Semantic-Spatial Feature Refinement Network for Road Extraction from Remote Sensing Images

Zhigang Yang, Huiguang Yao, Qiang Li, *Member, IEEE*, Weiping Ni, Junzheng Wu, Qi Wang, *Senior Member, IEEE*

***Abstract*—Extracting precise road information from remote sensing images remains challenging due to the interference from similar objects and occlusion from surroundings. To alleviate these issues, we propose a novel road extraction network to enhance both the precision and topological connectivity of extracted road networks, dubbed as *CRNet*. Specifically, a Global-Local Context Decoupling Module (GLCDM) is introduced to explicitly model long-range contextual dependencies while preserving fine-grained local road features, thereby improving the model's inference capability in occluded regions. Furthermore, a Semantic-Spatial Feature Refinement Module (SSFRM) is integrated into the skip connections, which leverages deep semantic features to guide the suppression of background noise in shallow feature maps across both channel and spatial dimensions, ensuring the decoder receives structurally accurate road representations. Experimental results on the remote sensing road datasets demonstrate that CRNet achieves state-of-the-art performance in terms of both segmentation accuracy and road connectivity. The source code is publicly available at https://github.com/CVer-Yang/CRNet.**

***Index Terms*—Remote sensing, road extraction, context decoupling, feature refine**

I. Introduction

DEEP learning-based road extraction methods for remote sensing images (RSI) [1]–[4] aims to extract road predictions from RSI in an automated manner, representing a significant research focus in RSI processing. The extracted geospatial information will serve as a crucial priori knowledge base for autonomous system, including high-precision navigation [5], path planning [6], and intelligent traffic decision-making [7].

Different from the common semantic segmentation task [8], road extraction from RSI mainly faces the following challenges. (1) ***Occlusion phenomenon.*** There are parts of the roads in the image that are occluded by the surroundings, such as the shadows of the buildings and trees, which leads to the missing phenomenon in the predicted roads. Therefore, the model must effectively capture the spatial correlations between roads and their surroundings, inferring occluded roads to improve the completeness of the predicted roads. (2) ***High similarity between road and background.*** There are some targets in the RSI that are similar to the road structure, such as railways, rivers, etc. This similarity introduces significant ambiguity, requiring models to incorporate contextual information to effectively distinguish roads from visually similar background elements.

Early approaches to road extraction rely on machine learning algorithms with manually crafted features. These methods exhibit limited adaptability to diverse scenarios, making road extraction both time-intensive and labor-intensive. They are commonly categorized into edge detection-based, morphology-based, and classical machine learning-based techniques. With the adoption of convolutional neural networks, researchers increasingly shift toward deep learning-based approaches, enabling the automation of road extraction tasks. These approaches focus on enhancing feature encoding, modeling contextual information, and incorporating attention mechanisms to improve the predicted results.

For the occlusion challenge, contextual information plays a critical role in road extraction by capturing the relationships between road pixels and their surrounding objects, such as adjacent roads, buildings, and cars. To enhance feature representations, techniques such as global attention mechanisms [9] and dilated convolutions [10] are widely utilized in feature encoders to model road-related contexts. However, the inherent mismatch between the shape of convolutional kernels and the elongated structure of roads, coupled with their limited capacity to capture long-range dependencies, poses substantial challenges to accurately and effectively modeling road-related context. Recently, transformer-based methods [11] and graph convolutional network (GCN) based approaches [12] have garnered significant attention for their potential in extracting contextual information. Although these methods show great potential, they often fail to effectively capture fine-grained road features, leading to an overemphasis on irrelevant contextual elements. This limitation reduces prediction accuracy and highlight the need for further improvements in road extraction tasks.

Encoder-decoder architectures commonly leverage skip connections to fuse shallow feature maps with deep feature maps of the same resolution, enhancing structural road features by combining fine-grained texture details with high-level semantic representations. However, this fusion process often introduces background noise, as shallow feature maps inherently contain irrelevant or misleading information from non-road regions. This makes it difficult to distinguish roads

This work was supported by the National Natural Science Foundation of China under Grant 62571437, Grant 62471394, and Grant U21B2041. 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) (Corresponding author: Qi Wang, Qiang Li.)

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

Junzheng Wu and Weiping Ni are with the department of remote sensing, Northwest Institute of Nuclear Technology, Xi'an 710072, P.R. China. (e-mail: niweiping@nint.ac.cn, wujunzheng@nint.ac.cn)

from similar background objects. Although attention-based methods [13] attempt to mitigate this issue by selectively emphasizing road-related features, they frequently fall short in achieving precise alignment between detailed road features and semantic information. This misalignment disrupts the effective integration of features, preventing detailed road features from being harmoniously aligned with high-level semantics. Consequently, it is crucial to design an accurate method that effectively integrates road structural features with semantic characteristics.

To alleviate the aforementioned challenges, we propose a novel method for road extraction from RSI, named CRNet. CRNet introduces two key innovations to enhance accuracy and connectivity in road extraction. Specifically, a global-local context decoupling module is designed to capture road-related context by explicitly modeling the interaction between fine-grained road details and long-range contextual dependencies. Furthermore, a semantic-spatial feature refinement module is introduced within the skip connections to enable precise alignment between low-level spatial features and high-level semantic representations. This refined alignment provides the decoder with structurally accurate information, ensuring the prediction of complete and well-connected road networks. The primary contributions of this work are summarized as follows:

- We propose a novel road extraction network that integrates global-local context and semantic-spatial feature refinement to achieve new state-of-the-art performance on publicly available road datasets. Comprehensive experiments validate the efficacy of the proposed approach.
- A global-local context decoupling module is introduced to adaptively capture the intricate relationships between fine-grained road details and their surrounding context. This module significantly enhances the feature representation for road-related structures by effectively decoupling and modeling global and local context, thereby improving reasoning capability in occluded road regions.
- We design a semantic-spatial feature refinement module to address the challenges posed by background noise and irrelevant information. By leveraging deep feature maps to refine shallow representations in both the channel and spatial dimensions, this module effectively suppresses interference and provides the decoder with highly accurate road-specific features, substantially enhancing discrimination capability between roads and shape-similar regions.

II. Related Work

In this section, we review the existing road extraction methods and the attention mechanism.

A. Road Extraction From RSI

Deep learning-based road extraction [14] from remote sensing images aims to autonomously delineate road networks from high-dimensional data, delivering unparalleled robustness and precision compared to traditional methodologies. Mnih and Hinton et al. [15] pioneer the application of deep learning to road extraction, setting a transformative precedent for subsequent advancements in the field. Since then, numerous improved methods are proposed to advance this field. For example, Zhang et al. [16] integrate residual networks into the UNet framework to deepen the encoder, which significantly improve the model feature extraction capability. Zhou et al. [17] future incorporate dilated convolutions with variable dilation rates into deep semantic feature maps, effectively capturing multi-scale contextual information and achieving superior prediction accuracy. Advanced strategies such as TransRoadNet [18] and SGCNet [12] incorporate long-range contextual information to enhance the ability to infer road structures in occluded regions. Zhu et al. [9] propose a global context-aware module that integrates global contextual information with Filter Response Normalization layers, enhancing the model robustness by capturing broader spatial dependencies and improving its performance in handling variations in road textures and occlusions. To tackle connectivity challenges, Mei et al. [19] design a road connectivity loss function to optimize the capacity for inferring robust road connections, particularly in occluded areas. To address the intricacies of multi-scale information, Yang et al. [20] employ strip convolutions with adaptive dilation rates, enabling detailed modeling of road contextual dependencies across varying scales. Building on feature integration, Qiu et al. [21] develop a multimodal framework that synthesizes GPS trajectory data with remote sensing features, significantly enhancing the ability to distinguish road structures. Similarly, Yang et al. [22] design a cross-spatial-scale feature fusion module to efficiently aggregate features from different stages of the encoder, ensuring that the decoder receives comprehensive and precise visual representations. Yang et al. [23] propose UGD-DLinkNet to improve road extraction under occlusion and noisy labels by integrating attention mechanisms and uncertainty modeling. Hua et al. [24] propose MADSNet to capture multiscale context and reduce false detection by adaptive feature selection and graph-based aggregation.

Although existing methods have demonstrated promising results in road extraction, they still have certain limitations. Specifically, these methods often struggle to accurately capture contextual information that adheres to road structures, leading to the incorporation of irrelevant features. Moreover, shallow features transmitted through skip connections often contain noise, resulting in the misclassification of non-road pixels as roads. To address these issues, We propose an attention mechanism based on grouping operations, which leverages the unique geometric properties of roads to capture relevant contextual information while ensuring precise alignment and fusion of deep and shallow feature maps. Consequently, the proposed method significantly enhances the ability of model to maintain road connectivity and delivers substantial improvements in prediction accuracy.

B. Attention Mechanism

Attention mechanisms [25], [26], with their dynamic capability to focus on critical features, play a vital role in enhancing the performance of segmentation models. By effectively modeling the complex information present in remote sensing

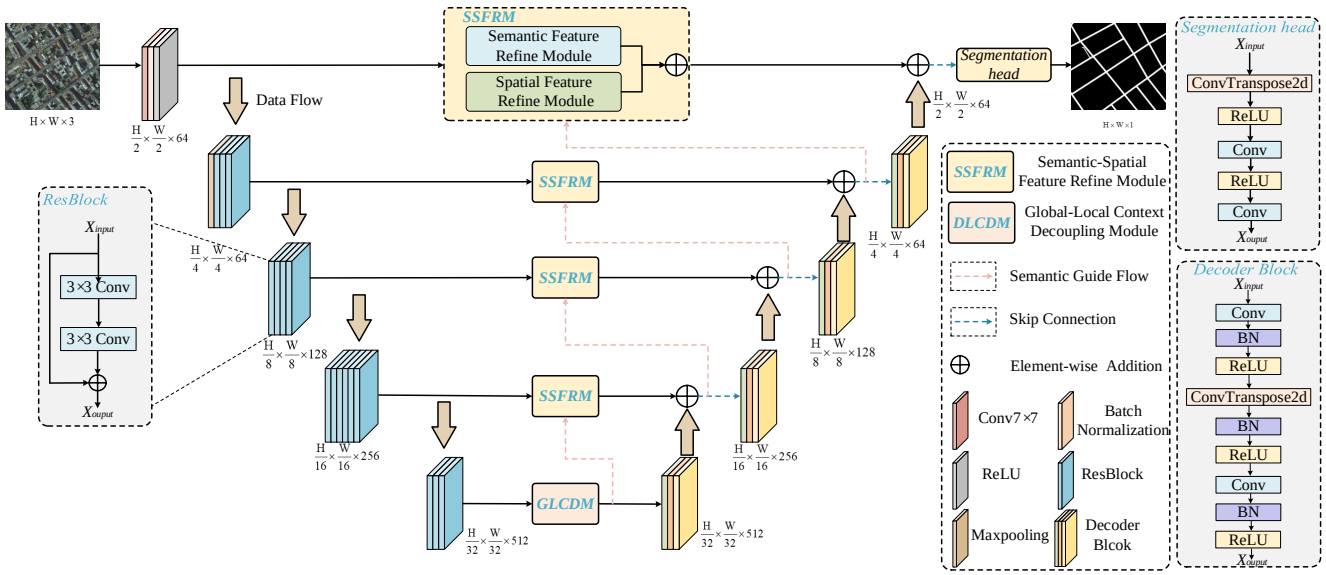

Fig. 1. The overall framework of the proposed CRNet, and it is divided into four sections: the encoder, GLCDM, SSFRM, and the decoder.

images, attention mechanisms significantly improve feature representation and context modeling. For instance, Woo et al. [27] integrates channel and spatial attention to jointly extract global and local features, facilitating precise interpretation of complex geospatial scenes. Wang et al. [28] utilize 1D convolution to efficiently capture inter-channel dependencies, achieving an optimal balance between computational efficiency and model performance. Hou et al. [29] decomposes feature aggregation into horizontal and vertical dimensions, enhancing the modeling of spatial positional information and effectively capturing long-range dependencies in remote sensing imagery. Dual-attention mechanisms, such as DANet [13], combine spatial and channel attention to further improve the understanding and representation of complex remote sensing scenarios. Similarly, Si et al. [30] propose a module integrating semantic spatial attention and progressive channel attention. This module aims to explore the interaction between spatial and channel attention mechanisms, achieving state-of-the-art performance across various visual tasks.

Transformer networks, known for their exceptional long-range feature modeling capabilities [31], have also been widely adopted in remote sensing tasks. However, their high computational cost poses challenges for high-resolution applications. To mitigate this, methods like agent attention [32] introduce agent tokens, enabling efficient global feature aggregation while maintaining high performance. Such advancements have broadened the applicability of lightweight attention mechanisms across diverse tasks, including image captioning [33], [34], infrared small target detection [35], and super-resolution [36].

Despite recent advancements, existing attention mechanisms continue to face significant challenges in high-resolution remote sensing applications. In particular, global modeling approaches, such as the positional attention mechanism in DANet, suffer from huge computational complexity, which limits their scalability in resource-constrained environments. We propose an attention mechanism based on grouping operations, this approach significantly reduces computational complexity and parameter overhead while preserving the ability to capture essential features, making it highly effective for road extraction and related tasks.

III. Proposed Method

In this section, we present the overall structure of our modeling approach as well as the details of the modules.

A. Overview

Fig. 1 illustrates the overall architecture of the proposed model, which comprises four components: encoder, global-local context decoupling module, semantic-spatial feature refinement module, and the decoder. Specifically, the input remote sensing image is processed by the encoder, which is based on a ResNet34 [37] network. The encoder consists of four stages, containing 3, 4, 6, 3 blocks, respectively. Each block includes convolutional layers (Conv), batch normalization (BN), and ReLU activation functions. This encoding process generates multi-scale feature maps with five different resolutions, denoted as E1, E2, E3, E4, and E5. The deepest feature map E5 is fed into the GLCDM to capture long-range context and enhance the overall representation of road features. Meanwhile, during the skip connections, shallow feature maps from the encoder are transferred to the SSFRM, which refines road feature representations in both channel and spatial dimensions under the guidance of road semantic features.

In the decoding stage, the feature map generated by the GLCDM is subsequently processed by the decoder. Firstly, a convolution operation reduces the channel dimensionality to one-fourth of its original size, followed by BN and ReLU activation, producing an intermediate feature representation. This intermediate feature is then upsampled via a ConvTranspose2d operation and fused with feature output from the SSFRM

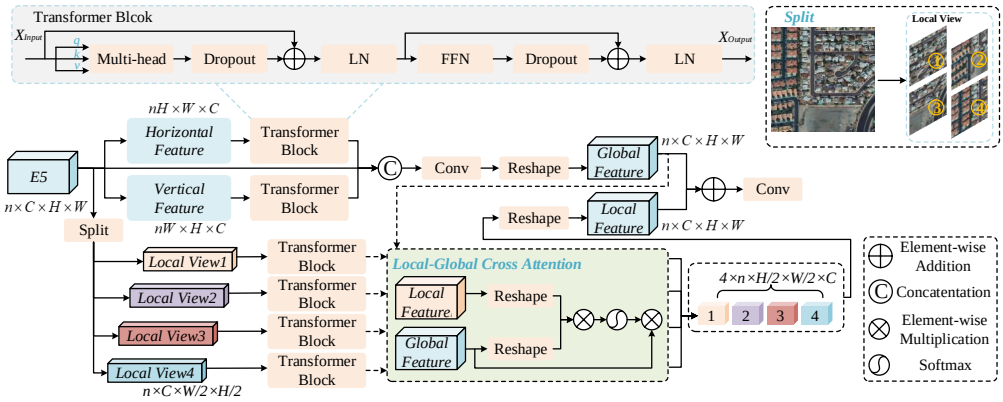

Fig. 2. The illustration of Global-Local Context Decoupling Module.It centers on capturing long-range spatial dependencies and preserving fine-grained road features, which enhance the inference capability of model for occluded regions.

through element-wise addition, resulting in the decoded feature map.

B. Global-Local Context Decoupling Module

Road are often blocked by building shadows or trees, leading to missing predictions. Existing methods based on dilated convolutions or transformers fail to capture relevant road context while filtering out irrelevant features, highlighting the need for more effective modeling to enhance extraction accuracy. To alleviate this issue, we propose a global-local context decoupling module, which models road contextual information at both global and local levels. This module effectively mitigates the negative impacts of occlusion by fusing global and local features. The structure of the module is shown in Fig. 2.

Global context branch: In the global context extraction branch, two specialized components, the horizontal context block and the vertical context block are designed to capture long-range contextual dependencies of roads in the horizontal and vertical directions, respectively. These components integrate these dependencies into precise global road context features. Specifically, for the horizontal context block, the input feature map $E_5 \in \mathbb{R}^{n\times c\times h\times w}$ is reshaped and flattened along the vertical axis to form a sequence $F_h \in \mathbb{R}^{(n\cdot h)\times w\times c}$, where each row represents the features along the horizontal direction. A Transformer block is applied to capture contextual dependencies in horizontal direction. The self-attention mechanism defined as, i.e.,

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q, K, V are the query, key, and value matrices computed from F_h, and d_k is the dimension of the query. Then a feed-forward network (FFN) is utilized to refine the learned representations. The updated horizontal feature representation is denoted as F'_h. Similarly, in the vertical context block, the feature map E_5 is flattened along the horizontal axis to produce $F_v \in \mathbb{R}^{(n\cdot w)\times h\times c}$. A Transformer block captures contextual dependencies along the vertical direction using the self-attention and FFN mechanisms as the horizontal branch. The refined vertical feature representation is denoted as F'_v. After extracting contextual features in both directions, the outputs F_g are fused via concatenation and a convolution operation.

Local context branch: In the local feature extraction branch, the primary goal is to enhance the fine-grained details of road targets. The input feature map $F \in \mathbb{R}^{n\times c\times h\times w}$ is divided into 2×2 spatial patches, each of size $\frac{h}{2}\times\frac{w}{2}$. Transformer blcoks are then applied independently to each patch to enhance the representation of local features, i.e.,

$$F^{(i,j)}_{\text{local}} = \text{Transformer}\left(F^{(i,j)}\right), \quad i,j \in \{1,2\}, \tag{2}$$

where $F^{(i,j)}$ is the feature patch located in the (i,j)-th quadrant of the spatial dimensions. The local and global feature maps are processed by the cross-attention mechanism, where the local feature map is treated as the query and the global feature map serves as the key and value. This design enables the learning of complementary information about mutual spatial relationships corresponding to local road features from the global feature map. These enhanced patches $F^{(i,j)}_{\text{local}}$ are then recombined through stacking and reshaping operations to form the complete enhanced local feature map $F_l \in \mathbb{R}^{n\times c\times h\times w}$. The operation can be expressed as, i.e.,

$$F_l = \text{Reshape}(\text{Stack}(F^{(i,j)}_{\text{local}})). \tag{3}$$

The $\text{Stack}(\cdot)$ operation concatenates all enhanced patches $F^{(i,j)}_{\text{local}}$ along the spatial dimensions, and the $\text{Reshape}(\cdot)$ operation reorders these patches into the original spatial layout to form F_l. Finally, the global road context F_g and the local features F_l are fused via element-wise addition and refined using a convolution operation. This process enables the model to integrate long-range contextual information from the global branch with fine-grained local details, resulting in improved road feature understanding and enhanced accuracy in road pixels prediction.

C. Semantic-Spatial Feature Refinement Module

Shallow feature maps contain rich detail information, which are essential for understanding local details and edge feature. However, they also include some background noise form

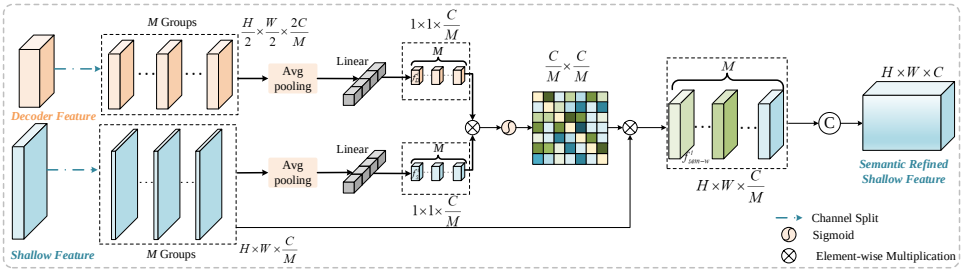

Fig. 3. The illustration of Semantic Feature Refinement Module. It splits shallow and decoder features into channel groups, computes semantic similarity, and generates attention weights for feature refinement.

shape-similar objects, which can compromise the accuracy of road predictions. To alleviate this problem, we propose a semantic-spatial feature refinement module that aligns and refines these feature maps across both semantic and spatial dimensions, enhancing road feature representation and improving network performance. Given the high resolution of feature maps, existing methods often incur significant computational and parameter costs. Therefore, we introduce a lightweight feature similarity modeling module based on group operations, enabling efficient feature alignment across channel and spatial dimensions while reducing computational overhead.

Semantic Feature Refinement Submodule: As shown in Fig. 3, in the semantic refinement stage, the module first divides the shallow feature maps ($F_{shallow} \in \mathbb{R}^{h\times w\times c}$) from skip connections and the intermediate feature maps ($F_D \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times 2c}$) from the decoder into M groups along the channel dimension. For each group, global semantic information is extracted using adaptive average pooling. The global features are subsequently processed through a convolutional layer to reduce their dimensionality into

$$f_s^i \in \mathbb{R}^{\frac{c}{M}\times 1}, f_D^i \in \mathbb{R}^{\frac{2c}{M}\times 1}, \quad i = 1, 2, \ldots, M. \tag{4}$$

The reduced features are reshaped to a unified dimension for feature alignment. A similarity matrix is then calculated between the grouped features of the shallow and intermediate feature maps to quantify their semantic correlation, i.e.,

$$S_i = \text{softmax}(\left(f_D^i\right)^\top \odot f_s^i), \tag{5}$$

where $\odot$ is element-wise multiplication operation, and $S_i \in \mathbb{R}^{\frac{2c}{M}\times\frac{c}{M}}$ represents the semantic similarity between the features. The above operations significantly reduce the computational complexity from $\mathcal{O}(2c^2)$ to $\mathcal{O}\left(\frac{2c^2}{M^2}\right)$ due to the division of channel features into M groups. These similarity scores are utilized to compute semantic-guided attention weights, which are applied to the corresponding shallow feature maps, i.e.,

$$f_{\text{sem-w}}^i = S_i \odot f_s^i. \tag{6}$$

Finally, the weighted feature maps from all groups are concatenated along the channel dimension to produce the semantically refined feature maps, i.e.,

$$F_{\text{refined}}^{Sem} = \text{Concat}(f_{sem-w}^1, f_{sem-w}^2, \ldots, f_{sem-w}^M). \tag{7}$$

Spatial Feature Refinement Submodule: As shown in Fig. 4, in the spatial refinement stage, the feature maps are divided into N spatial sub-regions, denoted as $F_s^j \in \mathbb{R}^{\frac{h}{N}\times\frac{w}{N}\times C}$ and $F_D^j \in \mathbb{R}^{\frac{h}{2N}\times\frac{w}{2N}\times 2C}$. Next, a 1×1 convolution and flatten operations are unilized to aligns these sub-region feature maps into $f_s^j \in \mathbb{R}^{\frac{h\times w}{N^2}\times 1}$ and $f_D^j \in \mathbb{R}^{\frac{h\times w}{N^2}\times 1}$. Then, a spatial-similarity matrix is computed between these sub-regions to generate precise spatial attention weights, i.e.,

$$S_j = \text{softmax}\left(f_s^j \odot (f_D^j)^\top\right). \tag{8}$$

The above operations significantly reduce the computational complexity from $\mathcal{O}\left(\frac{h^2w^2}{4}\right)$ to $\mathcal{O}\left(\frac{h^2w^2}{2N^4}\right)$ due to the division of spatial features into N^2 groups. These attention weights, $S_j \in \mathbb{R}^{\frac{H\times W}{N^2}\times\frac{H\times W}{N^2}}$, are applied to the corresponding sub-regions of the shallow feature maps, allowing the network to selectively emphasize spatially relevant areas, i.e.,

$$f_{\text{sp-w}}^j = S_j \odot f_s^j. \tag{9}$$

The adjusted sub-region feature maps are then merged by concatenated along the positional dimension and reshaped to reconstruct the original spatial resolution, i.e.,

$$F_{\text{refined}}^{\text{Sp}} = \text{flatten}(\text{Concat}(f_{\text{sp-w}}^1, f_{\text{sp-w}}^2, \ldots, f_{\text{sp-w}}^{N^2}). \tag{10}$$

The outputs of the two refinement submodules are integrated through element-wise addition. The combined features undergo further processing with Conv, BN, and ReLU activation to enhance the quality and representation of the features. By leveraging complementary semantic and spatial relationships, the proposed semantic-spatial feature refinement module effectively removes the noisy information in the shallow feature maps, which greatly result in substantial improvements in road extraction accuracy and overall model robustness.

IV. Experiments

In this section, extensive experiments on RSI road datasets are conducted to demonstrate the effectiveness of CRNet. We also conduct hyperparameter selection experiment in the Supplemental Material for Review.

A. Datasets

We conduct experiments on the SpaceNet, Paris and Massachusetts road datasets to evaluate the performance of our proposed model. The SpaceNet dataset consists of 2,780 pairs of remote sensing images and their corresponding road label. Following [22], the dataset is split into 2,224 image pairs for

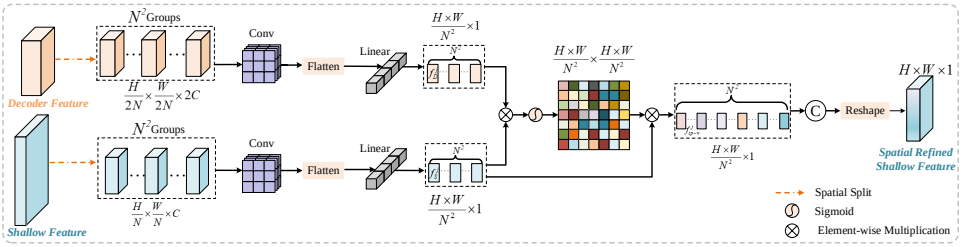

Fig. 4. The illustration of Spatial Feature Refinement Module. It divides features into spatial sub-regions, aligns and computes spatial similarity, then merges adjusted sub-regions to reconstruct spatial resolution.

training and 556 image pairs for testing. The Paris dataset [38] contains 625 pairs of RSI with three types of labels: road, building, and background. To simplify the task, building pixels are reclassified as background. The dataset is randomly divided into 500 image pairs for training and 125 image pairs for testing, maintaining a 4:1 split ratio. The Massachusetts road Dataset contains 1,171 aerial image pairs, with 1,108 pairs designated for training and 49 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

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

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

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

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

here, *TP*, *FP*, *TN* and *FN* represent the number of true positives, false positives, true negatives, and false negatives respectively. The Average Path Length Similarity (APLS) metric is used to measure the connectivity of predicted roads, i.e.,

$$\begin{aligned} S_{p\to T}(G, G') &= 1 - \frac{1}{M} \sum \min\left(1, \frac{|L(a,b) - L(a_1,b_1)|}{L(a,b)}\right), \\ S_{T\to P}(G', G) &= 1 - \frac{1}{M} \sum \min\left(1, \frac{|L(a,b) - L(a_1,b_1)|}{L(a_1,b_1)}\right), \\ APLS &= \frac{1}{N} \sum \left(\frac{1}{\frac{1}{S_{p\to T}} + \frac{1}{S_{T\to P}}}\right), \end{aligned} \tag{15}$$

where $L(a,b)$ and $L(a_1,b_1)$ are the shortest path length of nodes in the ground truth and the predicted graph. The *M* and *N* represent the number of unique paths and images. Among these indicators, IoU, F1-score and APLS can evaluate the quality of generated results comprehensively. All metrics are described as percentages(%).

C. Experimental Settings

The network is implemented using PyTorch 1.8 and trained on an NVIDIA GeForce RTX 4090 GPU with 24GB of memory. The Adam optimizer is utilized, while the loss function is defined as a combination of BCE and Dice coefficients. During training, the learning rate is reduced to one-fifth of its current value if the loss does not improve for three consecutive epochs. The batch size is set to 2. The model will end the training early if the loss does not decline in 6 consecutive losses. During testing, a test-time augmentation strategy is employed, including horizontal, vertical, and diagonal flipping of images.

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 public road datasets.

As shown in Table I, our model achieves the best overall performance on the SpaceNet dataset. U-Net, with its simple architecture, struggles to capture contextual information, making it insufficient for complex road scenarios where occlusion by surroundings and interference from shape-similar objects. DeepLabv3+ and DLinkNet, which incorporate dilated convolutions, improve IoU by 3.42% and 3.78%, respectively, but still fail to capture long-range contextual relationships to infer occluded road segments. Unet3+ aggregates multi-scale encoder features through concatenation, which enhances multi-scale representation but also introduces redundant information. In contrast, Scaleformer employs cross-scale feature fusion to address road scale variations, yet it struggles with narrow and intricate road structures. GCBNet improves global information handling but performs poorly in complex scenes requiring fine-grained local details. SGCNet enhances connectivity modeling but compromises recall due to insufficient local feature focus. In contrast, CRNet achieves a recall of 72.17% and precision of 64.68%, surpassing SGCNet by 4.79% in recall and 1.3% in precision. CRNet effectively balances global connectivity and local detail representation. CMLFormer, utilizing a transformer-based framework, improves global contextual understanding but struggles to balance global connectivity with local detail representation, resulting in both missing roads due to occlusion and misclassifications of similar background regions. OARENNet, with refined context modeling and improved decoding mechanisms, handles complex road scenarios with enhanced precision and structural completeness. Notably,

two recent methods UGDNet and MADSNet still exhibit clear limitations. UGDNet achieves a precision of 67.25%, slightly higher than that of CRNet. However, the recall of UGDNet is 6.00 % lower than that of CRNet. This pronounced imbalance between recall and precision leads to suboptimal F1-score and IoU, indicating limited structural completeness in road extraction. MADSNet adopts a multi-scale adaptive decoder with diverse feature selection to generate structurally complete roads, but it suffers from insufficient local detail preservation and thus achieves lower overall accuracy. Our proposed CRNet achieves superior performance by targeting the two core challenges of occlusion phenomenon and high similarity between roads and background via a dual-module architecture. Specifically, GLCDM models long-range topological dependencies to infer occluded road segments and preserves fine-grained local details, safeguarding road network completeness. Complementarily, SSFRM integrates into skip connections to eliminate redundant background noise and enhance discriminative road structural features via dual semantic-spatial refinement, effectively mitigating false positives from shape-similar background objects. CRNet synergistically balances global connectivity for occlusion inference and local feature discrimination for background suppression. It outperforms all existing methods across key metrics, highlighting robust adaptability in extracting complex road structures under challenging scenarios.

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

Method	Recall	Precision	IoU	F1-score	APLS
U-Net [39]	70.50	58.61	50.21	62.83	58.85
Deeplabv3+ [40]	71.33	63.11	53.63	65.77	63.37
DLinkNet [17]	**72.37**	62.70	53.99	65.96	65.26
Unet3+ [41]	67.06	64.73	52.24	64.88	58.86
GCBNet [9]	70.49	64.15	53.91	65.83	64.42
Scaleformer [42]	64.81	64.70	51.06	63.17	54.73
SGCNNet [12]	67.38	63.38	51.85	63.87	59.80
CMLFormer [43]	70.44	64.22	54.06	65.90	65.27
CU-dGCN [44]	69.76	64.49	53.46	65.49	63.62
OARENet [45]	70.84	64.52	54.50	66.32	65.25
UGDNet [23]	66.17	**67.25**	53.76	65.41	61.51
MADSNet [24]	71.83	64.26	54.75	66.66	64.99
CRNet	72.17	64.68	**54.85**	**66.85**	**65.47**

As shown in Table II, our model achieves the best IoU, F1, and competitive APLS, demonstrating its ability to effectively address the unique challenges of urban road extraction. Deeplabv3+ and DLinkNet improve IoU scores by employing dilated convolutions to capture broader contextual information. However, their performance is limited by their inability to accurately model complex road boundaries and connectivity. The integration of multi-stage feature aggregation in Unet3+ introduces significant information redundancy. In contrast, CMLFormer and GCBNet achieve competitive results by enhancing contextual information capture. However, both models fall short compared to our approach due to their suboptimal handling of local road details. OARENet exhibits a slight decline on the Paris dataset, with IoU and F1-score values

TABLE II
ROAD EXTRACTION RESULTS ON THE PAIRS DATASET.

Method	Recall	Precision	IoU	F1-score	APLS
Unet [39]	85.47	78.38	69.29	81.42	39.39
Deeplabv3+ [40]	86.04	83.88	73.84	84.75	46.60
DLinkNet [17]	87.10	83.66	74.46	85.13	47.87
Unet3+ [41]	85.56	77.99	68.96	81.17	37.08
GCBNet [9]	**88.59**	82.79	74.82	85.40	49.46
Scaleformer [42]	84.90	79.53	69.57	81.64	38.53
SGCNNet [12]	85.50	81.31	71.30	82.82	40.71
CMLFormer [43]	86.86	**84.49**	74.89	85.44	**51.03**
CU-dGCN [44]	87.26	80.84	72.31	83.64	43.77
OARENet [45]	87.04	83.61	74.30	85.03	45.26
UGDNet [23]	88.03	82.66	74.27	85.02	47.91
MADSNet [24]	87.63	83.27	74.52	85.21	48.50
CRNet	87.66	84.12	**75.23**	**85.65**	50.70

TABLE III
ROAD EXTRACTION RESULTS ON THE MASSACHUSETTS ROAD DATASET.

Method	Recall↑	Precision↑	IoU↑	F1-score↑	APLS↑
U-Net [39]	83.49	76.39	66.52	79.66	76.17
Deeplabv3+ [40]	83.12	76.83	66.43	79.65	75.11
DLinkNet [17]	81.87	79.57	67.61	80.44	75.41
Unet3+ [41]	81.89	79.63	67.73	80.52	74.37
GCBNet [9]	83.01	79.07	68.05	80.79	75.42
Scaleformer [42]	80.68	81.29	68.01	80.74	75.59
SGCNNet [12]	79.77	**82.98**	68.44	81.03	74.54
CMLFormer [43]	81.97	78.99	67.23	80.19	75.97
CU-dGCN [44]	79.65	80.95	67.13	80.14	75.75
OARENet [45]	80.69	80.32	67.31	80.24	75.10
UGDNet [23]	81.21	79.91	67.46	80.33	75.66
MADSNet [24]	82.24	79.44	67.74	80.52	76.36
CRNet	**83.54**	79.45	**68.71**	**81.29**	**76.80**

trailing ours by 0.93% and 0.62%, respectively. UGDNet employs attention mechanisms and uncertainty estimation to enhance road connectivity but compromises accuracy due to imprecise boundary localization. Our model distinguishes itself by effectively combining local feature refinement to suppress similar background interference with global contextual modeling to infer occluded roads, achieving an optimal balance between boundary precision and road connectivity. Moreover, as shown in Table III, CRNet also achieves the best performance across all metrics on the Massachusetts dataset. The results further highlight the robustness and adaptability of the proposed approach.

E. Visual Results

To analyze the effectiveness of our method, some predicted results are presented. We select six well-performing models for presentation ordered by IoU metric on SpaceNet and paris dataset.

Results on the Spacenet dataset: As shown in Fig. 5, CRNet demonstrates superior performance on the SpaceNet dataset, particularly in sparse and irregular road networks. In the first row, CRNet accurately captures the structure of intersections and branch roads, maintaining global continuity while ensuring boundary integrity. Other models, such as GCBNet,

Fig. 5. Qualitative evaluations between CRNet and comparison methods on SpaceNet road dataset. The red boxes mark the areas where CRNet outperforms other methods.

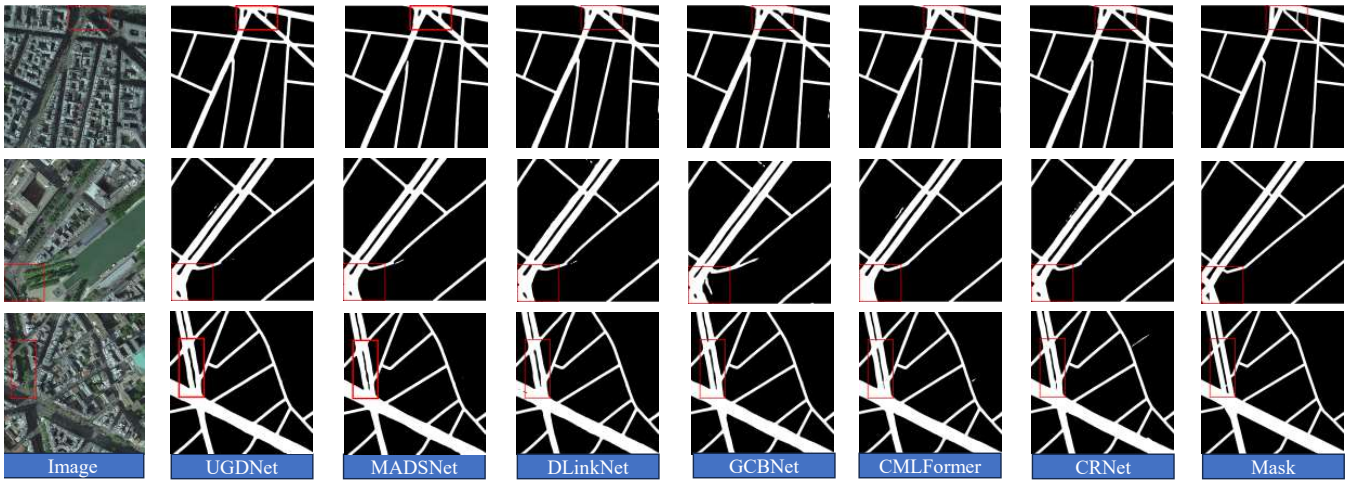

Fig. 6. Qualitative evaluations between CRNet and comparison methods on paris road dataset.

TABLE IV
ABALTION RESULTS ON THE SPACENET ROAD DATASET AND THE PARIS ROAD DATASET.

Components				SpaceNet					Paris				
Method	Baselline	GLCDM	SSFRM	Recall	Precision	IoU	F1-score	APLS	Recall	Precision	IoU	F1-score	APLS
ModelA	✓			**73.51**	61.05	53.21	65.52	61.87	87.66	82.13	73.57	84.54	45.76
ModelB	✓	✓		73.06	62.05	53.94	66.09	65.15	87.06	84.04	74.72	83.51	50.03
ModelC	✓		✓	69.47	**65.00**	53.87	65.67	65.15	**87.72**	82.81	74.13	84.91	46.93
CRNet	✓	✓	✓	72.17	64.68	**54.85**	**66.85**	**65.47**	87.66	**84.12**	**75.23**	**85.65**	**50.70**

DLinkNet, and CMLFormer, show noticeable disconnections in small intersections within the red box. These limitations arise from their inability to effectively model local details in sparse scenes and capture long-range contextual information. In the second row, some networks fail to produce coherent road predictions due to occlusions caused by surrounding objects. CRNet reconstructs the completeness of main roads and captures the intricate details of complex branch networks. Howerer, CRNet produces clear and coherent predictions for small branch roads, providing a significant advantage over other methods. The irregular road network in the third row further emphasizes the advantages of CRNet. With its semantic-spatial refining and global context modeling capabilities, CRNet captures the structures of narrow branches and complex intersections with high precision. The results closely align with the reference labels. In contrast, other models show issues such as disconnections, blurred boundaries, and missing details in narrow roads and intersections. These results demonstrate that CRNet effectively optimizes both local and global features, achieving significant improvements in detail accuracy and overall connectivity, outperforming other competing models in complex road scenarios.

Results on the Paris dataset: As shown in Fig. 6, the visualization results on the Paris road dataset further emphasize the superior performance and robustness of CRNet. In the first row, CRNet accurately reconstructs the structure of complex intersections, maintaining global road continuity while exhibiting precise detail modeling. In contrast, DLinkNet suffers from noticeable disconnections in intersection areas due to its limited ability to model local details, resulting in insufficient representation of complex regions. UGDNet and MADSNet demonstrate enhanced extraction capabilities for major arterial roads. However, these models exhibit significant false-positive errors, particularly in intersection regions where background pixels are erroneously classified as road surfaces. CMLFormer achieves superior global connectivity but compromises narrow boundary precision, which generates noticeable geometric distortions in road predictions. As illustrated in the second row of comparative results, CRNet continues to excel in extracting both main roads and smaller riverside branches. Within the red box, CRNet effectively captures small branch roads with complete topological structures, attributable to its robust feature refinement capabilities. In contrast, MADSNet exhibits considerable omissions, indicating weak performance in detecting small branch roads. DLinkNet and GCBNet show moderate improvements in predicting main roads but struggle with background misclassifications in local areas, revealing their vulnerability to redundant features in complex scenes. In the third row, CRNet also demonstrates its ability to accurately reconstruct the topology of complex intersections and narrow branches through its joint modeling of local details and global structures.

F. Ablation Study

To validate the effectiveness of the GLCDM and SSFRM, ablation experiments are conducted on the SpaceNet and Paris datasets.

Effect of GLCDM: The GLCDM plays a crucial role in capturing global contextual information, particularly in challenging scenarios such as occlusions, road fractures, and curved road segments. As shown in Table IV, removing GLCDM results in 0.98% reduction in IoU, 1.18% decrease in F1-score, and 0.32% drop in APLS on the SpaceNet dataset compared to the CRNet model. These declines highlight the importance of GLCDM in maintaining long-range connectivity in road predictions by modeling complex dependencies across spatial regions. Without GLCDM, the model struggles to produce coherent and structurally complete road predictions, as evidenced by the decreased performance in connectivity-related metrics such as APLS. On the Paris dataset, the removal of GLCDM leads to notable performance degradation. These results demonstrate the ability of GLCDM to maintain global consistency in road networks, allowing the model to address the complexity of interconnected road environments.

Effect of SSFRM: The SSFRM refines shallow feature maps by eliminating redundant information and enhancing the capture of detailed road boundaries. On the SpaceNet dataset, removal of the SSFRM results in 2.63% reduction in Precision, 0.91% decrease in IoU, and 0.32% drop in APLS. These declines highlight the critical role of the SSFRM in achieving accurate and consistent road predictions, particularly for fine-grained details. On the Paris dataset, which features more complex and densely connected road networks, the contribution of the SSFRM is even more significant. Exclusion of the SSFRM leads to 2.14% reduction in IoU and 0.67% decrease in APLS. Without the SSFRM, the ability to delineate narrow roads and complex intersections is impaired, which underscores its importance in urban road extraction tasks. The SSFRM ensures effective feature refinement and enables the model to adapt to challenging urban environments with greater robustness.

Synergistic Effect of GLCDM and SSFRM: Beyond individual module contributions the dual module design exhibits significant synergistic effects. On the SpaceNet dataset, combining both modules yields IoU and APLS improvements of 1.64% and 3.60% respectively gains exceeding the sum of individual contributions. This synergy stems from the targeted complementarity of the two modules GLCDM captures long range topological dependencies to aid occluded road inference providing structural connectivity guidance. SSFRM refines shallow features via dual semantic spatial refinement under deep semantic guidance suppressing shape similar background noise. Purified local features from SSFRM enable GLCDM to focus on meaningful contextual dependencies while global constraints from GLCDM guide SSFRM to prioritize road boundary refinement. This mutual reinforcement mitigates both occlusion induced missing segments and background similarity induced false positives validating the dual module framework rationality.

V. Conclusion

In this paper, we propose a novel approach for road extraction from remote sensing images. Specifically, we introduce a global-local context decoupling module to simultaneously capture long-range contextual dependencies and local detailed features for road extraction. To effectively bridge these two aspects, an interaction mechanism is designed to capture a robust connection between global and local contexts. Furthermore, we develop a semantic-spatial feature refinement module in during the skip connection stage, which utilizes deep feature maps as semantic guidance to suppress noise in shallow feature maps. This refinement process enhances the structural quality of road features provided to the decoder, significantly improving the accuracy of road predictions. Extensive experiments conducte on widely used road extraction datasets demonstrate that the proposed CRNet achieves satisfactory performance across diverse road scenarios. Additionally, comprehensive ablation studies validate the effectiveness and contributions of each module within the framework. In future research, we aim to explore the integration of model inference acceleration technologies to further enhance road extraction performance.

References

[1] J. Li, J. He, W. Li, J. Chen, and J. Yu, "RoadCorrector: A structure-aware road extraction method for road connectivity and topology correction," *IEEE Trans. Geosci. Remote Sensing*, 2024.

[2] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "RGB-induced feature modulation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–11, 2023.

[3] J. Zou, W. Zhang, Q. Li, and Q. Wang, "Mosaic-tracker: Mutual-enhanced occlusion-aware spatiotemporal adaptive identity consistency network for aerial multi-object tracking," *ISPRS-J. Photogramm. Remote Sens.*, vol. 229, pp. 138–154, 2025.

[4] Q. Li, Z. Yang, J. Cheng, and Q. Wang, "Spatial-frequency feature learning for infrared small target detection," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–11, 2026.

[5] J. Iqbal, A. Masood, W. Sultani, and M. Ali, "Leveraging topology for domain adaptive road segmentation in satellite and aerial imagery," *ISPRS-J. Photogramm. Remote Sens.*, vol. 206, pp. 106–117, 2023.

[6] C. Yang, K. Zhuang, M. Chen, H. Ma, X. Han, T. Han, C. Guo, H. Han, B. Zhao, and Q. Wang, "Traffic sign interpretation via natural language description," *IEEE Trans. Intell. Transp. Syst*, 2024.

[7] Y. Duan, D. Yang, X. Qu, L. Zhang, L. Chao, P. Gan, S. Yuan, H. Qin, and J. Qu, "Lcire-net: Lightweight cross-modal information interaction for road feature extraction from remote sensing images and gps trajectory/lidar," *IEEE Trans. Geosci. Remote Sensing*, 2024.

[8] Q. Li, M. Zhang, Z. Yang, Y. Yuan, and Q. Wang, "Edge-Guided perceptual network for infrared small target detection," *IEEE Trans. Geosci. Remote Sensing*, 2024.

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

[10] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 472–480.

[11] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] G. Zhou, W. Chen, Q. Gui, X. Li, and L. Wang, "Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–15, 2021.

[13] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[14] H. Bai, C. Ren, Z. Huang, and Y. Gu, "A dynamic attention mechanism for road extraction from high-resolution remote sensing imagery using feature fusion," *Scientific Reports*, vol. 15, no. 1, p. 17556, 2025.

[15] V. Mnih, *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.

[16] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, 2018.

[17] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. workshops*, 2018, pp. 182–186.

[18] 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 Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[19] 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.

[20] Z. Yang, D. Zhou, Y. Yang, J. Zhang, and Z. Chen, "Road extraction from satellite imagery by road context and full-stage feature," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2022.

[21] Y. Qiu, C. Lin, J. Mei, Y. Sun, H. Lu, and J. Xu, "Lightweight cross-modal information measure and propagation for road extraction from remote sensing image and trajectory/lidar," *IEEE Trans. Geosci. Remote Sensing*, 2024.

[22] 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 Trans. Geosci. Remote Sensing*, 2024.

[23] 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 J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2025.

[24] Z.-T. Hua, S.-B. Chen, W. Lu, J. Tang, and B. Luo, "Multi-scale adaptive decoder and diverse selection for road extraction in remote sensing images," *IEEE Trans. Geosci. Remote Sensing*, 2025.

[25] Y.-P. Song, X. Wu, W. Li, T.-Q. He, D.-F. Hu, and Q. Peng, "Highlightnet: Learning highlight-guided attention network for nighttime vehicle detection," *IEEE Trans. Intell. Transp. Syst*, 2025.

[26] L. Su, X. Ma, X. Zhu, C. Niu, Z. Lei, and J.-Z. Zhou, "Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer," *arXiv preprint arXiv:2412.14598*, 2024.

[27] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[28] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.

[29] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.

[30] Y. Si, H. Xu, X. Zhu, W. Zhang, Y. Dong, Y. Chen, and H. Li, "SCSA: Exploring the synergistic effects between spatial and channel attention," *arXiv preprint arXiv:2407.05128*, 2024.

[31] M. Kim, P. H. Seo, C. Schmid, and M. Cho, "Learning correlation structures for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 18 941–18 951.

[32] D. Han, T. Ye, Y. Han, Z. Xia, S. Pan, P. Wan, S. Song, and G. Huang, "Agent attention: On the integration of softmax and linear attention," in *European Conference on Computer Vision*. Springer, 2025, pp. 124–140.

[33] Z. Yang, Q. Li, Y. Yuan, and Q. Wang, "HCNet: Hierarchical feature aggregation and cross-modal feature alignment for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sensing*, 2024.

[34] Q. Wang, Z. Yang, W. Ni, J. Wu, and Q. Li, "Semantic-Spatial collaborative perception network for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sensing*, 2024.

[35] Q. Li, W. Zhang, W. Lu, and Q. Wang, "Multi-branch mutual-guiding learning for infrared small target detection," *IEEE Trans. Geosci. Remote Sensing*, pp. 1–1, 2025.

[36] Q. Li, Y. Yuan, and Q. Wang, "Multiscale factor joint learning for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–10, 2023.

[37] 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.

[38] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Trans. Geosci. Remote Sensing*, vol. 55, no. 11, pp. 6054–6068, 2017.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2015, pp. 234–241.

[40] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[41] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2020, pp. 1055–1059.

[42] 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:2207.14552*, 2022.

[43] 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 J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2024.

[44] 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.

[45] R. Yang, Y. Zhong, Y. Liu, X. Lu, and L. Zhang, "Occlusion-aware road extraction network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sensing*, 2024.