FSNet: Frequency-Spatial Joint Learning for Road Extraction From Remote Sensing Images

Huiguang Yao, Zhigang Yang, Qiang Li, *Member, IEEE* Qi Wang, *Senior Member, IEEE*

***Abstract*—Road extraction from remote sensing images (RSIs) plays an important role in a wide range of real-world applications. The primary challenges stem from the occlusions as well as the high visual similarity between road and background which complicate the identification particularly in complex scenes. Existing methods predominantly focus on enhancing the feature representations through sophisticated designs and tend to overlook the inherent sensitivity of spatial features. To alleviate this issue, we propose a novel network that jointly explores representations in both the *frequency* and *spatial* domains, termed FSNet. This network comprises two key components: the Joint-Domain Enhancement Module (JDEM) and Cross-Domain Hybrid Parser Module (CDPM). Specifically, the JDEM utilizes Mamba within the frequency domain to capture global relationship across different spectral bands, which helps alleviate road occlusions. Accordingly, the CDPM separately parses frequency and spatial features to fully leverage the strengths of each and then effectively integrates them to improve overall performance. Experimental results on publicly available datasets demonstrate that FSNet surpasses most previous methods in Intersection over Union(IoU) and F1-score, which indicate that our FSNet can generate road results with superior connectivity and accuracy. The source code is publicly available at https://github.com/jaybryant1/FSNet.**

***Index Terms*—Mamba, fourier transform, remote sensing, road extraction**

I. Introduction

ROAD extraction is a critical research topic in the field of remote sensing, which aims to automatically and accurately locate roads and distinguish them from background. This technology makes significant contributions to numerous facets of daily life, including traffic management, autonomous driving [1], intelligent transportation [2]. Nevertheless, road extraction from RSIs [3] still faces a series of severe challenges: 1) **Shape Diversity.** The width and orientation of roads vary greatly across different regions. For instance, highways typically feature broad, straight layouts whereas rural roads tend to be narrower and may even form complex intersecting or circular structures. This diversity requires the model to have a strong generalization ability to detect roads in various scenarios. 2) **Appearance Similarity.** The texture of roads often exhibit a high degree of similarity to that of surroundings such as building rooftops and rivers. This resemblance can result in false positives, which severely interferes with the precise detection of roads. 3) **Road Discontinuity.** Roads may be obscured by buildings and trees, which disrupts the continuity greatly. Therefore, the model needs to effectively incorporate contextual information to predict occluded regions properly. These factors make it challenging for traditional methods to achieve accurate and reliable results.

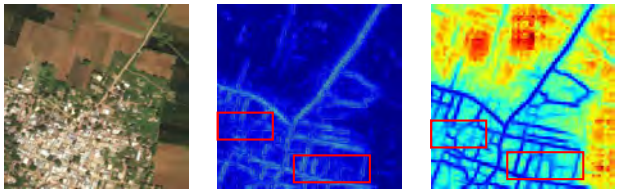
In recent years, the rapid advancement of deep learning has provided new approaches [4]–[10] to address aforementioned challenges. With their powerful feature extraction and contextual modeling capabilities [11]–[13], deep learning-based methods have boosted road segmentation to a higher level. Existing methods can be broadly categorized into two categories: segmentation-based methods and graph-based methods. The former regards road extraction as a binary semantic segmentation task, which leverages the encoder to capture deep feature representations and the decoder to progressively assign a class to each pixel. For instance, Zhou et al. [14] introduce dilated convolution to expand the receptive field while preserving spatial details, which improves the accuracy of road recognition effectively. To further improve long-range modeling ability, Yang et al. [15] propose a global-local context perception module which aims to alleviate the road discontinuity. In the decoder stage, Wang et al. [16] introduce an additional decoder to better recover the detailed information. Although segmentation-based methods focus on pixel-level accuracy, they still struggle with road connectivity. To address this issue, graph-based methods treat road as a graph structure, where intersections or endpoints are defined as nodes and road segments between these nodes are represented as edges. For example, Zao et al. [17] propose a framework with three branches: vertex, orientation and segmentation. The model can obtain more coherent results through the cooperation of these branches. We note that aforementioned methods primarily center on single spatial features. While these features are beneficial for road extraction, over-reliance on spatial features can cause interference due to the high similarity between road and complex background, which ultimately degrades the accuracy of recognition. Therefore, It's essential to develop novel methods for feature representation to overcome the inherent limitations of spatial features and achieve more accurate results.

As illustrated in the Fig. 1, surface textures of roads frequently exhibit repetitive patterns and show continuity in specific direction. For example, rural paths typically feature consistent low-frequency characteristics. In contrast, road boundaries display strong pixel variations, represented as high-frequency components in the image. Consequently, incorporat-

Zhigang Yang, Qiang Li and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. (e-mail: zgyang@mail.nwpu.edu.cn, liqmges@gmail.com, crabwq@gmail.com)

Huiguang Yao is with the School of Computer Science, and with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P. R. China. (e-mail: yhg2655@mail.nwpu.edu.cn)

Corresponding author: Qi Wang.



(a) (b) (c)

Fig. 1. The motivation of our proposed FSNet. (a) the input image, (b) the spatial feature map, (c) frequency-enhanced feature map. As highlighted by the red box, compared to (b), the feature map enhanced by frequency effectively predicts occluded road segments in complex scenarios, thereby resulting in a more continuous road structure. Overall, the road features in (c) appear more coherent and exhibit a stronger contrast with background.

ing frequency information into road extraction offers a more effective strategy. Recently, frequency features derived from Fourier transform have been proven to be beneficial for feature extraction [18], [19], which can break the bottleneck of spatial features and provide richer global information. For instance, Zhao et al. [20] leverage Fourier transform to extract frequency components and employ the cross-attention mechanism to enable high-low frequency interaction. This allows the model to focus on fine-grained details, which significantly improves the generation quality. However, this method incurs considerable computational overhead when handling high-resolution remote sensing images. Tatsunami et al. [21] introduce a dynamic global filter in the frequency domain, which helps the model to comprehend diverse image content. Nonetheless, its overemphasis on frequency information at the expense of spatial features limits overall performance to some extent.

To alleviate the above-mentioned issues, a road extraction framework based on frequency-spatial perception is proposed, termed FSNet. During the encoder stage, the network initially employs ResNet34 [22] to extract spatial features. Subsequently, the joint-domain enhancement module is utilized to strengthen the ability to model long-range dependencies in the frequency domain while maintaining linear complexity, which effectively combines the advantages of both spatial and frequency features. In the following decoder stage, the network decouples the encoded spatial-frequency components separately through the cross-domain hybrid parser module. These both characteristics are then integrated adaptively to yield accurate results.The main contributions of this paper are as follows:

- We propose a frequency-spatial perception network for road extraction, which effectively integrates global frequency and local spatial features to improve the performance of road extraction from RSIs. Our proposed FSNet achieves comparable results on publicly available road datasets.
- To enhance the feature representation of both frequency and spatial components, the joint-domain enhancement module is proposed. This module adaptively fuses spatial and frequency information, which facilitates a deeper extraction of road-specific features.
- To overcome the inherent limitations of spatial features, the cross-domain hybrid parser module is introduced. This module parses the frequency and spatial features, thereby enabling the model to identify long-range dependencies. The disentanglement facilitates the recover of fine-grained details and ultimately improves the continuity.

The rest of this paper is organized as follows. Section II discusses related work on Mamba and road extraction from RSIs. In Section III, we provide a detailed description of the proposed FSNet. Section IV presents an analysis of the experimental results. Finally, Section V offers the conclusion.

II. Related Work

In this section, we review state space models and existing methods for road extraction from remote sensing images.

A. Mamba

In more recent years, state-space models (SSMs) [23], [24] have garnered considerable attention due to linear complexity with respect to sequence length, demonstrating remarkable efficiency advantages in long-range modeling. Mamba [25] introduces a data-dependent SSM layer and establishes a backbone for general sequence model. Compared to mainstream Transformers [26], Mamba achieves superior performance across several large-scale real-world datasets while maintaining efficiency. Inspired by the success of Mamba in natural language processing (NLP), VMamba [27] utilizes the VSSBlock equipped with two-dimensional selective scanning (SS2D) module to develop a vision framework. This module bridges the gap between the sequential nature of one-dimensional selective scanning and the non-sequential structure of two-dimensional visual data. Moreover, PlainMamba [28] ensures that each visual token remains adjacent to the previously scanned token, which effectively preserves spatial consistency and semantic continuity. To date, Mamba has been successfully applied to various downstream tasks, including image classification [29], image segmentation [30], and image restoration [31], [32].

Although Mamba has shown excellent performance in processing natural images, it encounters challenges when applied directly to RSIs. This is largely due to variations in imaging angles, which make spatial features to be more dispersed in high-resolution remote sensing images, whereas features in natural images are predominantly aligned along the vertical direction. To address this issue, RS3Mamba [33] introduces the VSS module to construct an auxiliary branch, which provides additional long-range modeling ability to the convolution-based main branch. In contrast to RS3Mamba, CM-UNet [34] integrates Mamba into the decoder, which models long-distance correlations and multi-scale global contextual information, thereby facilitating efficient parsing of intricate details in. However, these methods mainly focus on utilizing Mamba to capture long-distance dependencies in the spatial domain. Due to the similarity in textures between road and background, spatial features often struggle to differentiate these subtle variations. Considering that frequency features can complement the shortcomings of spatial characteristics, we firstly utilize Fourier transform to convert the image into

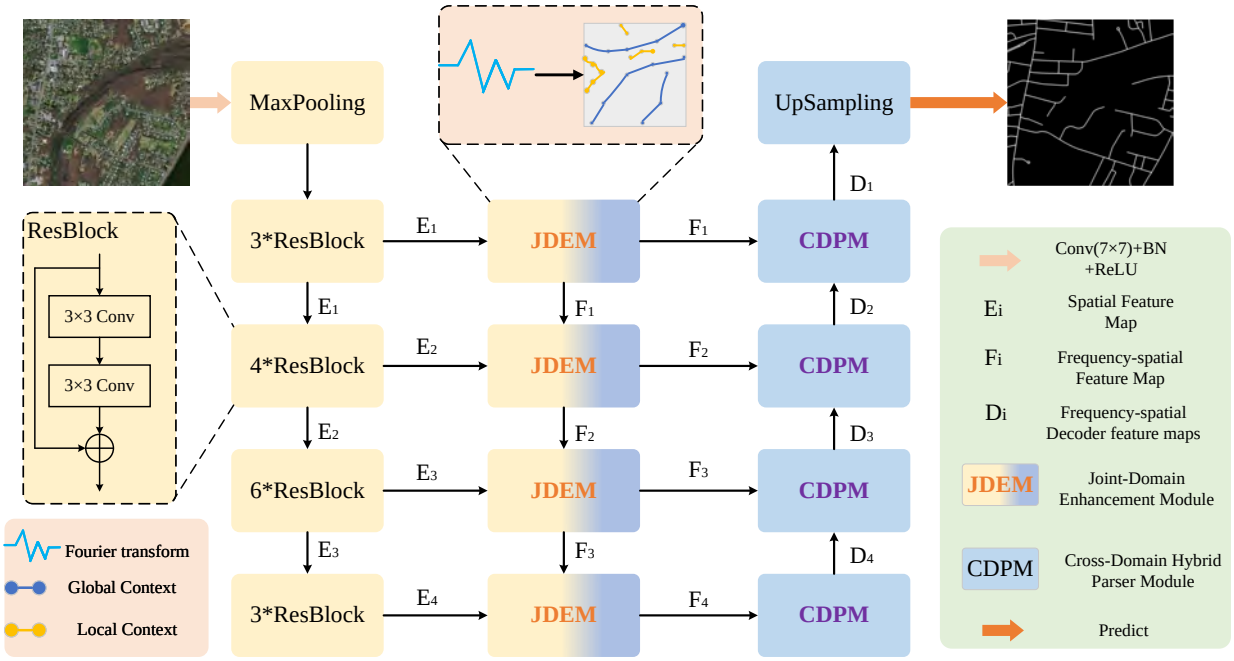


Fig. 2. The overall framework of the proposed FSNet.

the frequency domain, and leverage Mamba in the frequency domain to model global relationships, thereby resolving road occlusions effectively.

B. Road Extraction From RSIs

In recent years, there has been a surge in deep learning-based methods for road extraction from RSIs, which primarily focus on optimizing the encoder and introducing auxiliary tasks to raise the performance. For instance, He et al. [22] incorporate residual connection into the encoder, which facilitates the training of deeper neural networks. Inspired by this, Zhang et al. [35] combine the strengths of U-Net [36] and ResNet [22], thereby achieving higher accuracy while reducing the parameters. To expand the receptive field, Zhou et al. [14] employ multi-scale dilated convolutions at the final encoder stage which allows the model to capture broader contextual information. Building on this, Yang et al. [37] introduce dilated strip convolution to further boost road connectivity. Considering the distinctive geometric characteristics of roads, Sun et al. [38] replace traditional square convolution with one-dimensional convolution in four direction. Moreover, Yang et al. [39] employ self-attention module to enhance the ability to model long-range dependencies. Building upon this, Hu et al. [40] utilize deformable self-attention module for more comprehensive extraction of road features. In addition, Multi-task joint training helps in learning more robust representations. For example, Mei et al. [41] design a connectivity branch that predicts connections between adjacent pixels, which alleviates road fragmentation caused by occlusions. Beyond task-specific models, recent advancements have introduced foundation models and cross-domain benchmarks to enhance generalizability. For instance, SpectralGPT [42] explores generative pre-training for spectral data, providing powerful universal representations. While it excels in capturing general spectral characteristics, our FSNet focuses on decoupling frequency signals specifically to resolve the geometric continuity of linear road structures. Furthermore, benchmarks like CrossCity Matters [43] emphasize the importance of robust semantic segmentation across varying urban landscapes. Our frequency-domain approach aligns with this goal by leveraging domain-invariant frequency features to mitigate the texture variations highlighted in such cross-city scenarios.

Recently, frequency features based on Fourier transform has been widely applied in computer vision tasks [44], [45]. Road surfaces usually exhibit minimal variation and maintain coherence along a specific direction, primarily presented as low-frequency features. In contrast, the edges show significant differences from background, manifesting as high-frequency features. Based on above analysis, we integrate frequency information into road extraction and adaptively fuse features from both frequency and spatial domains to jointly represent the road target, which ultimately results in satisfactory performance.

III. Proposed Method

A. Framework

The overall framework of FSNet is shown in Fig. 2, which integrates the advantages of frequency and spatial features effectively. Specifically, we firstly employ the ResNet34 pre-trained on the ImageNet dataset as the encoder to extract multi-level features from the spatial domain, which generates four feature maps with different resolutions(E1-E4). Here, E1 focuses on the low-level texture details, while E4 emphasizes the high-level semantic features. These hierarchical feature maps are then sequentially fed into the joint-domain enhancement module, which innovatively integrates the Fourier

transform with the long-range dependency modeling capabilities of Mamba. The enhanced frequency features are fused with the original spatial features to acquire refined feature maps(F1-F4). In the decoder stage, the enhanced features are interpreted into the cross-domain hybrid parser module. Within this module, both local spatial features and global frequency information are parsed separately and then adopts an effective fusion strategy to merge their strengths. It ultimately generates a prediction map with the same resolution as the input image, which allows for comprehensive extraction of the road structure.

B. Joint-Domain Enhancement Module

Roads often exhibit high visual similarity to background such as lakes and rivers, while certain road pixels are occluded by surrounding trees or buildings. These obstacles make it difficult to extract road structure completely, especially when relying solely on shallow features like shape and texture. Note that boundaries typically exhibit notable pixel variations compared to adjacent regions, constituting high-frequency components in the image. In contrast, surfaces often feature repetitive patterns with subtle variations, represented as low-frequency components. Therefore, it is meaningful to capture the relationship between roads and their surroundings in the frequency domain, which can boost the feature representations and compensate for the drawbacks of spatial features.

Considering that convolutional neural networks expert in extracting local detailed features while Mamba is renowned for its outstanding performance in long-range modeling with linear complexity, we propose the joint-domain enhancement module to fully leverage the merits of both. As shown in Fig. 3, multi-scale low frequency submodule is applied to capture local details, while long-range high frequency submodule enhances the understanding of global context. Subsequently, high-frequency and low-frequency information are adaptively fused to produce comprehensive feature representations. This module not only benefits the overall performance but also maintains computational efficiency, which delivers an effective solution for road extraction in complex scenarios.

Long-range High Frequency Submodule: Given the spatial feature map, Fourier transform is leveraged to extract boundary features while filters out low-frequency components. Subsequently, Mamba is utilized to perform long-range modeling on the high-frequency features. Through modeling the global relationship between roads and their neighbors, the model can gain a more comprehensive understanding of object relationships in complex scenes. Moreover, long-distance context helps predict the occluded areas more precisely, which can improve the coherence in handling complex scenarios.

Multi-scale Low Frequency Submodule: Given the spatial feature map at a specific stage, it is firstly downsampled by the PixelUnshuffle operation to extract the low-frequency features F_n^{down} for road surface. Considering the special geometric shape of roads, we enhance the low-frequency features by using strip convolutions in horizontal and vertical directions. Moreover, since the width of roads varies across different regions, thereby incorporating multi-scale information can bolster the robustness. Therefore, we leverage convolution operations of diverse size in different channels to capture multi-scale information from the image, i.e.

$$F_{n,1}, F_{n,2}, F_{n,3}, F_{n,4} = split(F_n^{down}), \tag{1}$$

$$F_{n,i}^h = DWConv_{1\times kernel_i}(F_{n,i}), \tag{2}$$

$$F_{n,i}^v = DWConv_{kernel_i\times 1}(F_{n,i}), \tag{3}$$

$$F_{n,i}^E = F_{n,i}^h + F_{n,i}^v, \tag{4}$$

$$F_n^E = Concat\left(F_{n,1}^E, F_{n,2}^E, F_{n,3}^E, F_{n,4}^E\right), \tag{5}$$

where the $split(\cdot)$ operation divides F_n^{down} along the channel dimension into four groups. $DWConv(\cdot)$ refers to depthwise separable convolutions, where $kernel_i$ denotes the kernel size for the i-th group, with sizes of 3, 5, 7 and 9 respectively. $Concat(\cdot)$ represents the concatenation of the enhanced feature maps along the channel dimension.

Subsequently, we use the PixelShuffle operation to restore the resolution. Finally, the enhanced high-frequency and low-frequency features are adaptively fused with the spatial features, which can fully leverage the advantages of both spatial features and frequency information, i.e.

$$F_n = F_n^H \otimes E_n + F_n^L \otimes E_n, \tag{6}$$

where F_n^H represents high-frequency features, F_n^L represents low-frequency features, E_n denotes spatial features, $\otimes$ refers to element-wise multiplication operation.

C. Cross-Domain Hybrid Parser Module

As shown in Fig. 4, the feature map obtained from the joint-domain enhancement module contains both spatial and frequency information, both of which play a crucial role in representing features. Therefore, we propose the cross-domain hybrid parser module to parse these two types of features separately, which helps minimize mutual interference and preserve the continuity of roads better.

Spatial Parsing Submodule: Spatial features typically incorpate image details such as edges, textures and structures, which facilitates the accurate localization of road contours and boundaries. Therefore, we use 1×1 convolutions and transpose convolutions to adjust the spatial and channel dimensions of the feature map, obtaining the parsed spatial features:

Frequency Parsing Submodule: Spatial features are easy to be influenced by background noise and occlusions, while frequency information typically reveals global patterns and long-term relationships. For example, low-frequency components often represent the overall texture and coherent structure within an image, while high-frequency components are primarily utilized to capture edges and areas with notable changes. Moreover, frequency information provides global context, which makes it more effective to handle complex background and occlusion phenomena. Therefore, we parse high-frequency and low-frequency features separately. Specifically, we leverage average pooling on the spatial feature map F_i^s, which can preserve the overall structure and contour information while eliminate noise. In the meantime, we can

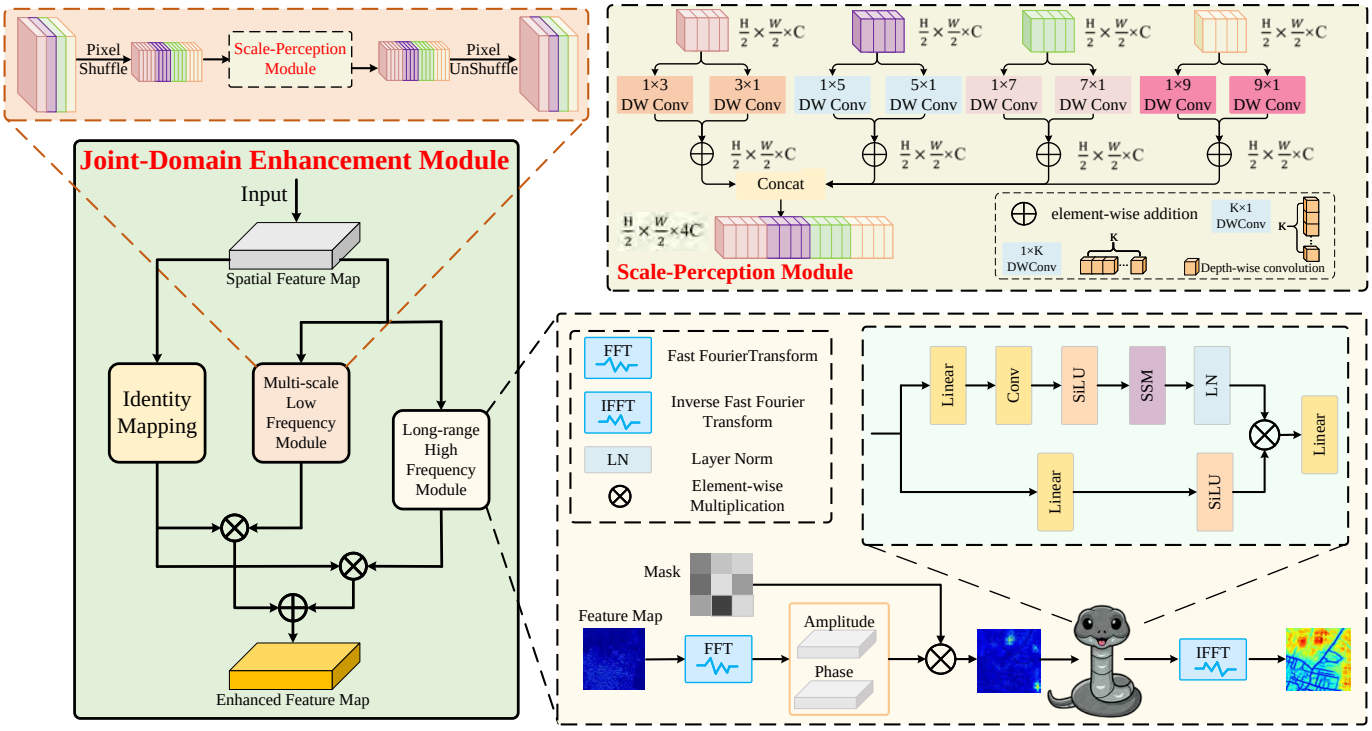


Fig. 3. Illustration of Joint-Domain Enhancement Module.

quickly derive high-frequency features by removing the low-frequency components from the input feature map, i.e.

$$F_i^L = Avg(F_i^s), \tag{7}$$

$$F_i^H = F_i^s - Up(F_i^L). \tag{8}$$

where $Avg(\cdot)$ represents average pooling operation, $Up(\cdot)$ represents Upsampling operation, bilinear interpolation is employed in this paper.

Compared to Fourier-based methods, the difference-based approach for high-frequency feature extraction preserves important boundary features and maintains computational efficiency. Since the spatial parsing submodule solely concentrates on local regions, the model faces limitations in capturing global context. Therefore, we leverage Mamba to perform global modeling for frequency features, which can ensure the full exploitation of contextual information. Finally, the parsed high-frequency and low-frequency features are fused with the spatial features, i.e.

$$F_i^{fusion} = Conv_{1\times1}(Concat(F_i^s, Up(F_i^L), F_i^H)), \tag{9}$$

$$F_i^{out} = Up(F_i^{fusion}), \tag{10}$$

where $Concat(\cdot)$ represents concatenation along the channel dimension, F_i^{fusion} denotes the feature map that integrates both spatial features and frequency information, F_i^{out} represents the output of this module. $Conv_{1\times1}(\cdot)$ represents point-wise convolution, $Up(\cdot)$ represents upsampling operation.

IV. Experiments

In this section, we conduct extensive experiments on road datasets to validate the effectiveness of our FSNet.

A. Datasets

We conduct experiments on the publicly available road datasets: DeepGlobe [46] and SpaceNet to evaluate the performance. As described in [15], the DeepGlobe dataset is divided into 5,500 image pairs for training and 726 image pairs for testing. The SpaceNet dataset is made up of 2,780 aerial images, and resized to 1024×1024 pixels for our experiments. According to [3], with 2,224 image pairs used for training and 556 for testing.

B. Evaluation Metrics

Some segmentation evaluations are adopted in this paper, including Recall, Precision, IoU, and F1-score. These metrics are computed by

$$Recall = \frac{TP}{TP + FN}, \tag{11}$$

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

$$\text{IoU} = \frac{TP}{TP + TN + FP}, \tag{13}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}, \tag{14}$$

where *TP*, *FP*, *TN* and *FN* represent the number of true positives, false positives, true negatives, and false negatives respectively.

Among these indicators, IoU and F1-score can evaluate the quality of generated results comprehensively. All metrics are described as percentages(%).

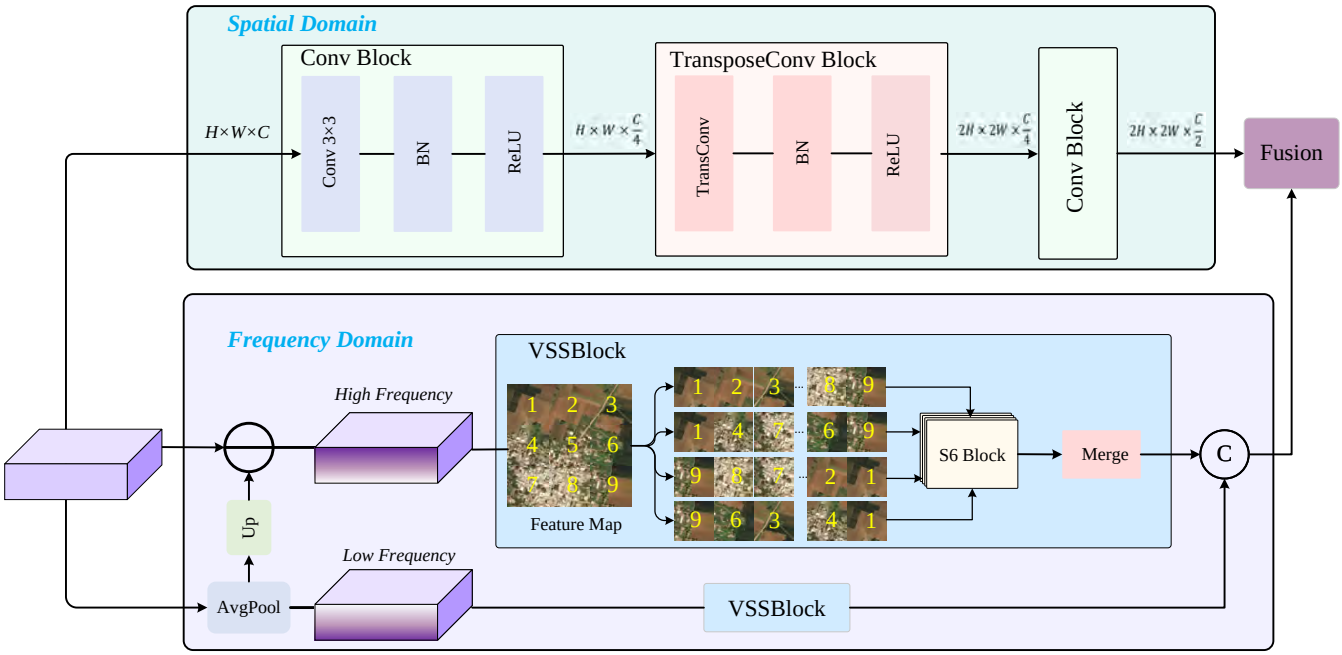


Fig. 4. Illustration of Cross-Domain Hybrid Parser Module.

C. Experimental Settings

The model is based on PyTorch framework, and all the experiments are conducted on NVIDIA 4090 GPU. The Adam is employed as the optimizer, and the combination of binary cross entropy with Dice coefficient is used as the loss function. The initial learning rate of the model is set to 2e-4. The batchsize is set to 4 on the DeepGlobe, and SpaceNet dataset. During the training process, if the loss does not decrease for continuous 3 epoch, the learning rate adjusts to one-fifth the current value. The model terminates early if the loss fails to decrease for continuous 6 epoch. The test images are processed by using the test time augmentation during the testing phase.

D. Comparison with Existing Methods

To illustrate validity of the designd method, several existing methods are selected and compared with the proposed model on three datasets, including Unet3+ [47] , GCBNet [48], SGCNNet [49], CMTFNet [50], CMLFormer [51], OARENet [52], Scaleformer [53], CU-dGCN [54] and UGDNet [55]

Results on the DeepGlobe dataset: Table I illustrates the performance of road extraction methods on the DeepGlobe dataset. It is evident that our proposed FSNet achieves significant improvements over previous approaches across all evaluation metrics except Recall. Specifically, while Unet3+ integrates multi-scale features, it also introduces information loss, ultimately leading to decline of 5.53% in IoU. Moreover, our method outperforms CMTFNet and SGCNNet by a considerable margin in Recall, which highlights the significance of interactions between frequency and spatial features in extracting comprehensive road topology. Furthermore, models with a larger receptive field obtain higher IoU, such as OARENet and CMLFormer. Meanwhile, the UGDNet achieves the second-best performance, which reflects the necessity of extracting long-distance dependencies. The proposed FSNet achieves satisfactory overall performance by incorporating JDEM and

TABLE I
ROAD EXTRACTION RESULTS ON THE DEEPGLOBE ROAD DATASET. THE BEST RESULTS ARE BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Method	Recall↑	Precision↑	IoU↑	F1-score↑
Unet3+ [47]	79.66	78.51	64.77	76.76
GCBNet [48]	<u>84.90</u>	78.64	68.65	80.53
SGCNNet [49]	84.57	76.00	66.45	78.73
CMTFNet [50]	76.03	78.71	61.68	74.76
CMLFormer [51]	84.84	79.31	69.40	80.94
OARENet [52]	84.35	79.54	68.98	80.76
Scaleformer [53]	81.01	79.10	66.20	78.63
CU-dGCN [54]	79.65	**80.95**	67.13	80.14
UGDNet [55]	**85.60**	79.49	<u>70.00</u>	<u>81.49</u>
Ours	84.81	<u>80.73</u>	**70.30**	**81.61**

TABLE II
ROAD EXTRACTION RESULTS ON THE SPACENET ROAD DATASET.

Method	Recall↑	Precision↑	IoU↑	F1-score↑
Unet3+ [47]	67.06	<u>64.73</u>	52.24	64.88
GCBNet [48]	70.49	64.15	53.91	65.85
SGCNNet [49]	67.38	63.38	51.85	63.87
CMTFNet [50]	**73.34**	62.73	54.49	<u>66.44</u>
CMLFormer [51]	70.44	64.22	54.06	65.90
OARENet [52]	70.85	64.52	<u>54.50</u>	66.33
Scaleformer [53]	64.81	64.70	51.06	63.17
CU-dGCN [54]	69.76	64.49	53.46	65.49
UGDNet [55]	66.17	**67.25**	53.76	65.41
Ours	<u>73.30</u>	62.86	**54.79**	**66.60**

TABLE III
ABALTION RESULTS ON THE DEEPGLOBE ROAD DATASET AND THE SPACENET ROAD DATASET.

	Components			DeepGlobe				SpaceNet			
Method	Baselline	FEEM	FEDM	Recall	Precision	IoU	F1-score	Recall	Precision	IoU	F1-score
ModelA	✓			82.42	**80.96**	68.75	80.42	**74.24**	61.10	53.48	65.72
ModelB	✓	✓		84.95	79.49	69.44	81.08	73.05	62.74	54.48	66.41
ModelC	✓		✓	**85.04**	79.49	69.38	80.97	73.60	61.77	54.03	66.09
FSNet	✓	✓	✓	84.81	80.73	**70.30**	**81.61**	73.30	**62.86**	**54.79**	**66.60**

CDPM, which can extract low frequency information and global high frequency features adaptively. It ensures that even occluded road segments are accurately identified and connected through global context modeling, thereby producing results with superior connectivity.

Results on the SpaceNet dataset: As shown in Table II, FSNet demonstrates superior performance compared to other road extraction methods, outperforming the second-best method by 0.30% in the IoU, which indicates the comprehensive performance of the model. Note that although FSNet fails to achieve the highest scores in Precision and Recall individually, it attains the best performance in the F1-score. This shows that the model strikes optimal balance between Precision and Recall, thereby providing a harmonious trade-off between prediction accuracy and completeness. Meanwhile, CMTFNet achieves the best performance in terms of the Recall metric, which fully underscores the necessity of integrating multi-scale information and long-range modeling to enhance prediction completeness. However, its poor Recall performance within the DeepGlobe dataset indicates that relying solely on spatial information is insufficient for better robustness and generalization. The comparison with UGDNet also reveals that FSNet maintains a higher Recall, indicating a lower rate of missed road segments.

E. Visual Results

To provide a more intuitive analysis of the effectiveness of the proposed method, we present several representative visualization results from two datasets separately.

Results on the DeepGlobe dataset: We analyze experimental results from the perspective of both road precision and completeness. As shown in Fig. 5, FSNet can produce complete and reasonable road results.

Results on completeness: The qualitative results demonstrate that, in comparison to competing methods, our proposed approach not only captures road features obscured by trees and buildings with higher fidelity but also generates more coherent and continuous road networks. This superior performance is primarily attributed to the integration of the Mamba architecture for long-range dependency modeling. By effectively leveraging global pixel-wise relationships, our model can precisely reconstruct occluded road information. Specifically, in the red-boxed regions (rows2-4), the roads are heavily obstructed by surrounding trees, which leads to discontinuities in prediction results. Notably, SGCN achieves commendable performance in these occluded areas due to integration of graph convolutional networks. Furthermore, the example in the bottom-left of the final row illustrates a distinct challenge: the high visual similarity between road surfaces and adjacent building rooftops. This ambiguity leads to erroneous predictions, characterized by redundant segments or missing sections, Therefore, it is crucial to integrate frequency information to obtain more accurate results.

Results on precision: In the first row of the images, the predictions generated by FSNet are most similar to the ground truth. Specifically, in the upper region of the image, the model is still able to extract accurate road results within the challenging scenario surrounded by numerous buildings and trees. Moreover, in the lower-right area of the image, some other methods mistakenly identify background information as roads. The aforementioned analysis demonstrates that our model can accurately extract road features even in highly complex scenarios.

Results on the SpaceNet dataset: As shown in Fig. 6, compared to existing methods, our model achieves more accurate and complete predictions of road in complex scenarios such as residential areas and parking lots. This superior performance is primarily attributed to the proposed JDEM. By leveraging Mamba for long-range modeling within the frequency domain, the JDEM effectively addresses the issue of road occlusions while also extracting hidden road regions, such as the spaces between vehicles in parking lots.

F. Ablation Study

We conduct ablation experiment on the Deepglobe and SpaceNet dataset to demonstrate the influence of JDEM and CDPM, As shown in Table III. Three variants that combination of different parts are defined as ModelA (only baseline), ModelB (combine baseline with JDEM), and ModelC (combine baseline with CDPM) respectively.

Effect of JDEM: The introduction of the joint domain enhancement module allows to perform long-range modeling on frequency features during the encoder stage, which effectively addresses challenges posed by road occlusions and compensates for the inherent locality limitations of purely spatial features. This results in more complete results on the DeepGlobe dataset. Specifically, compared to ModelA, our model achieves a 2.53% increase in Recall, though with a slight decline in Precision. Meanwhile, the F1-score improves by 0.66%, indicating a better balance between Precision and Recall. Additionally, the IoU increases by 0.69%, which demonstrates that the incorporation of JDEM contributes to the overall performance of the model. As for the SpaceNet dataset, both the IoU and F1-score show substantial improvements over

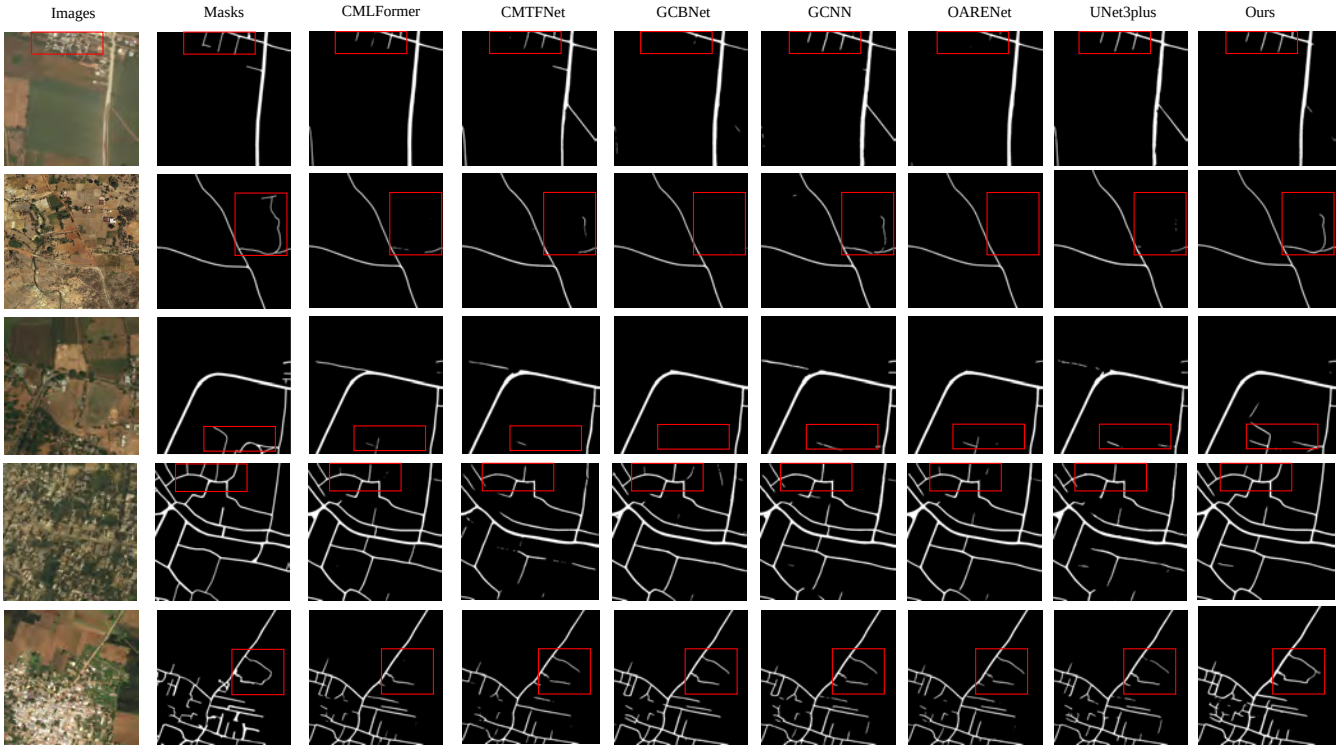


Fig. 5. Qualitative evaluations between FSNet and comparison methods on DeepGlobe road datasets. The red boxes mark the areas where FSNet outperforms other methods.

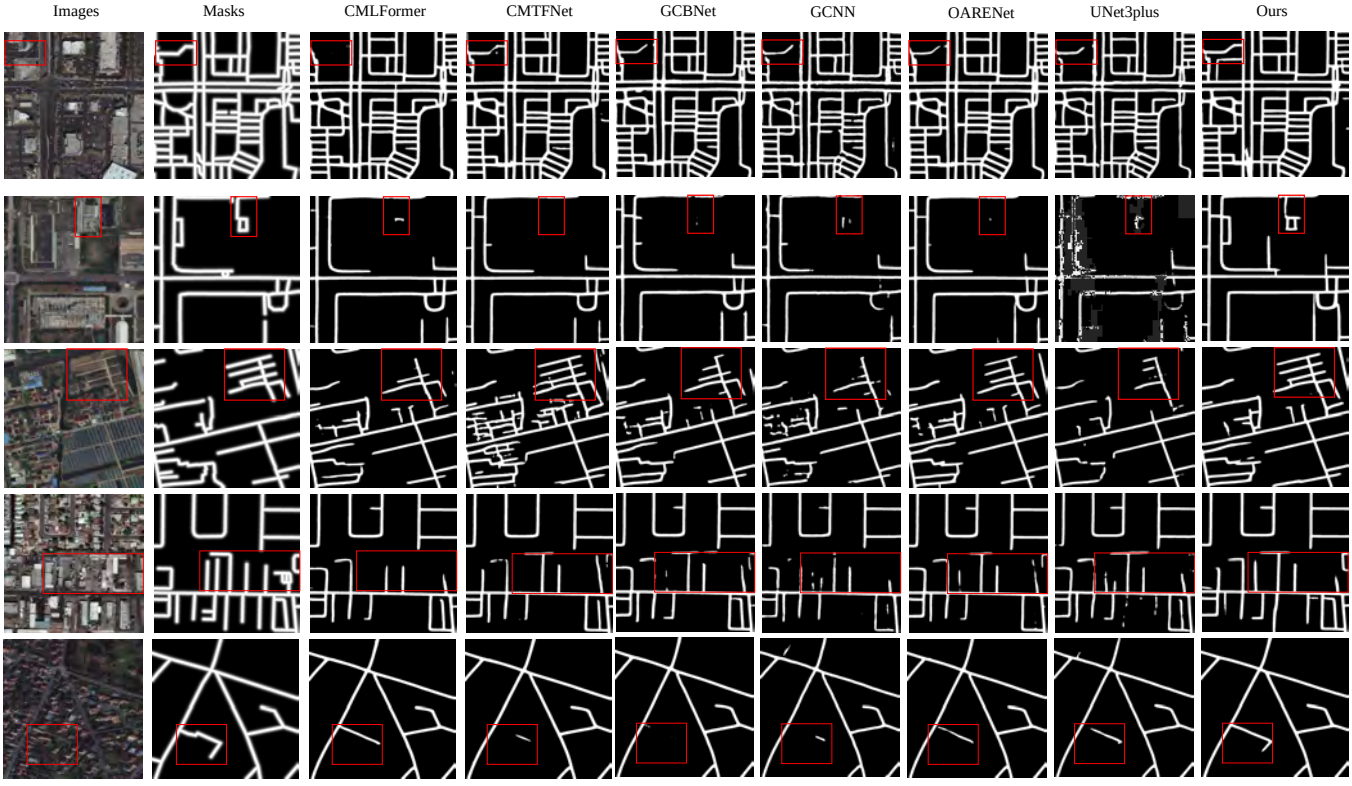


Fig. 6. Qualitative evaluations between FSNet and comparison methods on SpaceNet road dataset. The red boxes mark the areas where FSNet outperforms other methods.

ModelA. These gains reflect a significant boost in the overall effectiveness.

Effect of CDPM: On the DeepGlobe dataset, ModelC, equipped with our proposed Cross-Domain Parser Module, achieves notable improvements over the baseline (ModelA), with increases of 2.62%, 0.63%, and 0.55% in Recall, IoU, and F1-score respectively. These gains demonstrate that integrating frequency-domain information enhances the model's ability to capture global context. This is particularly beneficial for reconstructing occluded or fragmented road segments, thereby overcoming the locality constraints of purely spatial features and yielding more coherent road networks. Furthermore, the effectiveness of CDPM is further validated on the SpaceNet dataset, where ModelC outperforms the baseline by 0.55% in IoU and 0.37% in F1-score. The consistent performance gains across both datasets demonstrate the robustness and generalizability of our CDPM in improving overall performance.

V. Conclusion

In this paper, we propose an architecture for road extraction from remote sensing images. The key to this approach is to extract important features from both the frequency and spatial domains. To enhance the ability to capture road features in complex scenes, we propose the joint-domain enhancement module and its corresponding cross-domain hybrid parser module. The JDEM operates during the encoder stage by initially transforming images into the frequency domain to compensate for the limitations of spatial features. This approach not only endows long-range modeling capabilities but also incorporates multi-scale feature information, which can address road discontinuities caused by occlusions. The CDPM complements the JDEM during the decoder stage. It parses both the enhanced spatial features as well as high and low-frequency information separately, which can fully leverage the advantages of both. Our proposed FSNet achieves competitive performance on publicly available datasets, generating more complete and coherent predictions compared to other methods. Additionally, ablation experiments further validate the effectiveness of the proposed modules. In future research, we will focus on the following aspects: 1) integrating multi-modal inputs like LiDAR or SAR to handle extreme conditions; 2) exploring the integration of FSNet with large-scale foundation models to enhance performance in complex urban scenarios.

References

[1] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8340–8350.

[2] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-based segmentation and tracing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8919–8931, 2020.

[3] J. Hu, J. Gao, Y. Yuan, J. Chanussot, and Q. Wang, "LGNet: Location-Guided Network for Road Extraction From Satellite Images," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[4] Q. Li, M. Zhang, Z. Yang, Y. Yuan, and Q. Wang, "Edge-Guided Perceptual Network for Infrared Small Target detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[5] Q. Li, W. Zhang, W. Lu, and Q. Wang, "Multi-branch Mutual-guiding Learning for Infrared Small Target Detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[6] Q. Li, Y. Yuan, X. Jia, and Q. Wang, "Dual-stage approach toward hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, vol. 31, pp. 7252–7263, 2022.

[7] B. Xi, W. Zhang, J. Li, R. Song, and Y. Li, "Hypercasr: Spectral-spatial open-set recognition with category-aware semantic reconstruction for hyperspectral imagery," *IEEE Transactions on Image Processing*, vol. 34, pp. 7642–7655, 2025.

[8] B. Xi, M. Cai, J. Li, Z. Wang, S. Feng, Y. Li, and J. Chanussot, "Hyliosr: Staged progressive learning for joint open-set recognition of hyperspectral and lidar data," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[9] B. Xi, W. Zhang, J. Li, R. Song, and Y. Li, "Hypertafor: Task-adaptive few-shot open-set recognition with spatial-spectral selective transformer for hyperspectral imagery," *IEEE Transactions on Image Processing*, vol. 34, pp. 4148–4160, 2025.

[10] X. Zhao, Z. Yang, Q. Li, and Q. Wang, "Parameter-efficient transfer learning for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–12, 2025.

[11] Z. Yang, Q. Li, Y. Yuan, and Q. Wang, "HCNet: Hierarchical Feature Aggregation and Cross-Modal Feature Alignment for Remote Sensing Image Captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[12] Q. Wang, Z. Yang, W. Ni, J. Wu, and Q. Li, "Semantic-Spatial Collaborative Perception Network for Remote Sensing Image Captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[13] C. Sun, Y. Jia, H. Han, Q. Li, and Q. Wang, "A semantic-guided framework for few-shot remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[14] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 182–186.

[15] Z. Yang, W. Zhang, Q. Li, W. Ni, J. Wu, and Q. Wang, "C²Net: Road Extraction via Context Perception and Cross Spatial-Scale Feature Interaction," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[16] Y. Wang, Y. Peng, W. Li, G. C. Alexandropoulos, J. Yu, D. Ge, and W. Xiang, "DDU-Net: Dual-decoder-U-Net for road extraction using high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[17] Y. Zao, Z. Zou, and Z. Shi, "Topology-Guided Road Graph Extraction From Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[18] L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4479–4488, 2020.

[19] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–792.

[20] C. Zhao, W. Cai, C. Dong, and C. Hu, "Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8281–8291.

[21] Y. Tatsunami and M. Taki, "Fft-based dynamic token mixer for vision," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, 2024, pp. 15 328–15 336.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[24] J. T. Smith, A. Warrington, and S. W. Linderman, "Simplified state space layers for sequence modeling," *arXiv preprint arXiv:2208.04933*, 2022.

[25] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[26] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[27] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "VMamba: Visual State Space Model," *arXiv preprint arXiv:2401.10166*, 2024.

[28] C. Yang, Z. Chen, M. Espinosa, L. Ericsson, Z. Wang, J. Liu, and E. J. Crowley, "Plainmamba: Improving non-hierarchical mamba in visual recognition," *arXiv preprint arXiv:2403.17695*, 2024.

[29] J. Yao, D. Hong, C. Li, and J. Chanussot, "Spectralmamba: Efficient mamba for hyperspectral image classification," *arXiv preprint arXiv:2404.08489*, 2024.

[30] J. Ruan and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *arXiv preprint arXiv:2402.02491*, 2024.

[31] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, "Mambair: A simple baseline for image restoration with state-space model," in *European conference on computer vision*. Springer, 2025, pp. 222–241.
[32] J. Chu, K. Chi, and Q. Wang, "Rmmamba: Randomized mamba for remote sensing shadow removal," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.
[33] X. Ma, X. Zhang, and M.-O. Pun, "RS3Mamba: Visual State Space Model for Remote Sensing Image Semantic Segmentation," *IEEE Geoscience and Remote Sensing Letters*, 2024.
[34] M. Liu, J. Dan, Z. Lu, Y. Yu, Y. Li, and X. Li, "CM-UNet: Hybrid CNN-Mamba Unet for Remote Sensing Image Semantic Segmentation," *arXiv preprint arXiv:2405.10530*, 2024.
[35] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.
[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.
[37] Z. Yang, D. Zhou, Y. Yang, J. Zhang, and Z. Chen, "Road extraction from satellite imagery by road context and full-stage feature," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2022.
[38] T. Sun, Z. Di, P. Che, C. Liu, and Y. Wang, "Leveraging crowdsourced gps data for road extraction from aerial imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7509–7518.
[39] Z. Yang, D. Zhou, Y. Yang, J. Zhang, and Z. Chen, "TransRoadNet: A novel road extraction method for remote sensing images via combining high-level semantic feature and context," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
[40] P.-C. Hu, S.-B. Chen, L.-L. Huang, G.-Z. Wang, J. Tang, and B. Luo, "Road extraction by multi-scale deformable transformer from remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, 2023.
[41] J. Mei, R.-J. Li, W. Gao, and M.-M. Cheng, "CoANet: Connectivity attention network for road extraction from satellite imagery," *IEEE Transactions on Image Processing*, vol. 30, pp. 8540–8552, 2021.
[42] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia *et al.*, "Spectralgpt: Spectral remote sensing foundation model," *arXiv preprint arXiv:2311.07113*, 2023.
[43] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. X. Zhu, "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sensing of Environment*, vol. 299, p. 113856, 2023.
[44] T. Hu, Q. Yan, Y. Qi, and Y. Zhang, "Generating content for hdr deghosting from frequency view," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 732–25 741.
[45] Y. Zhang, T. Huang, J. Liu, T. Jiang, K. Cheng, and S. Zhang, "FreeKD: Knowledge Distillation via Semantic Frequency Prompt," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 931–15 940.
[46] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 172–181.
[47] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.
[48] Q. Zhu, Y. Zhang, L. Wang, Y. Zhong, Q. Guan, X. Lu, L. Zhang, and D. Li, "A global context-aware and batch-independent network for road extraction from vhr satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 353–365, 2021.
[49] G. Zhou, W. Chen, Q. Gui, X. Li, and L. Wang, "Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.
[50] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "CMTFNet: CNN and multiscale transformer fusion network for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
[51] H. Wu, M. Zhang, P. Huang, and W. Tang, "CMLFormer: CNN and Multi-scale Local-context Transformer network for remote sensing images semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
[52] R. Yang, Y. Zhong, Y. Liu, X. Lu, and L. Zhang, "Occlusion-aware road extraction network for high-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
[53] H. Huang, S. Xie, L. Lin, Y. Iwamoto, X. Han, Y.-W. Chen, and R. Tong, "Scaleformer: revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation," *arXiv preprint arXiv:2207.14552*, 2022.
[54] A. A. Vekinis, "Graph reasoned multi-scale road segmentation in remote sensing imagery," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 6890–6893.
[55] P. Yang, H. Xiao, C. Lin, and X. Xie, "Ugd-dlinknet: An enhanced network for occluded road extraction using attention mechanisms and uncertainty estimation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

Huiguang Yao is currently pursuing the M.S. degree at the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include semantic segmentation and deep learning.

Zhigang Yang is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing and computer vision.

Qiang Li (Member, IEEE) is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, machine learning, pattern recognition and remote sensing. For more information, visit the link (https://crabwq.github.io/)