DRENet: A Dual-branch Road Extraction Network for Enhanced Connectivity

Yao Yang, Pingfu Chao✉, Qiao Kun, and Junhua Fang

Soochow University, Soochow, Jiangsu, China
yyang11@stu.suda.edu.cn, pfchao@suda.edu.cn,
kqiao51643729@stu.suda.edu.cn, jhfang@suda.edu.cn

Abstract. The road network is the foundation of the transportation system. However, the availability and the correctness of road networks always face challenges due to new road construction and frequent road changes. Instead of conducting labor-intensive ground surveys for map construction and update, automatic road network extraction via satellite images and/or trajectory data becomes the new trend. Nevertheless, although existing methods can extract road networks with the correct shape and decent coverage, few studies focus on road network connectivity, which is a crucial indicator of road network usability. In this paper, we propose a novel Dual-branch Network that improves connectivity and achieves accurate road extraction. Our model incorporates a Shape Reshaper Module for enriching the connectivity information and an Attention-based Fusion Module that dynamically captures the relationships between modalities, enabling effective fusion. Furthermore, we propose a connectivity measurement metric for road networks and a data augmentation method to decrease the impact of occlusions in satellite images. Extensive experiments on datasets from Beijing and Porto demonstrate that our approach achieves new state-of-the-art results. The source code can be found at https://github.com/sallwe/DRENet.

Keywords: Road extraction · Road Connectivity · Multimodal.

1 INTRODUCTION

As the foundation of a transportation system, the road network is essential for many downstream tasks, such as navigation, urban planning, and map updates. Previously, map construction relied heavily on ground survey, an inefficient and highly labor-intensive method. This approach falls short in meeting the demands for timely map updates and new road construction, especially in distant areas, creating a need for automatic road network extraction methods. With rapid advancements in neural networks, automatic road network extraction has achieved substantial progress, primarily focusing on two key data sources: satellite images and trajectory data. Satellite images, captured through aerial photography, provide high-resolution images. Meanwhile, trajectories are collected from the

Fig. 1: The noise in the satellite images and trajectory data.

GPS data of numerous vehicles. However, both data sources contain noise when extracting roads, as shown in Figure 1. For instance, the cloud shadow in the top-left satellite image and the complex environment in the image below make extracting difficult. In the third column on the left, shadows, tall buildings, and planting further obscure the road areas. Additionally, certain regions of the trajectory data show sparse coverage due to a lack of vehicle visits. Each of these factors can negatively impact road extraction accuracy. Ideally, integrating these two modalities can achieve information complementarity, leading to accurate road extraction.

In general, deep learning-based road extraction methods can be classified into two categories: graph-based methods and segmentation-based methods. Graph-based methods achieve excellent road connectivity but may sacrifice details such as road width, whereas most segmentation-based methods overlook connectivity issues within the road network. Note that, road connectivity is a key indicator for measuring a road network's quality in terms of usability. Considering the aforementioned noise, we argue that extracting a reliable, high-quality road network requires two key factors: (1) achieving effective fusion of the two modalities, which necessitates dynamically handling the relationship between them; and (2) performing precise road segmentation while ensuring fine road connectivity. To achieve this goal, we propose a novel Dual-branch Road Extraction Network (DRENet). One branch is dedicated to precise road segmentation, while the other branch focuses on the skeleton structure of the road network depicting the road connectivity. Subsequently, we design the Road Reshaper Module to restore the shape of the skeleton branch and enrich the connectivity information on the other branch. The attention mechanism [11,31] has demonstrated its effectiveness in remote sensing tasks [10,25] due to its ability to capture relationships among features dynamically. Building on this capability, we develop an Attention-based Fusion Module that combines convolutional kernels with varying receptive fields to simulate multi-head attention, thereby facilitating effective fusion across modalities. Moreover, to avoid overfitting and enable the model to robustly learn road features even in the presence of noise, such as occlusions in satellite images, we introduce a data augmentation technique that randomly

masks patches in the area of roads. Finally, to assess the quality of the map as a road network, we propose a new metric for the evaluation of road network connectivity. Overall, the key contributions of this paper are as follows:

- We introduce a novel Dual-branch Road Extraction Network for enhanced connectivity based on satellite image and trajectory. The two branches focus on the road skeleton and precise road segmentation, respectively. Besides, a Road Reshaper Module is applied between the branches to enhance road connectivity information.
- We propose an Attention-based Fusion Module to capture relationships between modalities and effective fusion. Query, key, and value are computed by combining convolutional kernels with varying receptive fields to form a more comprehensive representation, working in a manner similar to a multi-head attention mechanism.
- We develop a new metric for road connectivity measurement and a data augmentation method to decrease the impact of occlusions in satellite images. Our method achieves state-of-the-art performance on real-world datasets from both road segmentation and connectivity perspectives.

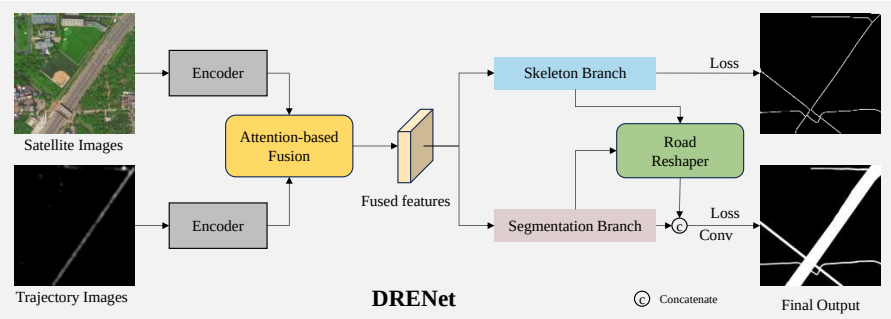


Fig. 2: The framework of DRENet.

2 RELATED WORK

2.1 Map Extraction based on Trajectory

The widespread use of navigation systems makes it easier to collect vast amounts of trajectory data. However, issues such as occlusion can cause positional inaccuracies, introducing significant noise to the data and posing challenges for map extraction from these trajectories. To accurately extract maps, many approaches aim to minimize the noise impact. Mostly they employ traditional methods, with

one category employing cluster-based algorithms [13,14]. These algorithms cluster scattered trajectory points to outline roads. Intersection-based algorithms [15,16] identify intersections by analyzing trajectory point features and then connect them to form the map. Another set of techniques applies Kernel Density Estimation [17] to delineate road networks according to trajectory density. However, due to inherent noise in the data and limited coverage, these methods often face challenges in generating dependable road networks.

2.2 Map Extraction based on Satellite Image

Recent advancements in neural networks have yielded various sophisticated feature extraction encoders such as ResNet [20], DenseNet [21], Swin Transformer [24], Vision Transformer [31], and MobileViT [22]. These encoders form the backbone of many methods used for extracting maps from satellite images, often utilizing an encoder-decoder architecture. Zhang et al [28] were the pioneers in adapting the U-Net[19] framework for map extraction. DlinkNet [18] leveraged dilated convolutions with varying dilation rates to enhance the receptive fields of their models. RoadFormer [25] introduced an Unet-like model that integrates Swin Transformer and Deformable Transformer within the encoder to efficiently extract roads. Sunet [26] incorporated a DULR block [29] in the skip connections, capitalizing on spatial-level continuous clues. Furthermore, Radanet [27] devised a deformable attention network designed to automatically learn the shapes of roads for better map extraction. Despite these innovations, accurately recognizing roads across all areas still poses challenges due to the complexity of the environments. All of the above are pixel-based segmentation models. In addition, there are graph-based road network models, such as the classic Sat2Graph[32]. Although graph-based models perform well in terms of connectivity, the extracted road networks lose information such as road width.

2.3 Map Extraction based on Multi-modal data

Considering the inherent noise in satellite images and trajectory data, exploring methods to fuse them for improved performance is logical. Numerous studies [1,2,3,4,5,6,7,8,10,23] have demonstrated that combining two modalities enhances performance over using just one in remote sensing. Among them, DeepDualMapper [5] introduced an encoder-decoder to extract features from satellite images and trajectory map images independently, using a Gate Fusion Module between decoders to enable mutual complementarity. Similarly, CMMPNet [6] and MoviNet [10] incorporate fusion modules within both encoder and decoder segments. CMMPNet's fusion approach merges convolution with Spatial Pyramid Pooling [30] to refine features from both inputs. MoviNet leverages MobileviT [22] as an encoder to draw local and global features from the modalities, which are then unified to reinforce each modality. DuARE [7] processes trajectory map images with thinning algorithms and adopts a coarse-to-fine strategy for satellite image feature extraction. Post-processing employs a cross-check-based strategy for fusion. Despite the diverse methods of modality fusion, there

remains potential for further improvements in managing the interplay between the two modalities.

3 Methods

3.1 Framework Overview

The framework of DRENet is shown in Figure 2. The input consists of a pair of satellite and trajectory images, where the trajectory image is a grayscale image generated by projecting trajectory points within the area of the satellite image. Then, the two modalities are fed into their respective encoders for road feature extraction. We choose pre-trained iFormer-S [9] as our encoders to extract features of two modalities, respectively. The feature set of the satellite images consists of four stages of information, denoted as: $[s1, s2, s3, s4]$, with size of $[(h, w, c1), (h/2, w/2, c2), (h/4, w/4, c3), (h/8, w/8, c4)]$, respectively. Similarly, the trajectory features are extracted in the same manner. After the encoding stage, the features are fed into an Attention-based Fusion Module, where the relationships between the two modalities are dynamically captured, enabling comprehensive fusion. The fused features denoted as $[f1, f2, f3, f4]$, have shapes identical to those of the satellite features. Next, the fused features are fed into the skeleton branch and the segmentation branch, which focus on the skeleton structure of the road and the precise road segmentation, respectively. The features from both branches will then enter the Road Reshaper Module to assist in restoring connectivity in the final output.

3.2 Attention-based Fusion Module

The detailed information of the Attention-based Fusion Module is shown in Figure 3. The features in corresponding areas of the satellite and trajectory images often exhibit similarities, especially in prominent regions such as major roads. However, there are also areas where the features are less distinguishable, such as on smaller roads. For example, satellite images may suffer from occlusions, while trajectory images may have a sparse distribution of data points. For effective fusion of these two modalities, it is crucial to dynamically assign appropriate weights across corresponding regions. Given that the attention mechanism excels at capturing the relationships between features, we introduce an Attention-based Fusion Module.

We choose to embed the fusion module after each encoder layer to achieve multi-scale fusion. Here, we use f_1 as an example to describe the fusion process. Since roads occupy a small portion of the map, dilated convolution is ideal for extracting sparse features, as it enlarges the receptive field without adding extra convolutional parameters, which has been proven in numerous road extraction works [8,18]. We combine convolutions with different receptive fields into a MultiConv Unit to hierarchically extract features, similar to the multi-head attention mechanism. A MultiConv Unit consists of a 1×1 convolution, a 3×3

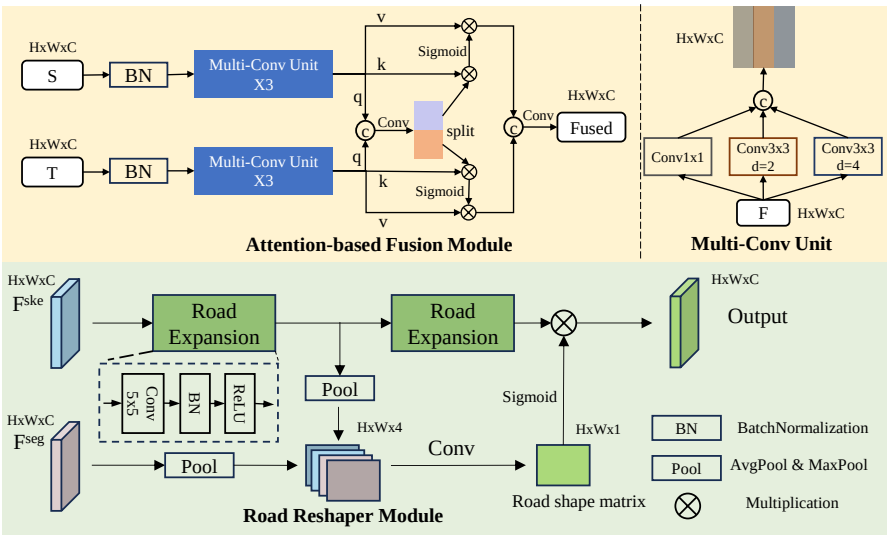


Fig. 3: The detailed information of Attention-based Fusion Module and Road Reshaper Module.

convolution with dilation 2, and a 3×3 convolution with dilation 4. Let F_s and F_t represent the features of the satellite and trajectory images, respectively. The entire process can be described as follows:

$$q_s, k_s, v_s = MultiConvUnit_{\times 3}(BN(F_s)) \tag{1}$$

where BN means BatchNormalization, q_t, k_t, and v_t are computed in the same way. To help the model focus on the most relevant areas, we choose to fuse the queries from both modalities and then split the fused query into two components, each paired with its corresponding key and value to obtain the results. Next, the results of the two modalities are concatenated and passed through a convolution layer to produce the final fused features. The process can be described as follows:

$$q_s^f, q_t^f = Split(Conv(Cancatenate(q_s, q_t))) \tag{2}$$

$$F_s^f, F_t^f = Sigmoid(q_s^f \times k_s) \times v_s, Sigmoid(q_t^f \times k_t) \times v_t \tag{3}$$

where q, k, and v keep the same shape. And fused features are computed as $f_1 = Conv(Cancatenate(F_s^f, F_t^f))$.

3.3 Dual-Branch Framework

Skeleton Branch. The skeleton branch can effectively predict skeleton pixels, as its prediction only requires extracting road features from regions adjacent to the satellite images and trajectories. This is a simpler learning task compared to

the segmentation branch, which needs to achieve precise segmentation of road boundaries. As a result, the skeleton branch exhibits better connectivity. The subsequent experiments will further clarify this point. To accurately extract the road skeleton, before being fed into the skeleton branch, $f1$, $f2$, and $f3$ are downsampled to the same shape as $f4$, i.e., $(h/8, w/8)$. Then, the input of the skeleton branch is performed as $F^{skeleton} = Conv(Concatenate(f1_{ske}, f2_{ske}, f3_{ske}, f4))$. Next, transposed convolutions are used for upsampling, progressively restoring the road skeleton. Between each upsampling layer, two residual blocks are inserted. The label of the skeleton branch is derived from the original label using the Zhang-Suen thinning algorithm [12].

Segmentation Branch. Similar to many segmentation models [34,35], $f2$, $f3$, and $f4$ are upsampled to the size of $f1$, with the same number of channels $c1$ and a size of (h, w). Then, the input of the segmentation branch is performed as $F^{segmentation} = Conv(Concatenate(f1, f2_{seg}, f3_{seg}, f4_{seg}))$. Then, $F^{skeleton}$ and $F^{segmentation}$ are fed into the Road Reshaper Module to help restore connectivity in the segmentation branch. The outputs of the two branches are upsampled 4 times for supervision, and the output of the segmentation branch serves as the final output.

3.4 Road Reshaper

The detailed information on the Road Reshaper Module is shown in Figure 3. Since the skeleton features only capture connectivity information and lack the actual shape of the road, they need to be transformed into a shape that accurately represents the road for precise prediction. This process begins by expanding the skeleton features, for which we choose to use large convolutional kernels, as they have proven effective in expanding the region of features [33]. To explicitly restore the road shape after expansion, it is essential to reference the features from the segmentation branch.

$F^{skeleton}$ with a size of (H, W, C) is first fed into a Road Expansion layer, which includes a 5×5 convolution layer, a batch normalization layer, and a ReLU activation. $F_e^{skeleton} = RoadExpansion(F^{skeleton})$. Next, the shape restoration of $F_{skeleton}$ needs to reference the information from $F^{segmentation}$ with a size of (H, W, C). For both features, average pooling and max pooling are applied along the height and width dimensions, forming a F_{pool} with a size of $(H, W, 4)$. Then, the road shape matrix map_{road} is computed through a convolutional and Sigmoid layer applied to F_{pool}. The process can be described as follows:

$$max_{ske}, avg_{ske} = MaxPool(F_e^{skeleton}), AvgPool(F_e^{skeleton}) \tag{4}$$

$$max_{seg}, avg_{seg} = MaxPool(F^{segmentation}), AvgPool(F^{segmentation}) \tag{5}$$

$$map_{road} = Sigmoid(Conv(Cancatenate(max_{ske}, avg_{ske}, max_{seg}, avg_{seg}))) \tag{6}$$

After map_{road} is obtained, for a more complete shape restoration, we choose to perform another road expansion on the features, as represented by $F_{ee}^{skeleton} = RoadExpansion(F_e^{skeleton})$. Then, map_{road} is multiplied with the expanded features to finally restore fine-shaped features F_{shape} with the size of (H, W, C),

which is served as the output of the Road Reshaper Module. The process can be described as: $F_{shape} = map_{road} \times F_{ee}^{skeleton}$.

3.5 Data Augmentation

Data augmentation enhances data diversity, improves model scalability, and reduces the risk of overfitting. In segmentation tasks, beyond standard augmentation techniques like flipping, rotation, translation, and scaling, we developed a specialized augmentation method to address occlusion issues often found in satellite images. Inspired by MAE [36], we randomly mask patches in road areas of satellite images, with road locations determined from labels during training. This technique encourages the model to learn to regain road features even when parts of them are obscured.

4 Experiments

4.1 Experimental Setup

Datasets. For our evaluation, we choose the BJRoad dataset [23] and the Porto dataset [23], both real-world datasets. The BJRoad dataset includes 348 pairs of high-resolution satellite images, trajectory images, and corresponding labels, each with a resolution of 1024×1024. We divide the BJRoad dataset into 250 training images, 28 validation images, and 70 test images. To speed up model training, the resolution is downsampled to 512×512. The Porto dataset consists of 1,006 pairs of satellite images, trajectory images, and corresponding labels, each with a resolution of 512×512, and this dataset is split into training, validation, and test sets in an 8:1:1 ratio.

Hyperparameters. The AdamW is used as the optimizer to update the model parameters, with a learning rate of 8e-5 and a batch size of 4. The loss function of two branches is a combination of binary cross-entropy loss and dice loss, expressed as Loss = $L_{dice} + L_{bce}$. Apart from MoviNet, Conats, DlinkNet, and RCAFNet which use the results reported in their papers, all other experimental results are based on this setup, taking the best results within 50 epochs.

Data augmentation techniques include horizontal and vertical flipping, rotation, translation, scaling, and random masking. The parameters for random masking include the proportion of masked areas within road regions and the patch size for masking. In this experiment, since the labels in the Porto dataset lack accurate road width information and appear wider than actual road widths in the satellite images, the masking parameters are set to $mask_{ratio} = (0, 0.05)$ and patch size $= (5, 5)$. In the BJRoad dataset, the parameters are set to $mask_{ratio} = (0.05, 0.15)$ and patchsize $= (5, 5)$. Subsequent ablation experiments will assess the effectiveness of these parameter choices. In Section 3, as mentioned, the parameters are set as follows: $(h, w) = (128, 128)$ and $(c1, c2, c3, c4) = (96, 192, 320, 384)$.

Metrics. Recall, IoU, and precision are chosen to evaluate the model's performance. The Iou is computed as $\frac{TP}{TP+FP+FN}$, precision is computed as $\frac{TP}{TP+FP}$, recall is computed as $\frac{TP}{TP+FN}$.

Connectivity is a key indicator of road network quality. To calculate the Average Connectivity Scores (ACS), we first apply a thinning algorithm [12] to obtain the road skeleton image from both predicted and actual binary images. In these images, pixels with a value of 1 indicate roads, while those with a value of 0 represent non-road areas. A graph is then generated from the road skeleton image, with pixels representing nodes that have specific spatial coordinates. Edges connect adjacent nodes, and their weights are assigned as either 1 or the square root of 2, depending on the spatial distance between the connected nodes. To calculate the ACS of predicted images, we initially select R pairs of connected nodes (m_s, m_e) from a mask's skeleton. Next, on the predicted image, we look for points within an α radius of each node pair (m_s, m_e). If no corresponding points are found, these nodes are deemed disconnected. When matching points (p_s, p_e) are located, we measure the shortest distance between both (m_s, m_e), denoted as D_{mask}, and between (p_s, p_e), denoted as D_{pred}. A connection is established if D_{pred} is less than or equal to $(1+\theta)$ times D_{mask}; otherwise, they're considered unconnected. The definition of ACS is in the following:

$$X(p,m) = \begin{cases} 1, & if D(p_i, m_i) < \alpha \text{ and } D(p_s, p_e) \leq D(m_s, m_e) \cdot (1+\theta) \\ 0, & \text{else} \end{cases} \tag{7}$$

$$ACS = \frac{\sum_{n=1}^{N} \sum_{r=1}^{R} X(p_n^r, m_n^r)}{R \cdot N} \tag{8}$$

where p and m denote the predicted graph and mask graph, i can be s or e. The function $D(a,b)$ means the minimum distance between node a and node b. N is the number of pairs of mask and predict graph and R is the number we select pairs of nodes in a mask. Note that we set α to 10 and the error threshold θ into several values for the experiment.

Baselines. We evaluate our DFNet against five state-of-the-art fusion models in map extraction and one representative segmentation model(the first one). DlinkNet [18]: An unet-like network that combines various dilated convolutions to learn features of roads. DeepDualMapper [5]: A gated fusion network for achieving automatic fusion between satellite image and trajectory. CMMPNet [6]: A CNN-based model that applies Spatial Pyramid Pooling [30] to learn both local and global information of satellite image and trajectory. Conats [8]: A road extraction network that adopts a strategy of combining multiple losses to enhance model performance, MoviNet [10]: A hybrid model based on CNNs and ViT, designed to effectively fuse both local and global information. RCAF-Net [1]: A channel attention fusion network that exploits the feature advantages of modalities.

Table 1: The performance of DRENet on a single modality, DRENet with different fusion methods on two real-world datasets.

Models	BJRoad			Porto		
	Recall	Prec	**IoU**	Recall	Prec	**IoU**
DRENet-trajectory	0.693	0.741	0.555	0.275	0.673	0.240
DRENet-satellite	0.755	0.787	0.626	0.831	0.845	0.722
DlinkNet	0.792	0.753	0.628	0.830	0.829	0.709
RCAF-Net	-	-	0.600	-	-	-
DeepDualMapper	0.732	0.815	0.624	0.766	0.871	0.688
CMMPNet	0.769	0.797	0.639	0.789	0.848	0.692
Conats	0.764	0.792	0.637	0.832	0.839	0.717
MoviNet	0.776	0.790	0.641	0.817	0.853	0.717
DRENet(ours)	0.789	0.805	**0.660**	0.846	0.842	**0.731**

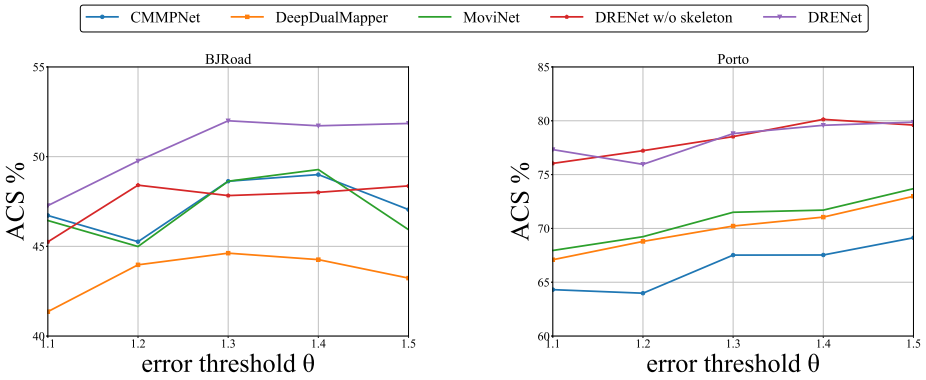
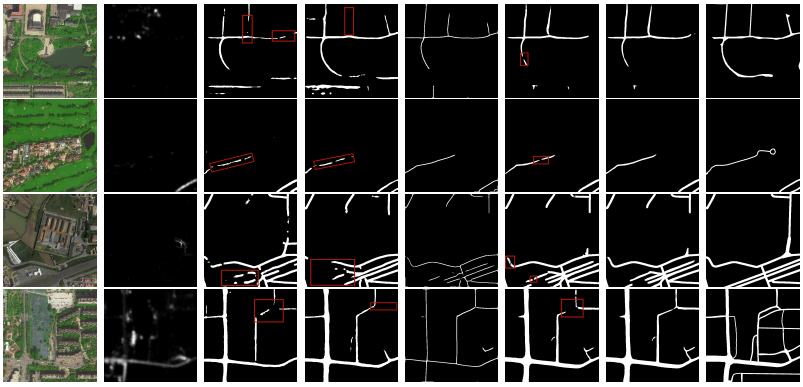


Fig. 4: The Average Connectivity Score of CMMPNet, DeepDualMapper, MoviNet, DRENet without skeleton branch, and DRENet. Note that, to maintain consistency in experiments, the outputs of DRENet without the skeleton branch and DRENet are derived from the same model parameters.

4.2 Quantitative Evaluation

We conduct experiments on the BJRoad and Porto datasets to evaluate the model's performance, and the results are listed in Table 1. Note that in map extraction tasks, the abundance of negative samples often leads to high precision. To validate the effectiveness of fusing the two modalities, we also design versions of DRENet for single modalities, named DRENet-satellite and DRENet-trajectory. The results reveal the distinct information of each dataset: The BJRoad contains a large amount of trajectory data, while the satellite images have a lot of noise. The Porto, on the other hand, contains fewer trajectories, and the road features in the satellite images are clearer.

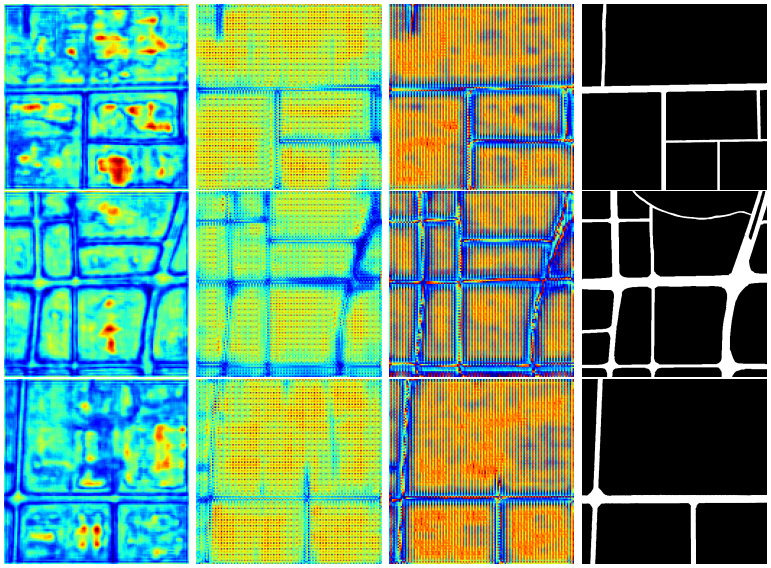
Recent state-of-the-art methods, such as DeepDualMapper, CMMPNet, Conats, MoviNet, and RCAF-Net, are largely designed with specialized fusion techniques or incorporate a global perspective to boost performance. However, most of these approaches primarily focus on segmentation performance, overlooking connectiv-



(a) sat (b) traj (c) 1 (d) 2 (e) 3 (f) 4 (g) Ours (h) map

Fig. 5: The visual results are presented, where 1, 2, 3, and 4 correspond to the outputs of CMMPNet, MoviNet, the skeleton branch of DRENet, and DRENet w/o ske, respectively. "map" refers to the ground truth. The outputs of the skeleton branch of DRENet, DRENet w/o ske, and DRENet are derived from the same model parameters to maintain consistency in experiments.

ity—a critical metric for measuring road network quality. Our model leverages a dual-branch framework, with segmentation and skeleton branches, allowing the predicted results to achieve both precise segmentation and robust connectivity. From Table 1, on the BJRoad dataset, our model significantly outperforms the latest model, MoviNet, with an IoU improvement of over 1.9%, while other models only hover around its baseline with minimal improvement. The performance gain on the Porto dataset is similarly notable, demonstrating the excellence of DRENet in road segmentation. To validate DRENet's effectiveness in connectivity performance, we selected three representative models: CMMPNet, DeepDualMapper, and MoviNet, for comparison. Additionally, we included a version of DRENet with only the single segmentation branch, referred to as DRENet without the skeleton branch (DRENet w/o ske). This setup enables us to verify the effectiveness of DRENet's dual-branch framework and the RoadReshaper module. The results are shown in Figure 4, due to the presence of significant occlusions in the satellite images, the average ACS of the BJRoad dataset is lower than that of the Porto dataset. Nevertheless, DRENet still outperforms other models in terms of ACS. On the other hand, DRENet w/o ske, lacking the additional connectivity information, shows similar performance to other models, with its metrics hovering around the baseline. In the Porto dataset, due to the minimal occlusion in the satellite image branch, connectivity is well-maintained even without the connectivity information from the skeleton branch. Compared to other models, it still achieves significantly higher ACS scores, which is attributed to the precise road extraction enabled by the Attention-based Fusion Module. The results align with the purpose of our idea and validate the rationality of our design.



(a) seg (b) ske (c) reshaped ske (d) map

Fig. 6: The feature visualization representation in the Road Reshaper Module. "seg," "ske," and "reshaped ske" represent the input from the segmentation branch, the input from the skeleton branch, and the output of the Road Reshaper Module, respectively. "map" refers to the ground truth.

4.3 Qualitative Evaluation

To validate the effectiveness of the visual results, we compare the visual outputs from CMMPNet, MoviNet, the skeleton branch of DRENet, DRENet without the skeleton branch (DRENet w/o ske), and DRENet. The results are shown in Figure 5. From the outputs of CMMPNet and MoviNet, although most areas of the roads are extracted, there are many discontinuities, especially in areas with occlusions in the satellite images. This issue makes the extracted road network challenging to apply to downstream tasks. From the outputs of the skeleton branch, this design allows the model to effectively capture road skeleton features and connectivity information, providing reliable connectivity to complement the segmentation branch. In the case of DRENet w/o ske, while fine segmentation performance is achieved, the segmentation branch struggles with road continuity compared to the skeleton branch. In contrast, the results from DRENet show that the Road Reshaper Module not only supplies connectivity information to the segmentation branch but also restores the original road width, highlighting the effectiveness of the dual-branch framework and the Road Reshaper Module in improving connectivity performance.

To verify the effectiveness of the Road Reshaper Module in restoring the shape of skeleton features, we visualize the input and output of the Road Reshaper Module. All features are summed along the channel dimension and then normalized, with cool colors representing low intensity and warm colors repre-

Table 2: Ablation studies are conducted on the Attention-based Fusion, Multi-Convolution, Dual-branch, RoadReshaper, and Random Mask methods.

Group	Atten-f	Muli-c	Dual-b	Road-r	Mask	Recall	Precision	**IoU**
a						78.2	79.27	64.58
b	✓					78.42	78.73	64.53
c	✓	✓				79.79	77.65	64.61
d	✓	✓	✓			80.33	77.77	65.00
e	✓	✓	✓	✓		78.08	80.33	65.28
f	✓	✓	✓	✓	✓	78.90	80.50	**66.02**

senting high intensity. The results are shown in the Figure 6. By comparing the input from the skeleton branch and the output of the Module, it shows that the skeleton feature has restored the shape of the road for the most part, which is attributed to the large convolutional kernels' ability to expand the features and referencing the features from the segmentation branch.

4.4 Ablation Studies

In this section, we use the BJRoad dataset to evaluate the effectiveness of each module and data augmentation technique in the model, as this dataset contains a large number of trajectories and fine-grained label segmentation. The results are shown in Table 2.

Group A serves as our baseline, using a single convolution layer in place of the fusion module. Building on this, Group B replaces the single convolution with an Attention-based Fusion Module but omits the Multi-Convolution Unit, resulting in a slight performance drop. In Group C, adding the Multi-Convolution Unit leads to improved results, especially in recall. Moving to Group D, the introduction of a Dual-branch framework allows the model to capture both segmentation and skeleton structures, enriching feature extraction during encoding and boosting overall performance. Group E further enhances connectivity in the model's output by adding the Road Reshaper, which contributes to additional performance gains. Finally, Group F incorporates random mask data augmentation, training the model to regain roads even when some features are occluded. Across Groups A to F, the IoU metric increases from 64.58% to 66.02%, validating the cumulative impact of each module on model performance.

The random mask has two parameters: the random mask ratio for the roads and the size of the mask patches. Since the proportion of occlusions in satellite images is very low in the Porto dataset, we set the parameters for this dataset to ratio=(0, 5%) and patch size=(5,5). For the BJRoad dataset, we experiment with different combinations of these two parameters to determine the optimal configuration. The results and process of the random mask in shown in Figure 7. Based on performance, we selected ratio=(5%, 15%) and patch size=(5, 5) as the optimal combination for the BJRoad dataset.

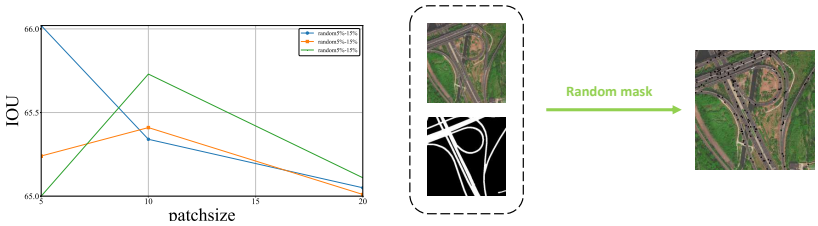


Fig. 7: Performance under different parameters and examples of random mask.

5 Conclusions

In this paper, we proposed a Dual-branch Road Extraction Network that achieves both precise road segmentation and fine connectivity. The model contains a segmentation branch and a skeleton branch, which focuses on fine-grained road segmentation and the skeleton structure of roads, respectively. An Attention-based Fusion Module is incorporated to dynamically capture the relationship between modalities for effective fusion, and a Road Reshaper Module supplements connectivity information from the skeleton branch to the segmentation branch. Our model demonstrated state-of-the-art performance in both road segmentation and connectivity on two real-world datasets.

Due to the limited availability of relevant public datasets, obtaining satellite images with corresponding raw trajectories and precise labels remains challenging. In the future, to utilize additional trajectory information such as speed and time, we plan to focus on automatic annotation for road network data, enabling the fusion of satellite images and raw trajectory data for broader applications.
Acknowledgement. This work was supported by the National Natural Science Foundation of China under Grant No. 62102277 and Natural Science Foundation of Jiangsu Province under Grant No. BK20210703.

References

1. Xu Y, Shi Z, Xie X, et al. Residual Channel Attention Fusion Network for Road Extraction Based on Remote Sensing Images and GPS Trajectories[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024.
2. Roy S K, Deria A, Hong D, et al. Multimodal fusion transformer for remote sensing image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-20.
3. Li J, Li Y, He L, et al. Spatio-temporal fusion for remote sensing data: An overview and new benchmark[J]. Science China Information Sciences, 2020, 63: 1-17.
4. Li J, Hong D, Gao L, et al. Deep learning in multimodal remote sensing data fusion: A comprehensive review[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 112: 102926.
5. Wu H, Zhang H, Zhang X, et al. DeepDualMapper: A gated fusion network for automatic map extraction using aerial images and trajectories[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(01): 1037-1045.

6. Liu L, Yang Z, Li G, et al. Aerial images meet crowdsourced trajectories: a new approach to robust road extraction[J]. IEEE transactions on neural networks and learning systems, 2022, 34(7): 3308-3322.
7. Yang J, Ye X, Wu B, et al. DuARE: Automatic road extraction with aerial images and trajectory data at Baidu maps[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 4321-4331.
8. Chen Z, Fang J, Chao P, et al. Conats: A novel framework for cross-modal map extraction[C]//International Conference on Web Information Systems Engineering. Cham: Springer International Publishing, 2022: 503-518.
9. Si C, Yu W, Zhou P, et al. Inception transformer[J]. Advances in Neural Information Processing Systems, 2022, 35: 23495-23509.
10. Chen Z, Fang J, Chao P, et al. MoviNet: A novel network for cross-modal map extraction by vision transformer and CNN[J]. Knowledge-Based Systems, 2023, 278: 110890.
11. Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
12. Zhang T Y, Suen C Y. A fast parallel algorithm for thinning digital patterns[J]. Communications of the ACM, 1984, 27(3): 236-239.
13. Chen C, Lu C, Huang Q, et al. City-scale map creation and updating using GPS collections[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016: 1465-1474.
14. Stanojevic R, Abbar S, Thirumuruganathan S, et al. Robust road map inference through network alignment of trajectories[C]//Proceedings of the 2018 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2018: 135-143.
15. Mariescu-Istodor R, Fränti P. CellNet: Inferring road networks from GPS trajectories[J]. ACM Transactions on Spatial Algorithms and Systems (TSAS), 2018, 4(3): 1-22.
16. Zourlidou S, Sester M. Intersection detection based on qualitative spatial reasoning on stopping point clusters[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; XLI-B2, 2016, 41: 269-276.
17. Terrell G R, Scott D W. Variable kernel density estimation[J]. The Annals of Statistics, 1992: 1236-1265.
18. Zhou L, Zhang C, Wu M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018: 182-186.
19. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015: 234-241.
20. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
21. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.
22. Mehta S, Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer[J]. arXiv preprint arXiv:2110.02178, 2021.

23. Sun T, Di Z, Che P, et al. Leveraging crowdsourced GPS data for road extraction from aerial imagery[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7509-7518.
24. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.
25. Jiang X, Li Y, Jiang T, et al. RoadFormer: Pyramidal deformable vision transformers for road network extraction with remote sensing images[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 113: 102987.
26. Shao R, Du C, Chen H, et al. SUNet: Change detection for heterogeneous remote sensing images from satellite and UAV using a dual-channel fully convolution network[J]. Remote Sensing, 2021, 13(18): 3750.
27. Dai L, Zhang G, Zhang R. RADANet: Road augmented deformable attention network for road extraction from complex high-resolution remote-sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-13.
28. Zhang Z, Liu Q, Wang Y. Road extraction by deep residual u-net[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(5): 749-753.
29. Pan X, Shi J, Luo P, et al. Spatial as deep: Spatial cnn for traffic scene understanding[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
30. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
31. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
32. He S, Bastani F, Jagwani S, et al. Sat2graph: Road graph extraction through graph-tensor encoding[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. Springer International Publishing, 2020: 51-67.
33. Li Y, Hou Q, Zheng Z, et al. Large selective kernel network for remote sensing object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 16794-16805.
34. Guo M H, Lu C Z, Hou Q, et al. Segnext: Rethinking convolutional attention design for semantic segmentation[J]. Advances in Neural Information Processing Systems, 2022, 35: 1140-1156.
35. Yuan Y, Xie J, Chen X, et al. Segfix: Model-agnostic boundary refinement for segmentation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer International Publishing, 2020: 489-506.
36. He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16000-16009.