C^{2}Net: Road Extraction via Context Perception and Cross Spatial-Scale Feature Interaction

Zhigang Yang, Wei Zhang, Qiang Li, *Member, IEEE*, Weiping Ni, Junzheng Wu, Qi Wang, *Senior Member, IEEE*

***Abstract*—Road extraction from remote sensing images holds significant application value in various aspects of daily scenarios. However, it is still challenging to extract high-quality road results from remote sensing images due to the interference of objects sharing similar structures with roads in the background, and the occlusion caused by surroundings. To alleviate these problems, a road extraction *network* based on the global-local *Context* perception and *Cross* spatial-scale feature interaction is proposed (C^{2}Net). First, a global-local context perception module is incorporated to capture the overall topology features of road, which aims to improve the ability of model to discriminate between roads and similar objects. Then, the cross spatail-scale feature interaction module is designed in the skip connection to effectively aggregate full-scale features without loss of feature information, which can provide rich and accurate road structural features for the decoder. Experiments conducted on public road datasets demonstrate that C^{2}Net outperforms existing methods in terms of comprehensive metrics such as Intersection over Union (IoU) and F1-score. The results indicate that C^{2}Net can produce road results with superior connectivity and quality. The source code will be publicly available at https://github.com/CVer-Yang/CCNet.**

***Index Terms*—Remote sensing, road extraction, context perception, cross spatial-scale, feature interaction**

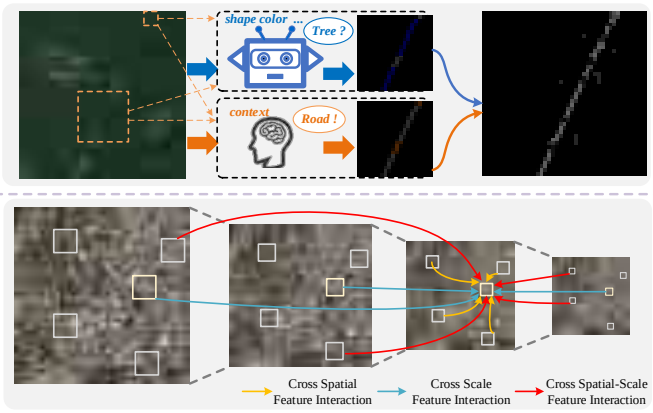


Fig. 1. The motivation of the proposed C^{2}Net. **Top:** There are some obscured road areas in the image, and these pixels will be determined as the background only by analyzing shallow features such as shape and color. However, humans can infer the occluded road region by capturing the road context. **Bottom:** Existing multi-scale feature fusion methods based on sampling operations cause information missing and semantic ambiguity. Therefore, we propose a feature interaction approach for hierarchical feature fusion without scale change. It can realize the effective integration of road details and semantic information in a cross spatial, cross scale, and cross spatial-scale manners.

I. Introduction

ROAD extraction [1] is an important research topic in the remote sensing images (RSI) process that aims to classify each pixel by drawing different colors to different categories of pixels. It plays an important role in a wide range of application fields, including urban planning, automatic driving [2], geographic information updating [3], etc.

Different from common semantic segmentation tasks, road extraction from RSI [4] faces some intricate challenges: 1) **Background complexity.** RSI contains complex feature information, with some objects in the image that are similar to the shape of roads, such as railroad tracks and rivers. The existence of these objects introduces interferences to the prediction of roads. 2) **The existence of occlusion.** There are lots of road pixels that are obscured by the surrounding trees or buildings, which makes it difficult to extract complete road results. 3) **The imbalanced ratio between foreground and background pixels.** It is difficult for the model to optimize due to the small rate of road pixels in the image and the complex structure of roads. The above challenges make it challenging for the model to extract complete and precise roads from the image.

As illustrated in the top of Fig. 1, the road extraction task benefits from capturing and introducing precise contextual information to reason the occluded road regions in the image. Consequently, many approaches [5] [6] [7] have been developed to obtain accurate context. For example, Zhou et al. [6] utilize dilated convolution with different rates to enlarge the receptive field of the convolution kernel and study multi-scale context. Nevertheless, the inherent local nature of convolutional operation limits the ability of model to extract the context over the long-distance range. Wang et al. [7] integrate the Non-local module to extract road context at long distances, enabling the model to study the global dependencies for each pixel and improving the quality of produced road results. Taking into account the natural advantages of Vision Transformer [8] in modeling long-distance contextual information, some networks [9] [10] [11] based on Transformer have appeared in road extraction field and exhibit superior performance, such as TransRoadNet. However, the network combined with Transformer cannot capture local features adequately, leading

This work was supported in part by the National Natural Science Foundation of China under Grant U21B2041, Grant 62471394, and Grant 62301385. 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.)

Wei Zhang is with School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, P.R. China. (e-mail:zhangwei707@mail.nwpu.edu.cn)

Junzheng Wu and Weiping Ni are with the department of remote sensing, Northwest Institute of Nuclear Technology, Xi'an 710024, P.R. China. (e-mail: niweiping@nint.ac.cn, wujunzheng@nint.ac.cn)

to the model being ineffective in predicting the road with complete structures. Indeed, these methods may introduce some irrelevant context, which restricts the discrimination ability of the network and produces false segmentation. Therefore, current methods are difficult to offer accurate contextual information, leaving room for further enhancements.

As shown in the bottom of Fig. 1, the combination of cross spatial-scale features fully is an effective way to accurately model road structural characteristics. Therefore, most researchers adopt multi-scale feature fusion mechanism [12] [13] to strengthen the feature representation of road. Numerous methods achieve feature aggregation by employing feature map deformation, such as Unet++ [14], Unet3+ [15]. These designs effectively improve the ability of the model to recognize segmented objects. However, they also bring some drawbacks. On the one hand, these methods fuse multi-feature by using some sampling and convolution operations, which introduce large parameters and limit the application in real-world scenes. On the other hand, downsampling operations on shallow feature maps inevitably result in the loss of crucial details, while upsampling operations on deep feature maps may cause the introduction of inaccurate semantic information. Hence, these methods fail to realize the effective aggregation of multi-scale features, and the generated fused feature cannot provide accurate road structural features for the decoder.

To solve these aforementioned challenges, we design a road extraction **network** from RSI based on global-local **Context** perception and **Cross** spatial-scale feature interaction, abbreviated as **C^{2}Net**. Specifically, the network utilizes ResNet34 [16] as encoder to extract features from the image. To improve the inference ability of model for occluded regions, a global-local context perception module (GLCPM) is designed to mine and fuse the precise context from the deep feature map. Meanwhile, a lightweight cross spatial-scale feature interaction module (CSFIM) is proposed in the skip connection, which can provide the decoder with rich road structural features. The feature map generated by the GLCPM, the feature maps delivered by the CSFIM and the skip connection are fed into decoder. The network finally produces road results through multiple decode stages. The main contributions of this work are summarized as follows:

- We propose a joint global-local road context perception and cross spatial-scale feature interaction method for road extraction from RSI, which achieves satisfying performance in publicly available road datasets.
- To improve the inference ability to predict the occluded regions, the global-local context perception module equipped with dual-attention mechanism is proposed. It can generate accurate dilated features and long-distance dependency simultaneously by combining convolution and Transformer networks.
- To produce desirable road results, the cross spatial-scale feature interaction module is designed in the skip connection stage, which supplies precise and comprehensive road structural features for the decoder by aggregating multi-stage feature maps within and across scales efficiently.

The rest of this paper is organized as follows. Section II provides related work in road extraction from RSI and the encoder-decoder framework. In Section III, we describe the details of our proposed C^{2}Net. The experimental results and limition are analyzed in Section IV. Finally, Section V offers the conclusion.

II. Related Work

In this section, we review the encoder-decoder methods and the existing road extraction from RSI.

A. The Encoder and Decoder Framework

The encoder-decoder structure is a popular feature extraction framework in deep learning. It is widely used in remote sensing image processing tasks [17] [18] [19] such as road extraction, building extraction, change detection, and image caption [20]. The framework is mainly divided into three parts: encoder, skip connection, and decoder. The encoder is responsible for encoding the input image and extracting the multi-level feature of the image. The features extracted by the encoder are fed into the decoder to generate the corresponding segmentation results. The skip connection delivers the shallow features extracted by encoder to decoder, which can supply structural information to the decoder and assist the model in generating accurate segmentation results.

Common encoder-decoder models include Unet [21], LinkNet [22], Deeplab [23] series, and some variants. They are usually used as baseline for road extraction tasks. Among them, Unet++ enables the model to effectively capture features at different scales and improve segmentation accuracy by designing skip connections in different resolution feature maps. The Unet3+ designs full-scale skip connections, which enables the decoder to more effectively utilize fine-grained detail features with coarse-grained semantic information. In addition, some encoder-decoder frameworks that incorporate novel networks have emerged, such as SwinUnet [24], scaleformer [25], manbaunet [26]. However, these improved methods that rely on dense connectivity cannot realize the effective interaction of full-scale features. Meanwhile, these models require high performance of the equipment, which limits the practical application scenarios of the model. Different from these methods, we propose a feature fusion module based on split-interaction-reverse to realize the aggregation of cross spatial-scale features without the loss of information. Therefore, the designed method can fully utilize encoder features across different scales and provide the decoder with comprehensive road structural features.

B. Road Extraction From RSI

In the past decade, numerous of RSI road extraction algorithms based on deep learning have been proposed, and these methods mainly include improving encoder, introduction of context, multi-task joint training, etc. Mnih et al. [27] firstly propose a method based on deep learning for road extraction. The limited feature extraction ability of method results in poor quality of generated road results. In [28], Zhu

et al. design a novel module to develop the road contextual information and further improve the performance of the model. As for the introduction of context, Zhou et al. [6] introduce dilated convolution with different rates, which aim to extract multi-scale context and improve accuracy of prediction road results. Later, some methods that aim to obtain precise context have been proposed, such as NLinknet [7], SGCNNet [29]. Considering the complex shape of the road, Dai et al. [30] combine deformable convolution with self-attention mechanism to model road features more effectively.

Multi-task learning [31] enables the model to acquire robust road features by designing auxiliary tasks for training, e.g., Mei et al. [32] design a connection loss that captures the relationship between neighboring pixel points to improve road connectivity. In the improvement of multi-branch network, Topology-Enhanced method [33] is proposed to extract road features at the pixel, edge, and region levels and ensure road connectivity by introducing direction-aware modules. Similar methods have [34] [35]. To increase the efficiency of computational, Hu et al. [36] present a road location auxiliary task to improve the missing components in forecast results. More recently, some novel road extraction methods have emerged, Qi et al. [37] put forward a dynamic snake convolution that incorporates continuity constraints into the design of the convolution kernel. Therefore this model can effectively adapt to the complex shape and extract the core features of the road. Inspired by large model, Zhang [38] proposes a SAM-based segmentation method that achieves satisfactory results in different scenarios.

Although the above methods can generate accurate road prediction results, they also exhibit certain limitations. Due to the existing methods exploring relationships between roads and surroundings only by utilizing Transformer or CNN network, these methods are difficult to precisely learn the relationship between roads and surroundings, and even produce some harmful context. The inaccurate context brings interference in the inference stage and leads to the generation of incorrect prediction results. Therefore introducing comprehensive and accurate context is crucial for road extraction tasks.

III. Proposed Method

In this section, we introduce the overall architecture of the C^{2}Net, and then describe the designed modules: GLCPM and CSSFIM, finally the detail of the decoder is provided.

A. Overview

As shown in Fig. 2, given a high-resolution remote sensing image, we adopt ResNet34 pre-trained on the ImageNet dataset as the encoder, which yields five feature with different resolutions, named E1, E2, E3, E4, and E5. The deep feature map E5 is fed into the GLCPM to capture the accurate road context from different scales, enhancing the ability of model to reason about the occluded regions effectively. Simultaneously, the full-scale feature maps are fed into the CSFIM, which can efficiently integrate road features at full scales to provide the decoder with supplementary feature maps that contain accurate semantic information and rich texture features. Additionally, the feature delivered by the skip connection, the feature output from the CSFIM, and the feature generated by the decoder in the previous stage are combined in decoder to produce the predicted feature map at the current stage. It ultimately generates predicted road results with the same resolution as the input image.

B. Global-Local Context Perception Module

A significant number of road pixels are obscured by the surrounding trees or the shadows in RSI, which makes it difficult to produce complete roads only by using the shape, color, and other shallow features. Moreover, certain objects in the image background are closely related to the road, the overall feature representation of the road can be effectively enhanced by capturing the relationship between the road and surroundings, thereby improving accuracy of predicted results.

Convolutional neural network excel in extracting detailed features, while Transformer network has excellent long-range feature modeling ability. Therefore, we merge the convolution and the Transformer network to utilize the advantages of both networks fully. This module can concurrently capture the detailed features and long-distance contextual dependencies of the pixel points around the road. The architecture of the GLCPM is illustrated in Fig. 3. The module utilizes dilated convolution with rates of 1, 3, 5, and 7 on deep feature E_5 to obtain the road detail feature at different scales. Meanwhile, the Swin-Transformer networks with window sizes of 2, 4, 8, and 16 are introduced after the dilated convolution respectively, which can analyze long-distance context across different scales. The residual connections are designed to fuse these feature maps, i.e.,

$$branch_i = ST_i(DC_i(E_5)) + DC_i(E_5), i \in [1, 4], \quad (1)$$

where $ST_i(\cdot)$ represents Swin-Transformer networks with different window sizes, and $DC_i(\cdot)$ represents dilated convolution with different rate. Features with different scales are fed into the spatial aware and semantic dependency submodules respectively and fused through the guidance of the attention mechanism.

Spatial Aware Submodule: By analyzing the positional correlation between feature maps at different scales, the module can accurately aggregate different features from a spatial perspective. The convolution operation is used to compress the channel information of different features, i.e.,

$$Spatial_i = Conv_i(branch_i), \quad (2)$$

where $Conv_i(\cdot)$ is the convolution with outchannel of 1. These features are fused through concatenation operation. Subsequently, the position weights of different features are obtained by using the Sigmoid and channel split operation, i.e.,

$$W^i_{spatial} = split(\delta(Cat(Spatial_i))), \quad (3)$$

where $split(\cdot)$ is channel split operation, $\delta(\cdot)$ is Sigmoid operation and $Cat(\cdot)$ is channel concatenation operation. The input features and the corresponding position weights are multiplied to weight the spatial information. Then the weighted

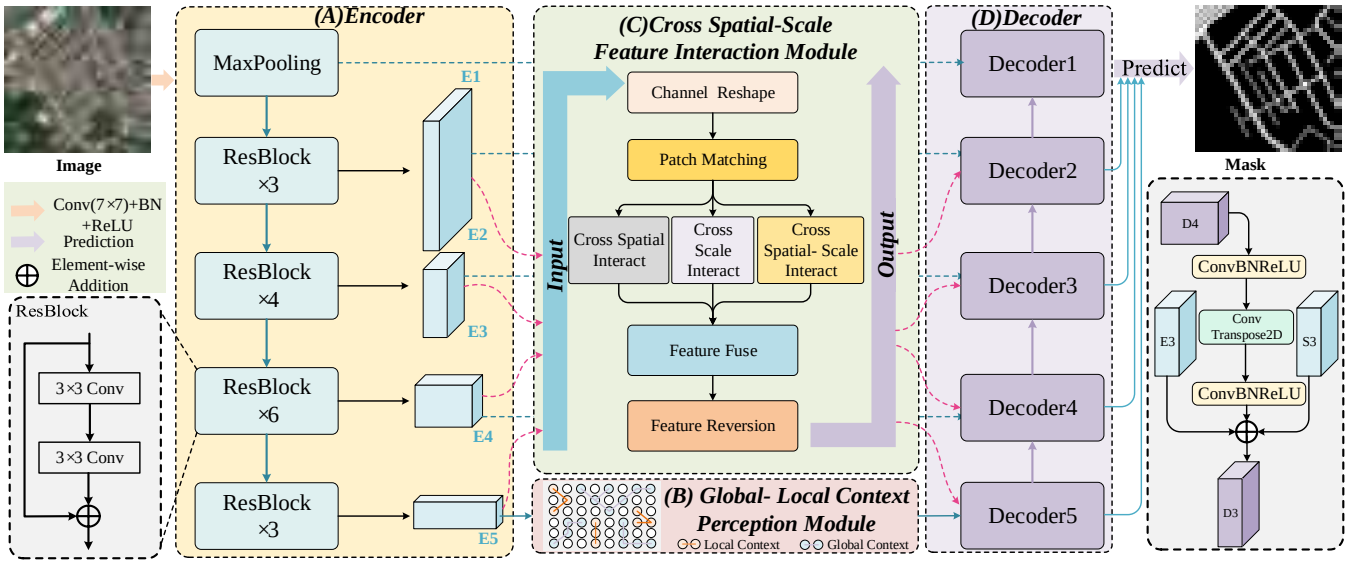


Fig. 2. The overall framework of the proposed C^{2}Net, and it is divided into four sections: the encoder, GLCPM, CSFIM, and decoder.

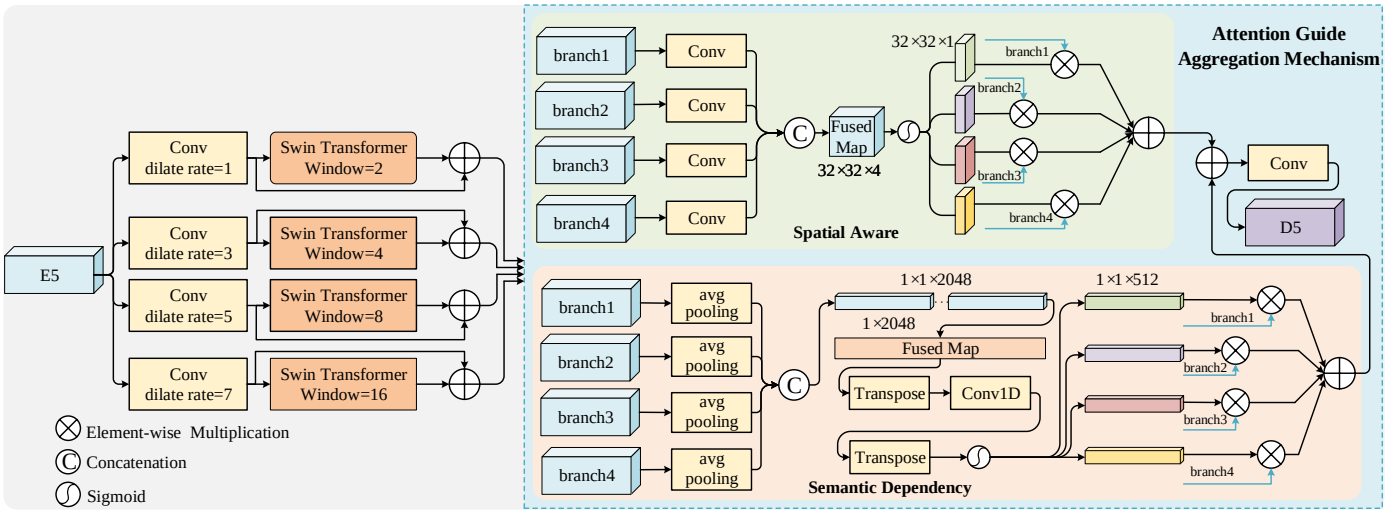


Fig. 3. Illustration of global-local context perception module (GLCPM).

feature maps are aggregated by sum operation and produce position fused map $F_{spatial}$. In this way, the submodule can perceive the correlation of spatial features at different scales, i.e.,

$$F_{spatial} = \sum_{i=1}^{4} W^i_{spatial} \otimes branch_i, \tag{4}$$

where $\sum$ is an element-wise sum operation. The $W^i_{spatial}$ is spatial weights corresponding to the feature map $branch_i$ and $\otimes$ is element-wise multiplication operation.

Semantic Dependency Submodule: By modeling the semantic dependencies between features at different scales, the model can effectively integrate different features from a channel perspective. The spatial information of feature map is compressed by using average pooling operation, i.e.,

$$Semantic_i = avg(brahch_i), \tag{5}$$

where $avg(\cdot)$ is average pooling. The compressed feature maps are concated to obtain a coarse fusion feature map $F_c \in \mathbb{R}^{2048\times32\times32}$. Flatten operation and 1D-convolution with the size of 11 are utilized to enhance the interaction between channels. Then, Sigmoid and channel split operations are used to obtain the channel weights of the branch feature. The overall process is defined as

$$W^i_{semantic} = split(\delta(Conv1d(f(Cat(Semantic_i))))), \tag{6}$$

where $Conv1d(\cdot)$ is Conv1d operation and $f(\cdot)$ is Flatten operation. The input feature maps and channel weights are multiplied to realize channel weighting, and the fusion of the feature map is realized by sum operation, i.e.,

$$F_{semantic} = \sum_{i=1}^{4} W^i_{semantic} \otimes branch_i, \tag{7}$$

where $W^i_{semantic}$ is channel weights corresponding to the feature map $branch_i$. By introducing this submodule, the model

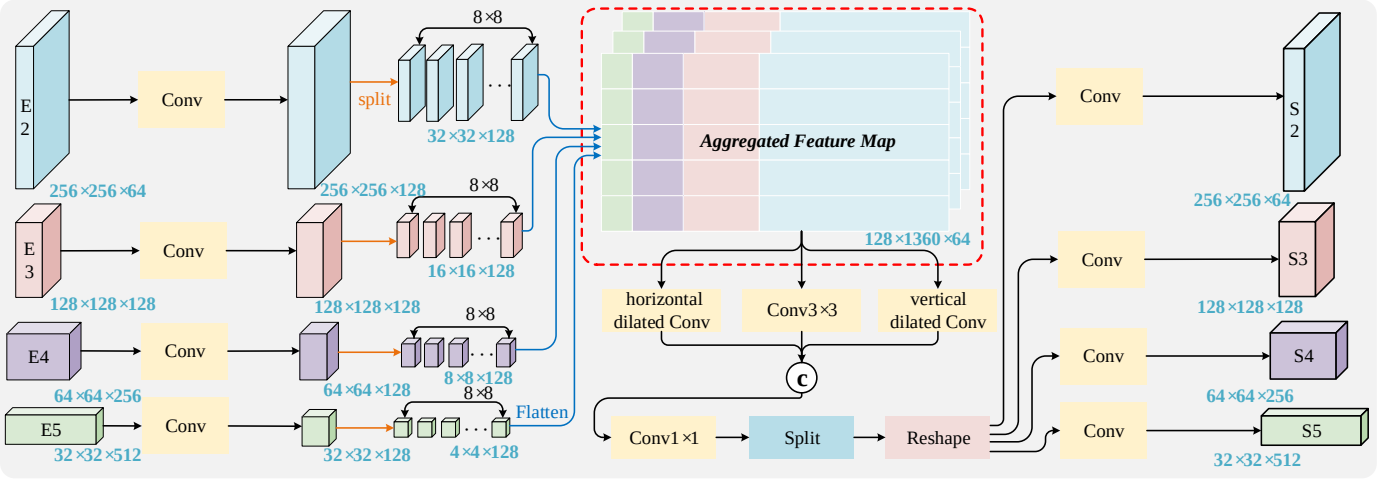


Fig. 4. Illustration of cross spatial-scale feature interaction module (CSFIM).

can perceive the dependencies between different channels and generate accurate semantic fused maps $F_{semantic}$. Finally, the spatial fused feature $F_{spatial}$ and the semantic fused feature $F_{semantic}$ are input into addition and 1×1 convolution operation to produce the fused feature map that contains rich detail characteristics and long-range contextual information.

C. Cross Spatial-Scale Feature Interaction Module

The skip connection plays a vital role in the encoder-decoder framework by supplementing rich detailed features for the decoder. The shallow feature contains rich road shape and texture features, which facilitate precise road edge detection. Besides, the deep feature carries accurate semantic information, it can assist the model locate road results accurately. However, existing methods typically resize the features through upsampling or pooling operations, which compromise information integrity across scales and hinder multi-scale feature interaction effectively. Unlike the existing dense connectivity methods, a novel cross spatial-scale feature interaction module that employs a centralized interaction strategy is designed. This novel approach design promotes both intra-scale and inter-scale feature interactions, that foster the provision of precise road features to the decoder. The architecture of the CSFIM is shown in Fig. 4.

Firstly, the channels of feature E_i from different scales are transformed to 128 by using 1×1 convolution to obtain H_i. Based on dimensions of the input feature maps, these feature maps H_i are reshaped into sequences with size of 8×8 through different slice operations $(h_i, w_i, 128) \rightarrow (\frac{h_i}{8}, \frac{w_i}{8}, 64, 128)$, followed by unfolding each slice feature to generate the feature $(\frac{h_i}{8} \times \frac{w_i}{8}, 64, 128)$. In the feature, the first dimension represents the position information of the slices, and the second dimension denotes the channel features of the slices.

The feature sequences of different scales are combined through a concatenation operation in the position dimension, thereby realizing the aggregation of features F_{fuse} from different scales. In this aggregated feature F_{fuse}, the row of the feature represents a set of different scale features at the same location, and the column of the fused feature denotes the set of the same scale features at different locations. Therefore, we use strip convolution to model the features from horizontal and vertical directions, and obtain the cross-scale features and cross-spatial features respectively. The 1×7 dilated convolution with rate of 129 is employed to obtain full-scale aggregated features with cross-scale feature interaction, and the 5×1 dilated convolution with a rate of 2 is employed to obtain features with cross-spatial interaction. Furthermore, the 3×3 convolution is adopted to obtain the feature at different locations and scales. These features are fused by concatenation and convolution operation, and ultimately generate cross spatial-scale fused features F_{csc}. The overall process is defined as

$$F_{csc} = Conv(Cat(HC(F_{fuse}), VC(F_{fuse}), Conv_3(F_{fuse}))), \tag{8}$$

where $HC(\cdot)$ is horizontal dilated convolution, $VC(\cdot)$ is vertical dilated convolution, and $Conv_3(\cdot)$ is convolution operation with kernel size of 3.

Then, the complementary feature S_i of the different decoder stages are obtained by employing channel split and the reshape operation$(\frac{h_i}{8} \times \frac{w_i}{8}, 64, 128) \rightarrow (h_i, w_i, 128)$. Finally, these features are deformed by a convolution operation with kernel size of 1 to align with the input feature map E_i. The overall process is defined as

$$S_i = Conv_1(R(split(F_{csc}))), \tag{9}$$

where $R(\cdot)$ is reshape operation and $Conv_1(\cdot)$ is convolution operation with kernel size of 1. By introducing this module, the network can integrate features of diverse scales adeptly while maintaining feature map unchanged. It can realize the profound fusion of road texture details with semantic information. Consequently, the model provides the decoder with comprehensive and precise road structural features.

D. Decoder

As shown in Fig. 2, the feature D_5 generated by GLCPM is input to the decoder to generate the road segmentation result. The input of decoder comprises three components:

the feature map D_{i+1} generated from the preceding decoder stage, the feature map E_i delivered by the skip connection, and the structural complementary feature S_i generated by the CSFIM. The feature map D_{i+1} is upsampled to the same resolution and channels as E_i by using 1×1 convolution and ConvTranspose2d operation. Then, these features are merged by addition operation. The feature maps D_i generated from multiple decoders are reshaped to match the resolution and channels of D_1 at the end of decoder. Finally, these feature maps are fused by an addition operation, which ultimately produces road prediction results of the same resolution as input image.

IV. Experiments

In this section, extensive experiments on RSI road datasets are conducted to demonstrate the effectiveness of C^{2}Net. Firstly, the dataset, evaluation metrics, and experiment settings are introduced. Then, a comparison and ablation study are conducted and analyzed. Finally, we provide some visual results.

A. Datasets

we conduct experiments on the DeepGlobe [39], Massachusetts [40], and Spacenet road datasets to validate the effectiveness. Referring to [9], the Deepglobe dataset is divided into 5,500 image pairs as a training set and 726 image pairs as a testing set. The Massachusetts road dataset consists of 1,171 pairs of 1500 × 1500 pixels aerial images. We resize the images to 1024 × 1024 pixels, and the experimental data consists of 1,108 pairs as the train set and 49 pairs as the test set. The SpaceNet dataset consists of 2,780 aerial images of 1300×1300 pixels, and these images are resized to 1024 × 1024 pixels. Referring to [36], the dataset is partitioned into two parts: 2,224 pairs for training and 556 pairs 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

$$Recall = \frac{TP}{TP + FN}, \tag{10}$$

$$Precision = \frac{TP}{TP + FP}, \tag{11}$$

$$\text{IoU} = \frac{TP}{TP + TN + FP}, \tag{12}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}, \tag{13}$$

where *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.,

$$S_{p\to T}(G, G^{'}) = 1 - \frac{1}{M}\sum \min(1, \frac{|L(a,b) - L(a_1,b_1)|}{L(a,b)}), \tag{14}$$

$$S_{T\to P}(G^{'}, G) = 1 - \frac{1}{M}\sum \min(1, \frac{|L(a,b) - L(a_1,b_1)|}{L(a_1,b_1)}), \tag{15}$$

$$APLS = \frac{1}{N}\sum (\frac{1}{\frac{1}{S_{p\to T}} + \frac{1}{S_{T\to P}}}), \tag{16}$$

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 model is based on Pytorch framework, and all the experiments are conducted on NVIDIA 3090 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, SpaceNet dataset, and 2 on the Massachusetts 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 two classical segmentation methods: Unet [21] and Deeplabv3+ [23], six methods introduce context: DLinkNet [6], GCBNet [28], SGCNNet [29], CMTFNet [41], CMLFormer [42] and CU-dGCN [43], two multi-scale feature fusion methods: Unet3+ [15] and Scaleformer [25], and a method based on multi-task joint training: CoaNet [32].

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↑	APLS↑
U-Net [21]	83.70	73.22	63.62	76.76	68.19
Deeplabv3+ [23]	84.74	72.68	63.73	75.71	67.91
DLinkNet [6]	82.93	78.81	67.66	79.43	71.72
Unet3+ [15]	79.66	78.51	64.77	77.35	65.63
GCBNet [28]	84.90	78.64	68.65	80.53	73.55
CoaNet [32]	80.63	77.42	65.13	78.69	64.62
Scaleformer [25]	81.01	79.10	66.20	78.63	66.16
SGCNNet [29]	84.57	76.00	66.45	78.73	69.98
CMTFNet [41]	76.03	78.71	61.68	74.76	60.36
CMLFormer [42]	<u>84.84</u>	<u>79.31</u>	<u>69.40</u>	<u>80.94</u>	<u>74.07</u>
CU-dGCN [43]	83.16	79.29	67.84	79.84	71.37
Ours	**85.69**	**79.39**	**70.00**	**81.50**	**74.45**

Results on the DeepGlobe dataset: Table I shows performances of road extraction methods on DeepGlobe dataset, it is evident that the proposed C^{2}Net has a significant advantage over the previous methods across all metrics. Specifically, the

TABLE II
ROAD EXTRACTION RESULTS ON THE MASSACHUSETTS ROAD DATASET.

Method	Recall↑	Precision↑	IoU↑	F1-score↑	APLS↑
U-Net [21]	**83.49**	76.39	66.52	79.66	76.17
Deeplabv3+ [23]	83.12	76.83	66.43	79.65	75.11
DLinkNet [6]	81.87	79.57	67.61	80.44	75.41
Unet3+ [15]	81.89	79.63	67.73	80.52	74.37
GCBNet [28]	83.01	79.07	68.05	80.79	75.42
CoaNet [32]	80.63	77.42	65.13	78.69	74.76
Scaleformer [25]	80.68	81.29	68.01	80.74	75.59
SGCNNet [29]	79.77	**82.98**	68.44	81.03	74.54
CMTFNet [41]	80.78	80.30	67.36	80.26	76.07
CMLFormer [42]	81.97	78.99	67.23	80.19	75.97
CU-dGCN [43]	79.65	80.95	67.13	80.14	75.75
Ours	81.57	81.13	**68.59**	**81.17**	**76.80**

TABLE III
ROAD EXTRACTION RESULTS ON THE SPACENET ROAD DATASET.

Method	Recall↑	Precision↑	IoU↑	F1-score↑	APLS↑
U-Net [21]	70.50	58.61	50.21	62.83	58.85
Deeplabv3+ [23]	71.33	63.11	53.63	65.77	63.37
DLinkNet [6]	72.37	62.70	53.99	65.96	65.26
Unet3+ [15]	67.06	64.73	52.24	64.88	58.86
GCBNet [28]	70.49	64.15	53.91	65.83	64.42
CoaNet [32]	65.02	**70.87**	55.06	66.52	54.16
Scaleformer [25]	64.81	64.70	51.06	63.17	54.73
SGCNNet [29]	67.38	63.38	51.85	63.87	59.80
CMTFNet [41]	**73.34**	62.73	54.49	66.44	**65.70**
CMLFormer [42]	70.44	64.22	54.06	65.90	65.27
CU-dGCN [43]	69.76	64.49	53.46	65.49	63.62
Ours	72.33	64.32	**55.10**	**66.98**	65.65

Unet network fails to capture the road context, resulting in inferior accuracy of generated road results. Models such as DLinkNet, GCBNet, and SGCNNet enhance the completeness of predicted roads by introducing context, thereby these methods achieve satisfactory performance in Precision metric. Compared with UNet3+, our method achieves improvements of 4.68% in Recall. It demonstrates the importance of cross-spatial and cross-scale feature interactions for extracting complete road topology. Moreover, the models that integrate road contextual information obtain higher APLS, such as CMLFormer and GCBNet. Meanwhile, the CMLFormer achieves the second-best performance, which reflects the importance of extracting long-distance road context. The proposed C^{2}Net achieves the best APLS by incorporating GLCPM, which can extract local details and global context and aggregate them adaptively, therefore the model can produce results with good connectivity.

Results on the Massachusetts dataset: As seen in Table II on the Massachusetts dataset, the proposed C^{2}Net obtains the best IoU, F1-score and APLS. It shows the effectiveness and robustness of the C^{2}Net. Notably, the CMLFormer method achieves competitive performance on Deepglobe dataset. However, it faces challenges when applied on the Massachusetts dataset, which is due to the limited quantity of images on Massachusetts datasets. This situation makes it challenging to optimize the Transformer-based approach. In contrast, our C^{2}Net combined dilated convolution with Transformer to capture comprehensive context, therefore the method also achieves satisfactory results on Massachusetts datasets.

Results on the Spacenet dataset: As shown in Table III, the C^{2}Net also obtains the best IoU and F1 as well as competitive APLS, which demonstrates the robustness of the designed method in different scenarios. Among them, the CoaNet method acquires the second-highest scores in both IoU and F1-score. However, it falls 11.49% lower than the C^{2}Net in APLS metrics. This significant decline is due to the multiple loss functions designed in CoaNet that guide the model to generate precise results. It also leads the model to determine some semantically ambiguous pixels as background. This situation inadvertently hinders the completeness of the predicted results.

E. Visual Results

To analyze the effectiveness of our method, some predicted results are presented. Based on the ranking of the models in the IoU metric on each dataset, we select seven well-performing predictions for presentation.

Results on the DeepGlobe dataset: We analyze experimental results in terms of both road completeness and precision. As shown in Fig. 5, C^{2}Net can produce complete and reasonable road results.

Results on completeness: The first line shows, compared with the existing method, our model not only can extract the road area blocked by the shadows, but also can produce the road results with excellent connectivity. This improvement is attributed to the designed GLCPM, which can accurately extract and fuse local details with global context, further improving the ability of model to discern occlusion regions. In contrast, there are lots of missing road regions in the road results predicted by DLinkNet and SGCNNet method. This is attributed to the factor that these models combined with the dilated convolution cannot capture the global context, which impairs the model to extract the overall topology feature of the road. Unlike these methods, the Scaleformer method demonstrates promising results by integrating feature maps across different scales, whereas the Transformer network struggles with obtaining sufficient road details. It results in the model producing thin road edges and poor performance at road junctions. The second and last rows also show the occlusion scene, and the proposed method achieves satisfactory results. This further demonstrates the robustness and effectiveness of our approach in handling complex scenarios, ensuring comprehensive road detection even in the presence of significant obstructions.

Results on precision: In the third row of images, the predicted result generated by C^{2}Net more closely approximates ground truth, while all other methods tend to identify the background as roads. This discrepancy underscores the superiority of the road results produced by C^{2}Net. The above results demonstrate that the method can effectively deal with the occlusion phenomenon.

Results on the Massachusetts dataset: As shown in Fig. 6, Massachusetts is primarily for urban road scenes, and C^{2}Net is also able to generate satisfactory road results. The

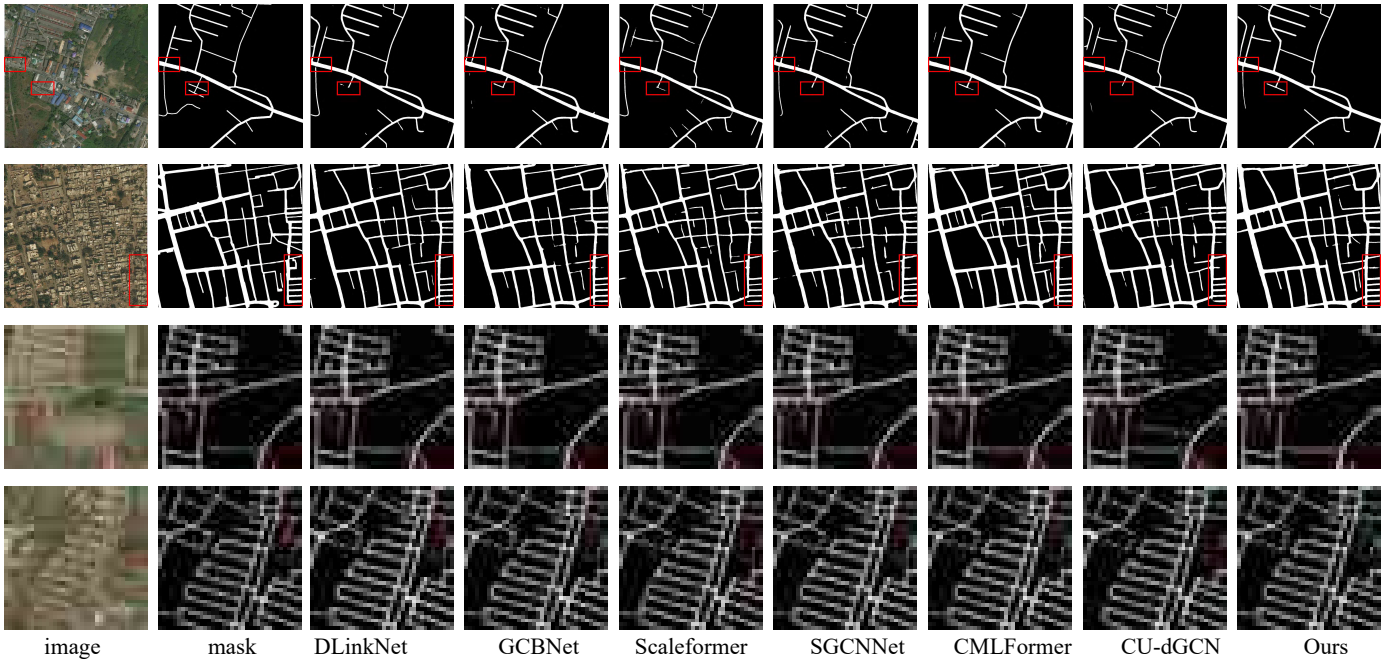


Fig. 5. Qualitative evaluations between C^{2}Net and comparison methods on DeepGlobe road datasets. The red boxes mark the areas where C^{2}Net outperforms other methods.

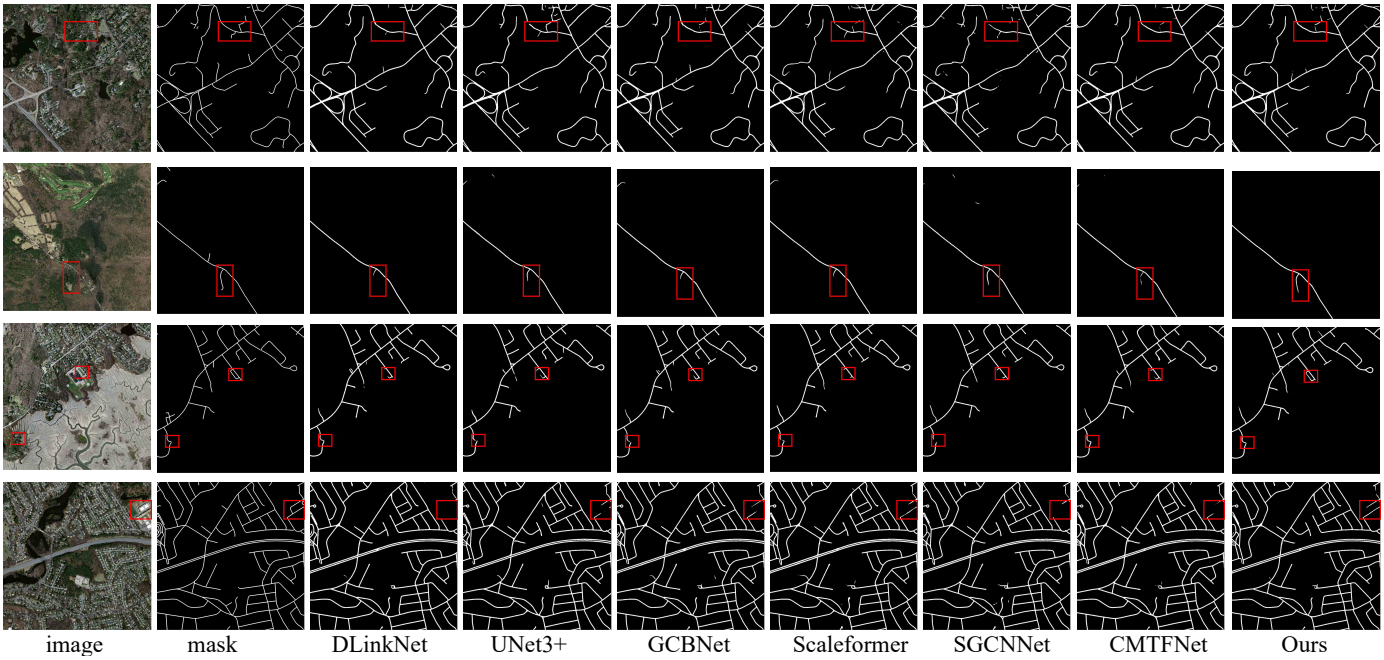


Fig. 6. Qualitative evaluations between C^{2}Net and comparison methods on Massachusetts road dataset. The red boxes mark the areas where C^{2}Net outperforms other methods.

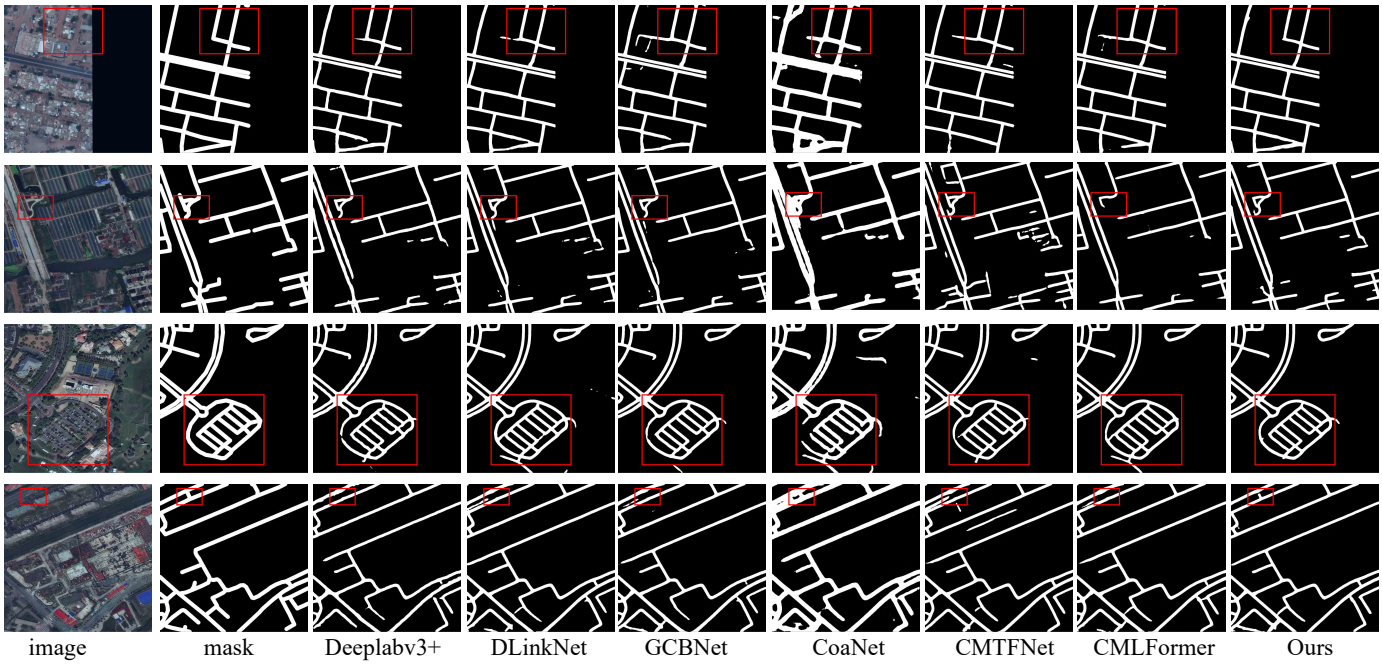


Fig. 7. Qualitative evaluations between C^{2}Net and comparison methods on Spacenet road dataset. The red boxes mark the areas where C^{2}Net outperforms other methods.

TABLE IV
PARAMETER COMPARISON AND INFERENCE SPEED OF DIFFERENT METHODS.

	U-Net	Deeplabv3+	DLinkNet	Unet3+	GCBNet	CoaNet	Scaleformer	SGCNNet	CMTFNet	CMLFormer	CU-dGCN	Ours
Params(M)	39.50	26.71	31.10	43.55	31.23	59.15	114.37	42.70	**22.64**	**22.64**	90.84	61.37
Time(s)	0.2260	0.2266	0.2268	0.3939	**0.1756**	0.6586	0.3083	0.4073	0.2316	0.2283	0.4577	0.3363

first and second rows of images show the situation where the road is occluded by surrounding trees, and this method is able to generate accurate and complete road results. The third and fourth images show, our method not only predicts complete road results in difficult-recognized road areas, but also generates predictions with good connectivity in the case where surrounding objects occlude roads.

Results on the Spacenet dataset: As shown in Fig. 7, whether in dense areas or open space, our proposed model can generate smooth and accurate road results compared to the comparative methods. This is attributed to the fact that our C^{2}Net can fully capture rich road details with accurate long-range context with the help of the GLCPM. Meanwhile, the designed method is able to realize cross spatial-scale feature interaction without information loss due to feature map degradation. The above results show that the C^{2}Net has better robustness in road extraction tasks and can effectively adapt to different scenarios.

F. Model Complexity and Time Complexity Analysis

As shown in Table IV, to comprehensively assess the practical value of our model, we compare it with other methods in terms of parameter count and inference speed. Specifically, we use the params to measure model complexity. Our C^{2}Net contains 61.37M parameters, which is on par with CoaNet. Additionally, we evaluate inference speed by measuring the time of producing road results on the SpaceNet dataset. Our method achieves approximately 0.3363 seconds per image, significantly outperforming Unet3+ in this regard. The increase in parameter count primarily stems from the extensive convolution operations involved in feature fusion within the CSFIM module. However, this module also contributes significantly to the overall performance improvement of the model. Taken together, these results highlight the superior efficiency and practical utility of our model in comparison to existing multi-scale feature fusion based road extraction methods.

G. Ablation Study

As shown in Table V, we conduct ablation experiment on the Deepglobe and Massachusetts dataset to demonstrate the influence of GLCPM and CSFIM. Three variants that combination of different parts are defined as ModelA (only baseline), ModelB (combine baseline with GLCPM), and ModelC (combine baseline with CSFIM) respectively.

Effect of GLCPM: Due to the introduction of GLCPM, model can capture accurate road context to predict difficult recognized road areas on the Deepglobe dataset. Therefore the Precision of the ModelB improves by 0.87% compared with ModelA. With the increase in Precision also causes decrease in Recall, and the model still achieves improvements in IoU and APLS respectively. It proves the ModelB produces better

TABLE V
ABALTION RESULTS ON THE DEEPGLOBE ROAD DATASET AND THE MASSACHUSETTS ROAD DATASET.

Components				DeepGlobe					Massachusetts				
Method	Baselline	GLCPM	CSFIM	Recall	Precision	IoU	F1-score	APLS	Recall	Precision	IoU	F1-score	APLS
ModelA	✓			83.56	80.17	69.11	80.76	73.62	81.29	80.25	67.69	80.34	76.15
ModelB	✓	✓		83.54	**81.04**	69.41	80.74	73.65	80.79	80.92	67.80	80.56	75.91
ModelC	✓		✓	84.90	79.49	69.61	81.05	74.43	**82.76**	79.13	67.93	80.67	76.56
C^2Net	✓	✓	✓	**85.69**	79.39	**70.00**	**81.50**	**74.45**	81.57	**81.13**	**68.56**	**81.17**	**76.80**

TABLE VI
ABALTION RESULTS ON THE DEEPGLOBE ROAD DATASET.

Components				DeepGlobe		
Method	Local	Global	Dual-att	IoU	F1-score	APLS
ModelD	✓		✓	69.12	**80.79**	**73.65**
ModelE		✓	✓	68.61	80.32	72.41
ModelB	✓	✓	✓	**69.41**	80.74	**73.65**

Fig. 8. The yellow box represents the result that the model predicts correctly in the occluded scene, and the red box shows the incorrect prediction in the occluded scene.

overall quality results. The experiment on the Massachusetts dataset also shows the conclusion.

We design experiments to demonstrate the necessity of dilated convolution with Transformer network. As shown in VI, IoU of the model decreases drastically when the dilated convolution and Transformer are removed in the GLCPM, respectively. It shows that the local and global context can assist the model in improving the completeness of produced results. Meanwhile, the performance of ModelE shows more decreases in all metrics. This shows that only relying on the Transformer network without integrating CNN for image feature extraction may introduce some irrelevant contextual information, thereby impacting the accuracy of prediction results. This is because the Transformer network may lack the ability to capture the visual details in the image. Therefore, it is necessary to combine CNN with Transformer network to extract precise context.

Effect of CSFIM: ModelC that combines CSFIM increases 1.34%, 0.5% and 0.29% in Recall, IoU, and F1-score respectively compared with ModelA on the DeepGlobe dataset. This demonstrates that it is important to improve the road target perception ability by integrating road features at different stages into the decoder. These enhancements are due to the cross-scale and cross-spatial features can be sufficiently aggregated. In this way, the model can obtain accurate semantic information and rich detail features. The ModelC shows a 0.33% improvement on F1-score and 0.41% on the APLS on the Massachusetts dataset, which indicates the model can generate higher quality road results with the help of the CSFIM.

H. Limition

Fig. 8 shows the scene where roads are blocked by surroundings. The designed C^2Net is able to generate complete and connected road results, as shown in the yellow box. This success is attributed to our method takes into account the introduction of complete context and the aggregation of cross spatial-scale features, which effectively alleviates the phenomenon of road occlusion. However, due to lack of semantic information about buildings, the model cannot fully explore the relationship between roads and buildings, which leads to the model being insensitive to the buildings. Consequently, there are some missing parts in the generated results, as shown in the red box. The correlation between roads and buildings is significant in RSI. By rationally modeling this correlation, the ability to analyze scenarios where roads are obscured by buildings can be more effectively supplemented to generate coherent road prediction results.

V. CONCLUSION

In this paper, we design a method for road extraction from remote sensing images. To enhance the capacity to distinguish roads from similar objects, the global-local context perception module is proposed to comprehensively capture and aggregate local detail features and global contextual information of roads hidden in the deep feature map, thereby improving the accuracy of predicted results. The cross spatial-scale feature interaction module is introduced to gather road features across various locations and stages, delivering the decoder with precise and comprehensive visual features. It can enhance the completeness of the generated results. The proposed C^2Net demonstrates competitive performance on publicly available datasets, which can produce accurate and well-connected road results compared with the existing methods. Besides, ablation experiments are conducted to analyze the effectiveness of the proposed module. In future research, we aim to develop a novel training strategy. It can extract precise context between roads and buildings to predict roads that are occluded by buildings.

REFERENCES

[1] Y. Du, Q. Sheng, W. Zhang, C. Zhu, J. Li, and B. Wang, "From local context-aware to non-local: A road extraction network via guidance of multi-spectral image," *ISPRS-J. Photogramm. Remote Sens.*, vol. 203, pp. 230–245, 2023.

[2] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17 853–17 862.

[3] J. Zhang, J. Lei, W. Xie, G. Yang, D. Li, Y. Li, and K. Seghouane, "Multimodal informative vit: Information aggregation and distribution for hyperspectral and lidar classification," *arXiv:2401.03179*, 2024.

[4] Z. Chen, L. Deng, Y. Luo, D. Li, J. M. Junior, W. N. Gonçalves, A. A. M. Nurunnabi, J. Li, C. Wang, and D. Li, "Road extraction in remote sensing data: A survey," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 112, p. 102833, 2022.

[5] 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.

[6] 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.

[7] Y. Wang, J. Seo, and T. Jeon, "NL-LinkNet: Toward lighter but more accurate road extraction with nonlocal operations," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[9] 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.

[10] Y. Wang, L. Tong, S. Luo, F. Xiao, and J. Yang, "A multi-scale and multi-direction feature fusion network for road detection from satellite imagery," *IEEE Trans. Geosci. Remote Sensing*, pp. 1–1, 2024.

[11] 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*, pp. 1–1, 2024.

[12] Q. Yu, X. Zhao, Y. Pang, L. Zhang, and H. Lu, "Multi-view aggregation network for dichotomous image segmentation," *arXiv:2404.07445*, 2024.

[13] S. Liu, Y. Ma, X. Zhang, H. Wang, J. Ji, X. Sun, and R. Ji, "Rotated multi-scale interaction network for referring remote sensing image segmentation," *arXiv:2312.12470*, 2023.

[14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[15] 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 *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 1055–1059.

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

[17] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "RGB-induced feature modulation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sensing*, 2023.

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

[19] Q. Li, Y. Yuan, and Q. Wang, "Multi-scale factor joint learning for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sensing*, 2023.

[20] 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.

[21] 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.

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

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[24] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2022, pp. 205–218.

[25] 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.

[26] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv:2401.04722*, 2024.

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

[28] 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.

[29] 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.

[30] L. Dai, G. Zhang, and R. Zhang, "RADANet: Road augmented deformable attention network for road extraction from complex high-resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–13, 2023.

[31] A. Batra, S. Singh, G. Pang, S. Basu, C. Jawahar, and M. Paluri, "Improved road connectivity by joint learning of orientation and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 385–10 393.

[32] J. Mei, R.-J. Li, W. Gao, and M.-M. Cheng, "CoANet: Connectivity attention network for road extraction from satellite imagery," *IEEE Trans. Image Process.*, vol. 30, pp. 8540–8552, 2021.

[33] Y. Zao, Z. Zou, and Z. Shi, "Topology-Guided road graph extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sensing*, 2023.

[34] X. Lu, Y. Zhong, Z. Zheng, D. Chen, Y. Su, A. Ma, and L. Zhang, "Cascaded multi-task road extraction network for road surface, centerline, and edge extraction," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–14, 2022.

[35] A. Mosinska, M. Koziński, and P. Fua, "Joint segmentation and path classification of curvilinear structures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1515–1521, 2019.

[36] J. Hu, J. Gao, Y. Yuan, J. Chanussot, and Q. Wang, "LGNet: Location-guided network for road extraction from satellite images," *IEEE Trans. Geosci. Remote Sensing*, 2023.

[37] Y. Qi, Y. He, X. Qi, Y. Zhang, and G. Yang, "Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6070–6079.

[38] J. Zhang, X. Yang, R. Jiang, W. Shao, and L. Zhang, "RSAM-Seg: A sam-based approach with prior knowledge integration for remote sensing image semantic segmentation," *arXiv:2402.19004*, 2024.

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

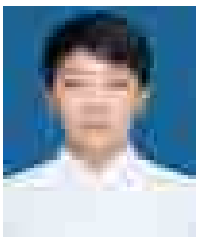
[40] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2010, pp. 210–223.

[41] 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 Trans. Geosci. Remote Sensing*, 2023.

[42] 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 Observ. Remote Sens.*, 2024.

[43] A. A. Vekinis, "Graph reasoned multi-scale road segmentation in remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.* IEEE, 2023, pp. 6890–6893.

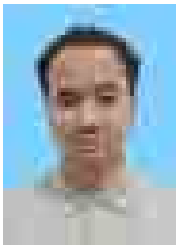
Zhigang Yang is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing and computer vision.



Wei Zhang is pursuing a Ph.D. in computer science and technology 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 computer vision, remote sensing, and 3D reconstruction.

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.

Weiping Ni received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2004, the M.S. degree from the National University of Defense Technology, Changsha, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent system from Xidian University, Xi'an, China, in 2016. Since 2014, he has been a Research Associate with the Northwest Institute of Nuclear Technology, Xi'an. His research interests include remote sensing image processing, automatic target recognition, and computer vision.



Junzheng Wu received the B.Sc. degree in automation from Tsinghua University, Beijing, China, in 2008, the M. Sc. degree in signal and information processing from Northwest Institute of Nuclear Technology, Xi'an, China, in 2011, and the Ph. D. in information and communication engineering from National University of Defense Technology, Changsha, China, in 2022, respectively. Currently, he is an Associate researcher with the department of remote sensing, Northwest Institute of Nuclear Technology, Xi'an, China. His research interests include processing of remote sensing images and machine learning.

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/)