AD-RoadNet: An Auxiliary-Decoding Road Extraction Network Improving Connectivity While Preserving Multiscale Road Details

Ziqing Luo, Kailei Zhou, Yumin Tan, Xiaolu Wang, Rui Zhu, and Liqiang Zhang

***Abstract*—Obtaining Road information from high-resolution remote sensing images is gaining attention in intelligent transportation systems. Existing road extraction methods tend to improve road connectivity with graph convolution or global attention, however, ignore the damage of introduced excessive effective receptive field (ERF) to multiscale road details. In this study, we propose an auxiliary-decoding road extraction network named AD-RoadNet, which decouples multiscale road representation and connectivity improvement based on two modules; the hybrid receptive field module (HRFM) and the topological feature representation module (TFRM). The HRFM is introduced in the encoder to emphasize target road features through adaptively matching the receptive field (RF) size for various scale roads, thus, beneficial for multiscale road representation. The TFRM is introduced in an auxiliary decoder to represent topological features with the position information encoded in the shared encoder and then helps the main decoder reason occluded roads, thus improving connectivity. Between the encoder and main decoder. The proposed model has a similar parameter scale as HRNetV2 and outperforms the state-of-the-art ResUnet, D-LinkNet, and HRNetV2 by 3.34%, 2.03%, and 1.53% in the mean intersection of union on DeepGlobe road dataset. Ablation analysis, inference size matter, and the robustness for unseen occlusion scenarios, low-quality labels, and various quality inference images are further presented to evaluate the proposed AD-RoadNet.**

***Index Terms*—Hybrid receptive field (RF), multiscale road extraction, road connectivity, semantic segmentation, topological feature.**

I. Introduction

OBTAINING road information is vital in many applications, such as autonomous navigation [1], autonomous driving [2], and intelligent transportation system [3]. High-resolution remote sensing images (HRSI) are widely used in producing rich road information and provide probability [4]. However, the following difficulties are still an immense challenge to extract roads from HRSI: First, there are diverse road scales and types, including urban trunk roads, urban overpasses, and rural roads. Not just the wide main roads but also trails and narrow rural roads are equally essential for road networks. Second, different lanes or adjacent roads are commonly confused due to the lack of clarity in lane marking and diverse median strips in HRSI. Third, the road connectivity in HRSI is easily affected by building shadows, green belts, trees, vehicles, etc. [5]

With the continuous development of deep learning technology, end-to-end semantic segmentation models have been the mainstream to extract roads from HRSI and have made remarkable achievements [6], [7]. The first attempt to use an encoder–decoder structure for semantic segmentation was with the fully convolutional network [8]. Then, a series of improvements for common semantic segmentation were presented from the *classical models* (Unet [9] and SegNet [10]) to the *modern models* (DeepLab families [11], [12], [13], [14] and HRNetV2 [15]), then to the current *transformer models* (SegFormer [16], UNetFormer [17]) designed for spatial information recovery, high-resolution semantic segmentation, and high-shape bias semantic segmentation, respectively. In addition, some domain adaptation methods are used to improve the robustness of road extraction models, such as [18].

To improve road connectivity based on these methods, road extraction models frequently design elaborate modules with graph convolution or global attention to encode more global contextual features, such as spatial and interaction space graph reasoning [19], global context-aware (GCA) block [20], spatial intensifier (DULR module) [21], and separable graph convolutional network (SGCN) [4]. However, these methods excessively enlarge the effective receptive field (ERF) and damage the multiroad details while improving connectivity. Taking a two-lane road as an example, a large ERF will blur the lane boundary and result in the confusion of lanes, as shown in Fig. 1. Although there are some few quality works taking their eyes to multiscale road details, such as DDU-Net [22] and Richer U-net [23], giving overall consideration on road connectivity and multiscale details remain a big challenge.

To overcome this problem, an auxiliary-decoding road extraction model termed AD-RoadNet is developed in this article to improve road connectivity while preserving the multiscale details. The essential ideas behind AD-RoadNet are as follows:

Manuscript received 3 March 2023; revised 16 May 2023 and 11 June 2023; accepted 21 June 2023. Date of publication 26 June 2023; date of current version 12 September 2023. This work was supported by the Sanxia Follow-up Project ‘The study on carbon neutrality Benefits and Contribution Accounting of Three Gorges Reservoir’ under Grant 10212622202027001908 1 and in part by the National Natural Science Foundation of China under Grant 41925006. *(Corresponding author: Xiaolu Wang.)*

Ziqing Luo, Kailei Zhou, and Xiaolu Wang are with the School of Transportation Science and Engineering, Beihang University, Beijing 100091, China (e-mail: zqluo@buaa.edu.cn; zhoukailei@buaa.edu.cn; zy2113207@buaa.edu.cn).

Yumin Tan and Rui Zhu are with the Civil Engineering, Beihang University, Beijing 100091, China (e-mail: tanym@buaa.edu.cn; buaa_zhurui@buaa.edu.cn).

Liqiang Zhang is with the School of Geography, Beijing Normal University, Beijing 100875, China (e-mail: zhanglq@bnu.edu.cn).

Code is available online at: https://github.com/zqluo-buaa/ADRoadnet

Digital Object Identifier 10.1109/JSTARS.2023.3289583

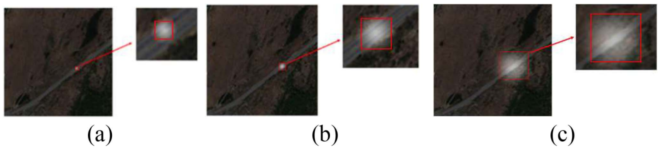

Fig. 1. Base map is a picture of a two-lane road. The brighter areas indicate the RFs. (a), (b), and (c) are feature maps with small, moderate, and large ERFs, respectively. For the extraction of the road, the semantic feature of the road is ambiguous on large-scale ERF, while is insufficient on small scale. (a) Small ERF. (b) Suitable ERF. (c) Large ERF.

1) Adaptively matching the RF size for various scale roads according to their surrounding features to optimize the extraction for multiscale and multilane roads.
2) Modeling the spatial location relation among pixels with the position information encoded through extensive zero-padding.
3) Decoupling connectivity improvement and multiscale road representation by adding an auxiliary decoder to help the main decoder reason the occluded multiscale roads.

Compared to other research studies on road extraction that focus on elaborate modules designed to learn long-range contextual information, our research contributions could be summarized as follows.

1) A novel auxiliary-decoding road extraction network named AD-RoadNet is proposed to give overall consideration to road connectivity and multiscale details improving the overall performance of road extractions, especially for intersections and multilane roads. In the proposed AD-RoadNet, an auxiliary decoder is embedded before the main decoder to help connect fragmented segments.
2) HRFM is introduced in the encoder to adjust the RF size for multiscale roads, which is beneficial for the detection of lanes and rural roads. In the HRFM, feature maps of branches with various RF sizes are obtained by stacking basic convolutions, and the weights for these maps are computed with respect to their surrounding features to match the suitable branch.
3) TFRM is introduced in the auxiliary decoder to represent topological information to help reason occluded roads. This module utilizes the position information encoded in the shared encoder to model spatial location correlation among pixels and does not disturb the multiscale semantic features fed from the encoder to the main decoder.

II. Related Work

A. Road Network Extraction

Numerous techniques have been developed in other works of the literature to extract road networks from remote sensing data. Traditional methods improve road connectivity with probabilistic models by incorporating contextual priors, such as spectral features [22], [23], road geometry [26], and marked point processes [27]. Song and Civco [24] used statistical spectral and geometric information as classification criteria to segment road pixels. Mena and Malpica [28] proposed a technique called texture progressive analysis to extract road networks in rural and semiurban areas. These methods utilized handcrafted features and required complex optimization techniques [29].

In recent deep learning-based techniques, road extraction is formulated as a segmentation problem [19], [30], [31], [32] using convolutional encoder–decoder structured models. Among them, Mnih and Hinton [33] made the first attempt to apply a convolutional neural network (CNN) in classifying roads, operating on the patches. Máttyus et al. [30] proposed an encoder–decoder structure model and used shortest path algorithms to improve the connectivity in the postprocessing step. Unet [9] and LinkNet [34] are two well-known encoder–decoder structures. There are many improvements based on these models. Zhang et al. [35] proposed the ResUnet that combines the strengths of residual learning and U-Net. Chen et al. [36] proposed a reconstruction bias U-Net, which increased the decoding branches to obtain multilevel semantic information in the up-sampling. Wang et al. [34] optimized the D-LinkNet with nonlocal blocks and gained better performance, with less computational cost as well as faster convergence. Zao and Shi [23] enhanced the U-Net with an enhanced detail recovery structure and edge-focused loss function to obtain complete and accurate results. The abovementioned methods perform well in segmentation, however, fail to detect roads obscured by trees or other objects and produce a lot of fragmented segments.

To improve road connectivity, one main idea is to enlarge the receptive field (RF) using dilated convolution [11] and detect occluded roads with extra contextual information. Zhou et al. [37] improved LinkNet with dilated convolution in DeepGlobe road extraction subchallenge [38]. Tao et al. [39] designed a spatial information inference structure to collect contextual information without introducing invalid context. Another great idea is to learn road orientation additionally, which was first proposed by Batra et al. [40]. Yi et al. [5] proposed the Efficient UNet multitask joint learning model, incorporating an orientation learning decoding branch to solve the discontinuity problem in road extraction. This idea shows commendable effectivity but requires extra effort for road orientation ground-truth. Currently, due to the great ability for extracting dependencies over distant regions of graph structure and attention mechanism, a lot of works [4], [20], [21], [32] utilize them in the encoder to improve road connectivity. Bandara et al. [19] introduced graph convolution modeling dependencies between different spatial regions and other contextual information to represent road connectivity. Zhu et al. [20] designed the GCA block to the encoder–decoder structure to effectively integrate global context features. Zhou et al. [4] proposed a split depth-wise (DW) SGCN to capture global contextual road information in channel and spatial features and extracted covered roads. These methods show their effectiveness in improving road connectivity. However, the blindly introduced excessive ERF destroys the multiscale road features and is not beneficial for multiscale and multilane roads which require accurate ERF.

B. Receptive Field in ConvNets

RF is a term originally coined by [41] to describe an area of the body surface where a stimulus can produce a reflex. For existing ConvNets, an RF can be described as a region of input affecting the value of an output unit [42]. The RF size affects the scope of extracted information as well as the expression of semantic information. Generally, it can be calculated layer by layer as

$$R_n = R_{n-1} + (k_n - 1)\prod_{i=1}^{n-1} S_i \tag{1}$$

where R_n and R_{n-1} are the RF size of the nth and $(n-1)$th layers, respectively; k_n is the kernel size; S_i is the stride size of the ith layer.

Two common operations to increase RF size are stacking more convolutional layers and subsampling. For large-scale image classification, a larger RF size means more effectiveness, which has been proven by a huge improvement from AlexNet [43], VGG [44] to ResNet [45]. For semantic segmentation (also called classification on pixel level), one extra requirement is preserving spatial information while enlarging the RF size. Atrous convolution [11], multilevel feature structured models, such as Unet [9], high-resolution models, such as HRNetV2 [15], and transformer models, such as SegFormer [16], were developed to deal with this problem. Another problem is the RF imbalance [46], [47], which means multiscale and multilevel architectures widely used [21], [23] provided unsuitable RF for some objects, negatively impacting the segmentation of objects of varying sizes. For this problem, Liu et al. [48] designed the scale-layer attention module and scale-feature attention module to weigh useful information after Atrous spatial pyramid pooling (ASPP) and skip connection, respectively. Li et al. [49] proposed an adaptive multiscale deep fusion residual network using the adaptive feature fusion module to emphasize useful information and suppress useless information during the multilevel feature fusion (MLFF). Wang et al. [50] designed the adaptive multiscale feature extraction module setting the RF according to feature map size to avoid introducing invalid information. These methods provide solutions from various perspectives but do not consider the impact of surrounding features on the required RF of target roads.

III. Methodology

To improve road connectivity while preserving great multiscale details, we propose an AD-RoadNet, which will be thoroughly introduced in this section. Specifically, we first illustrate the overview of AD-RoadNet for road extraction from HRSI. Then, the two designed modules, HRFM and TFRM, are introduced sequentially.

A. Pipelines of Proposed Model

The proposed AD-RoadNet comprises of the following four parts; an encoder, an auxiliary decoder, and the main decoder, our pipelines are shown in Fig. 2.

The encoder was designed for extracting multiscale semantic features with hybrid RFs. Specifically, each intermediate feature map is followed with a basic residual block [as shown in Fig. 3(a)] to provide a basic RF size. Where output_stride [10]> = 4, an HRFM is extra arranged in front of the residual block to guarantee the flexibility of RF size. This combination goes deeper with 1:2:4 to extract the superior road information. The ratio is determined experimentally, and the output_stride of the final feature map is 16, similar to the common semantic segmentation networks [50], [51]. In addition, we experimentally observe that two subsampling operations, max-pooling, and convolution with a stride of 2, have their own advantages in filtering invalid textures and refining road edges. We shall be combining the two operations using a parallel method to subsample the feature map, as shown in Fig. 3(b).

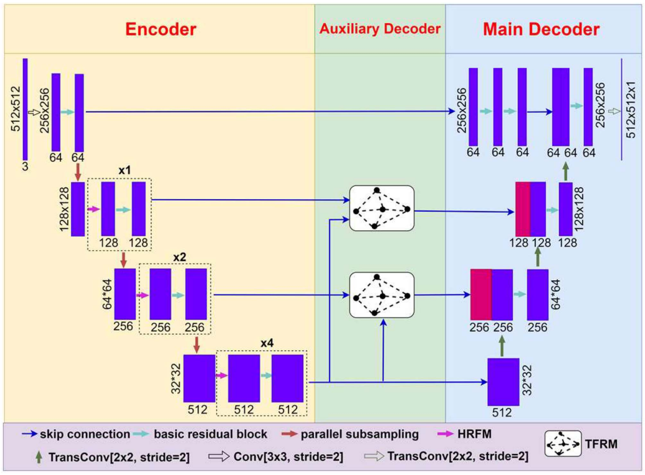

Fig. 2. Pipeline of the proposed networks. The HRFM is introduced in the encoder to adjust the RF size for target roads according to their surrounding features, while the TFRM is used in the auxiliary decoder to represent topological feature and the high-resolution feature is preserved.

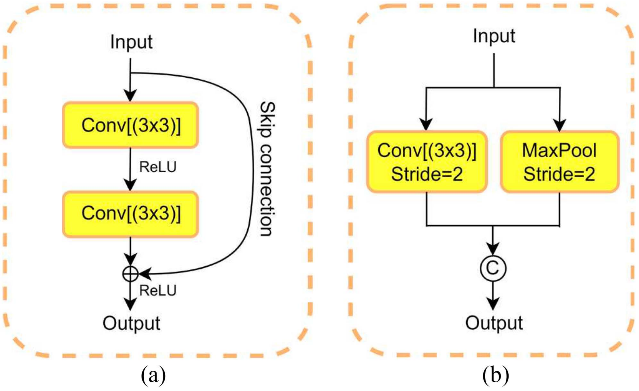

Fig. 3. Diagram of basic residual block and parallel down-sampling. (a) Basic residual block. (b) Parallel down-sampling.

The auxiliary decoder utilizes TFRM to represent road topological information, using the deep and various levels feature maps as input. This part allows extracting multilevel topological features without disturbing the multiscale semantic features fed to the main decoder.

We repeated residual blocks four times to replace the skip connection between high-resolution feature maps. With this connection, we optimize the extracted road details without the common MLFF to reduce the introduced invalid context information [48].

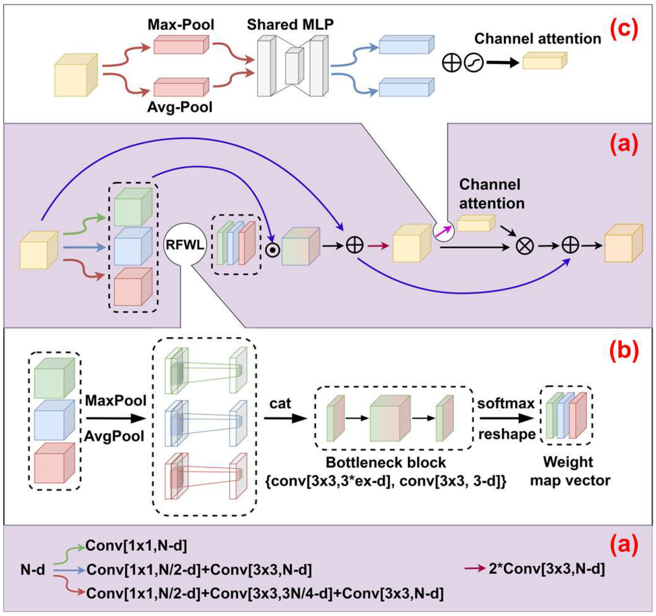

Fig. 4. (a) Diagram of HRFM. (b) Diagram of RFWL. (c) Diagram of channel attention. The lower figure (a) annotation of the upper figure (a). Conv[3 × 3, $N-$d] denotes a convolution operation with the filter size of 3 × 3 and the output channel of d. Similarly, the Conv[1 × 1, $N-d$], Conv[1 × 1, $N/2-d$], Conv[1 × 1, 3$N/4-d$] change the filter size and output channel to corresponding number.

The main decoder receives the following three features; the multiscale semantic feature, the multilevel topological feature, and the high-resolution feature, which are received from the encoder, auxiliary decoder, and high-resolution feature to corporately extract satisfactory results.

In terms of loss functions, we use a combination of binary cross entropy loss and dice loss, as shown in (2). Binary cross entropy is a common loss function for semantic segmentation, while the dice coefficient is widely used to highlight the foreground class (here is the road) [52]

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N} y^{(i)} \log \hat{y}^{i} + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{i}\right) + \left(1 - \frac{2\left|X \cap Y\right| + \text{smooth}}{\left|X\right| + \left|Y\right| + \text{smooth}}\right) \tag{2}$$

where N is the number of pixels; $y^{(i)}$ is the probability of predicting the pixel as road; $\hat{y}^{i}$ is the ground truth; X denotes a set of pixels predicted as roads; Y denotes a set of roads pixels; smooth is used to smooth the loss curve, here we set it as 1.

B. Hybrid RF Module

The HRFM matches the target roads' suitable RF size according to their surrounding features for extracting multiscale details. Compared to other multiscale techniques, such as spatial pyramid pooling (SPP) [53] and ASPP [13], HRFM produces no feature maps at an unsuitable scale, and each pixel in a feature map owns its customized RF. Fig. 4(a) illustrates its diagram. Given an intermediate feature map $x \in R^{C\times H\times W}$ as input, the HRFM feeds it to three branches simultaneously, which provide different RF sizes by varying the number of convolution stacks. Based on the surrounding features of each pixel in three output feature maps $F_i \in R^{C\times H\times W}$ (i = 1,2,3), a receptive field weighting learner (RFWL) is utilized to generate a weight vector $\mathbf{W} \in R^{3\times 1\times H\times W}$, which represents the weights of three RF sizes for each pixel. Then, we conduct dot product with the two vectors $\mathbf{F} \in R^{3\times C\times H\times W}$ and $\mathbf{W} \in R^{3\times 1\times H\times W}$ to mix the extracted features from three branches while using residual connection to avoid gradient vanishing [45]. Moreover, since each channel of feature maps could be considered as a feature detector [54], to avoid the mixed feature detector missing its focus, we then use a basic residual block with channel attention [55] to capture meaningful information. In short, the overall process of HRFM can be summarized as

$$\text{Vector } \mathbf{F} = (F_1(x), F_2(x), F_3(x)) \tag{3}$$

$$\text{Vector } \mathbf{W} = \text{RFWL}(\mathbf{F}) \tag{4}$$

$$F_O = f_C\ (\boldsymbol{F} \cdot \boldsymbol{W} + x) \tag{5}$$

where x is the input feature map; $\mathbf{F}$ consists of the output feature maps $F_i \in R^{C\times H\times W}$ (i = 1,2,3) from three branches; $\mathbf{W}$ is the output from RFWL using $\mathbf{F}$ as input; · denotes dot product; f_c denotes the basic residual block with channel attention. The following describes the details of RFWL and channel attention.

1) RFWL: Based on the distribution of each pixel in different branches of the surrounding features, we generate the weight maps of three RF sizes. Specifically, for each input $F_i \in R^{C\times H\times W}$ (i = 1,2,3), since applying pooling operations along the channel axis is proved to be effective in highlighting information [44], [56], we first aggregate feature information by using average-pooling and max-pooling operations along the channel axis and then utilize a 7 × 7 convolution layer to represent the surrounding feature of each pixel, generating an efficient feature descriptor $D_i \in R^{1\times H\times W}$ (i = 1,2,3). To learn the most suitable RF size for each pixel, three feature descriptors are then concatenated and subsequently feed to a bottleneck block containing two basic convolutions, producing the weight map vector $\mathbf{W} \in \mathrm{R}^{3*1*H*W}$. The computation process could be summarized as follows, and Fig. 4(b) shows the whole abovementioned process

$$D_i = f_1^{7\times7}([\text{AvgPool}(F_i), \text{MaxPool}(F_i)]) \tag{6}$$

$$\text{Vector } \mathbf{W} = \text{Reshape}\left(\sigma\left(f_3^{3\times3}\left(f_{3*ex}^{3\times3}(D_1, D_2, D_3)\right)\right)\right) \tag{7}$$

where σ denotes the softmax activation function; *ex* is an expansion ratio, here we set it as 16; $f_1{}^{7\times7}$ represents a convolution operation with the filter size of 7 × 7 and the output channel of 1; Similarly, $f_3{}^{3\times3}$ represents a convolution operation with the filter size of 3 × 3 and the output channel of 3; $f_{3\times ex}{}^{3\times3}$ represents a convolution operation with the filter size of 3 × 3 and the output channel of 3 × *ex*.

2) Channel Attention: The channel attention is proposed originally by [55] to capture meaningful information in the channel dimension. The processing process of channel attention is illustrated in Fig. 4(c). Specifically, after two basic convolutions, we aggregate spatial information of a feature map by using both average-pooling and max-pooling operations, generating two spatial information descriptors. Both descriptors are then fed to a shared multilayer perceptron with one hidden layer to produce the attention map [55]. The computation process could be summarized as follows:

$$F_m = \sigma\left(f_C^{3\times3}(f_C^{3\times3}(F_{\text{in}}))\right) \tag{8}$$

$$F_{\text{out}} = F_{\text{in}} \otimes (\sigma'(\text{MLP}(\text{AvgPool}(F_m)) + \text{MLP}(\text{MaxPool}(F_m)))) \tag{9}$$

where F_{in} is the input feature map; $f_C{}^{3\times3}$ represents a convolution operation with the filter size of 3 × 3 and the output channel of C; σ denotes the ReLU activation function; σ‘ denotes the sigmoid function; ⊗denotes element-wise multiplication; F_{out} is the output feature map.

C. Topological Feature Representation Module

Based on the shared encoder, the TFRM models the spatial location correlation among pixels to represent road topological information and help reason the occluded roads.

This module could be explained as follows: suppose $F_1 \in R^{C\times H\times W}$ is the final output from the encoder, and $F_e \in R^{C'\times H'\times W'}$ is the final feature map in a certain level. Given the extensive use of zero-padding, F_1 and F_e are proved to have encoded the position information of pixels [57]. $x_i \in R^{1\times C'}$ is a 1-D vector representing pixel *i* from F_e, which could be regarded as a feature descriptor [54]. We use (10) to model the global spatial location correlation (GSLC) on pixel *i*

$$x_i^{\text{new}} = f(\theta(x_i), \gamma(F_1)) \tag{10}$$

where θ and γ denote projection functions embedding x_i and F_1 to topological space; *f* is a function for relationship calculation; $x_i{}^{\text{new}}$ is the generated feature descriptor, which represents the GSLC on pixel *i*.

To convert the abovementioned equation into a computable neural network module, the functions θ, γ, and *f* need to be instantiated. Naturally, we set θ and γ as simple point-wise convolutions since linear transformations are enough. As for function *f*, the dot product is a common operation to represent the relation of vectors. Then, (10) can be written as

$$x_i^{\text{new}} = \text{Conv}_{C'}^{1\times1}(x_i) \bullet \text{Conv}_{C'}^{1\times1}(F_1) \tag{11}$$

where $\text{Conv}_{C'}^{1\times1}(x_i)$ represents a convolution operation with the filter size of 1 × 1 and the output channel of *C'*; • denotes dot product; $x_i{}^{\text{new}}$ is the generated feature descriptor, which represents the GSLC on pixel *i*.

For all pixels in F_e, we repeat the process to model their GSLCs. In practice, we compute all pixels simultaneously by matrix multiplication and finally generate a new feature map $F_S \in R^{(H\times W)\times(H'\times W')}$.

After modeling GSLC on all pixels, we utilize a bottleneck block to represent the topological relationship among pixels. Specifically, our bottleneck block first uses a 1 × 1 convolution to reduce the channels of new feature maps to *C'* as same as F_e, then two basic convolutions are used to extract topological features on each pixel. Moreover, to avoid gradient vanishing, here residual connection [45] is used. Fig. 5 shows the whole abovementioned process in TFRM.

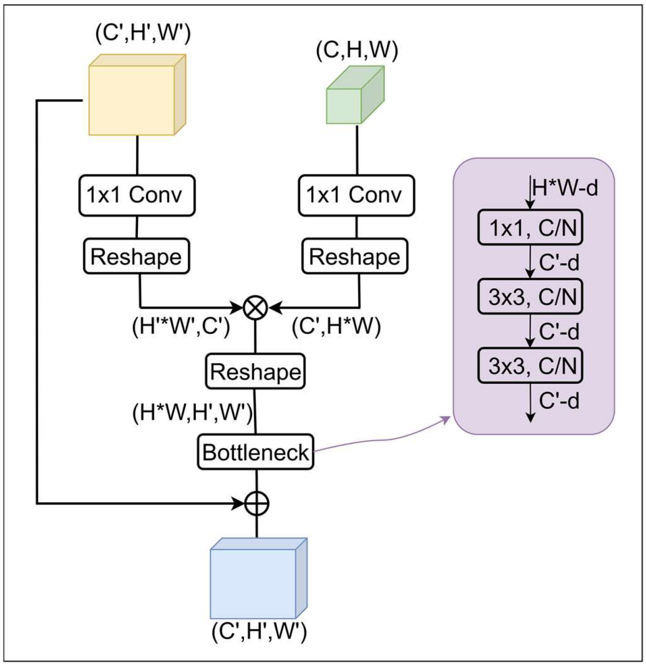

Fig. 5. Diagram of TFRM. Matrix multiplication is introduced to model the GSC on all pixels.

IV. Experiments

The proposed model is evaluated on two datasets: the Massachusetts road dataset [58] and the DeepGlobe road dataset [38]. In this part, the two datasets are first introduced, and then implementation details and evaluation metrics are given. Finally, a performance comparison between the proposed network with some state-of-the-art networks (SegNet [10], Unet [9], HRNetV2 [15], D-LinkNet [37], Residual Unet [35], DDU-Net [22], GAMSNet [59]) is made. Note that although the transformer-based models like UNetFormer [17] show excited performance, the high requirement for the memory of the graphic processing unit (GPU) limits their applications in many situations (In fact, there are some comparisons of calculated costs between the CNN-based model and the transformer based model, such as in [60], [61], and [62]), hence, the transformer families are not our competitive objects. We also make a brief comparison of model efficiency in Section V-A.

A. Dataset

1) The Massachusetts Road Dataset: The dataset consists of training, validation, and test sets with 1108, 14, and 49 images, respectively. There are a wide variety of road features in rural, suburban, and urban features in the dataset. The spatial resolution of these RGB images is 1 m. The annotations are road centerlines obtained from the OpenStreetMap, and all centerlines are converted to raster with a line thickness of 7 pixels [51]. The original image size is 1500 × 1500 pixels. Before feeding them to the segmentation model, images are cropped to a 512 × 512 pixel size with an overlap of 18. Moreover, we filter the images with heavily abnormal occlusion, since it could seriously disturb the performance of the network. After the abovementioned operations, the Massachusetts road dataset now contains 6856, 126, and 441 images of size 512 × 512 pixels, corresponding to the training, validation, and test set respectively.

2) The DeepGlobe Road Dataset: The DeepGlobe dataset consists of 6226 satellite images with a paired mask for road labels. These images have a size of 1024 × 1024 pixels and a spatial resolution of 50 cm/pixel [39]. Like in the Massachusetts road dataset, we crop these images to 512 × 512 pixels without overlaps before feeding them to the test network. Since the background (nonroad) pixels are much more than the road pixels in the satellite image, as the same as Tao et al. [39], we remove some images with an extremely small foreground ratio to alleviate the problem of class imbalance during optimization. After these preprocessing operations, a total of 11 350 images are obtained and divided into training, validation, and test sets with 9500, 350, and 1500 images, respectively.

B. Training Details

1) Avoid Overfitting: Given the relatively small size of the processed training data, we utilize several techniques to avoid overfitting, including online data augmentation, batch normalization [63], $L2$ regularization, and early stopping (to evaluate the model with validating dataset at every 200 iterations). Concretely, random flip, random rotate (by 90°), and random crop and resize are used as data augmentation.

2) Configuration of Hyperparameters: All models are trained with the same parameter settings and in the same environment. Specifically, we conduct all experiments with the Pytorch [64] tool and train the models using the AdamW [65] optimizer with one RTX3060 (memory 12 GB) that allows a batch size of 4 images. The weight decaying is set to 5e−4. We use a cosine annealing learning rate scheduler [66] with an initial learning rate of 0.001, while the warm-up and restart strategies are applied to avoid premature convergence. Concretely, 126 epochs are trained, in which the first epoch is used for warming up, and after 50 epochs with an initial learning rate of 0.001, the final learning rate is then reset to 0.0005 for another 75 epochs.

C. Evaluation Metrics

To quantitatively evaluate the performance of the proposed network architecture, five common and widely accepted metrics are utilized here, including Precision, Recall, F1 score[67], [68], intersection of union (IoU) [20], and mean IoU (mIoU) [39], [69]. Before introducing the definitions of these metrics, it is necessary to define the following four initials; TP, FP, TN, and FN, where TP represents the number of correctly classified foreground pixels, FP, TN, and FN represent the number of false positives, true negatives, and false negatives, respectively [48].

With these initials, precision and recall can be determined as shown in (12) and (13), respectively

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{13}$$

F1 score is the harmonic mean of precision (P) and recall (R), and it can be calculated by the following equation:

$$\text{F1} = 2 \times \frac{P \times R}{P + R}. \tag{14}$$

The IoU evaluates the ratio of the intersection value (TP) and the union value [the sum of FP, FN, and TP, as shown in (15)]. The mIoU is the average ratio of the correctly classified pixels in a class to the union of predicted pixels of this class and ground truth, and it can be calculated by (16)

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{15}$$

$$\text{mIoU} = \frac{1}{2} \times \left(\frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} + \frac{\text{FN}}{\text{TF} + \text{FN} + \text{FP}} \right). \tag{16}$$

D. Results

In this section, performance comparisons between the proposed network AD-RoadNet and some state-of-the-art (SegNet [10], Unet [9], HRNetV2 [15], D-LinkNet [37], Residual Unet [35], DDU-Net [22], and GAMSNet [59]) networks are conducted quantitatively and qualitatively on the abovementioned datasets.

Table I lists the accuracy assessment results on Massachusetts road dataset. All six methods present a good performance in road extraction. The SegNet [10] presents the lowest accuracy among all the six networks, while the proposed AD-RoadNet performs best with an IoU 1.02%–3.13% higher than the other methods. In addition, the rare improvement (from 77–78 to 79.44) on recall indicates the effectiveness of our HRFM and TFRM.

Fig. 6 displays the extracted road features on the Massachusetts road dataset. Overall, all methods can detect road features with relatively high accuracy, but the results differ when HRSI contains complex road structures and intersections. AD-RoadNet presents superiority over the other methods used in the extraction of multiscale roads and obscured roads, which require the ability to obtain road feature details and connection information. To further illustrate the advantages of AD-RoadNet, five typical scenes, which include trails close to main roads, urban overpasses, adjacent roads, a two-lane road covered by trees, and roads covered by shadows and vehicles, are selected as shown in Fig. 6.

To further verify the advantages of the proposed network AD-RoadNet, another comparison with the same five networks, including SegNet, Unet, Residual Unet, D-LinkNet, and HRNetV2 is conducted on the DeepGlobe dataset. The quantitative results are shown in Table II. The ground truth on the DeepGlobe dataset has true width, and scenes in the dataset are more comprehensive and complex. Results show that the proposed AD-RoadNet significantly outperforms others. Specifically, AD-RoadNet achieves the best results with an IoU 2.46%–6.03% higher than the other methods, and also gains a 1.61%–4.25% improvement on the F1 score.

Fig. 7 displays road extraction details from the DeepGlobe road dataset with the abovementioned networks. It could be seen that the proposed AD-RoadNet performs fairly well in dealing with multiscale road extraction of very complex satellite images.

TABLE I
ROAD EXTRACTION RESULTS OF APPROACHES ON THE MASSACHUSETTS ROAD DATASET

Method	Precision	Recall	F1	IoU	mIoU
SegNet	76.09	78.23	77.15	62.80	80.27
Unet	77.53	77.82	77.67	63.40	80.55
ResUnet	78.77	77.45	78.10	64.10	80.91
D-LinkNet	78.34	77.91	78.12	64.32	81.05
HRNetV2	79.01	78.22	78.61	64.91	81.45
DDU-Net	**82.54**	73.99	**79.98**	-	80.98
AD-RoadNet(ours)	79.35	**79.44**	79.39	**65.93**	**81.97**

The results in bold denote the best performance among different methods.

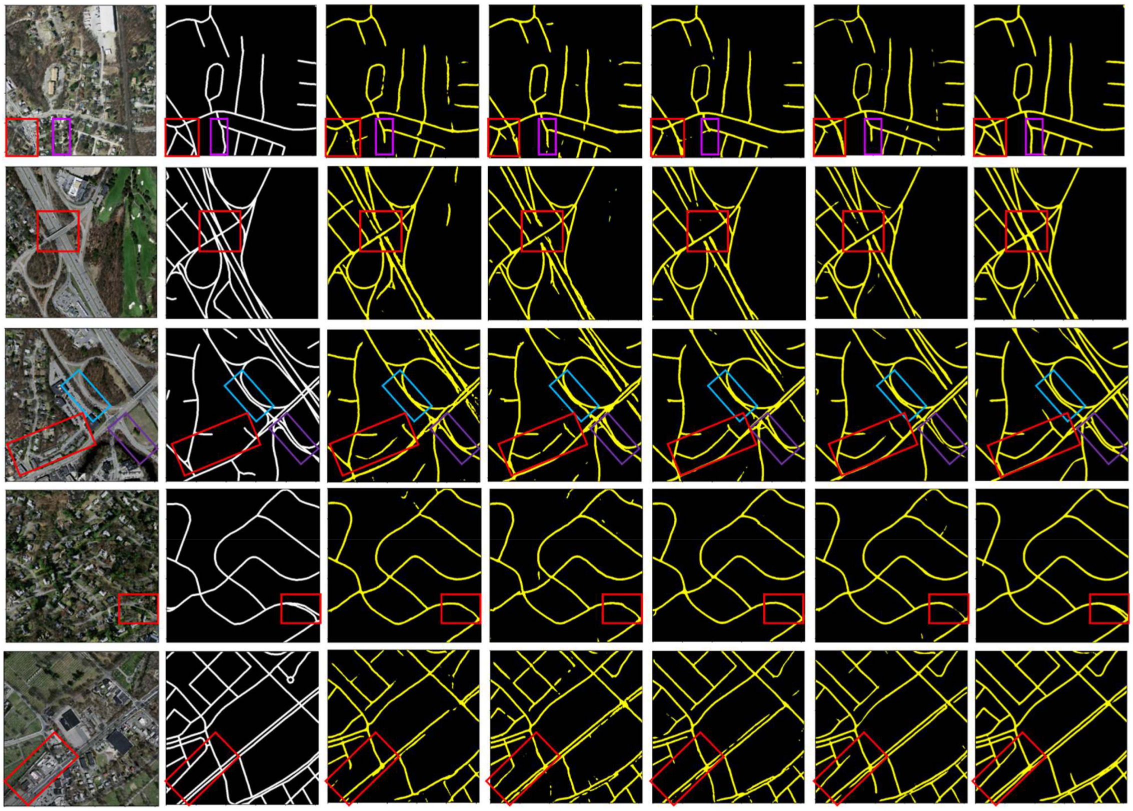

Fig. 6. Partial visualization of results on the Massachusetts road dataset. (a) Raw images. (b) Ground truth. (c) SegNet. (d) Unet. (e) D-linkNet. (f) HRNetV2. (g) Proposed AD-RoadNet.

TABLE II
ROAD EXTRACTION RESULTS OF APPROACHES USING THE DEEPGLOBE DATASET

Method	Precision	Recall	F1	IoU	mIoU
SegNet	81.19	76.87	78.97	65.25	79.44
Unet	79.60	79.86	79.73	66.29	79.96
ResUnet	78.77	79.32	79.04	65.85	79.70
D-LinkNet	82.28	79.48	80.86	67.87	81.01
HRNetV2	82.17	81.06	81.61	68.82	81.51
DDU-Net	77.86	62.23	69.17	-	75.32
GAMSNet	83.14	80.00	81.54	68,84	-
AD-RoadNet(ours)	**84.37**	**82.11**	**83.22**	**71.28**	**83.04**

The results in bold denote the best performance among different methods.

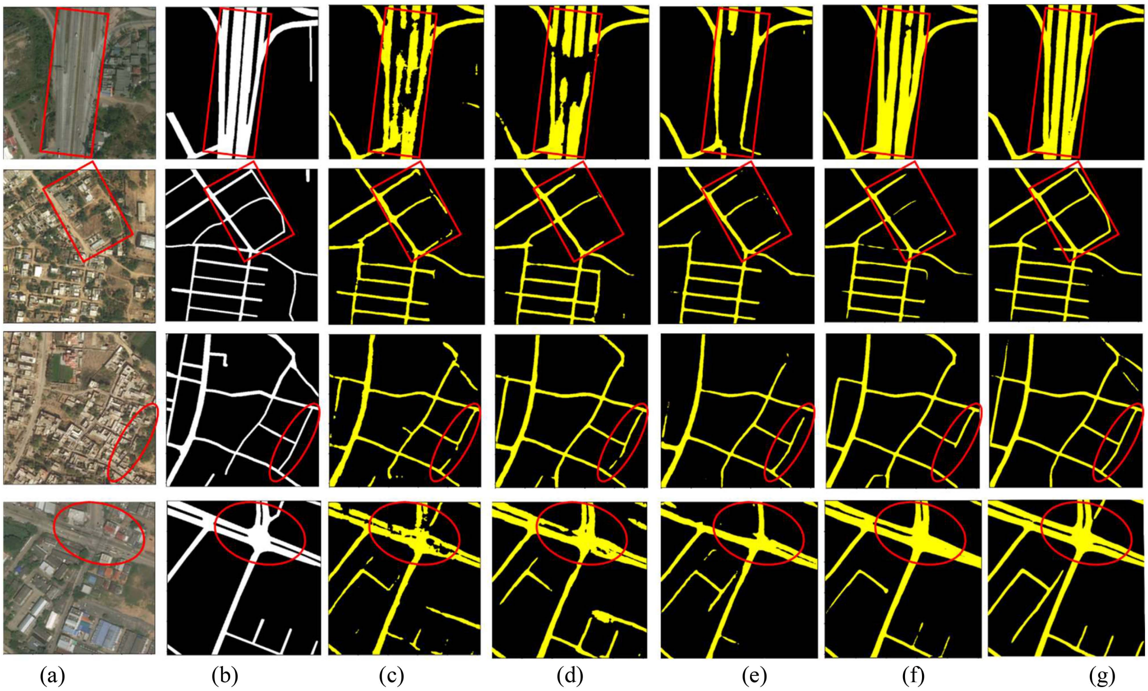

Fig. 7. Partial visualization of results on the DeepGlobe dataset. (a) Raw images. (b) Ground truth. (c) SegNet. (d) Unet. (e) D-linkNet. (f) HRNetV2. (g) Proposed AD-RoadNet.

To illustrate the effectiveness of AD-RoadNet on the DeepGlobe road dataset, four typical scenes, which include expressways with vehicles, median strips as well as lane markings, multiscale rural roads, roads obscured by trees, and intersections connecting roads of various types, are selected, as shown in Fig. 7.

V. Analysis

The abovementioned comparison experiments have demonstrated the effectiveness of the proposed AD-RoadNet for multiscale road extraction in HRSI. The results of the experiment can be interpreted based on the structure of AD-RoadNet. The encoder with hybrid RFs makes feature extraction achievable while still preserving road details, such as trails and lanes. While the main decoder decodes roads from extracted multiscale semantic features, TFRM in the auxiliary decoder is utilized to represent topological information, and help reason obscured roads. Moreover, the high-resolution feature prevents the disappearance of very narrow roads in the subsampling. The combination of these features guarantees rich multiscale road details as well as connectivity improvement.

This section further evaluates the proposed method in the following four parts. Section V-A analyzes the complexity of our model, Section V-B performs comprehensive ablation to validate proposed modules, Section V-C discusses the matter of inference patch size, and Section V-D evaluates the robustness of the proposed AD-RoadNet.

A. Parameter Scales and Model Complexity

The parameter scales, floating point operations (FLOPs), and inference time are widely used in evaluation of model complexity. The params and FLOPs are provided in articles [10], [15], [35] or calculated from Pytorch [64]. The FPS (inference time) are tested and calculated on our hardware platform.

TABLE III
Comparison of Calculation and Parameter Quantities

Method	Params(M)	FLOPs(G)	FPS
ResUnet	9.5	-	1.0
D-LinkNet	31.1	67.3	1.9
SegFormer (B4)	64.1	162.4	0.9
AD_RoadNet (ours)	70.1	180.0	12.5

The results in bold denote the best performance among different methods.

Fig. 8 shows the parameter scales and performances of methods. From Fig. 8, we can see the parameters of our model are similar to HRNetV2, while the performance achieves a remarkable improvement.

We also compare the efficiency of several models under our hardware (RTX3060 and i9-10900) and explain why we exclude the Transformer-based model from our analysis. Table III shows the approximate number of parameters, FLOPs, and FPS (inference times) for each model with an input size of 512 × 512.

We can see that the transformer-based model, Segformer (B4), which is a segmentation model based on transformers, has the longest inference time (lowest FPS) among all the models. Due to its larger epoch requirement and longer inference time, the Transformer-based model takes longer to train; this explains why we do not include it in our comparison.

B. Model Analysis

We perform comprehensive ablations to discuss the effectivity of the proposed HRFM and TFRM. Concretely, first train a

TABLE IV
QUANTITATIVE COMPARISONS (%) AMONG ABLATION STUDIES ABOUT PROPOSED MODULES ON MASSACHUSETTS DATASET AND DEEPGLOBE DATASET

Method	Massachusetts dataset		DeepGlobe dataset	
	F1	mIoU	F1	mIoU
baseline	77.84	80.74	80.58	80.73
baseline+H	78.47	81.19	81.48	81.36
baseline+T	78.74	81.33	81.59	81.43
baseline+H+T	**79.39**	**81.97**	**83.22**	**83.04**

The baseline is the same as AD-RoadNet, except has no HRFM and TFRM. H Is HRFM, T is TRFM.
The results in bold denote the best performance among different methods.

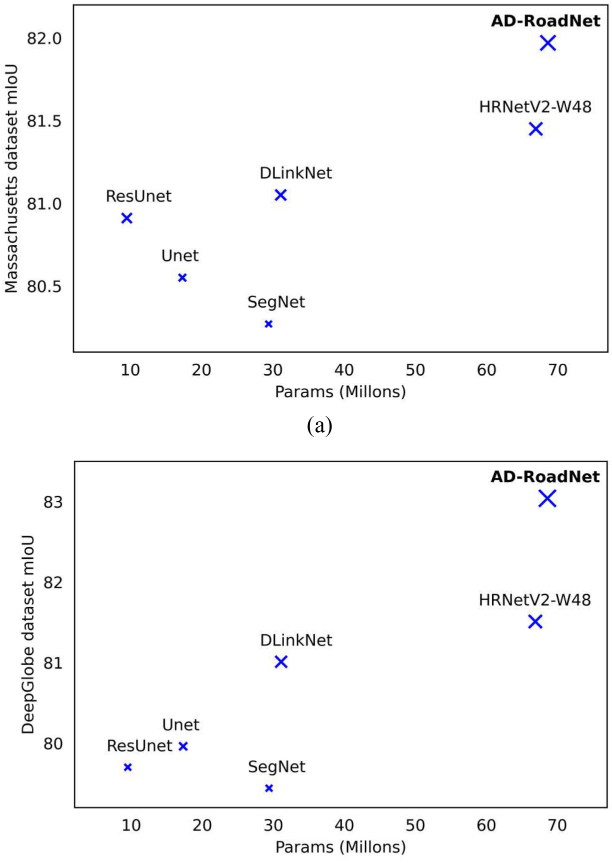

Fig. 8. Performance versus model params on (a) Massachusetts roads dataset and (b) DeepGlobe dataset. With a slightly higher params than HRNetV2, AD-RoadNet achieves the best 81.97% and 83.04% mIoU on two datasets.

baseline whose pipeline is the same as AD-RoadNet except without HRFM and TFRM, next gradually add HRFM and TFRM, then conduct experiments on the two datasets and finally report the performance via F1 score and mIoU.

Table IV lists all experimental results, where H is the abbreviation for HRFM and T is for TFRM. Compared to the baseline, HRFM helps the model increase the F1 and mIoU by 0.63% and 0.45% on the Massachusetts dataset, respectively, while the results achieved on the DeepGlobe dataset are 0.9% and 0.63%, respectively. Besides, TFRM helps improve the baseline's performance with 0.9% F1 and 0.59% mIoU on the Massachusetts dataset, while achieving a 1.01% F1 and 0.7% mIoU result on the DeepGlobe dataset. This indicates that when HRFM and TFRM are interpolated to the baseline independently, there is a limited improvement on performance. This is because if only HRFM is used, while various visible roads may be extracted, the overaccurate RF size for the target road weakens the network's robustness to obscured roads with large RF, which is commonly shown to be effective [19], [37], as shown in Fig. 9(a)–(c). On the other hand, if TRFM is used independently, the encoder with a unified RF size will not be good at extracting multiscale roads in an HRSI. Many feature details are missed, such as narrow roads and lane markings, resulting in incomplete and inaccurate extraction results, as shown in Fig. 9(d) and (e). Overall, HRFM and TRFM complement each other in the road extraction process, and the best accuracy achieved is with a complete AD-RoadNet (baseline+H+T), which further proves our hypothesis to some degree.

Table V lists experimental results of our model with different effective receptive filed in RFWL module. This experiment is to illustrate the effectivity of RFWL module. We force the three branch channels of the RFWL module to use the same size pooling operations; and in this way, we could control the size of the ERF of perception.

C. Inference Size Matters

In practical applications, roads need to be extracted in a large extensive region to indicate the overall distribution. However, limited to the memory of GPU, it is not feasible to directly infer such a big remote sensing image. Now we commonly crop the image into patches and merge them after inference. Various patch sizes for inference impact the performance [13], [48]. Thus, we discuss the optimal patch size by cropping patches with different sizes.

Fig. 10 displays the performance trend as inference patch size varies. We can see that precision shows an upward trend in the range of 128–768 and a backward trend in the range of 768–1532 while Recall shows a contrasting trend. The best performance (the highest value of IoU) occurs when the input size is 768, reaching 66.22%. This phenomenon may imply that when the input size is small, the road feature in the image is insufficient. Thus, distinguishing roads from objects with similar textures, such as parking lots, cement ground, and so on is difficult,

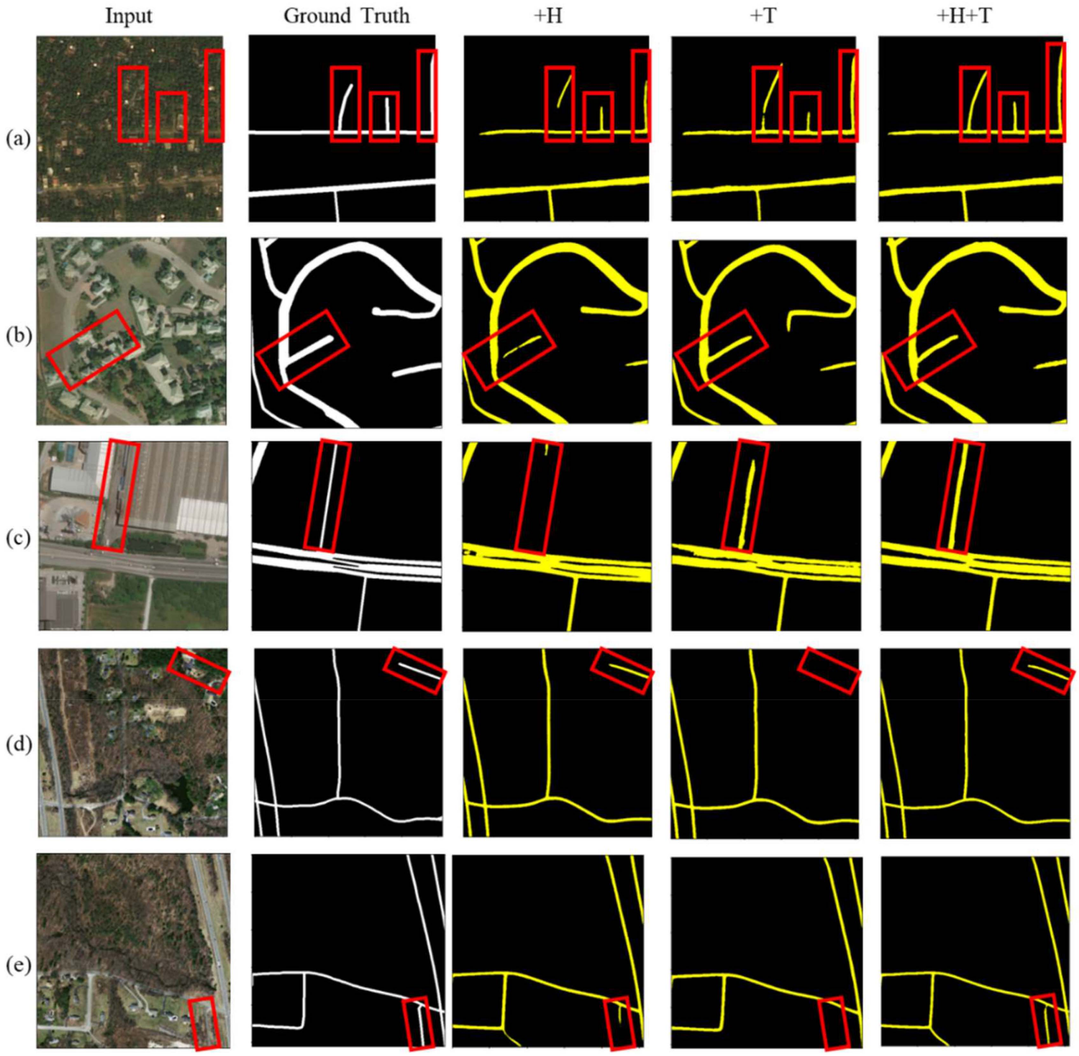

Fig. 9. Visualization of results for different combinations. From left to right are raw images, ground truth, the model only HRFM is interpolated, the model only TFRM is interpolated and complete AD-RoadNet. We can see that the HRFM is better at detecting narrow roads, as (d), (e), while the TFRM has advantages in the obscured roads, as (a), (b), (c), but they are all limited. The complete AD-RoadNet makes full use of their advantages.

TABLE V
Quantitative Comparisons (%) Among Ablation Studies About ERF in RFWL Module on the Massachusetts Road Dataset

Method	Precision	Recall	F1	IoU	mIoU
Small ERF	80.42	75.83	78.06	64.02	80.91
Large ERF	**85.27**	66.20	74.53	59.41	78.61
Suitable ERF	84.37	**82.11**	**83.22**	**71.28**	**83.04**

The results in bold denote the best performance among different methods.

resulting in a low precision. On the other hand, a small input size also decreases the variance of the widths of roads. Detecting roads with this input becomes easier, resulting in a high recall. When the input size is larger than 768, the road features the model could extract achieve saturation, while the more severe scale problem begins to degrade the detecting precision of our model. However, the long-range contextual information introduced proceeds a higher Recall. When the input size is 768, the road feature, scale problem, and long-range contextual information reach the best tradeoff.

D. Robustness of AD-RoadNet

We discuss the robustness of the proposed AD-RoadNet in the following three aspects: unseen occlusion scenarios, low-quality labels, and various quality of inference images.

1) Robustness Analysis in Unseen Occlusion Scenarios: To assess the robustness of AD-RoadNet and avoid just memorizing similar occlusion scenes from the trained samples, we manually attach some occlusions and test the robustness by strengthening them slowly. Specifically, three occlusion levels;

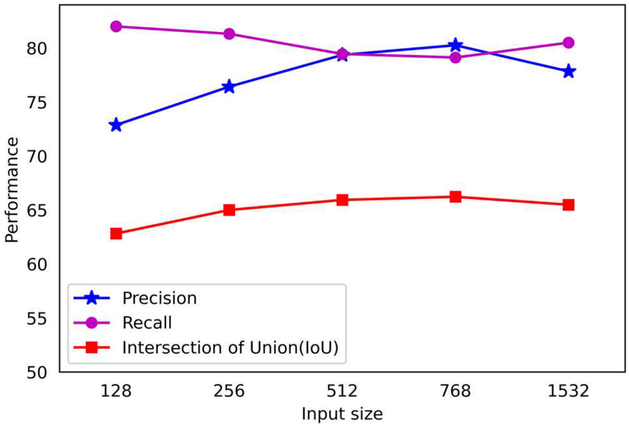

Fig. 10. Performance changes on Massachusetts road dataset as inference patch size grows up. The best performance (the highest value of IoU) occurs when the input size is 768, reaching 66.22%.

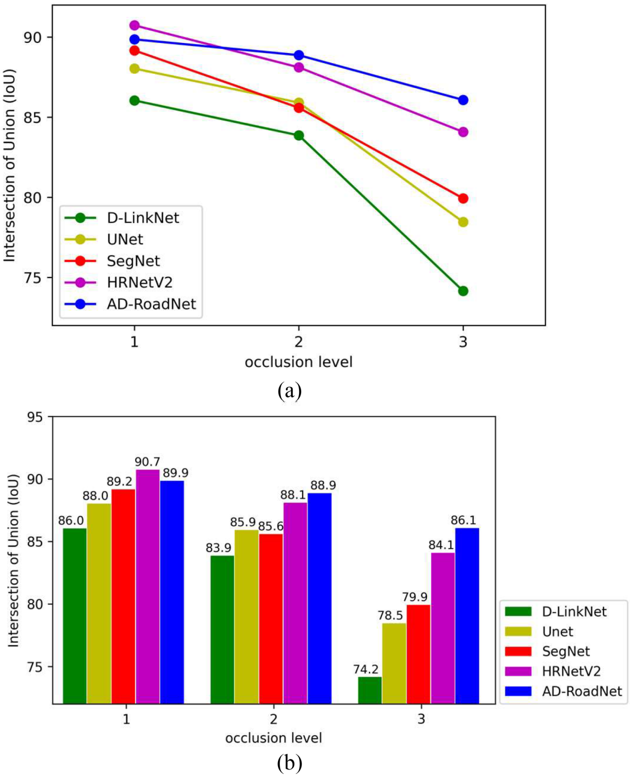

Fig. 11. Performance changes of all methods as the occlusion level grows up. (a) Changing trend of five methods. (b) Visually displays the obtained mIoU in each situation. Road extraction with proposed AD-RoadNet shows great robustness to the unseen occlusion.

the raw road image, road image with weak occlusion, and road image with strong occlusion are set in this part, and the performance changes of all methods are shown in Fig. 11. The visualization road extraction results are shown in Fig. 12, in which rows correspond to occlusion levels and the columns correspond to applied models. All five methods infer relatively high accurate results in the first row, while SegNet, Unet, and D-LinkNet are slightly disturbed with weak occlusion. Under the third occlusion level, the extraction results are all impacted greatly, but the one from AD-RoadNet still reserves basic road elements and their connectivity, which indicates its robustness.

2) Robustness Analysis for Low Quality Labels: Using correct ground truth to train and optimize road extraction networks could definitely release all their potential. However, it is not feasible since labeling images are error-prone and frequently has different standards for ambiguous situations. Moreover, recent research using open spatial vector data to create training datasets further introduced labelling errors [70]. Therefore, it is necessary to evaluate the robustness of the proposed model for low quality labels. In the road extraction dataset, most of incorrect labels mislabel the road as background, while there is almost no case where the background has been mislabeled as road. We randomly select some images with obvious labelling mistakes from our training dataset, and then predict the images again to test if the mistakes can be relabeled correctly. Fig. 13 shows these images and our corresponding prediction results. We can see that for the patterns, which have been labeled correctly in many samples, such as main roads, our AD-RoadNet could sufficiently correct the raw error mask as shown in the red box. However, for patterns that are often labeled incorrectly, such as the short roads in front of houses, the prediction results prefer to keep the incorrect label, as shown in the magenta box. The result implies that although the proposed model allows mislabeling, we could better ensure certain amount of correct labeling for each possible pattern, so as to get a great extraction result.

3) Robustness Analysis for Quality of Inference Images:: There are many data sources for HRSI, such as unmanned aerial vehicles, Google Earth, WorldView-4, and so on. Various acquisition conditions or equipment performance enlarge the variance of HRSI quality. To test the inference image quality robustness of our AD-RoadNet, we manually adjust an inference image quality by adding noise or perturbations in multidegrees. Specifically, we test 5 severities for 3 kinds of noise, Gaussian Blur, Pepper Noise and Contrast, and compare the performance of AD-RoadNet with D-LinkNet and HRNetV2. The performance changes are shown in Fig. 14. Overall, the compared three methods are at the same level in the robustness for the quality of inference images. But for the images with Gaussian Blur, the proposed AD-RoadNet suffers a relative degradation, getting a much lower performance than HRNetV2. This may be explained by the structure of HRFM. As the weights of various RF sizes for the target road are designed to consider its surrounding features, the Gaussian Blur smoothens the target and surrounding objects and passes a wrong message for RF size matching, thus resulting in weak feature extraction. The hypothesis could be proved to some degree by the similar performance changes between AD-RoadNet and baseline+H and the relatively better robustness of baseline+T, as shown in Fig. 14(d). From the perspective of training, it can be regarded as an overfit for the almost invariable ground sample distance (GSD). With this consideration, we retrain the AD-RoadNet with the same configures but stronger data augmentation adding a random blur. After that, the performance changes in 5 severities of Gaussian Blur as shown in Fig. 14(e). The AD-RoadNet with stronger data augmentation (AD-RoadNet-Aug) gets a significant improvement in its robustness for Gaussian Blur

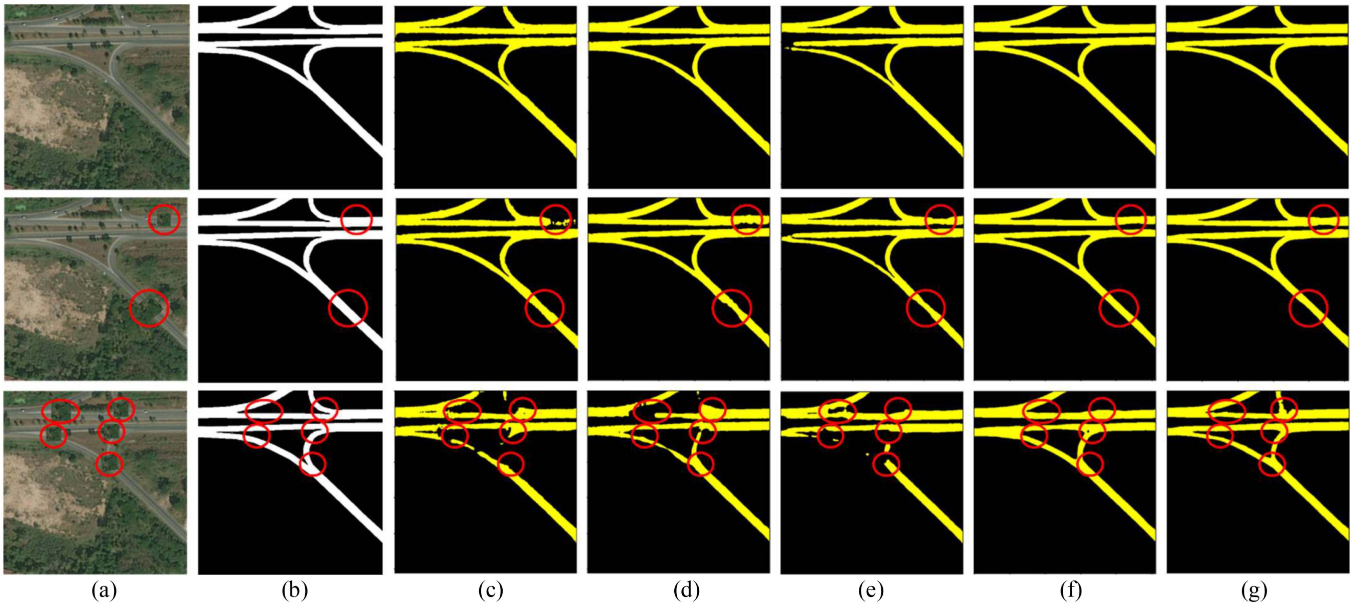

Fig. 12. Road extraction results with artificial occlusions (red circles). (a) Raw images. (b) Ground truth. (c) SegNet. (d) Unet. (e) D-linkNet. (f) HRNetV2. (g) Our AD-RoadNet. From top to down are raw scene, weak occlusion, and strong occlusion.

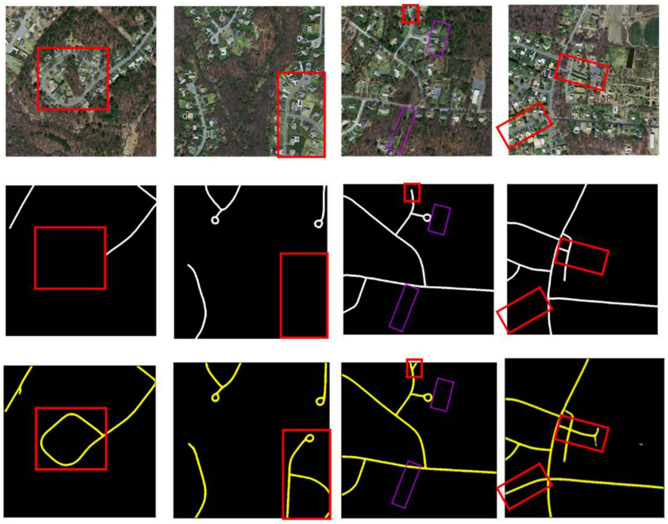

Fig. 13. Road extraction results for the mislabeled images. From first to third rows are image, raw label, and prediction result, respectively. The red box indicates the roads relabeled correctly with our model, while the magenta box indicates the roads fail to relabeled.

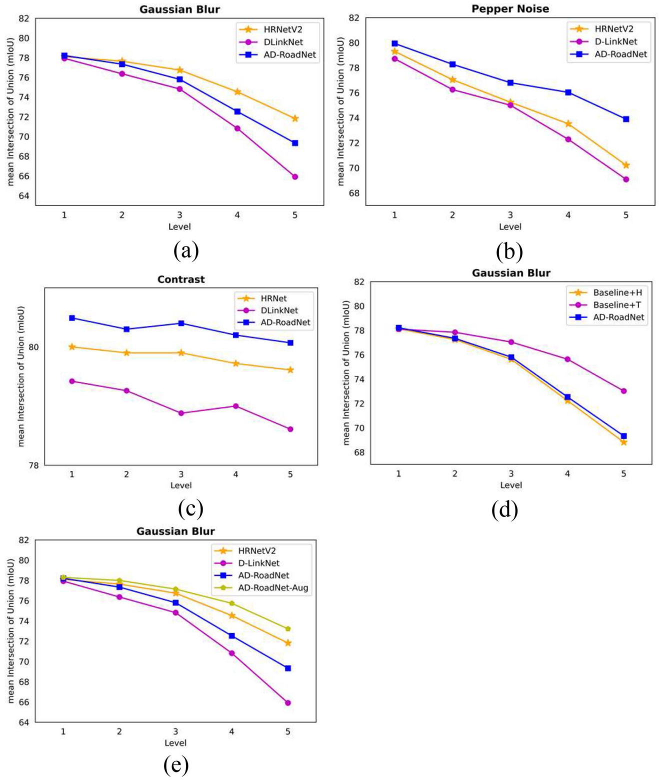

Fig. 14. Robustness analysis of AD-RoadNet for inference image quality. (a)–(c) Performance changes in various inference image qualities among HRNetV2, D-LinkNet, and AD-RoadNet. (d) Performance trend in Gaussian Blur among baseline+H, baseline+T, and the complete AD-RoadNet. H is the abbreviation of HRFM and T is TFRM. (e) Performance changes in Gaussion Blur among HRNetV2, D-LinkNet, AD-RoadNet, and AD-RoadNet-Aug. AD-RoadNet-Aug represents the model retrained with stronger data augmentation.

while retaining its initial performance. The abovementioned analysis may imply that since the HRFM dynamically adjusts target road RF size according to its surrounding features, there is a risk behind more targeted feature extraction, that is, the model only remembers the distribution of surrounding objects but does not recognize the corresponding pattern, thus resulting in an overfit for a certain GSD. Data augmentation techniques, which could change the GSD, such as Random Resized Crop and Random Blur, are essential to improve the robustness of proposed AD-RoadNet.

VI. Conclusion

In this article, to decouple and give an overall consideration to the representation and connectivity improvement of multiscale road details, the proposed AD-RoadNet performs well by introducing HRFM in the encoder to adaptively provide each unit with suitable RF size, high-resolution feature information is preserved by residual blocks to detect very narrow roads, and compared to some previous research works in multiscale road feature extraction (see [17], [21], [35], [69]), this study introduces a topological feature module (TFRM) that encodes the connectivity and directionality of roads as additional features, and experiment results demonstrate that the TFRM could improve the performance of the network by reducing false positives and enhancing the continuity and smoothness of extracted roads.

The proposed network has achieved state-of-the-art performance on two benchmark datasets, outperforming existing methods in terms of recall and IOU. The effectiveness of the proposed network is proven with solid experiments and achieves the SOTA performance. For future research, we suggest exploring more topological features that could enhance the road extraction performance, such as curvature, width, or intersection angles. We also recommend testing our network on different types of HRSI with varying spectral, spatial, and temporal resolutions to evaluate its adaptability and robustness. Furthermore, we would propose developing a more interpretable representation of topological features that could provide insights into how they affect the network's decision-making process.

Acknowledgment

The authors would like to thank the editors and anonymous reviewers for their instructive comments.

Declaration of Competing Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

References

[1] H. Huang, A. V. Savkin, and C. Huang, "Decentralized autonomous navigation of a UAV network for road traffic monitoring," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 4, pp. 2558–2564, Aug. 2021, doi: 10.1109/TAES.2021.3053115.

[2] G. Singh et al., "ROAD: The road event awareness dataset for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1036–1054, Jan. 2023, doi: 10.1109/TPAMI.2022.3150906.

[3] M. Yu, "Construction of regional intelligent transportation system in smart city road network via 5G network," *IEEE Trans. Intell. Transport. Syst.*, vol. 24, no. 2, pp. 2208–2216, Feb. 2023, doi: 10.1109/TITS.2022.3141731.

[4] G. Zhou, W. Chen, Q. Gui, X. Li, and L. Wang, "Split depthwise 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.

[5] F. Yi, R. Te, Y. Zhao, and G. Xu, "EUNetMTL: Multitask joint learning for road extraction from high-resolution remote sensing images," *Remote Sens. Lett.*, vol. 13, no. 3, pp. 258–268, Mar. 2022, doi: 10.1080/2150704X.2021.2019344.

[6] A. Abdollahi, B. Pradhan, N. Shukla, S. Chakraborty, and A. Alamri, "Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review," *Remote Sens.*, vol. 12, no. 9, May 2020, Art. no. 1444, doi: 10.3390/rs12091444.

[7] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 270–294, Jun. 2022, doi: 10.1109/MGRS.2022.3145854.

[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.-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.

[12] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous convolution for semantic image segmentation," Dec. 5, 2017. Accessed: Mar. 1, 2023, *arXiv:1706.05587*.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with Atrous separable convolution for semantic image segmentation," Aug. 22, 2018. Accessed: Mar. 1, 2023, *arXiv:1802.02611*.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," Jun. 7, 2016. Accessed: Mar. 1, 2023, *arXiv:412.7062*.

[15] K. Sun et al., "High-resolution representations for labeling pixels and regions," Apr. 9, 2019. Accessed: Feb. 28, 2023, *arXiv:1904.04514*.

[16] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, Red Hook, NY, USA, Curran Associates, Inc., 2021, pp. 12077–12090. Accessed: Mar. 1, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html

[17] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022, doi: 10.1016/j.isprsjprs.2022.06.008.

[18] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609413, doi: 10.1109/TGRS.2021.3104032.

[19] W. G. C. Bandara, J. M. J. Valanarasu, and V. M. Patel, "SPIN road mapper: Extracting roads from aerial images via spatial and interaction space graph reasoning for autonomous driving," in *Proc. Int. Conf. Robot. Autom.*, 2022, pp. 343–350, doi: 10.1109/ICRA46639.2022.9812134.

[20] 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.

[21] 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, doi: 10.1016/j.patcog.2022.108549.

[22] 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.

[23] Y. Zao and Z. Shi, "Richer U-net: Learning more details for road detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3003105, doi: 10.1109/LGRS.2021.3081774.

[24] M. Song and D. Civco, "Road extraction using SVM and image segmentation," *Photogramm Eng. Remote Sens.*, vol. 70, no. 12, pp. 1365–1371, Dec. 2004, doi: 10.14358/PERS.70.12.1365.

[25] J. Wang, Q. Qin, X. Yang, J. Wang, X. Ye, and X. Qin, "Automated road extraction from multi-resolution images using spectral information and texture," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2014, pp. 533–536, doi: 10.1109/IGARSS.2014.6946477.

[26] R. Stoica, "Road extraction in remote sensed images using stochastic geometry framework," in *Proc. AIP Conf. Proc.*, 2001, pp. 531–542, doi: 10.1063/1.1381915.

[27] O. Tournaire and N. Paparoditis, "A geometric stochastic approach based on marked point processes for road mark detection from high resolution aerial images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 64, no. 6, pp. 621–631, Nov. 2009, doi: 10.1016/j.isprsjprs.2009.05.005.

[28] J. B. Mena and J. A. Malpica, "An automatic method for road extraction in rural and semi-urban areas starting from high resolution satellite imagery," *Pattern Recognit. Lett.*, vol. 26, no. 9, pp. 1201–1220, Jul. 2005, doi: 10.1016/j.patrec.2004.11.005.

[29] 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, doi: 10.1109/TGRS.2012.2190078.

[30] G. Mattyus, W. Luo, and R. Urtasun, "DeepRoadMapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3458–3466, doi: 10.1109/ICCV.2017.372.

[31] A. Mosinska, P. Marquez-Neila, M. Kozinski, 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.

[32] Z. Zhang, C. Miao, C. Liu, and Q. Tian, "DCS-TransUperNet: Road segmentation network based on CSWIN transformer with dual resolution," *Appl. Sci.*, vol. 12, no. 7, Mar. 2022, Art. no. 3511, doi: 10.3390/app12073511.

[33] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 210–223, doi: 10.1007/978-3-642-15567-3_16.

[34] 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.

[35] 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.

[36] Z. Chen, C. Wang, J. Li, N. Xie, Y. Han, and J. Du, "Reconstruction bias U-net for road extraction from optical remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2284–2294, 2021, doi: 10.1109/JSTARS.2021.3053603.

[37] 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.

[38] 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. 171–182, doi: 10.1109/CVPRW.2018.00031.

[39] C. Tao, J. Qi, Y. Li, H. Wang, and H. Li, "Spatial information inference net: Road extraction using road-specific contextual information," *ISPRS J. Photogrammetry Remote Sens.*, vol. 158, pp. 155–166, Dec. 2019, doi: 10.1016/j.isprsjprs.2019.10.001.

[40] 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.

[41] C. Sherrington, *The Integrative Action of the Nervous System*. Cambridge, U.K.: CUP Archive, 1952.

[42] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, New York, NY, USA, Curran Associates, 2016. Accessed: Mar. 01, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/c8067ad1937f728f51288b3eb986afaa-Abstract.html

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Apr. 10, 2015. Accessed: Mar. 1, 2023, *arXiv:1409.1556*.

[45] 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.

[46] Y. Sun, Y. Tian, and Y. Xu, "Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning," *Neurocomputing*, vol. 330, pp. 297–304, Feb. 2019, doi: 10.1016/j.neucom.2018.11.051.

[47] X. Liang, Y. Zhang, and J. Zhang, "Water retrieval embedded attention network with multiscale receptive fields for hyperspectral image refined classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5509022, doi: 10.1109/TGRS.2021.3091985.

[48] R. Liu, L. Mi, and Z. Chen, "AFNet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7871–7886, Sep. 2021, doi: 10.1109/TGRS.2020.3034123.

[49] G. Li, L. Li, H. Zhu, X. Liu, and L. Jiao, "Adaptive multiscale deep fusion residual network for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8506–8521, Nov. 2019, doi: 10.1109/TGRS.2019.2921342.

[50] G. Wang, H. Liu, X. Yi, J. Zhou, and L. Zhang, "ARMS Net: Overlapping chromosome segmentation based on adaptive receptive field multiscale network," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102811, doi: 10.1016/j.bspc.2021.102811.

[51] W. Boonpook, Y. Tan, B. Bai, and B. Xu, "Road extraction from UAV images using a deep ResDCLnet architecture," *Can. J. Remote Sens.*, vol. 47, no. 3, pp. 450–464, May 2021, doi: 10.1080/07038992.2021.1913046.

[52] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8017505, doi: 10.1109/LGRS.2021.3098774.

[53] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.

[54] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833, doi: 10.1007/978-3-319-10590-1_53.

[55] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[56] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," Feb. 12, 2017. Accessed: Mar. 1, 2023, *arXiv:1612.03928*.

[57] M. A. Islam, S. Jia, and N. D. B. Bruce, "How much position information do convolutional neural networks encode?," Jan. 22, 2020. Accessed: Feb. 28, 2023, *arXiv:2001.08248*.

[58] V. Mnih, *Machine Learning For Aerial Image Labeling*. Toronto, ON, Canada: Univ. of Toronto, 2013.

[59] X. Lu, Y. Zhong, Z. Zheng, and L. Zhang, "GAMSNet: Globally aware road detection network with multi-scale residual learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 340–352, May 2021, doi: 10.1016/j.isprsjprs.2021.03.008.

[60] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[61] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 200:1–200:41, Sep. 2022, doi: 10.1145/3505244.

[62] Y. Zhao, G. Wang, C. Tang, C. Luo, W. Zeng, and Z.-J. Zha, "A battle of network structures: An empirical study of CNN, transformer, and MLP," Aug. 30, 2021. Accessed: May 15, 2023, *arXiv:2108.13002v2*.

[63] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

[64] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, New York, NY, USA, Curran Associates, Inc., 2019. Accessed: Mar. 1, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[65] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," Jan. 4, 2019. Accessed: Feb. 28, 2023, *arXiv:1711.05101*.

[66] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin, "Cyclical annealing schedule: A simple approach to mitigating," in *Proc. Conf. North, Minneapolis, Minnesota: Assoc. Comput. Linguistics*, 2019, pp. 240–250, doi: 10.18653/v1/N19-1021.

[67] A. Abdollahi and B. Pradhan, "Integrated technique of segmentation and classification methods with connected components analysis for road extraction from orthophoto images," *Expert Syst. Appl.*, vol. 176, Aug. 2021, Art. no. 114908, doi: 10.1016/j.eswa.2021.114908.

[68] F. Gao, J. Tu, J. Wang, A. Hussain, and H. Zhou, "RoadSeg-CD: A network with connectivity array and direction map for road extraction from SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3992–4003, 2022, doi: 10.1109/JSTARS.2022.3175594.

[69] Y. Hou, Z. Liu, T. Zhang, and Y. Li, "C-UNet: Complement UNet for remote sensing road extraction," *Sensors*, vol. 21, no. 6, Mar. 2021, Art. no. 2153, doi: 10.3390/s21062153.

[70] S. Glinka, T. Owerko, and K. Tomaszkiewicz, "Using open vector-based spatial data to create semantic datasets for building segmentation for raster data," *Remote Sens.*, vol. 14, no. 12, Jun. 2022, Art. no. 2745, doi: 10.3390/rs14122745.