# Dual-Task Network for Road Extraction From High-Resolution Remote Sensing Images

Yuzhun Lin, Fei Jin, Dandi Wang, Shuxiang Wang, and Xiao Liu

***Abstract*—In high-resolution remote sensing images, road scale diversity and occlusions caused by shadows, buildings, and vegetation often pose challenges for road extraction. Currently, end-to-end models constructed using deep convolutional neural networks are widely used in road extraction and have significantly improved the accuracy of this task. However, the connectivity and completeness of their results require improvement. This article proposes a dual task-driven deep convolutional neural network constructed by combining road shape patterns and scale differences. The mainline task is road-surface segmentation, the encoder of which employs residual convolution for feature extraction. The decoder comprises a multiscale and multidirection strip convolution module, the output of which is the final extraction result. The splitting task is road centerline extraction, the input features of which come from the coding layer of the road-surface segmentation branches. The intermediate features are incorporated into the decoder of the road-surface segmentation branches, to fully exploit the road centerline and thus improve the road-surface segmentation result connectivity. Implementation of the proposed method on the CHN6-CUG and DeepGlobe datasets reveals superior performance to comparative methods as regards quantitative evaluation metrics; evident advantages for road coverings, road intersections, and low-scale roads; greater model portability; and better small-sample learning capability.**

***Index Terms*—Convolutional neural network (CNN), deep learning, remote sensing image, road centerline, road extraction.**

## I. Introduction

ROAD extraction is essential for map updating, autonomous driving, urban planning, and vehicle navigation. Remote sensing images, which are obtained through noncontact acquisition, enable procurement of a large range of surface details in a short period of time. Hence, a road network can be displayed in a flat visual image. Moreover, the spatial and temporal resolutions of remote sensing images are continuously improving. Therefore, remote sensing images can form an effective database for road automation and real-time extraction.

Remote sensing-image road extraction techniques can be classified as traditional and deep learning methods [1], based on their development history. Traditional methods include template matching [2], [3], knowledge-driven [4], [5], and object-oriented [6], [7] methods, and mainly rely on the shape, spectral, and texture features presented by roads on remote sensing images, along with human-designed shallow combinations of such features. However, with the continuously improving spatial resolution of remote sensing images, their surface detail has been significantly enhanced. Thus, the "same material, different spectra–same spectra, different material" problem has become increasingly prominent, as roads are often insufficiently aggregated in the shallow feature space and intersect with other features. As a result, such methods have poor applicability and stability.

With the ongoing development of convolutional neural networks (CNNs), and especially the introduction of typical semantic segmentation networks such as fully convolutional networks [8], UNet [9], SegNet [10], and the DeepLab series [11], [12], [13], [14], CNN-based methods have recently been applied to pixel-level intelligent interpretation of remote sensing images. However, road extraction from remote sensing images is challenging, for the following reasons: 1) road width differences are evident and, thus, a small- and large-target coexistence phenomenon occurs; 2) buildings, trees, etc., shade the road surface; and 3) problematic similarities between roads and other targets (open spaces, ditches, etc.) exist. These difficulties often cause errors, omissions, and fragmentation of road extraction results. Therefore, researchers have improved the existing methods based on the typical "encoder-decoder" structure and the image characteristics of the road. Early improvements focused on two aspects of feature extraction and the supervisory principle. Feature extraction mainly revolves around network depth and convolutional field of view. For example, the residual module [15] is used as the basic unit of the network [16] avoiding the problem of network degradation during deep feature extraction. The introduction of multiscale dilation convolution, atrous spatial pyramid pooling [13] and nonlocal blocks [17] can enhance the network's ability to extract global and multiscale features [18], [19], [20], [21]. In terms of supervised mechanism, road extraction as a single element interpretation problem, focuses on a small percentage of targets covered on remote sensing images, which can cause the problem of positive and negative sample ratio imbalance. Based on this, weighted cross-entropy [22], balanced cross-entropy [23], focal loss [24], etc., have been explored and applied. In the literature [25], the authors provide a comparative analysis of the effectiveness of 12 loss functions widely used in the field of image segmentation for road extraction. Although the above-mentioned improvements do not significantly increase the complexity of the network, they mainly focus on the constituent units or supervisory units of

Manuscript received 3 February 2023; revised 9 May 2023 and 8 June 2023; accepted 21 June 2023. Date of publication 26 June 2023; date of current version 14 July 2023. This work was supported by the National Natural Science Foundation of China under Grant 42201443. *(Corresponding author: Fei Jin.)*

The authors are with the Information Engineering University, Zhengzhou 450001, China (e-mail: lyz120218@163.com; jf371@sina.com; wdd_93@163.com; shuxiang1007@163.com; liuxiao99919@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3289217

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

the network and do not involve much information about the attributes of the road and the overall framework of the network. To further improve the road extraction effect, the research scholars try to further realize the improvement by combining multiple methods, optimizing the extraction strategy, introducing road subsidiary information, and fusing multisource data. For example, the work in [26] and [27] combine CNNs with graph neural networks. In [28], an integrated reinforcement learning convolutional neural decision network is constructed. Decoder branches have also been added to the conventional framework to yield a double-decoder structure, with the detailed-information extraction performance being enhanced [29]. An extraction framework operating from coarse to refined features has also been constructed, with omissions and erroneous extractions in the coarse extraction process being corrected using the refined extraction results [30]. Taxi trajectories [31], geospatial data and street-level images [32], and radar images [33] are fused with remote sensing images to fully explore the dynamic patterns and static features presented by roads in different forms of data. In the above-mentioned methods, different perspectives on enhancing the accuracy and stability of road extraction results have been implemented. However, road shape patterns, scale differences, and connectivity relationships for high-resolution remote sensing images have not been simultaneously considered.

In this article, we propose a dual task-driven deep CNN that combines road shape patterns and scale differences. The contributions are as follows.

1) A multiscale and multidirectional strip convolution module (MSMD-SCM) is proposed to handle the strip-like characteristics of road shapes and the scale differences of different road classes.
2) Taking road-surface segmentation (RSS) as the basic framework, road centerline extraction (RCE) is introduced as a supplement to form a dual-task network structure.
3) In addition to the traditional accuracy comparison and ablation experiments, a detailed analysis is performed, which focuses on the method portability and road extraction capability in a small-sample environment.

## II. Related Work

### A. Multiscale and Multidirectional Strip Convolution Module

When road extraction tasks are performed on remote sensing images, roads of different scales must be handled. Unlike terrain elements such as buildings, vegetation, and lakes, roads often appear as strips. Thus, the feature space of concern is a key focus. In this context, the improvement proposed herein mainly involves two aspects: the convolutional kernel and supervision level. Relevant research on these two aspects is summarized in the following.

Regarding the convolution kernel, scale variability is generally achieved by setting different convolution kernel sizes, expansion rate sizes, and the location of the superimposed presence for feature extraction of differently sized perceptual fields. In [34], multiscale convolution attention was introduced, with three sets of horizontal and vertical strip convolutions of different scales being combined for multiscale feature extraction. In a similar study [35], channel separation was performed after the $1 \times 1$ convolution kernel, with a $3 \times 3$ convolution then being performed sequentially. As the superimposed $3 \times 3$ convolution expanded the perceptual field, feature extraction at different scales was achieved. In addition, atrous spatial pyramid pooling and multiscale feature aggregation modules [36] are also effective for extracting image features at different scales. With respect to the direction regularity, the main constraint on the feature-extraction direction is tuned by changing the shape of the convolution kernel. In various studies [37], [38], [39], a striped convolutional kernel that fit the road shape better than a square convolutional kernel was designed from the perspective of shape matching, that is, a striped convolutional kernel with four directions of 0°, 45°, 90°, and 135° was used for feature extraction.

Regarding the supervision level, the key technique in terms of scale variability is the simultaneous supervision of outputs at different scales using labeled data from RSS. In [30], [40], and [41], the results of each up-sampling process were output in the decoder process. As the output size varied with the level, the weights of roads at different scales (as foreground targets in the feature map) also varied. In other words, roads at different scales were awarded attention in a hierarchical manner in the supervision process. The direction regularity was primarily based on the existing road label for the road target, to provide the corresponding directional properties. In [42], [43], and [44], the concept of "direction learning" was applied; i.e., each road point in the image was assigned a direction label corresponding to the true direction. Hence, the road trend was constrained in the prediction process.

All the afore-described methods addressed road scale differences and shape patterns. However, most considered those aspects individually. However, road shape and scale features exist in random combinations in images. Therefore, the integration of multiscale feature-extraction capability to the strip target extraction process is more suitable for application to the actual road conditions contained in remote sensing images.

### B. Connectivity Module

As roads are an important transportation facility, correct connectivity corresponds to a correct route. Importantly, an incorrect route can result in longer driving time, entry to restricted areas, or even problems such as traffic accidents. Therefore, road extraction result connectivity relations, and particularly their correctness and completeness, are attracting increasing research attention. Current research on this topic is focusing on three main aspects: supervised data, task form, and loss function.

In terms of supervised data, connected labels are mainly constructed based on the true road value at the pixel level. In [38], a road connectivity label of the same size as the original image was generated, and a channel number of eight was assigned by determining whether the current point and the neighboring point in the specified direction were roads. A similar connected-label generation strategy was adopted in [45], with the difference that the label values represented the total number of pixels belonging to the road in the eight-neighborhood space. Thus, the

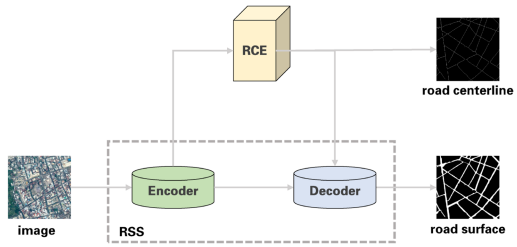


Fig. 1. Architecture of proposed method.

classification problem was converted to a regression problem. This type of method directly exploits the connectivity of the road itself. However, determination of the neighborhood space size often requires human empirical intervention.

As regards the task form, the main purpose is to build a multitask-driven road extraction network by combining road edges, centerlines, intersections, etc. This strategy was adopted in [41], [46], and [47]. Taking [41] as an example, following RSS, the output was combined with the original image and the combination was used as the input of the road edge and centerline networks. Thus, the RSS, road edge detection, and RCE processes were combined as a unified training network. This type of method fully exploits various road features. However, further research is needed to balance accuracy and efficiency.

Finally, in terms of loss functions, conventional calculations tend to be based on pixel-level differences and are insensitive to road topological changes. In [48] and [49], calculations were performed from the perspective of loss functions, so that the computational results would be more sensitive to road connectivity errors or omissions. For example, in [48], a pretrained VGG19 CNN [50] was used for deep feature extraction of prediction results and road labels. A loss function calculation based on the deep features was then performed. This method class can be seamlessly connected to existing network models. However, the image features of the roads themselves are not explored further.

## III. Method

In this section, the framework of the dual task-driven road extraction method is described in detail. RSS and RCE form the two branches of the framework, which performs simultaneous supervised learning using the corresponding labeled data. The RSS branch is the base and the output data are the final extraction result. The RCE branch is the supplement, the input features of which come from the RSS-branch encoder. The intermediate features are passed to the RSS-branch decoder to enhance the connectivity of the road extraction result, and the output data are the road centerline results (the auxiliary results). In addition, the proposed MSMD-SCM enhances the road feature capture capability in specified directions and at multiple scales, considering the road shape patterns and scale differences. The overall flow of the proposed framework is shown in Fig. 1.

### A. Network Structure

The network structure has a dual-task form and combines RSS and RCE. The training process is supervised based on the respective labeled data. Thus, the parameters of the two task lines are continuously updated during the backward propagation process to gradually improve the road prediction capability of the network model. The detailed structure of the network model is shown in Fig. 2.

*1) RSS Branch Network Structure:* This branch includes the encoder and decoder, the input and output data of which are the original images and the predicted RSS results, respectively. We employ ResNet34 [15] pretrained on ImageNet [51] as the encoder. Specifically, shallow feature extraction is first performed using $7 \times 7$ convolutional kernels and a $3 \times 3$ maximum pooling layer. Deep feature mining is then performed using four residual convolutional blocks with numbers 3, 4, 6, and 3; the final output feature-map size is 1/32 times the original image and the channel number is 512. In the decoder, four MSMD-SCMs are utilized for line feature extraction of roads at different scales, and to up-sample the feature maps to the appropriate size. Simultaneously, to alleviate the information loss that occurs during up-sampling, the output features of each of the four MSMD-SCMs are summed with the corresponding encoder feature maps. In addition, to enhance the connectivity of the road extraction results, the output features of the final MSMD-SCM are fused with the intermediate RCE results in the form of channel superposition. In the final decoder stage, RSS predictions consistent with the original image dimensions are obtained using operations such as up-sampling, convolution, and sigmoid activation.

*2) RCE Branch Network Structure:* The road centerline can visually reflect the road topology and directly promote road connectivity. Therefore, the RCE branch takes the design idea in [46] as a reference. The RCE branch is introduced as a supplement to the RSS branch to improve the RSS-result connectivity. The overall framework of the proposed network in this article differs from [46] in that the edge detection part is discarded. This is because the road edges are more likely to be obscured by other features on the remote sensing images, which can cause the remote sensing images to present the spectral information of other features at the road edge locations. In addition, the introduction of road edge detection will inevitably increase the workload of data preprocessing and the complexity of the network, influencing the overall efficiency of the method. In summary, only the RCE branch is retained in this article. The structure of the RCE branch is relatively simple compared to that of the RSS branch, and the output data are the road centerline prediction results, the corresponding truth values of which are obtained from the labeled data of the RSS results through the morphological thinning process. Considering the topological similarity between the road centerline and road surface, and the overall method complexity, the multiscale features in the RSS encoder are directly used as input data in this branch. Then, channel and scale unification are performed successively using $3 \times 3$ convolution and up-sampling. Channel superposition is performed on the feature maps of each scale on this basis. The superimposed fused feature-map size is $256 \times 256$ pixels, and the channel number is 64. After the above-mentioned processing, the obtained feature maps are further enhanced along two routes: Superimposition of the fusion results with the RSS decoder to enhance the RSS-result connectivity without over-suppressing

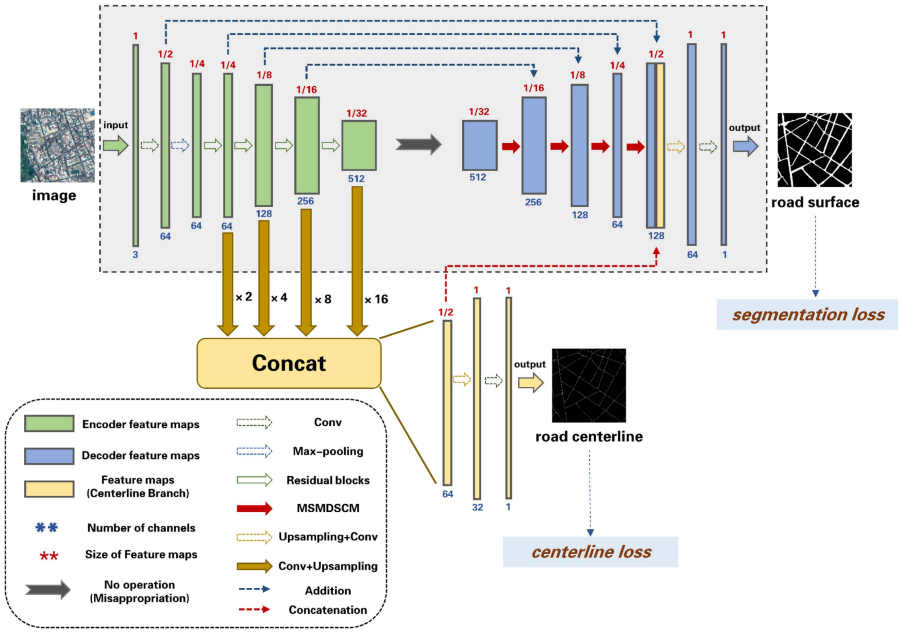


Fig. 2. Details of proposed network model.

the noncenterline regions of the road. The fusion results are up-sampled, convolved and sigmoid activated to obtain road centerline prediction results of the same size as the original image.

### B. MSMD-SCM

Roads have strict grade standards for both transportation and mapping, and different grades often correspond to different widths, which are expressed as different scales in remote sensing images. Therefore, when a dense prediction task such as RSS is performed, targets of different scales are encountered. If a fixed-size window is used for convolution, the target scale variability is often neglected. In addition, unlike terrain elements such as buildings, vegetation, and lakes, road shapes are often striped. Therefore, traditional square convolutional kernels inevitably capture more irrelevant information. In contrast, striped convolutional kernels can perform feature extraction in a specified direction, and their attention scopes are more congruent with road shape patterns.

On the basis of the afore-described analysis, we propose MSMD-SCM, which is based on a strip convolution module (the specific structure is shown in Fig. 3). In other words, the line features are extracted by using strip convolution kernels in the 0°, 45°, 90°, and 135° directions, and the feature-extraction results in each direction are fused using successive channel superposition, bilinear up-sampling, $1 \times 1$ convolution, and channel superposition.

The strip convolution in each direction contains multiple scales, and the specific multiscale fusion form is expressed as follows:

$$Y = Concat(X * W_{scale_i}) \quad i = 1, \ldots, k \tag{1}$$

where $X$ and $Y$ denote the input and output features, respectively; $Concat$ is the channel superposition operation; $W_{scale_i}$ is the linear convolution kernel at scale $i$; $k$ is the scale number; and $*$ denotes the convolution operation.

### C. Loss Function

The proposed network model has two branches that use RSS labels and road centerline labels for loss function calculation. The overall loss function is expressed as follows:

$$Loss = Loss_{seg} + Loss_{cen} \tag{2}$$

where $Loss$, $Loss_{seg}$, and $Loss_{cen}$ denote the total loss, RSS-branch loss, and RCE-branch loss, respectively. As the labels of both branches are dichotomous and an imbalance problem exists between the positive and negative samples, the sum of the binary cross-entropy (BCE) loss and dice coefficient loss are taken as the loss of each branch. The BCE loss treats each pixel equally. When the positive samples are small, the network is dominated by negative samples. Thus, the positive sample recognition is degraded. Dice coefficient loss focuses on information mining of positive samples (foreground region) and, thus, can better overcome the problem of positive and negative sample imbalance. However, the training loss easily becomes unstable. Therefore, a combination of these two losses can yield better results. The formulas for calculating the BCE loss and dice coefficient loss are given in (3) and (4), respectively.

$$Loss_{BCE}(P, Y) = -\sum_{i=1}^{W}\sum_{j=1}^{H}[y_{ij} \times \log p_{ij}(1 - y_{ij}) \times \log(1 - p_{ij})] \tag{3}$$

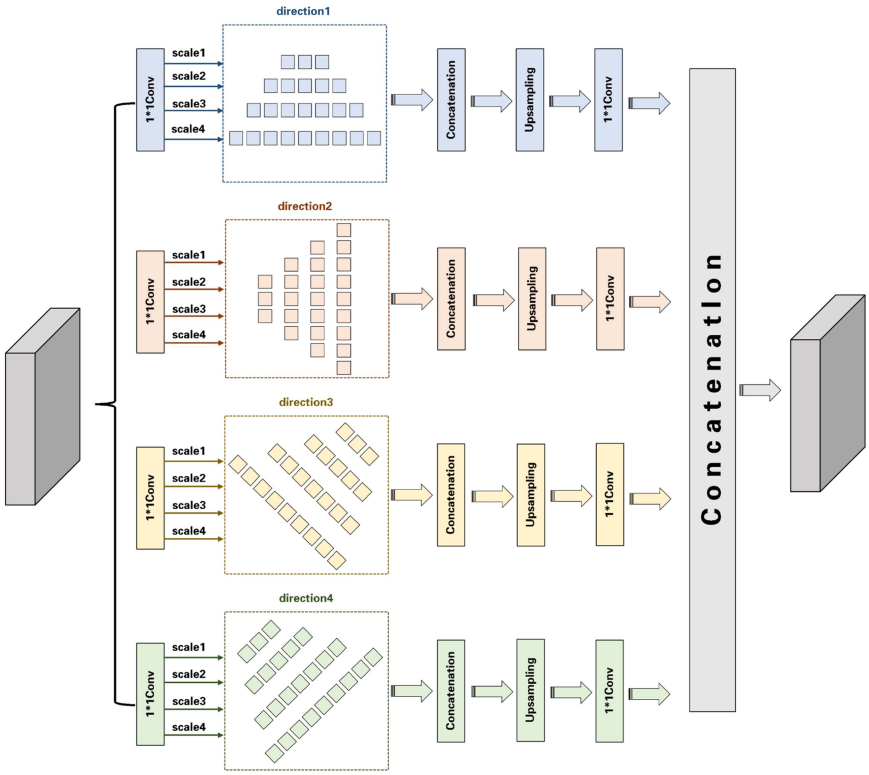


Fig. 3. MSMD-SCM flowchart.

$$Loss_{DCL}(P,Y) = 1 - \frac{2 \times |P \cap Y|}{|P| + |Y|} \tag{4}$$

where $P$ and $Y$ denote the prediction result and labeled data, respectively; $W$ and $H$ are the image width and height, respectively; and $p_{ij}$ and $y_{ij}$ are the prediction and label of position $(i, j)$ in the image, respectively.

## IV. Experiments

### A. Datasets

*1) CHN6-CUG Dataset [52]:* This dataset is sourced from Google Earth and includes highways, urban roads, and rural roads in Beijing, Wuhan, Shenzhen, Shanghai, Hong Kong, and Macau. There are 4511 labeled images in total, 3608 of which are for training while the remaining 903 are for testing. The ground sampling distance (GSD) for this dataset is 0.5 m per pixel. Each image has a size of 512 × 512 pixels.

*2) DeepGlobe Dataset [53]:* This dataset includes urban, suburban, and rural areas in Thailand, India, and Indonesia. A total of 6226 images are open access, with ground truth data. The GSD of this dataset is 0.5 m per pixel and each image has a size of 1024 ×1024 pixels. To improve the model training efficiency, we divided the original image and the corresponding labeled data in both the width and height directions synchronously, to generate a dataset with images of 512 × 512 pixels. We divided the training and test data according to a 3:1 ratio to obtain 18 784 training images and 6120 test images.

### B. Evaluation Metrics

*1) Pixel-level Evaluation Metrics:* To evaluate the performance of the proposed method with regard to RSS, we used the precision ($P$), recall ($R$), $F1$ score, overall accuracy ($OA$), and intersection over union ($IoU$) metrics. The formulas are as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2 \times \frac{P \times R}{P + R}, \quad OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{7}$$

where $TP$, $FP$, $TN$, and $FN$ represent the number of true positive, false positive, true negative, and false negative results, respectively.

*2) Connectivity Evaluation Metrics:* To verify the connectivity of the road extraction results, two evaluation metrics for specific measurements were designed: the completeness rate (*Com*) and error rate (*Eor*). In Fig. 4(a), the dark-colored buffer indicates the prediction result. The light-colored and red line segments are the morphological refinements of the labeled-data results, where the light-colored line segment is located inside the prediction result buffer with length $l_1$, and the red line segment is located outside the prediction result buffer with length $l_2$. In Fig. 4(b), the light-colored buffer represents the labeled data. The dark-colored and blue line segments are the morphological refinements of the predicted results, where the dark-colored line

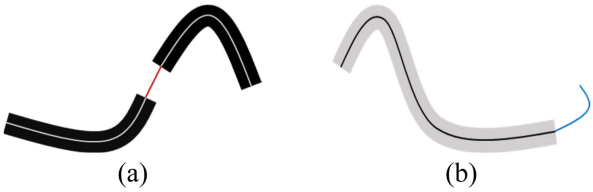


Fig. 4. Schematics of connectivity evaluation metrics. (a) *Com*. (b) *Eor*.

segment is located inside the labeled-data buffer with length $l_3$, and the blue line segment is located outside the labeled-data buffer with length $l_4$. The formulas for *Com* and *Eor* are as follows:

$$Com = \frac{l_1}{l_1 + l_2}, Eor = \frac{l_4}{l_3 + l_4}. \tag{8}$$

### C. Implementation Details

The experiments were implemented on 2 NVIDIA Tesla V100 GPUs with 64 GB memory. The Adam optimizer [54] with a batch size of 32 was adopted. The learning rate was initially set to 2*e*–4 and then reduced by a factor of 5 three times; the training loss was observed to decrease slowly. In all training experiments, the networks were trained for 150 epochs. In addition, for sample enhancement, vertical, horizontal, diagonal flip, and radial transformations were randomly applied to the training data (50%).

### D. Experiment Results

*1) Method Comparison:* In this stage of the experiment, the proposed method was applied to the above-mentioned two experimental datasets, and the extraction results and accuracy were compared with those given by seven typical semantic segmentation methods, namely UNet (2015) [9], D-LinkNet (2018) [19], DeepLabv3+ (2018) [14], ASPP-UNet (2019) [18], RoadNet (2018) [41], SGCN (2022) [26], and CoANet (2021) [38] when applied to the same datasets.

Fig. 5 shows RSS results for selected test images in the CHN6-CUG dataset. The five selected images are of different cities and scenes, and their features span those challenging for road extraction. Therefore, the comparative analysis is somewhat representative. The road in the lower right corner of Fig. 5(a) shows a heavily vegetated area; the overall road width is narrow and some sections are covered by vegetation. The extraction results show that U-Net, D-LinkNet, ASPP-UNet, RoadNet, and SGCN failed to extract this small section. DeepLabv3+ extracted a small portion. However, its road extraction results are evidently incomplete owing to the vegetation cover. In contrast, CoANet and the proposed method fully extracted the road section. However, it misidentified some of the nearby open spaces as roads. This problem must be addressed in future refinements of the proposed method. Fig. 5(b) shows a tall residential area in an urban area, and there is a large amount of shadow coverage on the road. The shadow has a darker shade. Thus, the spectral characteristics of the road surface itself do not correspond to the actual features. For example, the east–west road is heavily covered by building shadows. The extraction results of the seven comparison methods exhibited serious deficiencies in integrity for this section. However, the proposed method adapted to the shadow coverage phenomenon and effectively solves the problem of missed extraction. Fig. 5(c) shows an intersection of multiple roads, some of which have large widths, along with features such as parking lots with spectral characteristics close to those of the roads. The extraction results show that UNet, D-LinkNet, DeepLabv3+, and SGCN had serious omission extraction problems. The mis-extraction problems of ASPP-UNet and RoadNet were significant. The extraction results of CoANet and the proposed method were relatively positive. In Fig. 5(d), part of the east–west section is obscured by shadows, with evident interference from moving vehicles. The U-Net and RoadNet extraction results were almost blank for this road section. Although those of D-LinkNet, DeepLabv3+, and ASPP-UNet were slightly better, parts with continuous missed extractions were apparent. Thus, there were errors in road connectivity. In contrast, SGCN, CoANet, and the proposed method completely restored the road condition. Fig. 5(e) has an overall darker tone because of the building shadow, the acquisition environment, and large water and vegetation proportions. For this image, the extraction results of UNet, DeepLabv3+, ASPP-UNet, and RoadNet had obvious intermittent problems. D-LinkNet, SGCN, CoANet, and the proposed method completely restored the topology of the road. However, D-LinkNet, SGCN, and the proposed method had some mis-extraction, and CoANet had some missed extraction in the road edge part.

Fig. 6 shows surface segmentation results for test images from the DeepGlobe dataset, similar to those of Fig. 5. To achieve a representative comparative analysis, test images that featured current challenges for road extraction were selected. Fig. 6(a) shows farmland and contains rural-grade roads. Therefore, the road scale is small and some sections are covered by vegetation. For this image, D-LinkNet, DeepLabv3+, RoadNet, and SGCN extracted almost zero road sections. U-Net, ASPP-UNet, and CoANet extracted some of the road sections. However, the results were incomplete because of the effects of the vegetation cover. The proposed method overcame these difficulties and produced extraction results with superior completeness and correctness. In Fig. 6(b), the spectral features of the open space in the yard are essentially the same as those of the roads, and only small sections of some roads are included in the image. From the final extraction results, it can be seen that all methods exhibited different degrees of missed extraction. RoadNet, SGCN, CoANet, and the proposed method had relatively good performance. However, its extraction ability must be improved for fine roads with relatively short sections overall. Fig. 6(c) shows farmland. The overall tone is dark. However, the road running north–south is highlighted and has high contrast with other roads in the area. U-Net and CoANet defined this road as background. D-LinkNet, DeepLabv3+, and RoadNet recognized some road sections. SGCN recognized most road sections. However, the missed extraction problem was prominent. In contrast, the extraction results of ASPP-UNet and the proposed method had a high degree of completeness, with no evident

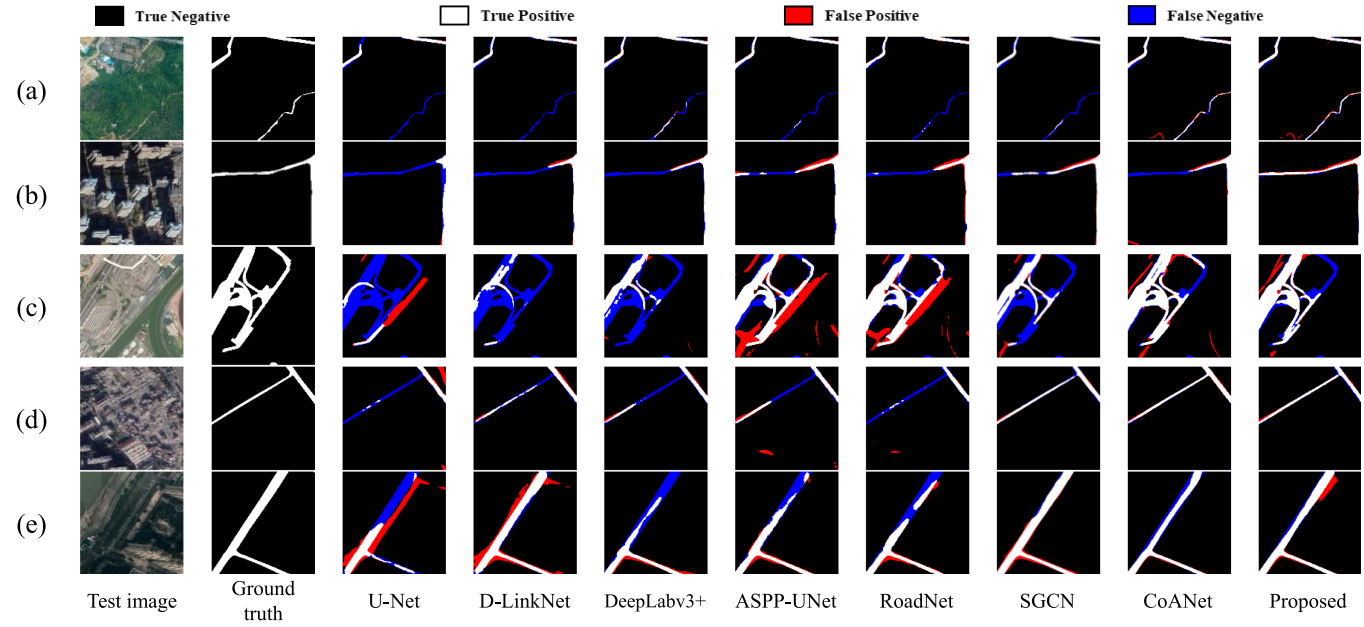


Fig. 5. Sample RSS results given by different methods (CHN6-CUG dataset).

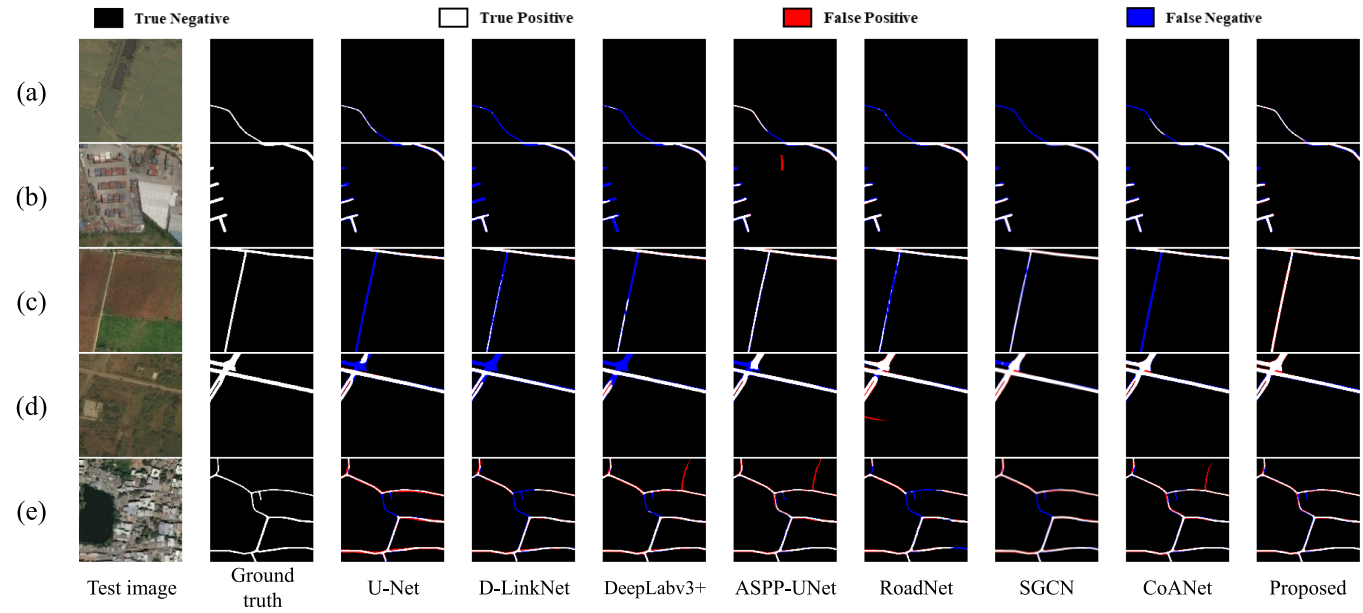


Fig. 6. Sample RSS results given by different methods (DeepGlobe dataset).

missed extractions. Fig. 6(d) contains two roads intersecting in both directions. Because a barrier is present, the intersection is a combination of an "L" intersection and a "T" intersection. UNet, D-LinkNet, DeepLabv3+, ASPP-UNet, and SGCN failed to correctly restore the actual characteristics of the road intersection. However, RoadNet, CoANet, and the proposed method effectively distinguished the road, the barrier, and other features. Fig. 6(e) is of a densely populated area with a highly complex road network environment. Buildings, vegetation, shadows, and even moving carriers all generate occlusions on the road surface, making correct road extraction difficult. From the extraction results, U-Net, D-LinkNet, DeepLabv3+, RoadNet, and SGCN had omission extraction problem. DeepLabv3+, ASPP-UNet, and CoANet incorrectly identified some other objects as roads. Overall, the proposed method had the best performance in recovering the connectivity of the road network.

In addition to the above-mentioned analysis of extraction performance for a typical sample image, the extraction capability of each method was further quantified comprehensively via a specific analysis using seven evaluation metrics, based on the road extraction difficulty. The values of each evaluation metric were the averages of those for all test images in the CHN6-CUG and DeepGlobe datasets.

TABLE I
COMPARISON OF RSS RESULT ACCURACY FOR DIFFERENT METHODS

| | *P*/% | *R*/% | *F1*/% | *OA*/% | *IoU*/% | *Com*/% | *Eor*/% |
|---|---|---|---|---|---|---|---|
| | | | | **CHN6-CUG** | | | |
| **U-Net** | 77.03 | 57.70 | 65.98 | 96.59 | 49.23 | 50.26 | 20.30 |
| **D-LinkNet** | 78.29 | 72.64 | 75.36 | 97.28 | 60.47 | 65.96 | 17.32 |
| **DeepLabv3+** | 80.39 | 72.62 | 76.31 | 97.42 | 61.70 | 68.32 | 15.98 |
| **ASPP-UNet** | 77.40 | 71.63 | 74.40 | 97.18 | 59.24 | 64.32 | 19.28 |
| **RoadNet** | 80.09 | 64.82 | 71.65 | 97.06 | 55.82 | 56.98 | 16.81 |
| **SGCN** | 82.51 | 72.41 | 77.13 | 97.54 | 62.77 | 65.87 | **13.73** |
| **CoANet** | **82.55** | 76.39 | 79.35 | 97.72 | 65.77 | 74.72 | 14.26 |
| **Proposed** | 81.57 | **77.98** | **79.74** | **97.73** | **66.30** | **76.89** | 14.96 |
| | | | | **DeepGlobe** | | | |
| **U-Net** | 79.88 | 76.98 | 78.40 | 98.19 | 64.48 | 79.54 | 14.30 |
| **D-LinkNet** | 82.06 | 78.39 | 80.18 | 98.35 | 66.92 | 82.29 | 13.25 |
| **DeepLabv3+** | 82.24 | 80.09 | 81.15 | 98.42 | 68.28 | 83.47 | 12.47 |
| **ASPP-UNet** | 79.85 | **81.79** | 80.81 | 98.35 | 67.80 | **85.53** | 16.18 |
| **RoadNet** | 79.74 | 76.38 | 78.03 | 98.17 | 63.97 | 78.97 | 14.00 |
| **SGCN** | 83.00 | 78.74 | 80.82 | 98.41 | 67.81 | 80.39 | **10.69** |
| **CoANet** | **83.99** | 81.26 | **82.60** | **98.54** | **70.36** | 85.20 | 11.89 |
| **Proposed** | 82.50 | 81.77 | 82.13 | 98.49 | 69.68 | 84.33 | 12.02 |

[a]Note: Bold font indicates the best metric, and underlined font indicates the second best metric.

Here, *P*, *R*, *F1*, *OA*, and *IoU* reflected pixel-level accuracy evaluations of the road extraction results. Larger *P* indicated a higher accuracy rate and larger *R* indicated a higher percentage of real roads extracted. Further, *F1*, *OA*, and *IoU* were comprehensive evaluation indexes combining positive and negative sample extraction results. Finally, larger *Com* indicated greater completeness of the road connectivity extraction and smaller *Eor* indicated a lower road connectivity extraction error rate.

Table I lists the results, from which the following conclusions can be drawn. 1) Most accuracy metrics of all methods in the DeepGlobe dataset were better than those in the CHN6-CUG dataset, and the specific metric values were closer in the DeepGlobe dataset. 2) The proposed method had the best metrics for both road datasets compared with U-Net, D-LinkNet, DeepLabv3+, and RoadNet. 3) Compared with ASPP-UNet, the proposed method had evident advantages when applied to the CHN6-CUG dataset. However, in the DeepGlobe dataset, the proposed method was better in most of the metrics, and only two metrics, *R* and *Com*, were slightly lower. It indicated that the completeness of the road extraction results of ASPP-UNet and the proposed method were close, but the error rates of ASPP-UNet were higher. 4) Compared with SGCN, the proposed method had evident advantages in most of the metrics, and only two metrics, *P* and *Eor*, were slightly worse. It indicated that the error rates of SGCN were lower, but the completeness of the road extraction results of the proposed method were slightly better. 5) Compared with CoANet, the proposed method had certain advantages in the CHN6-CUG dataset, while CoANet had slightly higher accuracy index in the DeepGlobe dataset, which proves that the proposed method had more outstanding ability to extract roads in remote sensing images of urban areas with smaller sample size and more complex environment. 6) Overall, the accuracy indexes of CoANet and the proposed method were higher, which also coincides with the final road extraction performance results shown in Figs. 5 and 6.

*2) Ablation Study:* In this section, the CHN6-CUG dataset was used as an example, and the dual-task form and MSMD-SCM were experimentally examined. Four specific cases, Situations 1–4 (S1–S4, respectively) were considered. In S1 and S3, the method contained the RSS branch only; in S1, MSMD-SCM in the decoder process was replaced with a $3 \times 3$ convolution kernel. S2 was based on S1 but the RCE branch was added, and S4 corresponded to the proposed method. The accuracy statistics for the four scenarios are listed in Table II.

From Table II, both the RCE branch and MSMC-SCM contributed significantly to the road extraction results. Specifically, comparing S1 and S3, and S2 and S4, we found that the addition of MSMD-SCM improved the metrics in the pixel-level evaluation. The improvement in *R* was particularly significant; this demonstrates that the module can extract road information more fully and completely and reduce the road extraction omission problem. This outcome also significantly improved the *Com* result of the connectivity evaluation index. Comparison of S1 and S2, and S3 and S4 revealed that the addition of the RCE

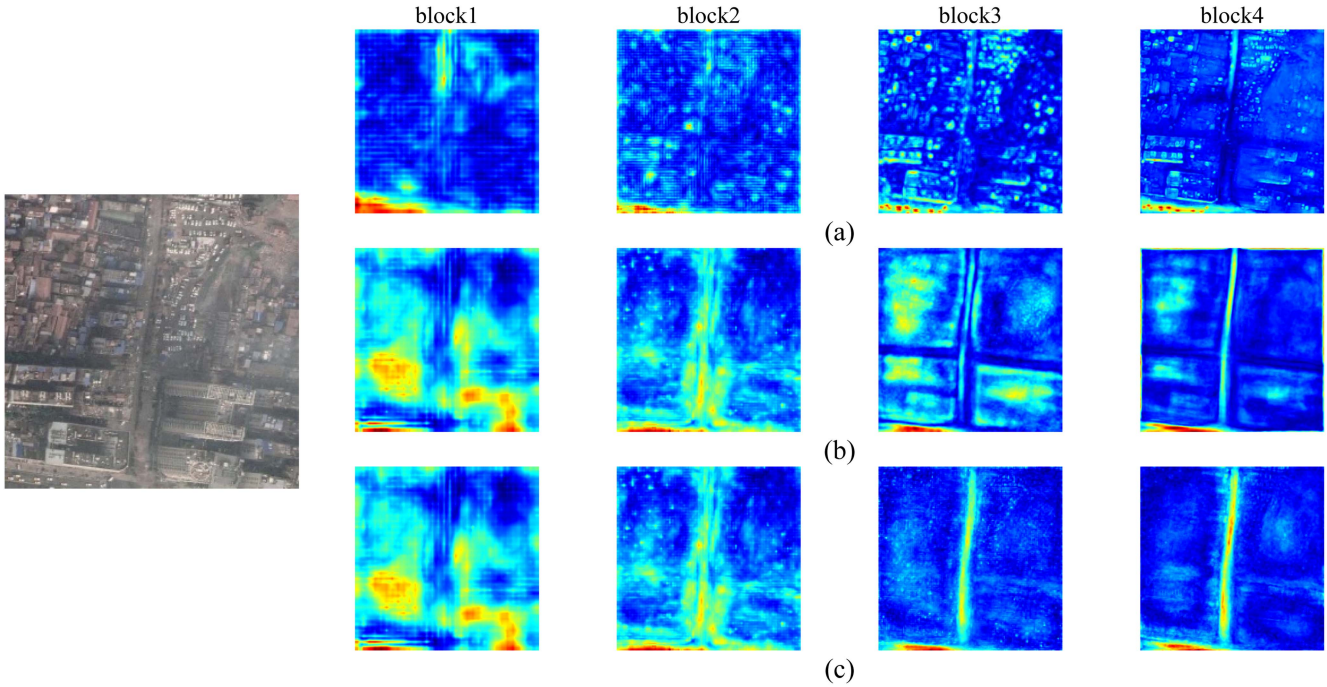


Fig. 7. Intermediate feature visualization results for various networks. (a)–(c) Baseline, baseline + MSMD-SCM, and baseline + MSMD-SCM + RCE, respectively.

TABLE II
ABLATION STUDY ACCURACY STATISTICS

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| **RCE branch** | | √ | | √ |
| **MSMD-SCM** | | | √ | √ |
| ***P* /%** | 78.16 | 79.49 | 82.17 | 81.57 |
| ***R* /%** | 72.18 | 72.34 | 76.27 | 77.98 |
| ***F1* /%** | 75.05 | 75.75 | 79.11 | 79.74 |
| ***OA* /%** | 97.25 | 97.35 | 97.69 | 97.73 |
| ***IoU* /%** | 60.06 | 60.97 | 65.44 | 66.30 |
| ***Com*/%** | 64.34 | 66.07 | 74.30 | 76.89 |
| ***Eor*/%** | 16.69 | 15.63 | 14.24 | 14.96 |

branch yielded improvements in most metrics (except for *P* and *Eor* in S3 and S4). The improvement in the *Com* result was particularly evident. This outcome proves that the RCE branch can improve the connectivity of the road extraction results and effectively suppress the problem of false extraction of negative samples.

In addition to the above-mentioned accuracy analysis, the output features of the first four decoder modules (block1–block4) were visualized to obtain a more intuitive representation of the role of MSMD-SCM and RCE in road extraction, as shown in Fig. 7. In the figure, "baseline" denotes the basic network framework of the proposed method, with MSMD-SCM and RCE excluded, and +MSMD-SCM and +RCE indicate addition of the corresponding module and branch, respectively.

From Fig. 7, the visualization results under all three conditions became closer to the actual road characteristics as the block1–block4 calculation progressed. In addition, the north–south road was more prominent following addition of MSMD-SCM, and the surrounding small, faceted buildings were somewhat suppressed in block4. Following further addition of RCE, the road feature separation from the other features was significantly accelerated, based on comparison of the block3 results. The highlighted features in block4 were essentially only roads, with interference from the other features further excluded. In summary, MSMD-SCM and RCE help improve the efficiency and accuracy of road separation from other objects in a feature space. Hence, the final road extraction results are optimized.

## V. DISCUSSION

### A. Evaluation of Model Transferability Performance

At present, the main factor restricting the full-scale application of deep learning is a limited supply of samples. However, the powerful portability of network models can provide a basic learning framework for migration learning, etc., thus reducing the dependence on samples and improving the reliability of "cross-domain supervision." Therefore, this subsection analyzes the portability of each method using the training data selected from the DeepGlobe dataset. Two experiments are conducted. In Experiment 1, the test data are from the Massachusetts road dataset [55], and in Experiment 2, the test data are from the CHN6-CUG dataset. As large differences in ground rules and background characteristics existed between the training and test data, the test results could be used as evaluation criteria for the model portability. To visually and comprehensively evaluate

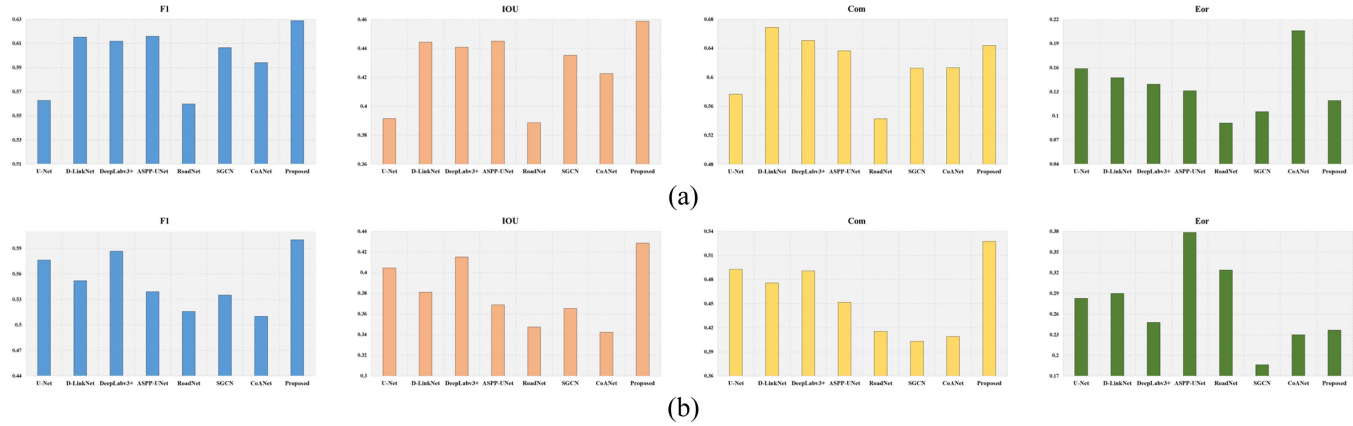


Fig. 8. Model portability experiment results for (a) and (b) experiments 1 and 2, respectively.

the model portability, the comprehensive pixel-level evaluation metrics *F1* and *IoU*, and the connectivity evaluation metrics *Com* and *Eor*, were selected. The results are shown in Fig. 8.

From Fig. 8, based on the pixel-level evaluation metrics, the proposed method yielded optimal results in both Experiments 1 and 2. As regards the connectivity evaluation metrics, the proposed method was in the top 3. The *Eor* metrics of the proposed method were higher than RoadNet and SGCN in Experiment 1, and higher than SGCN and CoANet in Experiment 2. However, the compared methods had lower *F1*, *IoU*, and *Com* in the corresponding experiments, indicating that the combined effect of road extraction results was worse, especially the topology integrity was low. Therefore, the lower *Eor* does not represent the advantage of extraction ability. Moreover, the proposed method achieved the best *Com* metrics in Experiment 2. However, for Experiment 1, this result was slightly poorer than those for D-LinkNet and DeepLabv3+. This outcome may have been related to the poorer representation of feature details (lower spatial resolution) and the relatively concentrated regional focus of the Massachusetts road dataset. In summary, the proposed method had the best portability when there were significant differences between the training and test data. Thus, the proposed deep learning network can provide a more reliable model framework with better generalization ability for migration learning than those of the comparison methods.

### B. Evaluation of Small-Sample Performance

Similar to improved model portability, the use of small samples as a form of weakly supervised learning can effectively alleviate the need for deep learning samples, thus enhancing the automation and intelligence of the entire process. This subsection reports an analysis of the accuracy and stability of each method for different sample sizes, using the DeepGlobe dataset as an example. The results are shown in Fig. 9, in which the term "original training sets (OTS)" indicates that the training and test data reported in Section IV-A were used in the experiments. Further, 8000, 6000, 4000, 2000, and 1000 denote the number of samples randomly drawn from the OTS training data (the test data were the same as the OTS). Because the experimental

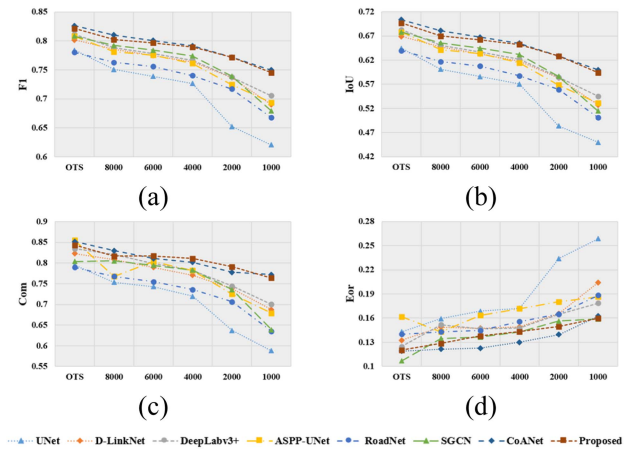


Fig. 9. Statistical charts of precision results for different sample sizes. (a)–(d) *F1*, *IoU*, *Com*, and *Eor*, respectively.

variables of this analysis were the method and sample number only, the road extraction ability of each method for different sample sizes could be measured directly, and the selected evaluation indexes were consistent with those of Section V-A.

From Fig. 9, *F1*, *IoU*, and *Com* gradually decreased with decreasing sample size, whereas *Eor* exhibited an increasing trend. In terms of the change degree, U-Net exhibited the largest changes in the four evaluation metrics. The flattest change trends were observed for CoANet and the proposed method. Therefore, these methods had the strongest ability to maintain road extraction efficiency as the sample size decreased. In addition, the gaps between the values of the four evaluation indexes for the proposed method and other six comparison methods showed a widening trend from OTS to the sample sizes of 8000, 6000, 4000, 2000, and 1000. In particular, when the sample size was 1000, CoANet and the proposed method had a clear advantage. In summary, CoANet and the proposed method have better extraction ability than other comparison methods when there are fewer samples.

As the key concept of the proposed method is the dual-task form, outstanding efficiency is not obtained under the same

TABLE III
TRAINING EFFICIENCY COMPARISON TABLE

| Method | Num | *F1*/% | *IoU*/% | *Com*/% | *Eor*/% | *Time*/s |
|---|---|---|---|---|---|---|
| **U-Net** | 8000 | 75.08 | 60.10 | 75.35 | 15.94 | 84.98 |
| **D-LinkNet** | 4000 | 76.38 | 61.79 | 77.08 | 14.91 | 67.25 |
| **DeepLabv3+** | 4000 | 76.69 | 62.19 | 78.13 | 14.77 | 136.70 |
| **ASPP-UNet** | 4000 | 76.13 | 61.46 | 78.18 | 17.18 | 142.36 |
| **RoadNet** | 6000 | 75.58 | 60.75 | 75.49 | 14.47 | 221.05 |
| **SGCN** | 4000 | 77.43 | 63.18 | 78.30 | 14.30 | 529.25 |
| **CoANet** | 2000 | 77.19 | 62.85 | 77.77 | 13.95 | 265.67 |
| **Proposed** | 2000 | 77.19 | 62.86 | 79.12 | 14.95 | 68.12 |

training conditions. However, when small-sample analysis results are considered, the proposed method has efficiency advantages when obtaining extraction results with approximately the same accuracy. Table III presents training efficiency results. In Table III, *Num* denotes the number of training-set samples and *Time* is the average training time for each epoch in the corresponding training set. (The experimental environment cannot support SGCN and CoANet to run at a batch size of 32. Therefore, the batch size of the method in the experiment was 16.) The seven methods achieved almost the same accuracy for the selected number of samples (based on IoU). However, the proposed method had roughly the same training time as D-LinkNet, but outperformed UNet, DeepLabv3+, ASPP-UNet, RoadNet, SGCN, and CoANet. In addition, in order to accurately compare the efficiency of the proposed method with SGCN and CoANet, the batch size of the proposed method was set to 16 for training, and the results show that the average training time for each epoch of the proposed method under this condition is 77.48 s, so the proposed method is better than SGCN and CoANet in terms of efficiency. Therefore, from a comprehensive perspective, the proposed method can balance accuracy and efficiency with higher practical value.

## VI. CONCLUSION

As important topographic elements, roads have their own shape and scale irregularities, and road image features can be extracted accurately and with good detail from high-resolution remote sensing images using established rules. In addition, as roads form the basic transportation framework of a given location, the connectivity relationships constructed from road extraction results directly reflect the topology of the targeted transportation network. Thus, these relationships are important for practical applications of extracted data to transportation. Here, a targeted study of road extraction was performed considering road shape patterns and scale differences, as well as connectivity, and a dual task-driven road extraction method was proposed. In this approach, the newly developed MSMD-SCM was added and the extraction strategies were improved, with end-to-end networks being used as the basic framework. Hence, the proposed method was shown to have superior performance to comparable typical networks in terms of quantitative evaluation metrics, model portability, and small-scale learning capability. However, as road extraction is an intensive prediction task, the generation of appropriate training data requires excessive human intervention. Therefore, future research should focus on the introduction of multiple data sources (OpenStreetMap data, trajectory data, etc.) for automatic sample collection, along with weakly supervised learning.

## REFERENCES


[1] Z. Chen et al., "Road extraction in remote sensing data: A survey," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102833, doi: 10.1016/j.jag.2022.102833.

[2] S. Leninisha and K. Vani, "Water flow based geometric active deformable model for road network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 102, pp. 140–147, Apr. 2015, doi: 10.1016/j.isprsjprs.2015.01.013.

[3] G. Fu, H. Zhao, C. Li, and L. Shi, "Road detection from optical remote sensing imagery using circular projection matching and tracking strategy," *J. Indian Soc. Remote Sens.*, vol. 41, no. 4, pp. 819–831, Jul. 2013, doi: 10.1007/s12524-013-0295-y.

[4] Y. Shao, B. Guo, X. Hu, and L. Di, "Application of a fast linear feature detector to road extraction from remotely sensed imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 3, pp. 626–631, Sep. 2011, doi: 10.1109/JSTARS.2010.2094181.

[5] K. Treash and K. Amaratunga, "Automatic road detection in grayscale aerial images," *J. Comput. Civil Eng.*, vol. 14, no. 1, pp. 60–69, Jan. 2000, doi: 10.1061/(ASCE)0887-3801(2000)14:1(60).

[6] Z. Miao, W. Shi, P. Gamba, and Z. Li, "An object-based method for road network extraction in VHR satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 10, pp. 4853–4862, Oct. 2015, doi: 10.1109/JSTARS.2015.2443552.

[7] M. Maboudi, J. Amini, M. Hahn, and M. Saati, "Road network extraction from VHR satellite images using context aware object feature integration and tensor voting," *Remote Sens.*, vol. 8, no. 8, Aug. 2016, Art. no. 637, doi: 10.3390/rs8080637.

[8] 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, doi: 10.1109/CVPR.2015.7298965.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[10] 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, doi: 10.1109/TPAMI.2016.2644615.

[11] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Representations*, 2015.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[13] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[14] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818, doi: 10.1007/978-3-030-01234-2_49.

[15] 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, doi: 10.1109/CVPR.2016.90.

[16] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018, doi: 10.1109/lgrs.2018.2802944.

[17] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803, doi: 10.1109/cvpr.2018.00813.

[18] H. He, D. Yang, S. Wang, S. Wang, and Y. Li, "Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss," *Remote Sens.*, vol. 11, no. 9, Apr. 2019, Art. no. 1015, doi: 10.3390/rs11091015.

[19] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 192–1924, doi: 10.1109/CVPRW.2018.00034.

[20] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-based segmentation and tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8919–8931, Dec. 2020, doi: 10.1109/TGRS.2020.2991733.
[21] Y. Wang, J. Seo, and T. Jeon, "NL-LinkNet: Toward lighter but more accurate road extraction with nonlocal operations," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3000105, doi: 10.1109/LGRS.2021.3050477.
[22] V. Pihur, S. Datta, and S. Datta, "Weighted rank aggregation of cluster validation measures: A Monte Carlo cross-entropy approach," *Bioinformatics*, vol. 23, no. 19, pp. 1607–1615, Jul. 2007, doi: 10.1093/bioinformatics/btm158.
[23] S. Xie and Z. Tu, "Holistically-nested edge detection," *Int. J. Comput. Vis.*, vol. 125, no. 1-3, pp. 3–18, Dec. 2017, doi: 10.1007/s11263-017-1004-z.
[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/tpami.2018.2858826.
[25] H. Xu, H. He, Y. Zhang, L. Ma, and J. Li, "A comparative study of loss functions for road segmentation in remotely sensed road datasets," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, Feb. 2023, Art. no. 103159, doi: 10.1016/j.jag.2022.103159.
[26] G. Zhou, W. Chen, Q. Gui, X. Li, and L. Wang, "Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5614115, doi: 10.1109/TGRS.2021.3128033.
[27] S. He et al., "RoadTagger: Robust road attribute inference with graph neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10965–10972, doi: 10.1609/aaai.v34i07.6730.
[28] F. Bastani et al., "RoadTracer: Automatic extraction of road networks from aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4720–4728, doi: 10.1109/CVPR.2018.00496.
[29] Y. Wang et al., "DDU-Net: Dual-decoder-U-Net for road extraction using high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412612, doi: 10.1109/TGRS.2022.3197546.
[30] M. Zhou, H. Sui, S. Chen, J. Wang, and X. Chen, "BT-RoadNet: A boundary and topologically-aware neural network for road extraction from high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 168, pp. 288–306, Oct. 2020, doi: 10.1016/j.isprsjprs.2020.08.019.
[31] Y. Li, L. Xiang, C. Zhang, and H. Wu, "Fusion taxi trajectories and RS images to build road map via DCNN," *IEEE Access*, vol. 7, pp. 161487–161498, 2019, doi: 10.1109/access.2019.2951730.
[32] A. Grillo, V. A. Krylov, G. Moser, and S. B. Serpico, "Road extraction and road width estimation via fusion of aerial optical imagery, geospatial data, and street-level images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2413–2416, doi: 10.1109/igarss47720.2021.9554540.
[33] E. Khesali, M. J. V. Zoej, M. Mokhtarzade, and M. Dehghani, "Semi automatic road extraction by fusion of high resolution optical and radar images," *J. Indian Soc. Remote Sens.*, vol. 44, no. 1, pp. 21–29, Feb. 2016, doi: 10.1007/s12524-015-0480-2.
[34] M. Guo, C. Lu, Q. Hou, Z. Liu, M. Cheng, and S. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1140–1156.
[35] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/TPAMI.2019.2938758.
[36] X. Chen, Q. Sun, W. Guo, C. Qiu, and A. Yu, "GA-Net: A geometry prior assisted neural network for road extraction," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 114, Nov. 2022, Art. no. 103004, doi: 10.1016/j.jag.2022.103004.
[37] T. Sun, Z. Di, P. Che, C. Liu, and Y. Wang, "Leveraging crowdsourced GPS data for road extraction from aerial imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7501–7510, doi: 10.1109/CVPR.2019.00769.
[38] J. Mei, R.-J. Li, W. Gao, and M.-M. Cheng, "CoANet: Connectivity attention network for road extraction from satellite imagery," *IEEE Trans. Image Process.*, vol. 30, pp. 8540–8552, 2021, doi: 10.1109/TIP.2021.3117076.
[39] C. Wang, R. Xu, S. Xu, W. Meng, and X. Zhang, "DA-Net: Dual branch transformer and adaptive strip upsampling for retinal vessels segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2022, pp. 528–538, doi: 10.1007/978-3-031-16434-7_51.
[40] Z. Zhang, X. Sun, and Y. Liu, "GMR-Net: Road-extraction network based on fusion of local and global information," *Remote Sens.*, vol. 14, no. 21, Oct. 2022, Art. no. 5476, doi: 10.3390/rs14215476.
[41] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, "RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2043–2056, Apr. 2019, doi: 10.1109/TGRS.2018.2870871.
[42] A. Batra, S. Singh, G. Pang, S. Basu, C. V. Jawahar, and M. Paluri, "Improved road connectivity by joint learning of orientation and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10377–10385, doi: 10.1109/CVPR.2019.01063.
[43] M. Zhou, H. Sui, S. Chen, J. Liu, W. Shi, and X. Chen, "Large-scale road extraction from high-resolution remote sensing images based on a weakly-supervised structural and orientational consistency constraint network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 193, pp. 234–251, Nov. 2022, doi: 10.1016/j.isprsjprs.2022.09.005.
[44] L. Ding and L. Bruzzone, "DiResNet: Direction-aware residual network for road extraction in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10243–10254, Dec. 2021, doi: 10.1109/TGRS.2020.3034011.
[45] X. Li, Y. Wang, L. Zhang, S. Liu, J. Mei, and Y. Li, "Topology-enhanced urban road extraction via a geographic feature-enhanced network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8819–8830, Dec. 2020, doi: 10.1109/TGRS.2020.2991006.
[46] X. Lu et al., "Cascaded multi-task road extraction network for road surface, centerline, and edge extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621414, doi: 10.1109/TGRS.2022.3165817.
[47] Y. Wei and S. Ji, "Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602312, doi: 10.1109/TGRS.2021.3061213.
[48] A. Mosinska, P. Marquez-Neila, M. Koziński, and P. Fua, "Beyond the pixel-wise loss for topology-aware delineation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3136–3145, doi: 10.1109/CVPR.2018.00331.
[49] D. Oner, M. Koziński, L. Citraro, N. C. Dadap, A. G. Konings, and P. Fua, "Promoting connectivity of network-like structures by enforcing region separation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5401–5413, Sep. 2022, doi: 10.1109/TPAMI.2021.3074366.
[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
[52] Q. Zhu et al., "A global context-aware and batch-independent network for road extraction from VHR satellite imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 353–365, May 2021, doi: 10.1016/j.isprsjprs.2021.03.016.
[53] I. Demir et al., "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 172–181, doi: 10.1109/CVPRW.2018.00031.
[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[55] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Graduate Dept. Comput. Sci., Univ. of Toronto, Toronto, ON, Canada, 2013.

**Yuzhun Lin** received the B.S. and M.S. degrees in photogrammetry and remote sensing in 2015 and 2018, respectively, from the Institute of Geospatial Information, Information Engineering University, Zhengzhou, China, where he is currently working toward the Ph.D. degree in remote sensing image processing and machine learning.

He is currently a Lecturer with Information Engineering University. His research interests include remote sensing image processing and machine learning.

**Fei Jin** received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from the Institute of Geospatial Information, Information Engineering University, Zhengzhou, China, in 2006, 2009, and 2013, respectively.

He is currently an Associate Professor with Information Engineering University. His research interests include remote sensing image analysis and machine learning.

**Dandi Wang** received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from the Institute of Geospatial Information, Information Engineering University, Zhengzhou, China, in 2015, 2018, and 2022, respectively.

She is currently a Lecturer with Information Engineering University. Her research mainly focuses on remote sensing data processing and machine learning.

**Shuxiang Wang** received the B.S. degree in photogrammetry and remote sensing from Information Engineering University, Zhengzhou, China, in 2005, and the M.S. degree in photogrammetry and remote sensing from Hohai University, Nanjing, China, in 2009. She is currently working toward the Ph.D. degree in remote sensing image processing and machine learning with Information Engineering University.

She is currently an Associate Professor with Information Engineering University. Her research mainly focuses on remote sensing image processing and machine learning.

**Xiao Liu** is currently working toward the M.S. degree in machine learning and its application with the Institute of Geospatial Information, Information Engineering University, Zhengzhou, China.

Her research interests include machine learning and its application.