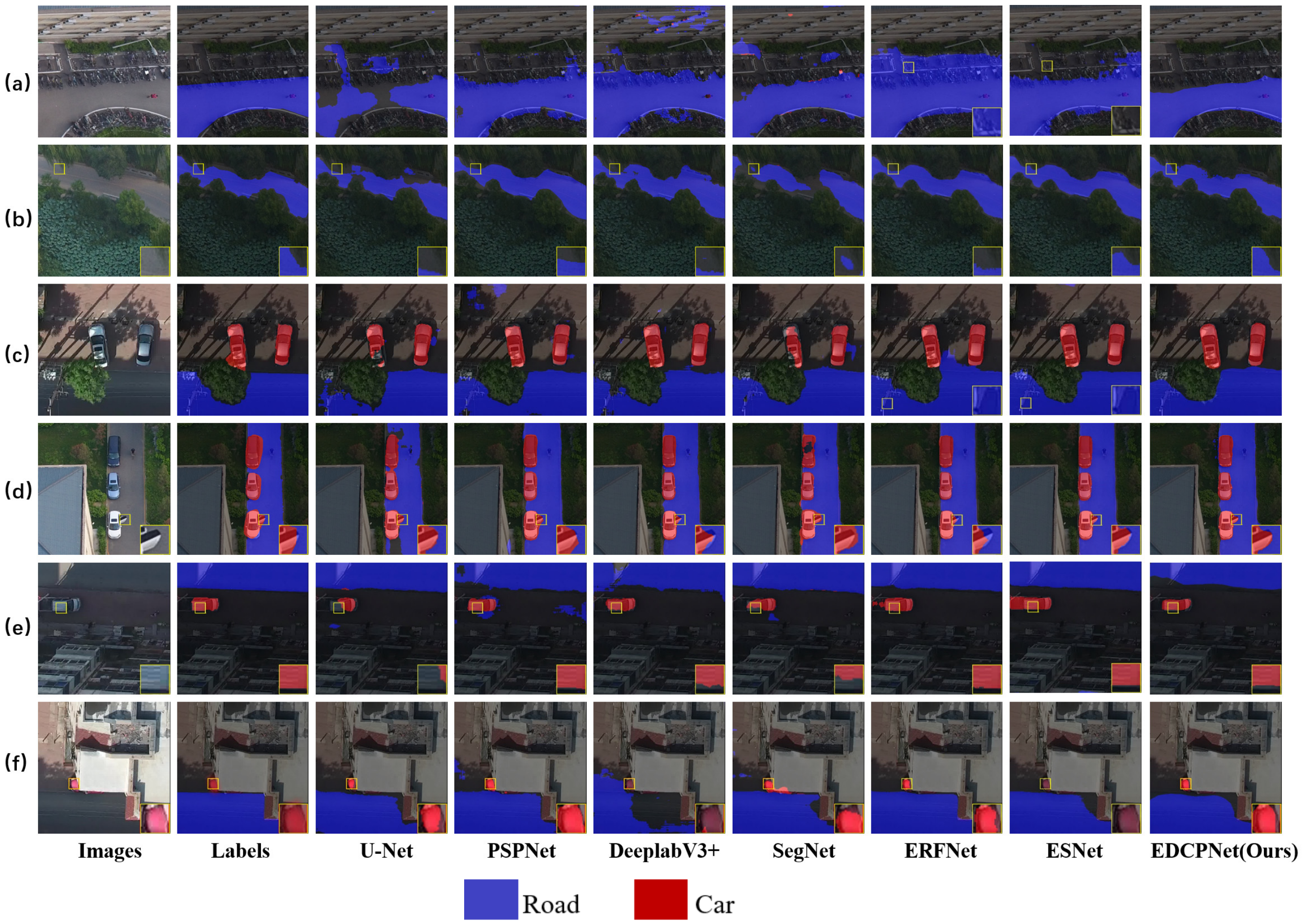


Fig. 7. Semantic segmentation results on the UDD dataset with U-Net, PSPNet, Deelabv3+, SegNet, ERFNet, ESNet, and the proposed EDCPNet.

supervision is required. Fig. 6(b) shows the problem that often occurs in urban traffic supervision, where urban roads are encroached on by irrelevant elements, bringing great traffic safety hazards. For this complex scene, we notice that our method can extract road surfaces in an accurate manner, largely overcoming the interference of irrelevant road information. Fig. 6(c) and (d) reflects the performance of all methods when dealing with tree occlusion issues. It can be seen from the enlarged part of the figure that the proposed EDCPNet performs better in terms of extracting cars surrounded and covered by trees compared with the other six methods. For the extraction of roads obscured by trees, road extracted by PSPNet, DeepLabv3+, ESNet, and SegNet is fragmented [see the enlarged part in Fig. 6(e)], suggesting that these methods lose attention to the details of lane lines, thus leading to the lack of road extraction accuracy. Conversely, ERFNet incorrectly identifies the surrounding open space as a road. From Fig. 6(f), we observe that DeepLabv3+, ESNet, and PSPNet misclassify the green belt into the road, while SegNet, ERFNet, and U-Net misclassify the shaded car into the road. In comparison, the proposed EDCPNet method addresses these issues well. The above qualitative analysis, agreeing with the quantitative evaluations, verifies the superiority of the EDCPNet method for road and car extraction on the UAVid dataset.

Considering the number of parameters, our EDCPNet (2.4M parameters) has much fewer parameters compared to other deep networks, such as PSPNet (46.7M parameters) and DeepLab-V3+ (41.2M parameters). The smaller model facilitates its deployment on edge computing devices with low memory constraints, which can be easily applied to UAVs. On the other hand, EDCPNet has better car and road extraction outcomes compared to other lightweight networks, such as ESNet (1.66M parameters) and ERFNet (2.07M parameters). In summary, our EDCPNet achieves very competitive performance with low computational cost, can well overcome the large-scale variation problem caused by the tilted photographic view of UAVs, and can well overcome various issues that include tree occlusions and rich background noises.

2) Experiments on UDD: To validate the effectiveness and superiority of EDCPNet, we further conduct additional experiments on the UDD dataset. Compared with UAVid, the purpose of experiments on the UDD is to evaluate the extraction performance of the proposed EDCPNet in urban living areas (mainly residential and campus areas), where complex backgrounds and building shadows occur. Fig. 7 shows selected detection results on the UDD test dataset. It can be seen that the overall extraction results from EDCPNet are, again, superior to the ones from competing algorithms. Fig. 7 (a) shows a scenario with complex foreground/background and densely placed car

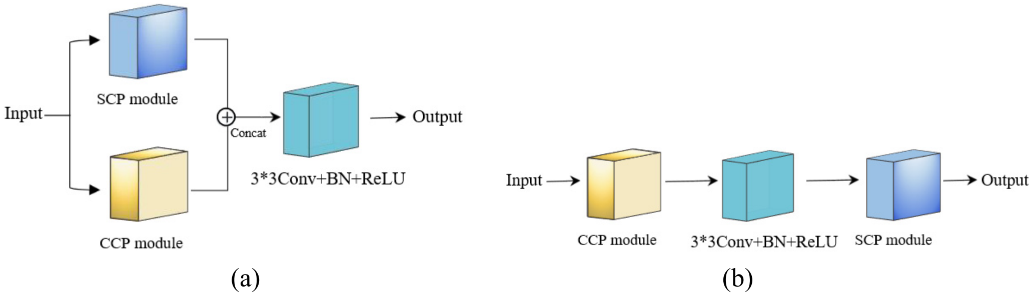


Fig. 8. Structure of (a) EDCP-P and (b) EDCP-C modules.

TABLE II
QUANTITATIVE EVALUATION RESULTS ON UDD DATASET COMPARED WITH OTHER COMPETING METHODS

Methods	IoU (%)		F1 (%)		MIoU (%)	MPA (%)	Parameters
	road	car	road	car			
U-Net [15]	58.47	50.13	73.79	66.78	66.57	72.13	24.79M
PSPNet [43]	63.37	70.62	77.58	82.78	75.24	83.72	46.66M
DeepLabv3+ [44]	56.36	66.74	72.09	80.05	71.02	79.59	41.15M
SegNet [17]	63.2	39.03	77.45	56.14	64.52	71.00	11.44M
ESNet [45]	63.69	58.64	77.82	73.92	71.27	81.76	**1.66M**
ERFNet [46]	63.67	55.94	77.8	71.75	70.31	81.85	2.07M
EDCPNet (Ours)	**64.93**	**72.35**	**78.74**	**83.96**	**76.62**	**83.51**	2.37M

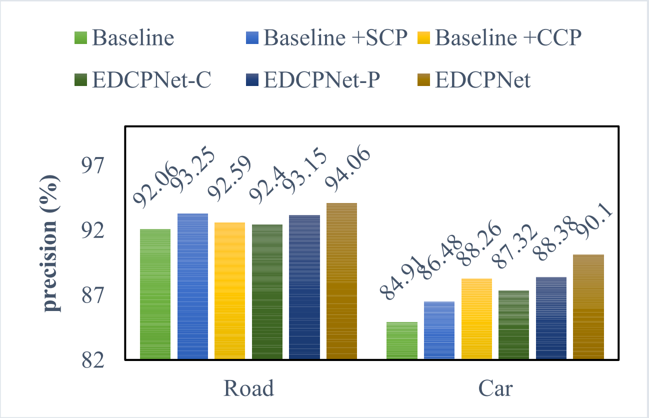


Fig. 9. Precision (%) on the UAVid dataset with Baseline, Baseline + SCP, Baseline + CCP, EDCPNet-P, EDCPNet-C, and EDCPNet.

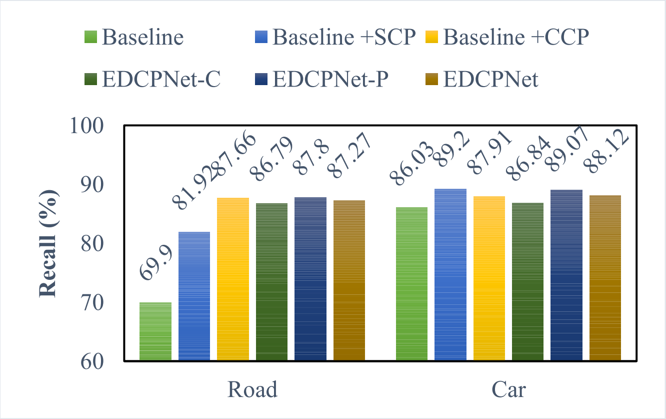


Fig. 10. Recall (%) on the UAVid dataset with Baseline, Baseline + SCP, Baseline + CCP, EDCPNet-P, EDCPNet-C, and EDCPNet.

beside the road. We observe that our method can extract accurate road surfaces with complete structures, while other methods tend to misclassify road-independent elements or buildings into the road. Fig. 7(b) and (c) demonstrates the performance of methods in terms of overcoming tree shades. Compared with the other six methods, the proposed EDCPNet presents a considerably better performance in terms of extracting roads heavily obscured by trees. The road extracted by U-Net and SegNet is almost discontinuous and very fragmented. Despite that PSPNet, DeepLabv3+, ESNet, and ERFNet can also extract approximate road outlines, their extracted road details are unsatisfactory, given the existence of many discontinuities in the extraction results. For car extraction [see Fig. 7(d)], our method also presents better performance than other competing algorithms. The road in the living areas tends to be narrow and mostly distributed between buildings; therefore, the shadows of the buildings have a great impact on the extraction results. From Fig. 7(e) and (f), DeepLabv3+ and PSPNet tend to misclassify the shadows of the buildings into the road, while SegNet and U-Net tend to misclassify the shaded car into the road. ESNet and ERFNet have false and missed inspections of the car under the interference of shadows. In comparison, the proposed EDCPNet method shows high robustness when dealing with shadow interference.

Table II shows quantitative evaluation results comparing EDCPNet and other algorithms on the UDD dataset. We notice that the proposed EDCPNet achieves the highest accuracy in road and car categories, with MIoU and MPA reaching 76.62%

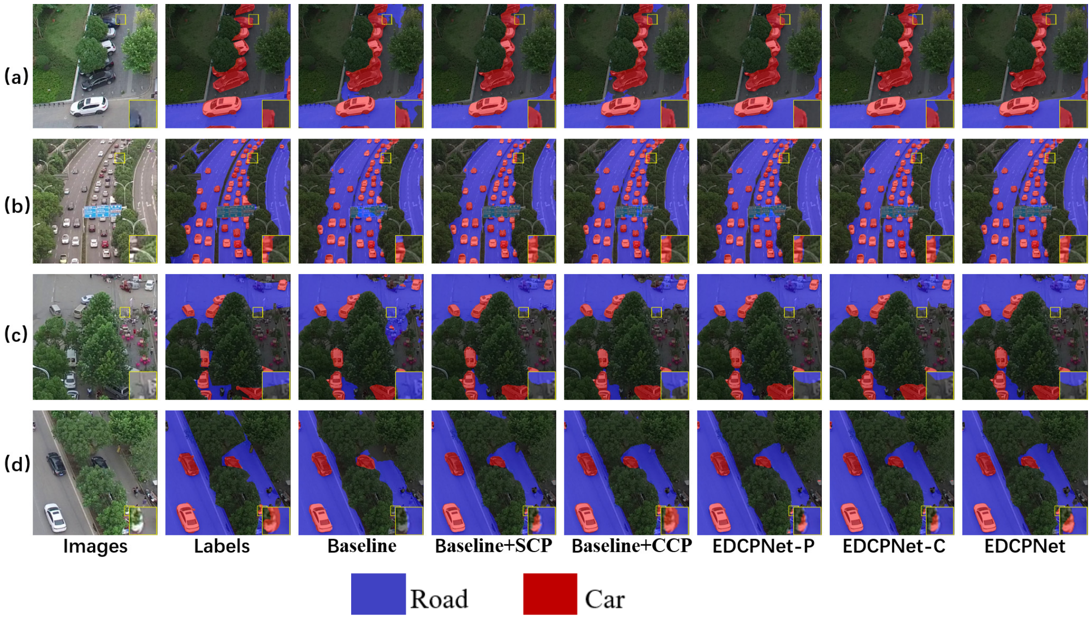


Fig. 11. Semantic segmentation results on the UAVid dataset with Baseline, Baseline + SCP, Baseline + CCP, EDCPNet-P, EDCPNet-C, and EDCPNet.

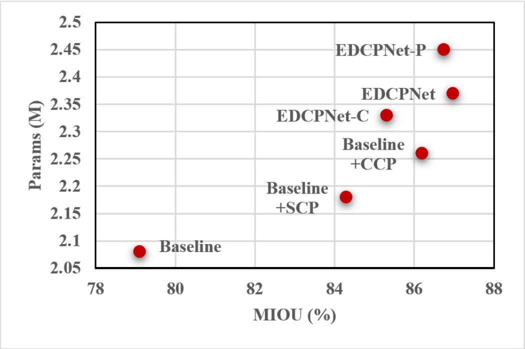


Fig. 12. Relationships between accuracy and model complexity of the proposed model and its variations.

and 83.51%, respectively. Compared with the PSPNet method, which has the highest accuracy among other competing methods, the proposed EDCPNet leads to an improvement of 1.73% and 1.18% on IoU and F1 in the car category, and 1.56% and 1.16% of IoU and F1 in the road category. The above results uniquely verify the superiority of EDCPNet in terms of extracting both road and car objects on the UDD dataset.

IV. Discussion

In this section, we conduct ablation experiments to verify the effectiveness of the EDCP module proposed in EDCPNet. The EDCP module aims to acquire contextual and global information in both channel and spatial dimensions, adaptively recalibrate attention weights, perceive salient features in images, and suppress the importance of irrelevant elements. We define the EDCPNet with the removed EDCP module as the Baseline method. In addition, we perform an in-depth analysis of the two modules, SCP and CCP, and their combination methods. In addition to the EDCP model with spatially prioritized tandem order, the model where the SCP and CCP modules are arranged in parallel before channel integration is termed EDCP-P [see Fig. 8(a)]. The model where the CCP modules are tandem first and channel integration is followed by tandem SCP modules in channel priority order is termed EDCP-C [see Fig. 8(b)]. EDCPNet-P refers to the usage of the EDCP-P module, EDCPNet-C indicates the usage of the EDCP-C module, and EDCPNet denotes the usage of the EDCP module.

From Table III and Figs. 9 and 10, the introduction of the two modules, i.e., SCP and CCP, can further improve the quality of the UAV-based car and road extraction results. Both Baseline + SCP and Baseline + CCP and their combinations EDCPNet-P, EDCPNet-C, and EDCPNet present better extraction results than the Baseline. The involvement of the SCP module leads to improvement in MIoU, MPA, IoU of road, IoU of car, F1 of road, and F1 of the car by 5.18%, 3.23%, 11.41%, 3.11%, 7.76%, and 2.35%, respectively. The involvement of the CCP module leads to improvement in MIoU, MPA, IoU of road, IoU of car, F1 of road, and F1 of the car by 7.09%, 3.85%, 15.99%, 3.54%, 10.6%, and 2.61%, respectively. The CCP module has a greater impact on the improvement in extraction accuracy. From Fig. 10, EDCPNet-P achieves the best recall value in both road and car extraction. Moreover, we notice that the combination and order of these two modules directly affect the overall performance of the network, given their different functionalities. Based on the experimental results, we notice that the EDCP module with spatial priority order in series outperforms the EDCP-P module with parallel order in series and the EDCP-C module with channel priority order in series. On the UAVid dataset, our EDCPNet using the EDCP module improves 7.86% and 4.88% on MIOU and MPA, respectively, and 16.79% and 11.08% on

TABLE III
ABLATION EXPERIMENTAL RESULTS ON UAVID DATASET

methods	IoU (%)		F1 (%)		MIoU (%)	MPA (%)	Parameters
	road	car	road	car			
Baseline	65.93	75.17	79.46	85.47	79.11	89.33	**2.08M**
Baseline+SCP	77.34	78.28	87.22	87.82	84.29	92.56	2.18M
Baseline+CCP	81.92	78.71	90.06	88.08	86.2	93.18	2.26M
EDCPNet-P	82.47	79.73	90.39	88.72	86.74	93.41	2.45M
EDCPNet-C	81.0	77.11	89.5	87.08	85.31	92.76	2.33 M
EDCPNet	**82.72**	**80.34**	**90.54**	**89.1**	**86.97**	**94.21**	2.37M

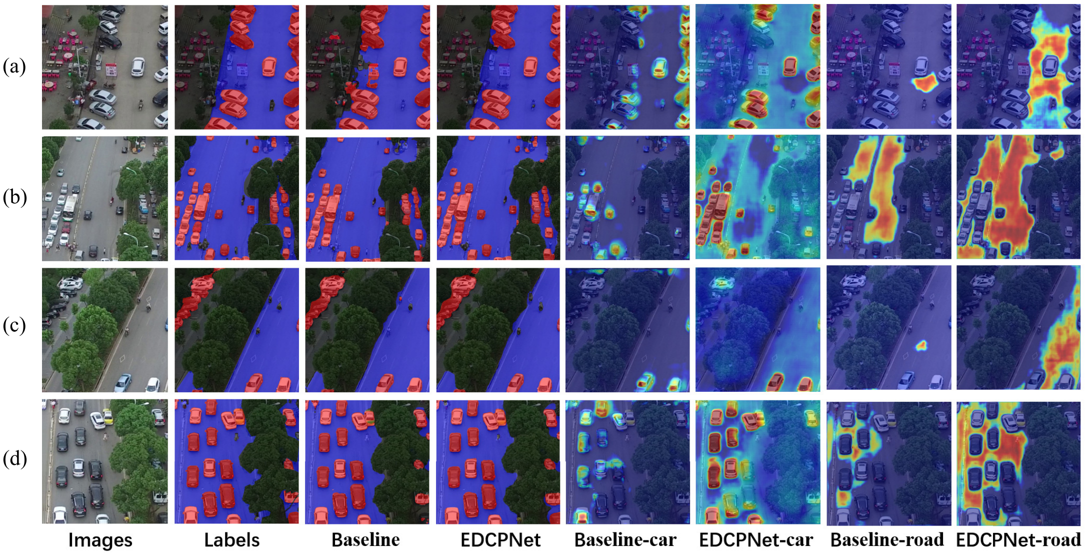


Fig. 13. Semantic segmentation results and the heat maps on the UAVid dataset with Baseline and the proposed EDCPNet.

IOU and F1 of road, compared to the Baseline. The above results indicate that the usage of the EDCP module achieves the best performance and can effectively improve the accuracy of car and road extraction in our proposed EDCPNet.

Fig. 11 shows the selected results of car and road extraction on the UAVid dataset. From Fig. 11(a) and (b), it can be seen that the introduction of the EDCP module benefits the acquisition of contextual features, which leads to a wider perceptual field and richer semantic information. The involvement of EDCP leads to better distinguishment between road surfaces and parking lots and accurate identification of cars under tilted views. From Fig. 11(c) and (d), we can see that the Baseline + EDCP significantly reduces the misclassification and omission of road and car, presumably due to the fact that the EDCP module can emphasize target features and suppress the weight of irrelevant pixels. Given the contribution of EDCP, we notice that the extraction accuracy of road objects is significantly improved after the introduction of the EDCP module (see Table III). The above results again demonstrate the importance of the proposed EDCP module in improving the accuracy of car and road extraction.

We present the relationships between the accuracy and model complexity of the proposed model and its variations in Fig. 12. The EDCPNet-C (2.33M parameters) and EDCPNet (2.37M parameters) arranged in series have fewer parameters compared to the EDCPNet-P (2.45M parameters) arranged in parallel. Among them, EDCPNet achieves the best extraction accuracy with an MIoU of 86.97%. Thus, we can conclude that the proposed EDCPNet achieves the best balance between extraction accuracy and network complexity.

In order to better highlight the effectiveness of EDCPNet, we added a comparison of the experimental results of EDCPNet and the Baseline on the UAVid dataset. The heat map drawn by the Grad-CAM [47] method is used to analyze the attention area of the network for different targets, and the visualization results are shown in Fig. 13.

Fig. 13 shows that EDCPNet extracts more accurate car and road results with less false detection rate and less missed detection rate. It can be clearly seen by the Grad-CAM method that EDCPNet can acquire rich contextual features, pay more attention to the car and road surroundings, and get a wider range of perceptual fields and richer semantic information. This results in a significant reduction in the cases of roads and cars being misclassified and missed. In the color value of the heat map, the redder color represents the higher attention of the network. EDCPNet perceives the salient features of the target well, pays more attention to the parts belonging to car and road features, and is more accurate, which helps to improve the accuracy of car and road extraction.

In summary, it can be clearly demonstrated that EDCPNet can acquire rich contextual features and perceive local salient features of the target, thus obtaining better classification results in complex scenes.

V. Conclusion

In this study, we propose a lightweight, EDCPNet for car and road extraction from UAV remote sensing images, which greatly benefits large-scale traffic monitoring. The proposed EDCPNet can effectively acquire the global and contextual information of extracted targets, perceive the salient features of targets in images, and suppress the importance of irrelevant elements, such as trees and shadows for contextual parsing, thus leading to improved performance and adaptability that facilitate the practical applications of large-scale urban traffic monitoring in complex urban scenes. We conduct experiments on two datasets, i.e., UAVid and UDD, and validate the effectiveness of proposed modules via extensive ablation studies. The experimental results show that the proposed network achieves great performance in car and road extraction from UAV remote sensing images with a balanced computational cost. Compared with other competing methods, i.e., U-Net, PSPNet, Deelabv3+, SegNet, ESNet, and ERFNet, the proposed EDCPNet model presents notable superiority, evidenced by its capability in addressing large changes in target scales, complex foregrounds and backgrounds, and severe shadow occlusions.

References

[1] A. Tahir, J. Boling, M. H. Haghbayan, H. T. Toivonen, and J. Plosila, "Swarms of unmanned aerial vehicles—A survey," *J. Ind. Inf. Integr.*, vol. 16, pp. 100–106, Dec. 2019.

[2] J. Shahmoradi, E. Talebi, P. Roghanchi, and M. Hassanalian, "A comprehensive review of applications of drone technology in the mining industry," *Drones*, vol. 4, no. 3, pp. 34–59, Sep. 2020.

[3] F. Al-Turjman, H. Zahmatkesh, I. Al-Oqily, and R. Daboul, "Optimized unmanned aerial vehicles deployment for static and mobile Targets' monitoring," *Comput. Commun.*, vol. 149, pp. 27–35, Jan. 2020.

[4] M. Hegarty-Craver et al., "Remote crop mapping at scale: Using satellite imagery and UAV-acquired data as ground truth," *Remote Sens.*, vol. 12, no. 12, pp. 1984–1999, Jun. 2020.

[5] T. Chowdhury, M. Rahnemoonfar, 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 (Big Data)*, Dec. 2020, pp. 3904–3913.

[6] X. Y. Zhu, J. W. Liang, and A. Hauptmann, "MSNet: A multilevel instance segmentation network for natural disaster damage assessment in aerial videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2022–2031.

[7] E. Vattapparamban, I. Guvenc, A. I. Yurekli, K. Akkaya, and S. Uluagac, "Drones for smart cities: Issues in cybersecurity, privacy, and public safety," in *Proc. Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Sep. 2016, pp. 216–221.

[8] B. Alzahrani, O. S. Oubbati, A. Barnawi, M. Atiquzzaman, and D. Alghazzawi, "UAV assistance paradigm: State-of-the-art in applications and challenges," *J. Netw. Comput. Appl.*, vol. 166, pp. 102706–102753, Sep. 2020.

[9] N. Mohamed, J. Al-Jaroodi, I. Jawhar, A. Idries, and F. Mohammed, "Unmanned aerial vehicles applications in future smart cities," *Technol. Forecasting Social Change*, vol. 153, pp. 119293–119308, Apr. 2020.

[10] F. Outay, H. A. Mengash, and M. Adnan, "Applications of unmanned aerial vehicle (UAV) in road safety, traffic and highway infrastructure management: Recent advances and challenges," *Transp. Res. A, Policy Pract.*, vol. 141, pp. 116–129, Nov. 2020.

[11] Z. Ullah, F. Al-Turjman, L. Mostarda, and R. Gagliardi, "Applications of artificial intelligence and machine learning in smart cities," *Comput. Commun.*, vol. 154, pp. 313–323, Mar. 2020.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2017.

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

[14] Y. Xie, F. Miao, K. Zhou, and J. Peng, "HsgNet: A road extraction network based on global perception of high-order spatial information," *Isprs Int. J. Geo-Inf.*, vol. 8, no. 12, pp. 571–585, Dec. 2019.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, vol. 9351, Nov. 2015, pp. 234–241.

[16] M. D. Zeiler, and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 8689. Springer, 2014, pp. 818–833.

[17] 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.

[18] R. B. Lian, W. X. Wang, N. Mustafa, and L. Q. Huang, "Road extraction methods in high-resolution remote sensing images: A comprehensive review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5489–5507, 2020.

[19] Z. X. Zhang, Q. J. Liu, and Y. H. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[20] J. Wan, Z. Xie, Y. Xu, S. Chen, and Q. Qiu, "DA-RoadNet: A dual-attention network for road extraction from high resolution satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6302–6315, 2021.

[21] Y. F. Ren, Y. T. Yu, and H. Y. Guan, "DA-CapsUNet: A dual-attention capsule U-Net for road extraction from remote sensing imagery," *Remote Sens.*, vol. 12, no. 18, pp. 2866–2883, Sep. 2020.

[22] R. Kestur, S. Farooq, R. Abdal, E. Mehraj, O. Narasipura, and M. Mudigere, "UFCN: A fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle," *J. Appl. Remote Sens.*, vol. 12, Feb. 2018, Art. no. 016020.

[23] N. Varia, A. Dokania, and J. Senthilnath, "DeepExt: A convolution neural network for road extraction using RGB images captured by UAV," in *Proc. IEEE Symp. Comput. Intell. (SSCI)*, Nov. 2018, pp. 1890–1895.

[24] J. Senthilnath, N. Varia, A. Dokania, G. Anand, and J. A. Benediktsson, "Deep TEC: Deep transfer learning with ensemble classifier for road extraction from UAV imagery," *Remote Sens.*, vol. 12, no. 2, pp. 245–264, Jan. 2020.

[25] N. Ammour, H. Alhichri, Y. Bazi, B. Benjdira, N. Alajlan, and M. Zuair, "Deep learning approach for car detection in UAV imagery," *Remote Sens.*, vol. 9, no. 4, pp. 312–327, Apr. 2017.

[26] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.

[27] A. Bouguettaya, H. Zarzour, A. Kechida, and A. M. Taberkit, "Vehicle detection from UAV imagery with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 24, 2021, doi: 10.1109/TNNLS.2021.3080276.

[28] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6699–6711, Nov. 2018.

[29] N. Audebert, B. Le Saux, and S. Lefevre, "Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images," *Remote Sens.*, vol. 9, no. 4, pp. 368–386, Apr. 2017.

[30] M. K. Masouleh and R. Shah-Hosseini, "Development and evaluation of a deep learning model for real-time ground vehicle semantic segmentation from UAV-based thermal infrared imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 155, pp. 172–186, Sep. 2019.

[31] M. K. Masouleh and R. Shah-Hosseini, "A hybrid deep learning–based model for automatic car extraction from high-resolution airborne imagery," *Appl. Geomatics*, vol. 12, no. 2, pp. 107–119, Jun. 2020.

[32] Y. Qian, J. M. Dolan, and M. Yang, "DLT-Net: Joint detection of drivable areas, lane lines, and traffic objects," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4670–4679, Nov. 2020.

[33] L. C. Lo Bianco, J. Beltrán, G. F. López, F. García, and A. Al-Kaff, "Joint semantic segmentation of road objects and lanes using convolutional neural networks," *Robot. Auto. Syst.*, vol. 133, Nov. 2020, Art. no. 103623.

[34] M. Teichmann, M. Weber, M. Zollner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1013–1020.

[35] L. P. Osco et al., "A review on deep learning in UAV remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 102, Oct. 2021, Art. no. 102456.

[36] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

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

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

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

[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

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

[42] 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. (PRCV)*, pp. 347–359, 2018.

[43] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[44] 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. (ECCV)*, 2018, pp. 801–818.

[45] Y. Wang, Q. Zhou, J. Xiong, X. Wu, and X. Jin, "ESNet: An efficient symmetric network for real-time semantic segmentation," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, 2019, pp. 41–52.

[46] 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, Oct. 2017.

[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Yueming Sun received the B.S. degree in remote sensing science and technology from Chang'an University, Xi'an, China, in 2020. She is currently pursuing the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China.

Her research interests include high spatial resolution remote sensing image classification, remote sensing applications, and image processing.

Zhenfeng Shao received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2004.

Since 2009, he has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. He has authored or coauthored over 50 peer-reviewed articles in international journals. His research interests include high-resolution image processing, pattern recognition, and urban remote sensing applications.

Dr. Shao was a recipient of the Talbert Abrams Award for the Best Paper in Image Matching from the American Society for Photogrammetry and Remote Sensing in 2014 and the New Century Excellent Talents in University from the Ministry of Education of China in 2012. Since 2019, he has been serving as an Associate Editor for the *Photogrammetric Engineering and Remote Sensing* (PE & RS) specializing in smart cities, photogrammetry, and change detection.

Gui Cheng received the B.S. degree in geographic information science from Chang'an University, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China.

His research interests include image processing and computer vision.

Xiao Huang is currently an Assistant Professor with the Department of Geosciences, University of Arkansas, Fayetteville, AR, USA. His research area covers GeoAI, remote sensing, spatial machine learning, social sensing, and disaster mitigation.

Zhongyuan Wang (Member, IEEE) received the Ph.D. degree in communication and information systems from Wuhan University, Wuhan, China, in 2008.

He is currently a Professor with the School of Computer Science, Wuhan University. He has been directing four projects funded by the National Natural Science Foundation of China. He has authored or coauthored over 80 refereed journal articles and conference papers, and has been granted more than 30 invention patents. His research interests include biometrics and identification, computer vision and pattern recognition, and image processing.