IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 September 2023, accepted 26 September 2023, date of publication 29 September 2023,
date of current version 4 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3320684

RESEARCH ARTICLE

RFE-LinkNet: LinkNet with Receptive Field Enhancement for Road Extraction from High Spatial Resolution Imagery

HUA ZHAO, HUA ZHANG, AND XIANGCHENG ZHENG
School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China

Corresponding author: Hua Zhang (zhhua_79@163.com)

This work was supported in part by the National Natural Science Foundation, China, under Grant 41971400, Grant 41974039, and Grant U22A20569; and in part by the Fundamental Research Funds for the Central Universities under Grant 2019ZDPY09.

ABSTRACT Extracting roads from high spatial resolution imagery (HSRI) has been a hot research topic in recent years. Particularly, the fully convolutional network (FCN)-based methods have shown promising performance in accurately extracting roads from HSRI. However, most existing FCN-based approaches suffer from such deficiencies of convolution in spatial detail loss, inadequate fusion of multi-scale features, and lack of consideration for long-range dependencies, making road extraction from HSRI remain a challenging task. To address the above challenges, based on LinkNet architecture, this paper provided a novel neural network named RFE-LinkNet, which employs a U-shaped framework and integrates several receptive field enhancement modules and dual attention modules. In the RFE-LinkNet, in order to enhance the spatial information perception and capture long-range dependencies, the multiple receptive field enhancement module is devised to expand the receptive field while preserving the spatial details of feature maps. And dual attention module is provided to capture accurate features for road extraction by refining multi-scale features from the different-level feature maps in the view of their relative importance. Experiments on Massachusetts road dataset and DeepGlobe road dataset are conducted to evaluate the performance of RFE-LinkNet, respectively. Experimental results show that the proposed method achieves superior performance compared to previous road extraction, establishing its state-of-the-art effectiveness. The code of RFE-LinkNet is available at https://github.com/zhengxc97/RFE_LINKNET.

INDEX TERMS Road extraction, high spatial resolution imagery (HSRI), receptive field.

I. INTRODUCTION

Timely and complete road information plays an important role in urban planning, traffic navigation, digital map updates, and autonomous driving, *etc* [1], [2]. In the last few years, the field of remote sensing has experienced a rapid evolution, leading to significant improvements in the spatial and spectral resolution of remote sensing images. Especially, high spatial resolution imagery (HSRI) can provide rich semantic and spatial details information for ground objects and has gradually become one of the main data sources for road extraction [3]. Although manual interpretation can capture accurate road from HSRI, it is time-consuming and inefficient. Compared with medium to low resolution remote sensing images, HSRI contains more complex background information, and road extraction is easily disturbed by background information such as shadows, buildings, and railways, which makes this task challenging. Therefore, how to achieve efficient and accurate road extraction under the interference of complex background information has become a research focus [4], [5], [6].

Fortunately, numerous methods have been introduced in the past few decades to extract roads automatically from images [5], [6]. These methods can be broadly categorized into two main types: traditional handcrafted feature-based approaches and deep learning-based approaches.

The associate editor coordinating the review of this manuscript and approving it for publication was Amin Zehtabian.

The former primarily aims to develop a proficient classifier that can extract roads from images by utilizing manually extracted features including geometry, texture, spectrum, and shadow, etc. [7], [8], [9], [10], [11], [12], [13]. Although these methods have achieved certain progress in road extraction from images, most of them were designed based on low-level road features, and had complex extraction processes, poor stability and weak generalization ability, *etc*. Thus, these methods are difficult to meet the accuracy and time requirements of road extraction tasks, and are not suitable for application on large-scale datasets.

Recently, convolutional neural network (CNN), especially fully-convolutional network (FCN) architecture, had achieved great success in image semantic segmentation [14], [15], [16], [17], [18], [19]. Compared with traditional algorithms, based on low-level road feature design, CNN has powerful learning and feature expression abilities, and can extract information layer by layer from pixel level, showing greater advantages in the accuracy and automation of road extraction from images. Several research efforts in the past have utilized CNN to tackle the road segmentation task [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34]. For example, Mnih and Hinton [20] utilized restricted Boltzmann machines to extract road from images. Cheng et al. [21] designed a cascaded end-to-end CNN (CasNet) to simultaneously perform the road area and centerline extraction tasks from remote sensing images. A road extraction network with fewer parameters, however, better performance, which combines the strengths of U-Net and residual learning, was proposed [5]. Cheng et al. [22] proposed a novel framework for segmentation of the general road region from one single image based on the road vanishing point estimation by using the Locally Adaptive Soft-Voting (LASV) algorithm. To achieve road centerline and smooth and complete segmentation, the MSMT-RE algorithm was provided by Lu et al. [23]. To solve such problems as occupied proportion of road area in UAV images and constant size of convolutional kernel, Li et al. [24] presented a method for road extraction from UAV images by combing the GANs and multiscale context aggregation. To solve the problem of road connectivity exists in road vectorization, a road vectorization mapping network (RVMNet) framework was proposed, in which a node proposal network (NPN) module and a node connectivity-based road refinement module were designed [25]. Ding and Bruzzone [26] designed a direction-aware residual network (DiResNet) to enhance the road topology and local directions feature. Tan et al. [27] designed a point-based iterative aerial image exploration method characterized by usage of flexible step and segmentation cues, experimental results showed the proposed method can achieve a significant improvement on the road graph alignment and connectivity compared to state-of-the-art methods. Wei and Ji [28] proposed the ScRoadExtractor for road surface extraction, in which a road label propagation algorithm was designed to propagate semantic information from sparse scribbles to unlabeled pixels. A BT-RoadNet was proposed to improve the learning ability of boundary and topological structure using a composition of two U-Net-like networks [29]. In summary, despite these CNN-based approaches had achieved good performance on road extraction, most of them simply focused on multi-scale encoder architectures or multiple branches in neural networks, but ignored some inherent characteristics of road surface, and they failed to capture long-range dependencies of roads, resulting in insufficient contextual information and spatial details loss during the process of road extraction.

Indeed, roads in HSRI are characterized by their long distances and natural interconnections. Consequently, road extraction networks need to incorporate a large receptive field that covers the entirety of image. Furthermore, roads within the HSRI dataset commonly display attributes of being slim, intricate, and occupying a limited portion of the overall image. Given the characteristics of narrow and complex roads in HSRI, preserving detailed spatial information becomes a crucial aspect. To address this problem, network designs have been tailored to enhance performance in road extraction tasks [30], [31], [32], [33], [34]. D-LinkNet was specifically proposed to tackle the task of road extraction from HSRI by addressing the intricacies, interconnections, narrowness, and extensive span of roads in a particular scenario [30]. Wang et al. [31] designed a nonlocal LinkNet (NL-LinkNet) containing differentiable nonlocal operations to capture long-range dependencies. By introducing the GCA block, Zhu et al. [32] designed the GCB-Net to improve the accuracy and robustness of road extraction task. Xie et al. [33] presented HsgNet for road extraction by capturing long-distance information and global-context semantic information. A dual attention dilated-LinkNet (DAD-LinkNet) was designed to extract road by integrating local features with global dependencies using image and floating vehicle trajectory data [34]. An attention-based hybrid multiple attention network (HMANet) was proposed to achieve spatial context long-range dependencies [35]. Although the above methods done well in achieving long-distance dependencies, global information and local information were not being well considered at the same time. Especially, these methods acquired local features by learning part of the spatial information, while the detailed spatial information is lost, which is not conducive to the road extraction with complex background, long span, and natural connectivity, *etc.*

In summary, despite the significant advancements made by the aforementioned methods, accurately extracting complete and uninterrupted roads from images remains a challenge. In this paper, a novel RFE-LinkNet was designed to further improve completeness and smoothness of road extraction from HSRI. The contributions of this work can be summarized as

1) An innovative semantic segmentation network, named RFE-LinkNet, was designed for the automated extraction of roads from HSRI.

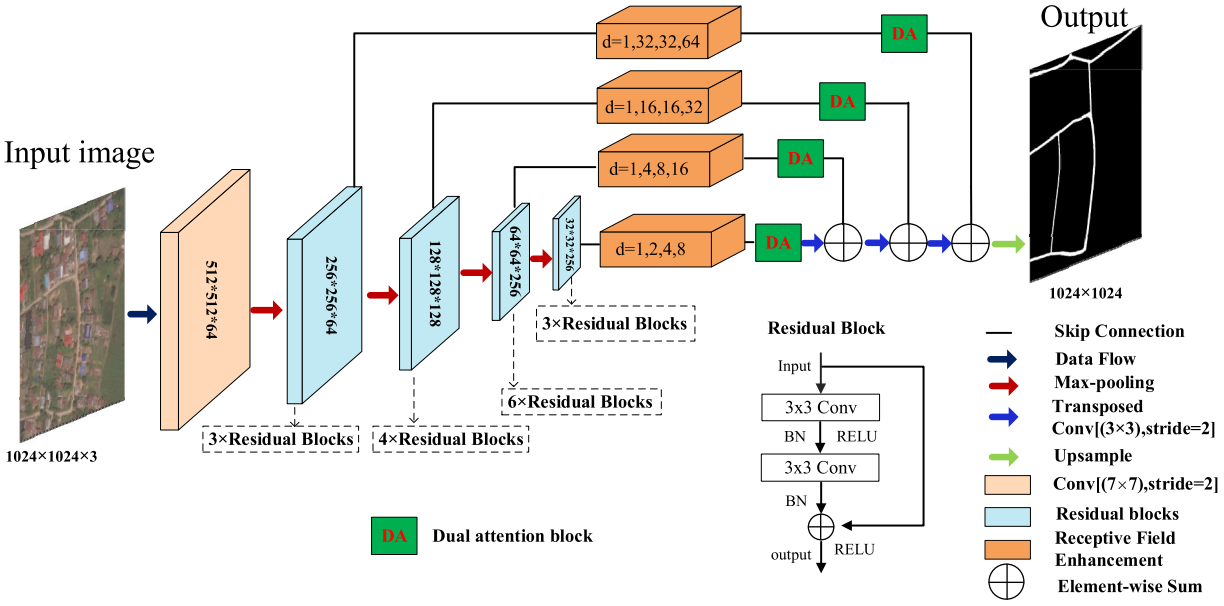

FIGURE 1. The architecture of the proposed RFE-LinkNet which is composed of four modules. (a) Multi-scale feature extraction module. (b) Multiple receptive field enhancement module. (c) Feature optimization module, and (d) Multi-scale features fusion module.

2) In order to enhance the extracted multi-scale feature representation ability of long-range dependencies and detailed spatial information, the receptive field enhancement modules with different structures were embedded in the output parts of the feature maps at different levels in the proposed network.
3) To capture representative features and effectively aggregate feature maps from multiple hierarchical layers, a dual attention block was developed. This block aims to optimize the feature maps based on their respective degrees of importance.
4) The proposed RFE-LinkNet achieved 2.21% and 4.85% F1 and 3.03% and 6.78% intersection over union (IoU) improvements compared with DlinkNet on the Massachusetts road dataset [36] and DeepGlobe road dataset [37]. and outperforms other three SOTA methods on the two datasets.

II. METHODOLOGY

The object of this study is to explore a method for road extraction from HSRI, especially to improve completeness and smoothness of the extracted roads. By embedding the proposed multiple receptive field enhancement modules and dual attention modules into the traditional LinkNet, a novel U-shape network is designed to automatically extract roads from HSRI. As shown in Fig. 1, the proposed approach comprises four key components: multi-scale feature extraction module, multiple receptive field enhancement module, feature optimization module and multi-scale features fusion module. Firstly, the images are fed into the multi-scale feature extraction module to obtain multi-scale feature map of the road. Then, in order to enhance the extracted multi-scale feature representation ability, the receptive field enhancement modules with different structures are embedded in the output parts of the feature maps at different levels. Secondly, the dual attention block is formulated to refine the feature maps from various hierarchical levels based on their individual degrees of importance. Finally, the optimized multi-scale features are aggregated and are used to produce the final road map through upsampling method.

A. MULTI-SCALE FEATURE EXTRACTION MODULE

Variations of the encoder-decoder architecture, such as U-Net, have gained significant popularity for their ability of extracting multi-scale features from images. This is primarily due to the utilization of skip connections, which facilitates the fusion of deep, semantic, coarse-grained feature maps from the decoder sub-network with shallow, low-level, fine-grained feature maps from the encoder sub-network. In particular, LinkNet is a highly efficient neural network designed for semantic segmentation tasks. It leverages the benefits of skip connections, residual blocks, and the encoder-decoder architecture [17]. In this research, the proposed RFE-LinkNet model utilizes LinkNet as the backbone with a pretrained encoder. As shown in Fig. 1, the multi-scale feature extraction module includes a convolutional block with kernel size of 7 × 7, stride=2 and four residual blocks used in ResNet34 [38]. Indeed, ResNet34 was initially designed for image classification tasks on images of size 256 × 256. However, in this study, the objective is to extract roads from HSRI with a larger size of 1024 × 1024. Therefore, it is necessary to adapt the encoder layers of the network to accommodate this new input domain. Firstly, the Conv [(7 × 7), stride=2] is used to down-sample the features and extract feature maps with 64 channels and decrease the

spatial resolution of the input image to 1/2, and subsequently to 1/4 by using a 2 × 2 maximum pooling layer, then the size of output feature map is 256 × 256 and can be fed into the ResNet34 network to achieve multi-scale features by cascaded convolutions and pooling operations. Lastly, the feature maps are downscaled to 1/4, 1/8, 1/16, and 1/32 of the original input image's resolution, respectively. The corresponding channels in each level are 64, 128, 256, and 256.

B. RECEPTIVE FIELD ENHANCEMENT MODULE

As described in section II-A, multi-scale features for road extraction are captured through the multi-scale feature extraction module, and the highest scaling rate of the feature map reaches 1/32, which can ensure a large receptive field for the network to some extent. However, firstly, to address the issue of large road span in the images, it is important for the road extraction network to have a large receptive field that can cover the entire image. Secondly, preserving detailed spatial information is crucial for accurately extracting narrow and complex roads that cover only a small part of the whole image. Thirdly, considering the natural connectivity and characteristics of roads, it is important to design a network that addresses the need for enlarging the receptive field, extracting multi-scale features, and preserving detailed information.

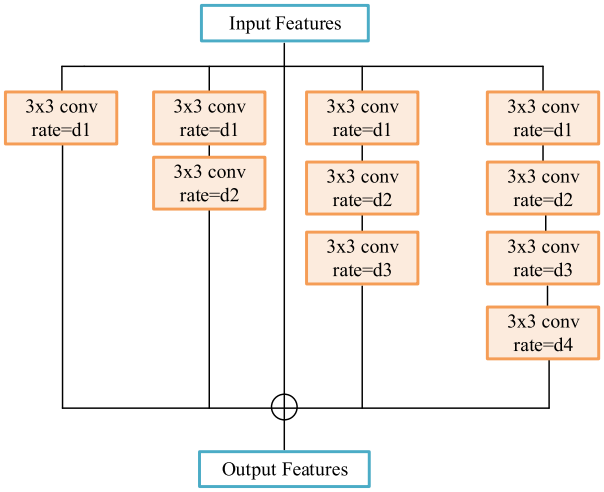

FIGURE 2. Receptive field enhancement module which is composed of four multi-scale features branches and a residual branch similar to residual mapping. The values of d1, d2, d3 and d4 represent different dilation rates, which are determined by the values designed in the Fig. 1, for example, 'd=1, 2, 3, 4' denotes d1=1, d2=2, d3=4, and d4=8, respectively.

Indeed, Atrous convolution is a powerful technique for enlarging the receptive field of a network without reducing the image resolution or increasing the number of network parameters. Generally, convolution with large receptive field can capture more abstract features for large objects, while convolution with small receptive field are better for small objects. By combining atrous convolutions with different dilation rates, multi-scale features can be extracted for road extraction. In the paper, inspired by the atrous convolution and inception structure [39], the receptive field enhancement (RFE) module is designed to accomplish preserving spatial details of the road, larger receptive field and multi-scale features aggregation. As shown in Fig. 2, the atrous convolutions are stacked in a cascaded manner. The proposed RFE module contains five cascaded branches, four branches are designed to extract multi-scale features map, in which, atrous convolution with different dilation rates and numbers are stacked to obtain four different branches of receptive fields, respectively. To mitigate the gradient disappearance issue in multiple convolution stacks during network training and address the grid effect caused by cascaded dilated convolutions, incorporating ideas from residual networks [38], we designed a residual mapping branch. Finally, by fusing the multi-scale features obtained from the five branches, the RFE module can enhance multi-scale features discriminability and robustness, also enlarge the receptive field of RFE-LinkNet in a certain.

C. FEATURE OPTIMIZATION MODULE

As shown in Fig. 2, multi-scale features can be captured from five branches of the RFE module, it is essential to aggregate them to get multi-scale context information for precise road extraction. In typical encoder-decoder architectures, multi-scale features are fused by simple sum or concatenation operations, leading to importance of features at different scales is ignored. The reasons are as follows: for the obtained feature maps, the low-level feature maps have small receptive fields and thus contain rich spatial location information but poor semantic understanding. On the other hand, the high-level feature maps have larger receptive fields and capture more semantic information but less precise or weak spatial location information. By combining both low-level and high-level feature maps, networks achieve a balance between capturing semantic information and preserving spatial location information. In addition, as the extracted features are often influenced by similar patterns and noisy backgrounds, it is necessary to emphasize important parts and suppress unimportant parts. In this article, to obtain representative features by aggregating feature maps of different levels, we will assign different weights to feature maps of different levels as their relative importance in the channel dimension and as sub features in the spatial dimension.

Considering the characteristics of road, inspired by reference [40], we designed feature optimization module shown as Fig. 3, the output features from the RFE module are taken as input features Fin, and then S is produced through channel attention module (CAM) and spatial attention module (SAM). In addition, the residual connection is added to enhance the fitting ability of the module.

1) CHANNEL ATTENTION MODULE

The CAM explores 'which channel' is important in the feature map by exploiting the inter-channel relationship of features using squeeze-and-excitation operations.

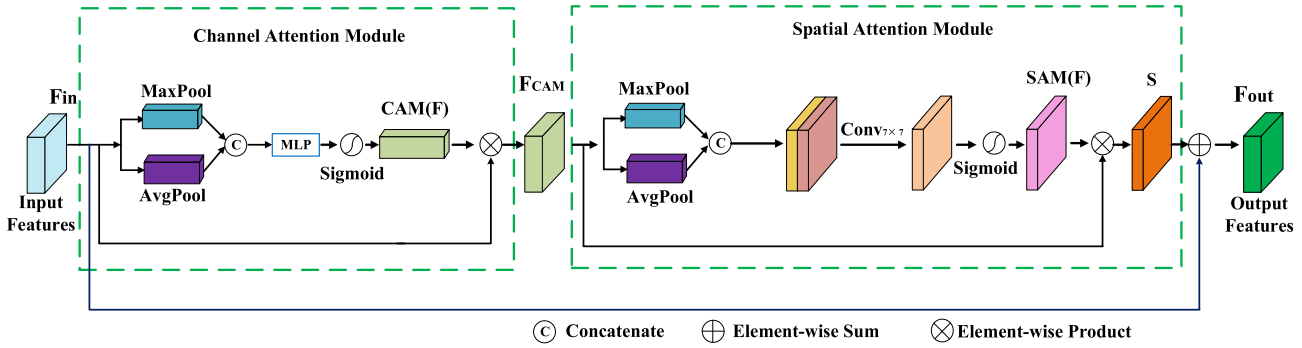

FIGURE 3. Feature optimization module, which consists of two modules: channel attention module (CAM) and spatial attention module (SAM).

Firstly, to reduce the interference of spatial location information and the computational complexity, the maximum pooling and average pooling are applied simultaneously, then the generated features are concatenated and normalized using the sigmoid function. Next, the channel attention weight vector CAM (F) can be obtained. lastly, the output feature map F_{CAM} is achieved through aggregating CAM (F) and F by the element-wise product. Here, the size of hidden activation layer is $R^{c/r\times 1\times 1}$, and r is 16. The definition of CAM is described as follows:

$$F_{CAM} = \sigma(\text{MLP}(\text{MaxPool}(\text{Fin})) + \text{MLP}(\text{AvgPool}(\text{Fin}))) \otimes \text{Fin} \quad (1)$$

where $\text{Fin} \in R^{C\times H\times W}$ is the input features, C and H denote the size of image, and *C* is the dimension of feature map, σ represents the sigmoid function, MLP is a multi-layer perceptron network, AvgPool(Fin) *and* MaxPool(Fin) are the generated features by the average pooling and maximum pooling, respectively, $\otimes$ denotes the element-wise product operation, and F_{CAM} is the intermediate output features through the CAM.

2) SPATIAL ATTENTION MODULE

The SAM explores 'where' is a useful part, which focuses on informative regions and suppresses the ineffective regions. First, based on the CAM, average-pooling and max-pooling operations are utilized to obtain refined feature F_{CAM}, $AP(F_{CAM})$ and $MP(F_{CAM})$, respectively. $AP(F_{CAM})$ and $MP(F_{CAM})$ are concatenated and then processed by a 7×7 convolution layer, SAM(F) is obtained by normalizing them. At last, feature map S is obtained by aggregating SAM(F) and F_{CAM} by element-wise product. The SAM is defined as

$$S = \sigma(\text{Conv}_{7\times 7}(\text{Concat}(AP(F_{CAM}), MP(F_{CAM}))) \otimes F_{CAM} \quad (2)$$

where S denotes the output feature map produced by the SAM, $\text{Conv}_{7\times 7}$ is a convolution operation with a kernel of 7×7, Concat denotes the element-wise sum operation.

Inspired by the residual networks, the final output feature map F_{out} is achieved by aggregating S and Fin using element-wise sum.

$$F_{out} = S + \text{Fin} \quad (3)$$

D. MULTI-SCALE FEATURES FUSION MODULE

As illustrated in the right part of Fig. 1, multi-scale feature maps are achieved through different levels through different residual block, RFE module and DA module. It is essential to aggregate them to achieve multi-scale context information and restore the spatial information lost in the encoder parts. Inspired by the U-Net architecture, the obtained multi-scale feature maps are progressively fused from neighbor branches in a bottom-up manner to shrink the semantic and resolution gaps. As shown in Fig. 4, the high-level feature F_H is upsampled to the same resolution as the low-level feature F_L, and concatenated with F_{HL}. Finally, the fused feature F_{HL} is achieved by using a residual block.

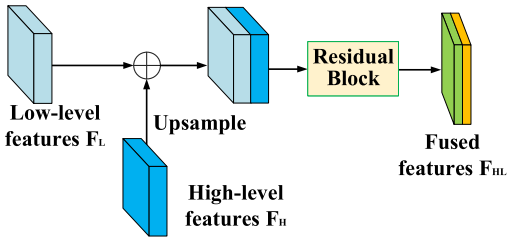

FIGURE 4. Multi-scale features fusion module.

E. LOSS FUNCTION OF RFE-LINKNET

The role of loss function is designed to determine the inconsistency between the predicted value by the model and the true value. Generally, the smaller the loss function, the better the performance of the model. Usually, in the field of image segmentation, Binary Cross Entropy Loss (BCE) and Dice Coefficient Loss (DICE) are often used to construct the loss function. BCE can satisfy the optimization to reduce the error, while DICE can be used to compare the similarity between two samples. Considering the particularity of the road extraction task and the characteristics of the two

functions, a hyperparametric loss function is constructed as follows:

$$L_{\text{Total}} = \lambda_1 L_{BCE} + \lambda_2 L_{DICE} \tag{4}$$

where L_{Total} is the hyperparametric loss, L_{BCE} is the BCE loss, and L_{DICE} is the DICE loss, they are weighted by λ_1 and λ_2, respectively. Usually, the weight terms λ_i are set to be equal or found through expensive grid searches. Here, both λ_1 and λ_2 are set to 1.

III. EXPERIMENTS AND RESULTS

A. DATASETS

To evaluate the performance of RFE-Link, we applied two different datasets in the experiment, i.e., Massachusetts road dataset [36] and DeepGlobe road dataset [37].

1) MASSACHUSETTS ROAD DATASET

The Massachusetts roads dataset covers an area of approximately 2600 square kilometers, encompassing a wide range of urban, suburban, and rural areas. It consists of 1171 aerial images with a size of 1500 × 1500 and a resolution of 1 meter per pixel. The images were split into 1108 for training, 14 for validation, and 49 for testing. In which 342 images were severely flawed or mislabeled (333 for training set, 3 for validation set and 6 for test set). Finally, 827 images were selected as the experimental data after cleaning the dataset, and then the images were cropped into 512 × 512 tiles. The ground truth of images are binary images that contain two classes: roads and non-roads.

2) DEEPGLOBE ROAD DATASET

The DeepGlobe Road dataset contains images collected from three different areas: Thailand, Indonesia and India, which covers a variety of scenes such as cities, villages, wilderness, seashores, and tropical rainforests. It consists of 6,226 images with a spatial size of 1024 ×1024 pixels and resolution of 0.5m per pixel. We further divided the images into 512 × 512 pixels and randomly split the entire dataset into training set, validation set, and test set with a ratio of 8:1:1, respectively.

B. NETWORK CONFIGURATIONS AND TRAINING

In the two experiments, configurations of each network were samely set as follows: the adaptive moment estimation (Adam) optimizer was utilized with an initial learning rate of 0.0001, batch-size was 4, and decreased the learning rate by 0.5 times every 30 epochs. 100 epochs were conducted for all models on the two datasets. Besides, the segmentation loss function adopted the hyperparametric loss function constructed by BCE loss and the DICE loss. Networks used in two experiments were carried out with python 3.7.9 and Pytorch 1.2.0, and were implemented with a single GPU NVIDIA Quadro RTX 6000 24 GB and 128 GB memory. To assess the proposed RFE-LinkNet, taking the architecture of RFE-LinkNet into account, some state-of-the-art (SOTA) methods, including U-Net [15], DeeplabV3+ [18], HRNet [41] and DlinkNet [30], were adopted for comparations on the two datasets. Among them, U-Net belongs to typical encoder-decoder architecture with skip connection, in which, low-level features and high-level features are aggregated progressively to achieve high resolution. The Deeplabv3+ combines dilated convolution with spatial pyramid pooling to aggregate multi-scale features in a parallel manner. HRNet learned semantically strong and spatially precise representations by a high-resolution convolution stream. The DlinkNet embed atrous convolution layers to enhance receptive fields and multi-scale features and reserved the detailed information simultaneously.

C. EVALUATION METRICS

To quantitatively evaluate the performance of the proposed RFE-LinkNet, recall, precision, F1 score and IoU are applied. They can be described as follows:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{8}$$

where, TP is true positives (pixels correctly extracted as road), FP is false positives (pixels misclassified as road), TN is true negatives (correctly labeled as non-road pixels), FN is false negatives (pixels mislabeled as non-road or can be classified as missed road pixels).

D. EXPERIMENTAL RESULTS

1) RESULTS ON MASSACHUSETTS ROAD DATASET

Fig. 5 illustrates the road extraction results of the Massachusetts road dataset produced by the U-Net, DeeplabV3+, HRNet, DlinkNet and RFE-Link, respectively. Most roads can be extracted correctly by the comparison of methods. The proposed RFE-Link achieves the best results, it shows the best completeness and connectivity of extracted roads. The reason may be that the RFE and DU feature optimization modules are introduced into the RFE-Link, which can not only capture rich multi-scale features information and long-range dependencies, but also retain local details. Moreover, the continuity of the road can be guaranteed by the extracted context information. It can be observed from the areas A-E indicated by red circles, U-Net achieved the worst performance compared with the other methods, whose results contained a lot of break roads and omission roads. Although U-Net can aggregate the low-level features and high-level features using the skip connection, while the maximum downsampling rate is only 1/16 of the original input image, leading to an insufficient receptive field for the long span roads. Compared with U-Net, DeeplabV3+ and DlinkNet produce better results due to the embed atrous convolution layers

which enhances receptive fields and multi-scale features, while retains many bread roads for that many noise features are extracted when fusing feature maps from different layers without considering importance of features at different scales. In comparison with U-Net, DeeplabV3+ and DlinkNet, we can find the HRNet produces a more complete roads by learning semantically strong and spatially precise representations by a high-resolution convolution stream without dilated convolution which may lead to the loss of precise spatial information. However, the integrity and connectivity of the road still need to be further improved.

TABLE 1. Quantitative comparison of Precision, Recall, F1 and IoU of the massachusetts road dataset.

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)
U-Net	77.36	75.62	76.48	61.92
DeepLabV3+	79.29	79.67	79.48	65.95
HRNet	78.42	**80.21**	79.31	65.71
DlinkNet	75.89	79.92	77.86	63.74
RFE-Link Net	**80.88**	79.29	**80.07**	**66.77**

The quantitative evaluation results of the Massachusetts road dataset are reported in Table 1. DeeplabV3+ and HRNet yield higher extraction accuracies than that of U-Net and DlinkNet. RFE-Link obtains the best accuracy. Compared to DlinkNet, by introducing the RFE and DU modules, the segmentation accuracy is significantly improved, RFE-Link obtains 80.8%, 79.29%, 80.07%, and 66.77% with respect to Precision, Recall, F1 and IoU index, respectively, the accuracy gains of RFE-Link over DlinkNet are 4.99%, 2.21% and 3.03% with respect to Precision, F1 and IoU index, respectively, and achieved approximately 4.85%, 0.82%, 1.06% and 3.03% IoU improvement compared with the other four models.

2) RESULTS ON DEEPGLOBE ROAD DATASET

Fig. 6 presents the results based on the DeepGlobe road dataset using U-Net, DeeplabV3+, HRNet, DlinkNet and RFE-Link, respectively. With the improvement resolution of the image, the interference factors of complex background are greatly increased. Visually, DeeplabV3+ and HRNet generate more uniform segmentation maps compared to U-Net and DlinkNet, and perform better in road accuracy, while RFE-Link has the best performance. However, there are still a large number of disconnections and leaks in the results of U-Net, DeeplabV3, HRNet and DeeplabV3+. This can also be illustrated in areas A-E. While RFE-Link shows better performance. The main reason may be that using only plain skip connections of U-Net may lead to underutilization of feature maps of different layers.

Though DeeplabV3+ and DlinkNet applied atrous convolution to enhance receptive fields and multi-scale features, HRNet proposes a parallel multipath architecture with high-resolution representation to extract multi-scale features, many noisy features can be extracted without considering the importance of features at different scales when aggregating these multi-scale features. Compared with U-Net, DeeplabV3+, HRNet and DlinkNet, RFE-Link introduces RFE module to enhance the receptive field and multi-scale features for the long span road, and DU feature optimization modules are designed to adaptively refine feature maps extracted by RFE modules according to their contributions. Thus, RFE-Link can handle roads' properties such as narrowness, connectivity, complexity and long span to some extent.

As shown in Table 2, RFE-Link obtains the greatest accuracy with 82.60%, 82.85% and 70.72% with respect to Recall, F1 and IoU index, respectively. Especially, compared with DlinkNet, the Precision, Recall, F1 and IoU and increased by 4.55%, 5.15%, 4.85%, 6.78%, respectively, and obtains approximately 11.65%, 5.67%, 5.43% and 6.78 IoU improvement compared with U-Net, DeeplabV3+, HRNet and DlinkNet, respectively.

TABLE 2. Quantitative comparison of precision, Recall, F1 and IoU of the DeepGlobe road dataset.

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)
U-Net	87.68	64.41	74.27	59.07
DeepLabV3+	**88.26**	71.21	78.82	65.05
HRNet	88.14	71.57	79.00	65.29
DlinkNet	78.54	77.45	78.00	63.94
RFE-Link Net	83.09	**82.60**	**82.85**	**70.72**

3) ABLATION STUDY

In order to evaluate the contributions of different modules contained in RFE-LinkNet, the ablation experiments were performed on the Massachusetts road dataset and the DeepGlobe road dataset, respectively. As shown in Table 3, we take the DlinkNet as the baseline. The baseline was improved with the receptive field enhancement module, which is denoted as baseline + RFE. Samely, Baseline + DU denotes the DU feature optimization modules were introduced into the baseline, and Baseline + RFE + DU denotes both RFE and DU are imbed in the baseline.

TABLE 3. Ablation comparison of OA, Precision, Recall, F1, IoU and mIoU of massachusetts road dataset.

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)
Baseline	75.89	79.92	77.86	63.74
Baseline + RFE	78.16	78.44	78.30	64.25
Baseline + DU	74.98	82.12	78.38	64.45
Baseline + RFE + DU	80.88	79.29	80.07	66.77

Table 3 lists the ablation results on the Massachusetts road dataset, we can find that the baseline achieved a 63.74% in IoU and 77.86% in F1-score. When the RFE module is introduced, the segmentation achieves a 0.51% improvement in IoU and a 0.44% improvement in F1-score, this indicates that more rich semantic information and accurate spatial information can be captured by the RFE module rather than that by original encoder-decoder network with hollow convolution layer in the central area. Embedding the DU module to the network can achieved 0.52% and 0.71%

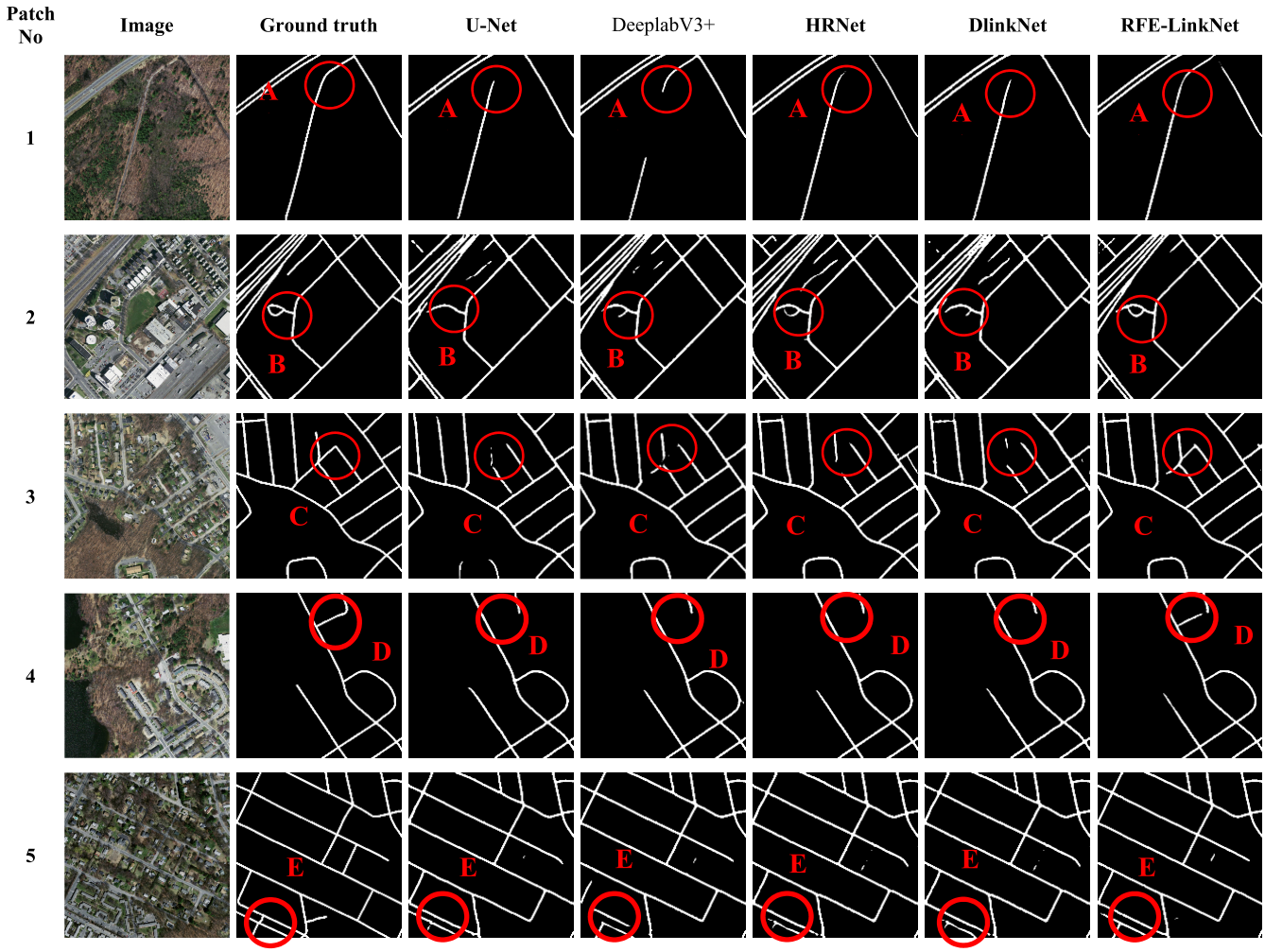

FIGURE 5. Segmented patches from U-Net, DeeplabV3+, HRNet, DlinkNet and RFE-LinkNet based on the Massachusetts road dataset. The white color in the segmented patches represents pixels belonging to roads.

improvements on F1-score and IoU, respectively. This proves that the DU can enhance feature fusion by reducing the sematic and resolution gaps between features learned different level and considering different importance of different level features. At last, baseline + RFE + DU, namely RFE-LinkNet, achieves the best accuracies, and improvements on IoU and F1-score over baseline are 3.03% and 2.21%, respectively. Which indicates MFE module can enlarge the receptive field of the network and capture more multi-scale features simultaneously.

As list in Table 4, after adding each module one by one, compared with the baseline, it has increased by 6.78%, 4.27% and 4.24% on IOU, respectively. The combination of baseline + RFE + DU achieves the highest scores on IoU and F1-score on the DeepGlobe road dataset. The results show the effectiveness of our proposed methods.

TABLE 4. Ablation comparison of OA, Precision, Recall, F1, IoU and mIoU of deepglobe road dataset.

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)
Baseline	78.54	77.45	78.00	63.94
Baseline + RFE	81.17	78.42	79.77	66.35
Baseline + DU	80.93	78.82	79.86	66.48
Baseline + RFE + DU	83.09	82.60	82.85	70.72

4) COMPLEXITY ANALYSIS

To evaluate the tradeoff between the performance and complexity of RFE-LinkNet, we also compared FLOPs, trainable parameter, and IoU score of related methods on the DeepGlobe road dataset.

As shown in Table 5, the U-Net has the lowest complexity but with poor performance. HRNet is the most complicated model due to the most numbers of convolutional layers and the highest numbers of channels among the related models. By introducing RFE modules which contain more number and depth of convolution layers, compared with the basic network, the number of model parameters and FLOPs of "Baseline + RFE" have increased by 3.1 M and 28.97, respectively, but achieved a 2.41%

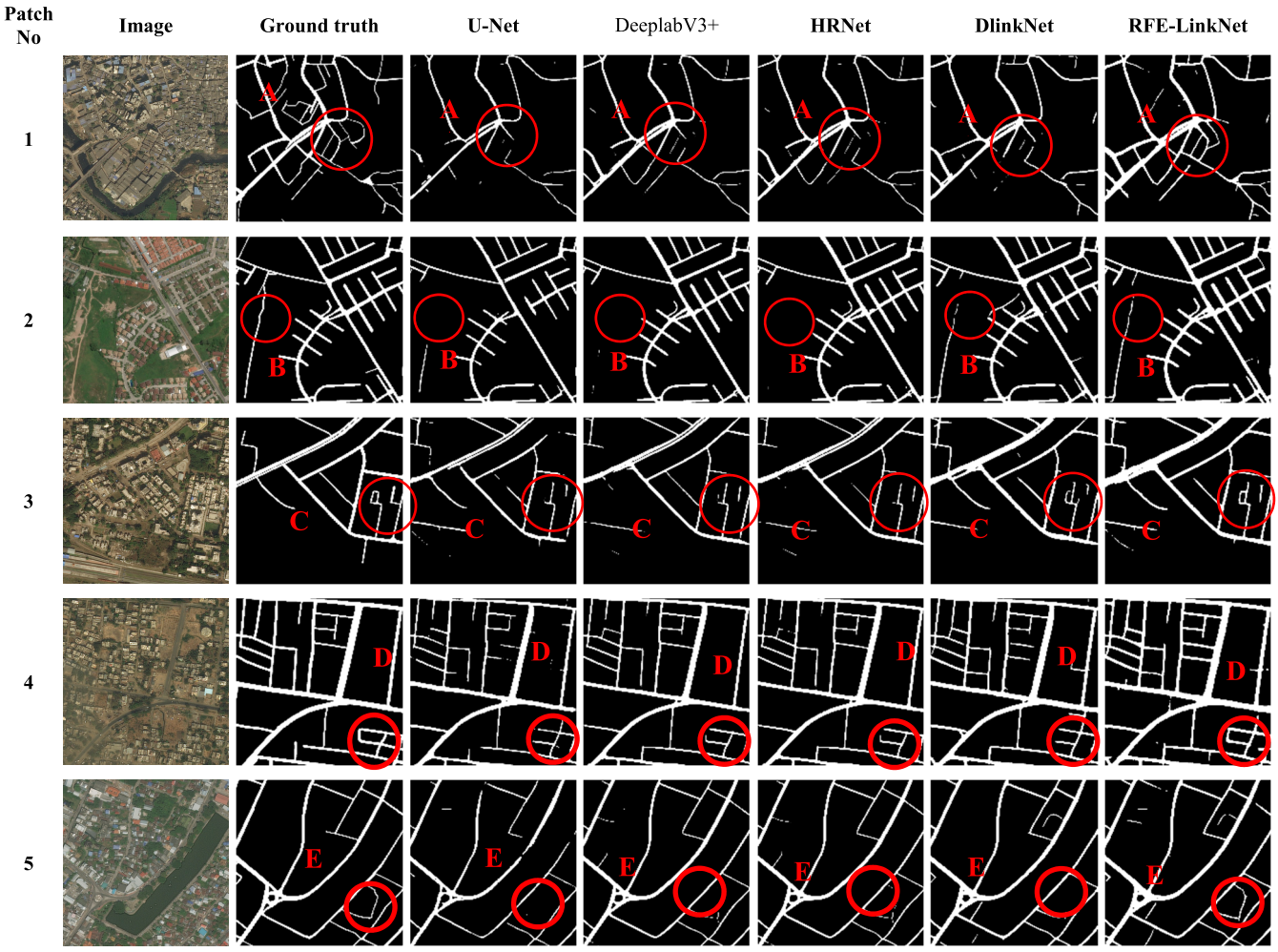

FIGURE 6. Segmented patches from U-Net, DeeplabV3+, HRNet, DlinkNet and RFE-LinkNet based on the DeepGlobe road dataset. The white color in the segmented patches represents pixels belonging to roads.

TABLE 5. Comparison of related methods on FLOGs, Trainable parameters, and IoU on the DeepGlobe road dataset.

Method	Params (M)	FLOPs	IoU（%）
U-Net	9.85	322.54	59.07
DeepLabV3+	17.78	366.50	65.05
HRNet	65.57	370.37	65.29
Baseline (DlinkNet)	31.09	134.35	63.94
Baseline + RFE	34.19	163.33	66.35
Baseline + DU	21.75	124.70	66.48
Baseline + RFE + DU (Ours)	34.29	163.35	70.72

improvement in IoU. "Baseline + DU" can achieve 2.54% improvement in IoU with computation decreases of 9.34 M and 9.65 on FLOPs since the representative features refined by DU module. The RFE-LinkNet has slightly more parameters and FLOPs than the baseline, but it yields an 6.78% improvement on IoU. From above, we can draw the conclusion that our proposed methods maintain higher accuracy and lower complexity compared with other related methods.

IV. CONCLUSION

This paper proposes a novel network based on receptive field enhanced LinkNet, named as RFE-LinkNet, to improve completeness and connectivity of road extraction from HSRI. RFE-LinkNet takes advantage of the prominent ability of feature encoding based on RFE enhanced residual blocks and refined spatial context information to predict the precise roads.

This method can overcome the drawbacks of road extraction interruptions, omissions, and incorrect labels to some extent. This can be attributed to the introduction of the DU feature optimization module, which is used to fuse and refine multi-scale features based on the relative importance in different levels of feature maps, and imbedding the RFE module to obtain richer multi-scale feature information while preserving local spatial details. To evaluate the performance of RFE-LinkNet, experiments on Massachusetts road dataset and DeepGlobe road dataset were conducted, compared with existing state-of-the-art networks according to the metrics of F1-score and IoU, RFE-LinkNet is more accurate by visual and quantitative evaluation. Therefore, RFE-LinkNet

can provide an effective method for road extraction from HSRI.

However, RFE-LinkNet still has issues with error recognition and road connectivity, we will perform more researches on these issues in the future, such as introducing prior knowledges, multi-task learning, etc. Besides, the proposed RFE-LinkNet architecture was originally proposed for road extraction tasks, it could also be useful in other segmentation tasks, and we will investigate this in future research.

ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers whose insightful suggestions have significantly improved this article.

REFERENCES

[1] W. Shi, Z. Miao, and J. Debayle, "An integrated method for urban main-road centerline extraction from optical remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3359–3372, Jun. 2014.

[2] 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.

[3] M. Yang, Y. Yuan, and G. Liu, "SDUNet: Road extraction via spatial enhanced and densely connected UNet," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108549.

[4] G. Máttyus, W. Luo, and R. Urtasun, "DeepRoadMapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3458–3466.

[5] 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.

[6] 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.

[7] S. Hinz and A. Baumgartner, "Automatic extraction of urban road networks from multi-view aerial imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, nos. 1–2, pp. 83–98, Jun. 2003.

[8] X. Huang and L. Zhang, "Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines," *Int. J. Remote Sens.*, vol. 30, no. 8, pp. 1977–1987, Apr. 2009.

[9] Z. Miao, W. Shi, H. Zhang, and X. Wang, "Road centerline extraction from high-resolution imagery based on shape features and multivariate adaptive regression splines," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 583–587, May 2013.

[10] C. Unsalan and B. Sirmacek, "Road network detection using probabilistic and graph theoretical methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4441–4453, Nov. 2012.

[11] J. Hu, A. Razdan, J. C. Femiani, M. Cui, and P. Wonka, "Road network extraction and intersection detection from aerial images by tracking road footprints," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4144–4157, Dec. 2007.

[12] S. K. Choy, S. Y. Lam, K. W. Yu, W. Y. Lee, and K. T. Leung, "Fuzzy model-based clustering and its application in image segmentation," *Pattern Recognit.*, vol. 68, pp. 141–157, Aug. 2017.

[13] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 911–923, Feb. 2019.

[14] 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.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Munich, Germany: Springer, 2015, pp. 234–241.

[16] 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.

[17] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[19] 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.

[20] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. ECCV*, 2010, pp. 210–223.

[21] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3322–3337, Jun. 2017.

[22] H. Kong, J.-Y. Audibert, and J. Ponce, "General road detection from a single image," *IEEE Trans. Image Process.*, vol. 19, no. 8, pp. 2211–2220, Aug. 2010.

[23] X. Lu, Y. Zhong, Z. Zheng, Y. Liu, J. Zhao, A. Ma, and J. Yang, "Multi-scale and multi-task deep learning framework for automatic road extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9362–9377, Nov. 2019.

[24] Y. Li, B. Peng, L. He, K. Fan, and L. Tong, "Road segmentation of unmanned aerial vehicle remote sensing images using adversarial network with multiscale context aggregation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2279–2287, Jul. 2019.

[25] D. Chen, Y. Zhong, Z. Zheng, A. Ma, and X. Lu, "Urban road mapping based on an end-to-end road vectorization mapping network framework," *ISPRS J. Photogramm. Remote Sens.*, vol. 178, pp. 345–365, Aug. 2021.

[26] 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.

[27] Y.-Q. Tan, S.-H. Gao, X.-Y. Li, M.-M. Cheng, and B. Ren, "VecRoad: Point-based iterative graph exploration for road graphs extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8907–8915.

[28] 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.

[29] 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. Photogramm. Remote Sens.*, vol. 168, pp. 288–306, Oct. 2020.

[30] 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 (CVPRW)*, Jun. 2018, pp. 182–186.

[31] 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, pp. 1–5, 2022.

[32] Q. Zhu, Y. Zhang, L. Wang, Y. Zhong, Q. Guan, X. Lu, L. Zhang, and D. Li, "A global context-aware and batch-independent network for road extraction from VHR satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 353–365, May 2021.

[33] 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, p. 571, Dec. 2019.

[34] L. Gao, J. Wang, Q. Wang, W. Shi, J. Zheng, H. Gan, Z. Lv, and H. Qiao, "Road extraction using a dual attention dilated-LinkNet based on satellite images and floating vehicle trajectory data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10428–10438, Sep. 2021.

[35] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603018.

[36] V. Mnih, "Machine learning for aerial image labeling," 2013.

[37] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.

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

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[40] H. Zhang, X. Zheng, N. Zheng, and W. Shi, "A multiscale and multipath network with boundary enhancement for building footprint extraction from remotely sensed imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8856–8869, Feb. 2022.
[41] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

HUA ZHAO received the M.S. degree in cartography and geographical information engineering from the China University of Mining and Technology, Xuzhou, China, in 2004. He is currently a Researcher with the School of Environment and Spatial Informatics, China University of Mining and Technology. His research interests include computer vision, deep learning, and digital image processing.

HUA ZHANG received the Doctoral degree in cartography and geographical information engineering from the China University of Mining and Technology, Xuzhou, China, in 2012. He is currently a Professor in GIS and remote sensing with the School of Environment and Spatial Informatics, China University of Mining and Technology. He has authored or coauthored more than 50 peer-reviewed articles in international journals, such as IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *ISPRS Journal of Photogrammetry and Remote Sensing*. His current research interests include multi/hyperspectral and high-resolution remotely sensed images processing, uncertainty in classification, pattern recognition, and remote sensing applications.

XIANGCHENG ZHENG received the B.S. degree in electronic engineering from the Shandong University of Science and Technology, Qingdao, Shandong, China, in 2020. He is currently pursuing the master's degree with the School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou, China. His research interests include deep learning, object detection, and image understanding.