IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 October 2023, accepted 7 November 2023, date of publication 13 November 2023,
date of current version 16 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3332121

RESEARCH ARTICLE

DRCNet: Road Extraction From Remote Sensing Images Using DenseNet With Recurrent Criss-Cross Attention and Convolutional Block Attention Module

DEBIN WEI, PINRU LI, HONGJI XIE, AND YONGQIANG XU
Communication and Network Laboratory, Dalian University, Dalian 116622, China

Corresponding author: Debin Wei (weidebin@163.com)

This work was supported by the National Natural Science Foundation of China under Grant 61931004 and Grant U21B2003.

ABSTRACT Extracting road networks from remote sensing images holds critical implications for various applications including autonomous driving, path planning, and road navigation. Despite its importance, the task remains arduous due to the complex backdrops in remote sensing imagery, intricate road geometries, and the challenges arising from vegetation and structural obstructions. To address these multifaceted issues, we introduce a specialized model for road extraction in remote sensing images, termed DRCNet. This model employs a pre-trained DenseNet-121 as its encoder and is fortified with both Recurrent Criss-Cross Attention (RCCA) and Convolutional Block Attention Module (CBAM). RCCA facilitates the capture of global contextual information across all pixels, thereby enriching the model's understanding of global image relationships. Simultaneously, CBAM is integrated within the skip connections to optimize the network's focus on significant road features. Comprehensive experiments conducted on both the DeepGlobe Road and Massachusetts Road datasets substantiate that DRCNet outperforms other benchmark models in road detection tasks.

INDEX TERMS Road extraction, remote sensing image, DenseNet-121 network, attention module.

I. INTRODUCTION

Owing to advancements in remote sensing technology, the availability of high-resolution imagery has surged, finding applications across a spectrum of fields such as environmental monitoring, natural resource management, urban planning, and geographic information systems (GIS) [1]. Among these applications, the extraction of road information from high-resolution remote sensing images serves as a pivotal component for GIS mapping and updates [2]. Despite growing interest and research efforts, automated road extraction remains a complex challenge due to the inherent complexities presented by roads in high-resolution imagery.

Road features in high-resolution remote sensing images mainly include geometric features, spectral features, topological features, and texture features [3]. According to these features, many road extraction methods have been proposed, which can be divided into two categories: traditional road extraction methods and deep learning-based road extraction methods. Traditional road extraction methods are based on experience, which mainly uses the shallow features of the image, such as gray level, edge, texture, and geometric shape. Deep learning-based methods mainly use deep convolutional neural networks to implicitly extract deep abstract features from original images, and use these features to automatically extract road information.

Traditional techniques for road extraction predominantly encompass methods such as threshold segmentation, template matching, region growing, edge detection, and morphological algorithms. For instance, Singh and Garg [4] developed an automated road extraction technique that employs adaptive global thresholding followed by morphological operations

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

to eliminate small, non-road regions. Similarly, Xie et al. [5] introduced an adaptive variable bandwidth template matching algorithm aimed at extracting linear roads from remote sensing images. Lu et al. [6] devised a rapid road network extraction strategy based on region growth, tailored for remote sensing images of varying resolutions. Darweesh et al. [7] proposed an algorithm reliant on Canny edge detection using the Canny operator, while Zhou et al. [8] advanced an automatic road extraction algorithm founded on topological derivatives and mathematical morphology. Despite the degree of success achieved by these traditional methods, they come with notable limitations, including but not limited to, a lack of robustness in handling complex terrains and road occlusions, the necessity for substantial manual intervention and post-processing, limited extraction accuracy, and an inability to support real-time and large-scale data processing.

With the development of deep learning, deep convolutional neural networks (DCNN) have achieved significant success in tasks such as remote sensing data classification [9], object detection [10], [11], [12] and semantic segmentation [13], [14], [15], [16]. Concurrently, deep learning-based road extraction methods have overcome the inherent limitations of traditional approaches, leading to a substantial surge in the advancement of road extraction technology. Various deep learning architectures have demonstrated their efficacy in producing high-quality road extraction outcomes. For example, Ronneberger et al. [17] introduced the U-Net architecture, an extension of fully convolutional networks (FCNs) [18]. U-Net utilizes transpose convolutions for upsampling and employs skip connections to integrate features from both the encoder and decoder segments of the network. This integration enables effective fusion of information across multiple layers, thereby aiding in the recovery of finer spatial details crucial for segmentation. Zhang et al. [19] melded deep residual learning [20] with U-Net to develop the ResUnet framework specifically for road area extraction. Similarly, Chen et al. [21] presented DeepLabV3, which incorporates an atrous spatial pyramid pooling (ASPP) module, employing dilated convolutions at varying rates to expand the receptive field and to better capture multi-scale contextual information. This significantly enhances segmentation performance, particularly for smaller objects. Zhou et al. [22] unveiled D-LinkNet, featuring a LinkNet [23] architecture with a pretrained encoder and an embedded ASPP module, effectively boosting the network's ability to connect road features. Finally, Wang et al. [24] proposed the DDU-Net model, which amalgamates dilated convolutions and attention mechanisms to enhance the extraction of global contextual semantic features. Moreover, DDU-Net employs a dual-decoder structure to preserve a higher proportion of low-level features, thereby offering more detailed detection for smaller roads. While these DCNN-based methods have demonstrated good performance in road extraction, there are still several challenges in road extraction:

1) Roads in high-resolution remote sensing images can often be affected by a multitude of intricate environmental and contextual interferences, such as buildings, trees, and vehicles. These disruptive elements share similar textures and colors with the roads, making it a formidable challenge for road extraction algorithms to distinguish roads from surrounding objects effectively.
2) Roads may be partially obscured by buildings, trees, or other objects, and the presence of shadows can further introduce variations in brightness and color within the road areas, compounding the complexity of road extraction. These occlusions and shadows result in discontinuities in the road network, necessitating algorithms capable of identifying and reconstructing these missing road segments.
3) Roads exhibit a diverse range of shapes and topological structures, including straight segments, curves, intersections, and roundabouts, among various other forms and connectivity patterns. This implies that road extraction algorithms must be endowed with a high degree of flexibility, enabling them to adapt to roads of differing shapes and topological configurations.

In order to identify approaches for tackling the aforementioned challenges, we conducted an extensive review of existing deep learning techniques [25], [26], [27]. During this review, we observed that many of these methods commonly employ ResNet as the backbone for feature extraction. The use of atrous spatial pyramid pooling (ASPP) is also prevalent, as it broadens the receptive field of the network and enhances its ability to accumulate multi-scale contextual data. Additionally, the incorporation of upsampling operations with skip connections is widely adopted to enable the network to amalgamate both low-level and high-level features. However, these methods are not without drawbacks: (1) they tend to lose crucial road features during the image downsampling process; (2) they often struggle to capture long-range dependencies in remote sensing imagery, leading to suboptimal extraction accuracy; and (3) while skip connections do facilitate the blending of deep and shallow semantic features, this straightforward fusion approach neglects the spatial and channel-wise distribution of road information, thereby constraining the network's segmentation capabilities.

To address these challenges, we draw upon the strengths of DenseNet [28] and attention mechanisms in deep learning [29], [30] to introduce a novel encoder-decoder network optimized for high-resolution road segmentation in remote sensing images, which we designate as DRCNet. Our network employs a pre-trained DenseNet-121 architecture as the encoder to extract feature maps endowed with varying levels of semantic information from its dense blocks. To capture rich, global contextual data, we incorporate a Recurrent Criss-Cross Attention (RCCA) module at the core of both the encoder and decoder segments of the network. Moreover, in the skip connections, we deploy the Convolutional Block Attention Module (CBAM) [31] to weight both spatial and

channel-wise information within the feature map, thereby enhancing the ability to precisely locate and extract road information.

The main contributions of this paper are as follows:

1) A specialized DRCNet model has been introduced for the automatic extraction of intricate road networks from high-resolution remote sensing images. By incorporating a pre-trained DenseNet-121 as the network's encoder, it effectively enhances the model's capacity to represent features, resulting in an improved precision in remote sensing road extraction.
2) We enhance the network's ability to capture long-range dependencies between pixels and improve its understanding of image context information by introducing the RCCA module into the network. This ensures that the model can better recognize the interaction between roads and their surrounding environment, ultimately improving the accuracy and continuity of road extraction.
3) By incorporating the CBAM module into skip connections, it facilitates the provision of precise road structure information while minimizing background interference, effectively mitigating challenges related to inaccurate information and misclassification in road prediction.
4) By conducting a thorough analysis of the results obtained from comparative experiments and ablation studies on two publicly available road extraction datasets, our findings underscore the compelling necessity of the proposed DRCNet model. These experiments clearly demonstrate that the DRCNet not only markedly enhances the quality of road extraction results but also exhibits robust generalization capabilities. In the realm of road extraction, DRCNet's superior performance surpasses that of some cutting-edge models.

The remainder of the paper is organized as follows: Section II provides a comprehensive discussion of relevant literature. In Section III, the proposed method is introduced in detail. Section IV meticulously outlines the experimental setup, encompassing information about the dataset, experimental environment, and evaluation metrics. Section V conducts a thorough analysis of the experimental results using a variety of metrics. Finally, Section VI summarizes the paper's key findings and contributions and offers insights into future prospects.

II. RELATED WORKS

A. ROAD SEGMENTATION MODEL

Road segmentation networks based on deep learning typically employ an encoder-decoder architecture. The encoder serves as the feature extractor for the entire network, and architectures such as VGG [32] and ResNet [20] are commonly used in road segmentation networks to gradually extract feature maps containing high-level semantic information from input images. The role of the decoder is to upsample the high-level features extracted by the encoder and perform pixel-wise classification to obtain prediction results at the same spatial level as the input image. Zhang et al. [19] introduced the ResUnet network for road area extraction, which combines the advantages of residual learning and U-Net to achieve excellent road segmentation results. Zhou et al. [22] proposed D-LinkNet, which uses LinkNet as its backbone network and incorporates the ASPP module in the central part. By enlarging the receptive field and fusing multi-scale features in the central part while preserving detailed information, D-LinkNet can address issues related to narrow, connectivity, and long-span roads to some extent. Ding and Bruzzone [33] presented DiResNet, which introduces directional supervision into the network. This imparts the model with directional awareness, thereby enhancing the detection of linear features and improving the integrity and connectivity of road extraction. Ge et al. [34] introduced a deep feature-review transmit network to review and promote contour features back to the encoder network. This effectively mitigates road fragmentation and missing connection points commonly encountered in road extraction. Jiang et al. [35] proposed a road semantic segmentation model that integrates attention mechanism, gated decoding block, and dilated convolution. The central part of the network combines serial and parallel dilated convolutions with coordinate attention modules, effectively expanding the network's receptive field while enhancing its feature extraction capabilities in spatial and domain channel information. Additionally, gated convolutions are introduced in the decoder part to enhance the model's ability to extract road edges. Wang et al. [36] introduced NL-LinkNet, which effectively captures contextual information by incorporating a non-local neural network, leading to more accurate road segmentation. Dai et al. [37] combined deformable convolutions with spatial self-attention mechanisms and proposed a road enhancement deformable attention network to learn long-range dependencies for specific road pixels, thereby enhancing road segmentation results.

B. ATTENTION MECHANISMS

The attention mechanism is a method that mimics the human visual system and is widely cited in the field of computer vision. Mnih et al. [38] designed the attention mechanism on the RNN model for image classification and achieved good performance. SENet proposed by Hu et al. [39] uses attention weights to adaptively calibrate channel features to improve the representation ability of neural networks. Woo et al. [31] proposed CBAM, which realizes the adaptive selection of channel and spatial features of convolutional neural networks by combining channel attention and spatial attention modules; and improves the representation and discrimination ability of the network. SKNet proposed by Li et al. [40] is an attention module based on feature combination. SKNet can adaptively adjust the receptive field size of the convolution kernel, to better capture features of different scales. Fu et al. [41] proposed a Dual Attention

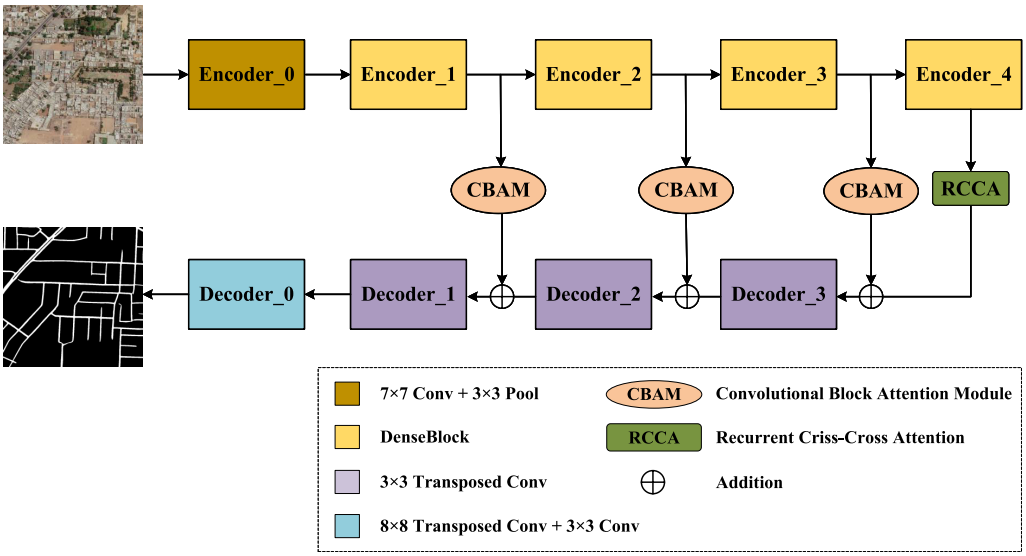


FIGURE 1. The overall structure of the DRCNet network.

network (DANet) to adaptively fuse local features and their global dependencies, showing good performance on image segmentation tasks. Huang et al. [42] proposed a criss-cross attention network (CCNet) to obtain the context information of the whole image in a very effective and efficient way. All of these modules are almost plug-and-play and can be embedded into any existing network to improve speed, result quality, and generalization ability.

The attention mechanism is currently widely applied in various remote sensing image interpretation tasks. For example, inspired by the attention mechanism, Yao et al. [14] proposed a new multimodal deep learning framework called ExViT for land use and land cover classification tasks. Li et al. [43] proposed a synergistic attention perception neural network (SAPNet) for semantic segmentation of remote sensing images. To jointly model spatial and channel affinity, they designed a synergistic attention module (SAM) that allows for the extraction of channel affinity while preserving spatial details. Wan et al. [44] constructed a dual-attention road extraction network (DA-RoadNet) using a shallow encoder framework. This approach explores and analyzes the correlation of road features in both spatial and channel dimensions, allowing the extraction of road information from complex environments.

III. DRCNET NETWORK STRUCTURE

As shown in Figure 1, DRCNet adopts an encoder-decoder structure, where the encoder completes the encoding function of the sample features, and the decoder realizes the restoration of the feature decoding. First, we use the pre-trained DenseNet-121 to build an encoder to extract dense features and ensure accurate road segmentation. Then, the recurrent criss-cross attention module (RCCA) was used as the connection part of the encoder and decoder to learn the long-range dependencies of road pixels from the feature map generated by the DenseNet encoder; and provide the global information to the feature decoding module. Finally, the shallow feature map generated by the dense block is passed through the CBAM module to obtain the channel attention features and spatial attention features, which enhances the attention degree of road information while suppressing the background information, and provides more accurate road feature information for the decoder module.

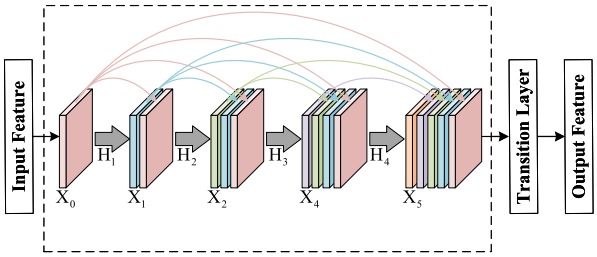


FIGURE 2. A 5-layer dense block.

A. ENCODER:DENSENET-121

As verified by several experiments, the pre-trained DenseNet-121 achieves better results with fewer parameters. Therefore, it is chosen as the encoder for road feature extraction. Dense block is the core module of DenseNet-121. DenseNet-121 consists of multiple dense blocks, each of which consists of multiple convolutional layers. Inside each dense block, the input to each convolutional layer is the concatenation of the feature maps of all previous layers. This densely connected design makes DenseNet perform better in feature reuse, effectively reducing the number of parameters of the model and improving the overall computational efficiency. In addition, this design also helps alleviate the vanishing

TABLE 1. Layer configuration of DenseNet-121 in DRCNet.

Layers	Output size	DenseNet-121
Convolution	512 × 512	7 × 7conv, stride2
Pooling	256 × 256	3 × 3maxpool, stride2
Dense Block (1)	256 × 256	[1 × 1conv, 3 × 3conv] × 6
Transition Layer (1)	256 × 256	1 × 1conv
	128 × 128	2 × 2avgpool, stride2
Dense Block (2)	128 × 128	[1 × 1conv, 3 × 3conv] × 12
Transition Layer (2)	128 × 128	1 × 1conv
	64 × 64	2 × 2avgpool, stride2
Dense Block (3)	64 × 64	[1 × 1conv, 3 × 3conv] × 24
Transition Layer (3)	64 × 64	1 × 1conv
	32 × 32	2 × 2avgpool, stride2
Dense Block (4)	32 × 32	[1 × 1conv, 3 × 3conv] × 16

gradient problem. Figure 2 illustrates a 5-layer dense block, which is implemented as follows:

$$X_l = H_l([X_0, X_1, \ldots, X_{l-1}]), \tag{1}$$

where X_l represents the feature map of the layer. $[X_0, X_1, \ldots, X_{l-1}]$ represents the feature map that is connected to the layer and $H_l(\cdot)$ is a composite function that includes batch normalization (BN), rectified linear units (ReLU), and convolution.

However, as the number of layers of the network becomes deeper, the number of channels becomes larger, and the number of parameters becomes larger, which makes it difficult to train a deeper network. To this end, DenseNet also contains an important Transition Layer to connect two dense blocks. The transition layer reduces the number of channels of the obtained feature map to half of the originaland performs downsampling to halve the size, thereby simplifying the calculation and improving the calculation efficiency.

The detailed network configuration of DenseNet-121 in DRCNet is shown in Table 1, which lists the size and number of convolution kernels in each convolutional layer, as well as the output size of the feature map.

B. RECURRENT CRISS-CROSS ATTENTION (RCCA)

Recurrent criss-cross attention (RCCA) is an enhanced version of Criss-cross attention (CCA), which uses the attention mechanism to facilitate the information exchange between different spatial locations in the feature map to capture long-range dependencies and enhance the ability of feature representation. RCCA consists of two CCA submodules with shared parameters. The calculation process of CCA is illustrated in Figure 3.

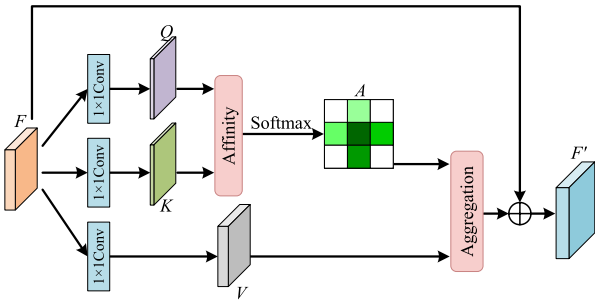


FIGURE 3. The calculation process of criss-cross attention.

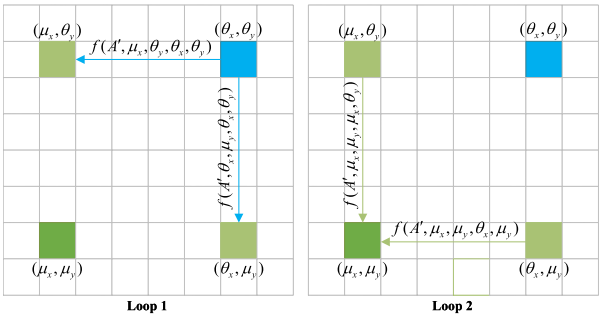


FIGURE 4. An example of information propagation in the RCCA module.

In CCA, the input feature map $F \in R^{C\times H\times W}$ is first processed by three parallel 1×1 convolution operations to obtain the feature maps $Q \in R^{C'\times H\times W}$, $K \in R^{C'\times H\times W}$, and $V \in R^{C\times H\times W}$ ($C' < C$). At each position u in the spatial dimension of Q, a feature vector $Q_u \in R^{C'}$ can be obtained. Simultaneously, extract the feature vectors of $H+W-1$ pixels in both the horizontal and vertical directions at the position of u in K, resulting in the set $\Omega_u \in R^{(H+W-1)\times C'}$. Then, the feature map $D \in R^{(H+W-1)\times(W\times H)}$ is generated by affinity operation, and the softmax operation is performed on D to transform it into a new feature layer A with the same size.The computation of affinity operation is shown as follows:

$$d_{i,u} = Q_u \Omega_{i,u}^T, \tag{2}$$

where $\Omega_{i,u} \in R^{C'}$ is the ith element in Ω_u, and $d_{i,u} \in D$ is the degree of correlation between features Q_u and $\Omega_{i,u}$.

Based on the aforementioned generation process of Ω_u, create the set $\Phi_u \in R^{(H+W-1)\times C}$ on V. Then, perform an aggregation operation on the feature map A and V, resulting in the feature map $F' \in R^{C\times H\times W}$ enriched with vertical and horizontal spatial context information. The aggregation operation is computed as follows:

$$F'_u = \sum_{i\in|\Phi_u|} A_{i,u}\Phi_{i,u} + F_u, \tag{3}$$

where F'_u is the feature vector at position u in F', $A_{i,u}$ is the value at channel i and position u in A, and $\Phi_{i,u}$ is a feature vector at position i of Φ_u.

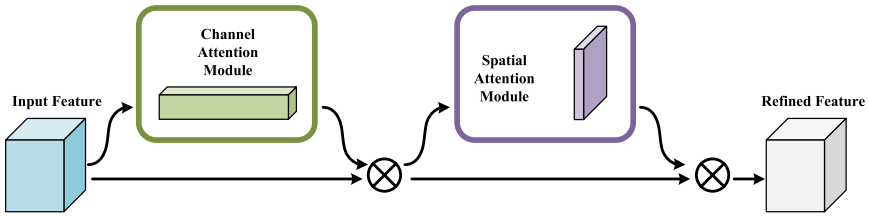


FIGURE 5. Convolutional block attention module.

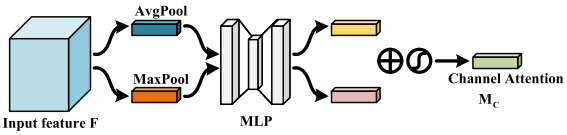


FIGURE 6. Channel attention module.

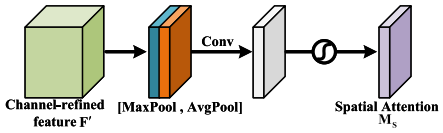


FIGURE 7. Spatial attention module.

However, the CCA module captures remote context information only in horizontal and vertical directions, while the connections between pixels and surrounding pixels are still sparse. For ease of understanding, we visualize the information propagation in RCCA as shown in Figure 4. In Loop1, the pixel at position (θ_x, θ_y) first passes the information to (μ_x, θ_y) and (θ_x, μ_y). Then, in the Loop2, (μ_x, θ_y) and (θ_x, μ_y) transfer information to (μ_x, μ_y). Therefore, the RCCA module, which consists of two CCA submodules, is able to obtain remote dependencies from all pixels and generate new feature layers with dense, rich context information.

C. CONVOLUTIONAL BLOCK ATTENTION MODULE (CBAM)

The overall structure of the convolutional block attention module (CBAM) is illustrated in Figure 5. It consists of a channel attention module (CAM) and a spatial attention module (SAM). Specifically, the CAM, as shown in Figure 6, is utilized to enhance the feature representation capability between different channels. It achieves this by learning channel weights to adaptively select important channel features. On the other hand, the SAM, depicted in Figure 7, is employed to enhance the feature representation capability across different spatial positions. It accomplishes this by learning spatial weights to adaptively select significant spatial regions. This design enables the CBAM module to flexibly perform feature selection and enhancement within convolutional neural networks, effectively improving the network's feature extraction performance across various dimensions.

Compared to SENet, CBAM not only focuses on the relationships between channels of the input image but also emphasizes the spatial relationships within the image. The channel attention module highlights the significance of eachfeature channel within the input feature map. It achieves this by computing the internal relationships between channels and ranking their importance to allocate specific weights for each channel. Firstly, average pooling and max pooling are applied on the feature map $F \in R^{C\times H\times W}$ to obtain feature maps F^C_{avg} and F^C_{max}. Subsequently, F^C_{avg} and F^C_{max} are fed into a weight-shared network composed of multiple perceptrons (MLP) to generate the channel attention map $M_C \in R^{1\times 1\times C}$. The mathematical expression for this process is as follows:

$$\begin{aligned} M_C(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^C_{avg})) + W_1(W_0(F^C_{max}))), \end{aligned} \tag{4}$$

where σ represents the sigmoid function, and W_0 and W_1 denote the shared weights in the MLP.

The spatial attention mechanism explores the internal relationships of the feature map at the spatial level, determining which regions of features are crucial and complement the channel attention mechanism. Firstly, average pooling and max pooling are conducted along the channel dimension on the input features, resulting in F^S_{avg} and F^S_{max}. Subsequently, F^S_{avg} and F^S_{max} are concatenated and convolved by a standard convolution layer. Finally, the updated feature map $M_S \in R^{1\times H\times W}$ is obtained through the spatial attention mechanism. The entire process can be represented as follows:

$$\begin{aligned} M_S(F) &= \sigma(f^{7\times 7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times 7}([F^S_{avg}; F^S_{max}])), \end{aligned} \tag{5}$$

where σ represents the sigmoid function, $f^{7\times 7}$ denotes a convolution operation with a kernel size of 7×7.

The complete computation formula for the CBAM module is as follows:

$$F' = M_C(F) \otimes F, \tag{6}$$

$$F'' = M_S(F') \otimes F', \tag{7}$$

where F' represents the input features after undergoing the channel attention operation, F'' is the output features after refinement, and $\otimes$ denotes pixel-wise multiplication.

IV. EXPERIMENT

A. DATASET

To verify the effectiveness of the proposed method, two different remote sensing-image road datasets are selected to train and validate the proposed model. One of the datasets is the DeepGlobe Road dataset and the other is the Massachusetts Road dataset. Let's briefly introduce two datasets:

(1) The DeepGlobe Road dataset is a pixel-level labeled dataset containing pixel-level annotated data from Thailand, India, and Indonesia. The images include various scenes such as urban, rural, and suburban. The image spatial resolution is 0.5 m/pixel and the image size is 1024×1024 pixels. The dataset contains 6226 pairs of training images and image labels. All images are randomly divided into training set and test set according to the ratio of 4:1, resulting in a total of 4981 images in the training set and 1245 images in the test set.

(2) The Massachusetts Road dataset is a publicly available dataset for road extraction research. It covers an area of more than 500 square kilometers in Massachusetts and contains a variety of different landscape types. The dataset includes 1108 training images, 49 testing images, and 14 validation images. The size of the images is 1500×1500 pixels with a resolution of 1.2 m. To facilitate experimentation and ensure consistent experimental conditions, we uniformly crop the original images to a size of 1024×1024 pixels using random cropping.

B. IMPLEMENTATION DETAILS

This experiment utilizes the PyTorch deep learning framework on a Linux system. The environment versions include torchvision = 0.14.1, torch = 1.13.1, and python = 3.7.16. The model is trained and tested on two NVIDIA GeForce RTX 3090 (24GB) GPUs. We use BCE (Binary Cross entropy) + dice coefficient loss as the loss function and choose Adam [45] as our optimizer. The initial learning rate is set at 2e-4 and train our network with a batch size of 8. When the training loss stagnates, the learning rate is reduced to one-fifth of its current value. Training is stopped when the learning rate drops below 5e-7. The model with the lowest loss on the training set is chosen for testing. During the prediction phase, test time augmentation (TTA) is applied, involving horizontal, vertical, and diagonal flips of the images. Each image is predicted 8 times, and the outputs are then aligned with the original images. The averaged probability of each prediction is computed, using a threshold of 0.5 to generate binary output results.

C. EVALUATION METRIC

In order to prove the effectiveness of the proposed method, Precious, Recall, F1-score, and IoU are selected as the evaluation metrics to represent the quality of road extraction. Precious represents the ratio of pixels correctly predicted as roads to all pixels predicted as roads. Recall represents the ratio of pixels correctly predicted as roads to the total pixels of roads. F1-score represents the harmonic mean of precision and recall. IoU is the ratio of the intersection and union of the predicted and true values. It is calculated as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{8}$$

$$Recall = \frac{TP}{TP + FN}, \tag{9}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{10}$$

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

where TP signifies pixels correctly predicted as road, FP denotes pixels erroneously predicted as road while they are background. TN represents pixels accurately predicted as background, and FN indicates pixels mistakenly predicted as background when they are road.

TABLE 2. Comparative results on the DeepGlobe Road dataset.

Method	Recall	Precision	IoU	F1-score
U-net	0.7765	0.7946	0.6467	0.7854
DeeplabV3+	0.8107	0.8394	0.7019	0.8248
LinkNet	0.8042	0.8405	0.6978	0.8220
D-LinkNet	0.8128	**0.8443**	0.7068	0.8282
NL-LinkNet	0.8038	0.8360	0.6943	0.8196
MACU-Net	0.7885	0.8172	0.6703	0.8026
RCFSNet	0.8120	0.8299	0.6961	0.8208
DRCNet	**0.8329**	0.8350	**0.7153**	**0.8340**

TABLE 3. Comparative results on the Massachusetts Road dataset.

Method	Recall	Precision	IoU	F1-score
U-net	0.7642	0.7877	0.6337	0.7758
DeeplabV3+	0.7760	0.7937	0.6458	0.7847
LinkNet	0.7699	0.8002	0.6458	0.7848
D-LinkNet	0.7637	**0.8126**	0.6494	0.7874
NL-LinkNet	0.7588	0.8105	0.6445	0.7838
MACU-Net	0.7811	0.7937	0.6493	0.7873
RCFSNet	0.7762	0.7926	0.6451	0.7843
DRCNet	**0.7983**	0.7999	**0.6655**	**0.7991**

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. COMPARATIVE RESULTS ON THE DEEPGLOBE ROAD DATASET

To comprehensively evaluate the effectiveness of the proposed DRCNet model, we conducted extensive comparative experiments on the DeepGlobe Road dataset. We implemented several mainstream remote sensing image segmentation methods based on encoder-decoder architectures, including U-Net [17], LinkNet [23], DeepLabv3+ [46], D-LinkNet [22], NL-LinkNet [36], MAC-UNet [47], and RCF-SNet [48]. Table 2 presents the evaluation metrics

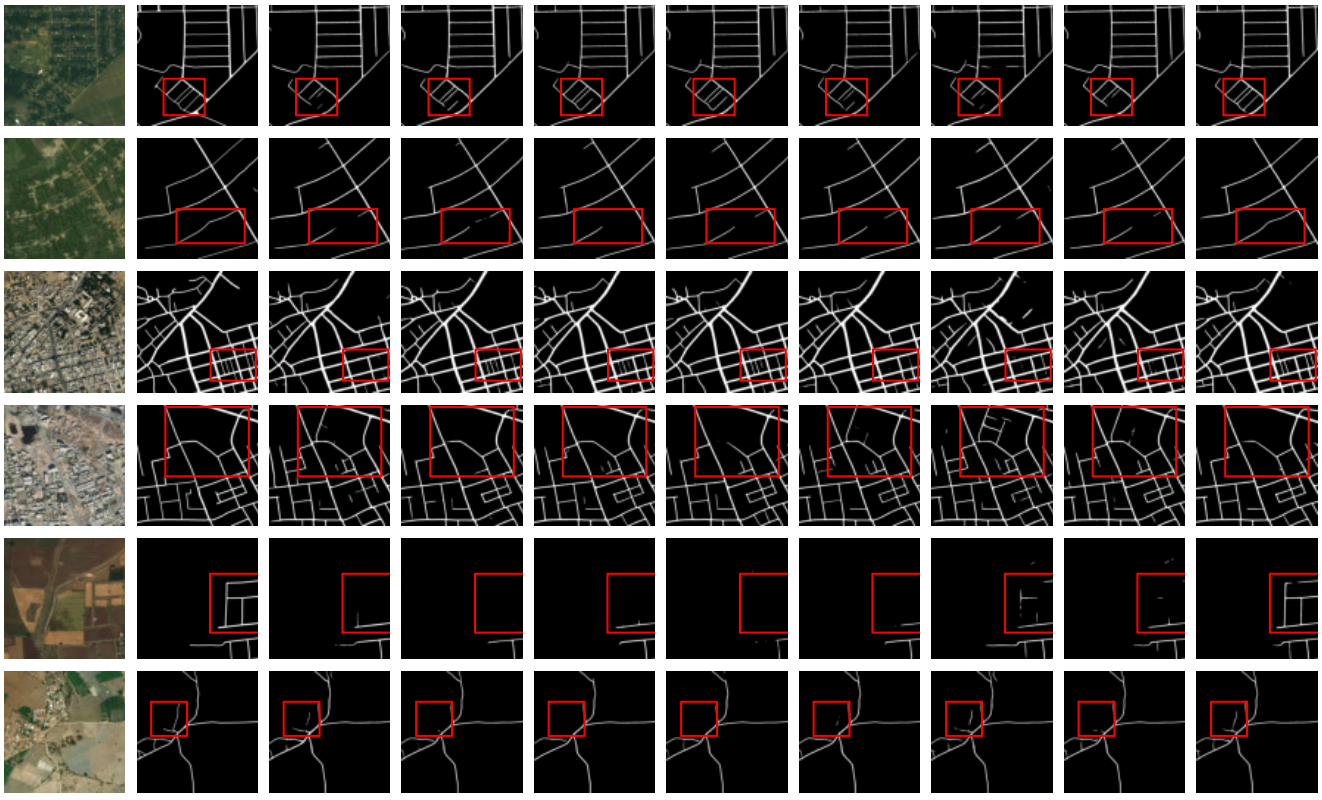


FIGURE 8. Visual experimental results on the DeepGlobe Road dataset. (a) input image, (b) ground truth, (c) U-net, (d) DeepLabv3+, (e) LinkNet, (f) D-LinkNet, (g) NL-LinkNet, (h) MACU-Net, (i) RCFSNet, and (j) DRCNet. The red boxes highlight roads that are more difficult to identify, indicating a significant improvement of our model over the reference ones.

TABLE 4. Comparison of parameters and computational complexity.

Method	Params(M)	FLOPs(G)
U-net	39.50	**34.96**
DeeplabV3+	26.71	165.74
LinkNet	21.64	109.50
D-LinkNet	31.10	134.35
NL-LinkNet	21.82	125.98
MACU-Net	**5.15**	134.34
RCFSNet	58.23	729.18
DRCNet	10.61	213.61

for the performance of different segmentation models on the DeepGlobe Road dataset. Among these models, U-Net stood out by introducing skip connections to enhance segmentation through the incorporation of finer details during the upsampling process, achieving IoU and F1-score values of 64.67% and 78.54%, respectively. Additionally, models that incorporated the Dilated Spatial Pyramid Pooling (ASPP) module, such as DeepLabV3+ and D-LinkNet, also demonstrated strong performance with IoU values of 70.19% and 70.68%, respectively. We attribute this performance to the ASPP module's ability to extract features across multiple scales effectively. NL-LinkNet, which introduced Non-Local operations to capture long-range dependencies and model contextual information, did not yield particularly ideal results. We speculate that the inclusion of Non-Local operations might introduce excessive background information, potentially leading to unnecessary interference and performance degradation. MACU-Unet and RCFSNet, as recently proposed road extraction models, sought performance improvement by introducing attention mechanisms but still fell short compared to some mainstream models. DRCNet outperformed all other models, achieving the highest values for IoU and F1-score. Compared to U-Net, it exhibited significant improvements of 6.86% and 4.86%, respectively, highlighting its capacity for accurate pixel-level classification and localization. We attribute DRCNet's remarkable performance to several key features: firstly, the use of DenseNet-121 as the backbone network for extracting richer features; secondly, the introduction of the RCCA module, which aids in capturing dense global context information; and lastly, the incorporation of the CBAM module within skip connections, effectively suppressing background information and emphasizing road structural details. DRCNet takes into consideration the strengths and weaknesses of other road segmentation models, leading to significant improvements in road segmentation results.

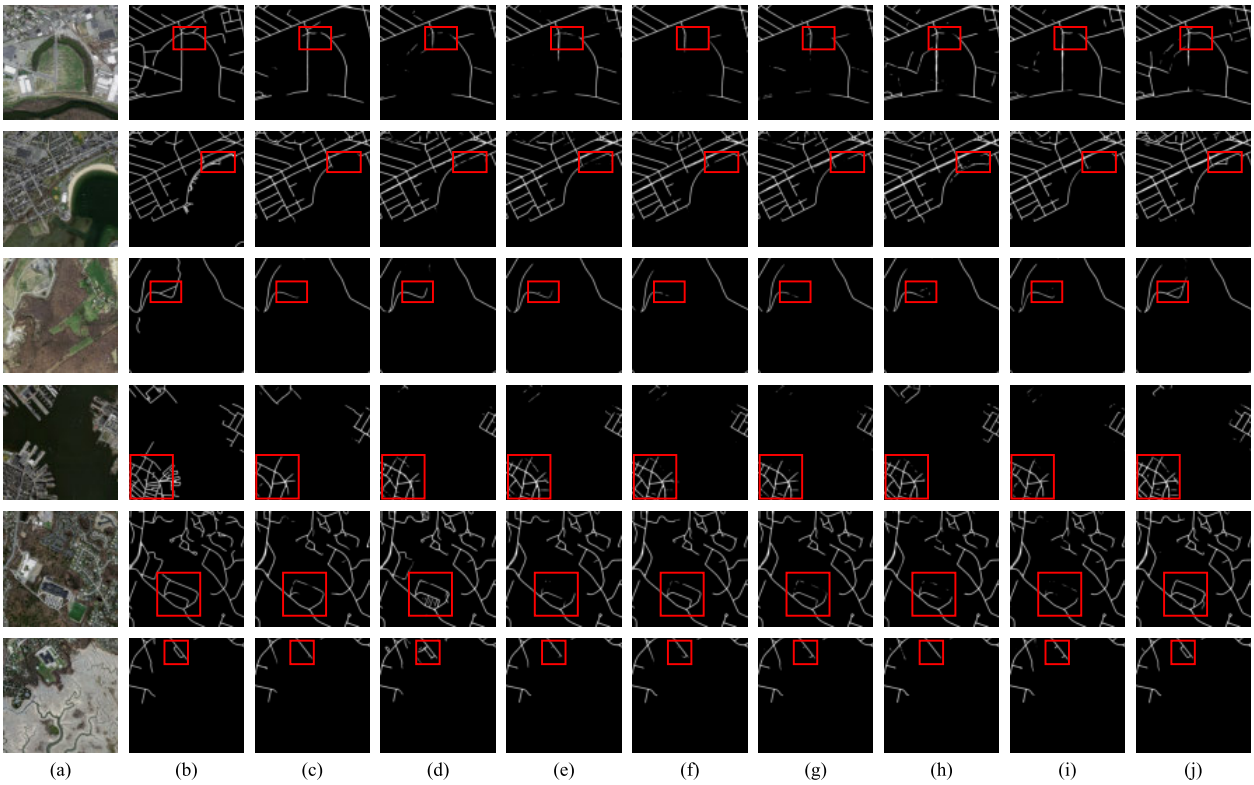


FIGURE 9. Visual experimental results on the Massachusetts Road dataset. (a) input image, (b) ground truth, (c) U-net, (d) DeepLabv3+, (e) LinkNet, (f) D-LinkNet, (g) NL-LinkNet, (h) MACU-Net, (i) RCFSNet, and (j) DRCNet. The red boxes highlight roads that are more difficult to identify, indicating a significant improvement of our model over the reference ones.

For the purpose of a more direct comparison of road extraction performance between different models, Figure 8 presents some visual experiment results, with red boxes highlighting areas prone to omission due to vegetation cover and small size. From these results, it can be observed that DRCNet outperforms the reference models in detecting small-sized roads and roads with complex background information. For instance, some intermittently covered small roads by vegetation are challenging (Rows 1 and 2). In such cases, other reference models fail to effectively detect these obscured roads, while DRCNet provides near-perfect detection results. Similarly, compared to the reference models, DRCNet demonstrates higher detection rates and lower false positive rates for small-sized roads in complex urban scenes (Rows 3 and 4). In rural scenes, some classical road extraction models struggle to distinguish between dirt roads and main roads among fields, while DRCNet exhibits better interference resistance, particularly for dirt roads (Row 5). Furthermore, for roads with complex shapes and irregular outlines (Row 6), DRCNet also delivers more comprehensive road extraction results compared to the reference models.

B. COMPARATIVE RESULTS ON THE MASSACHUSETTS ROAD DATASET

The Massachusetts Road dataset lacks pixel-level annotations, with only centerline information and road width included. Consequently, annotation errors may introduce instability in the prediction results. As presented in Table 3, DRCNet notably outperforms other road segmentation models in terms of F1-score and IoU, achieving scores of 79.91% and 66.55%, respectively. While D-LinkNet attains an IoU of 64.94%, DRCNet exhibits a noteworthy improvement of 1.61%. We also observed significant performance disparities among models such as DeepLabV3+ and RCFSNet on both the Massachusetts Road dataset and the DeepGlobe Road dataset. Notably, DRCNet consistently demonstrates superior overall performance on both datasets, underscoring its robust generalization capabilities and suitability for remote sensing road extraction tasks. These comprehensive results affirm that our proposed DRCNet model excels in accurately delineating road regions and exhibits remarkable resilience to interference.

Figure 9 shows DRCNet's strong performance on test samples. It excels at detecting roads hidden by trees, improving road extraction continuity (Rows 1 and 2). Moreover, DRCNet exhibits greater accuracy in extracting small-sized roads, ensuring the completeness of the extraction process (Rows 3 and 4). Additionally, DRCNet effectively harnesses contextual information, allowing for a more precise differentiation between road and non-road areas (Rows 5 and 6).

C. COMPLEXITY ANALYSIS

In order to evaluate the performance of road segmentation models more comprehensively, Params and FLOPs are also

TABLE 5. Ablation experiment on the DeepGlobe Road dataset with LinkNet as baseline model.

Method	DenseNet121	CBAM	RCCA	Recall	Precision	IoU	F1-score
LinkNet				0.8042	0.8405	0.6978	0.8220
DRCNet1	✓			0.8124	0.8399	0.7035	0.8259
DRCNet2	✓	✓		0.8261	0.8308	0.7071	0.8284
DRCNet3	✓		✓	0.8236	**0.8426**	0.7138	0.8330
DRCNet	✓	✓	✓	**0.8329**	0.8350	**0.7153**	**0.8340**

TABLE 6. Ablation experiment on the Massachusetts Road dataset with LinkNet as baseline model.

Method	DenseNet121	CBAM	RCCA	Recall	Precision	IoU	F1-score
LinkNet				0.7762	0.7926	0.6451	0.7843
DRCNet1	✓			0.7906	0.7900	0.6533	0.7903
DRCNet2	✓	✓		0.7871	0.7938	0.6535	0.7904
DRCNet3	✓		✓	0.7802	**0.8036**	0.6552	0.7917
DRCNet	✓	✓	✓	**0.7983**	0.7999	**0.6655**	**0.799**1

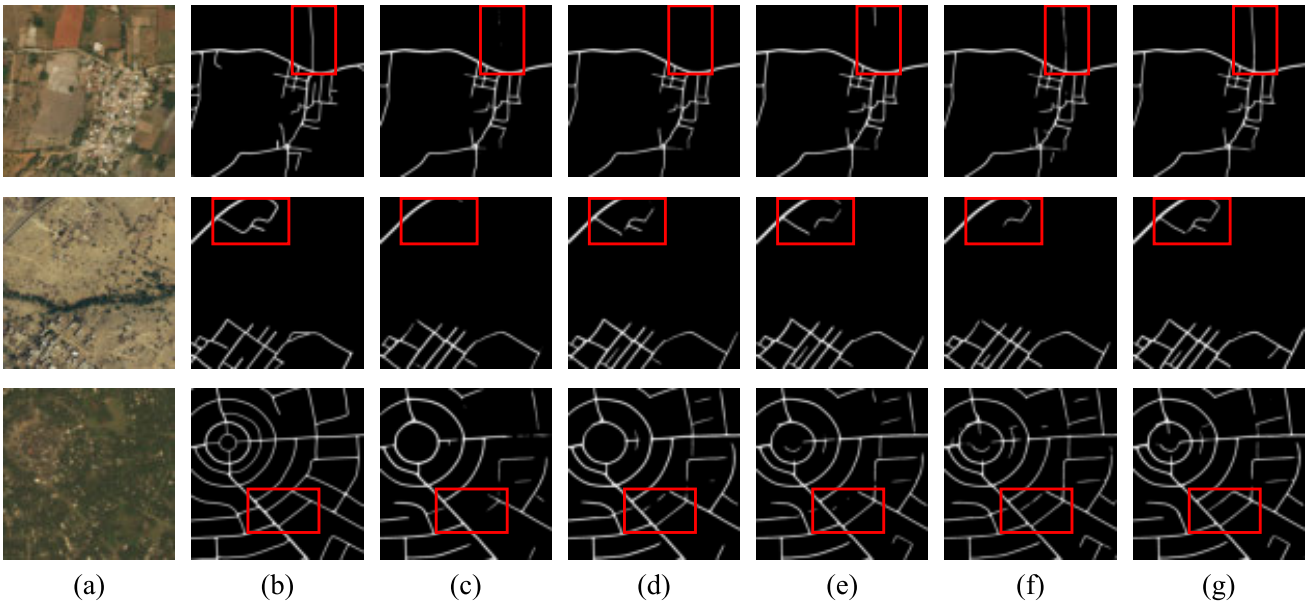


FIGURE 10. Visual comparison of ablation experiments on the DeepGlobe Road dataset.(a) input image, (b) ground truth, (c) baseline(LinkNet), (d) DenseNet-121, (e) DenseNet-121 with CBAM, (f) DenseNet-121 with RCCA, (g) DRCNet. Red boxes highlight the hard regions to extract.

considered as important indicators, and these two metrics are used to measure the number of parameters and computational complexity of the model, respectively. As shown in Table 4, the Params and FLOPs of DRCNet are 10.61(M) and 213.61(G), respectively. DRCNet has fewer parameters, only 5.49(M) more than MACU-Net, but significantly lower than the other six models. While our model's computational complexity is relatively high, only smaller than RCFSNet, considering the excellent performance of DRCNet in terms of IoU and F1-score, the slightly higher FLOPs are not deemed unacceptable.

D. ABLATION STUDY

To verify the effectiveness of Densenet-121, CBAM, and RCCA, we analyze the performance of Densenet-121, Densenet-121 with RCCA, DenseNet with CBAM, and the proposed DRCNet. The results shown in Tables 5 and 6 report quantitative results on the DeepGlobe Road dataset and Massachusetts Road dataset, respectively. Visual comparisons of ablation experiments performed on DeepGlobe Road dataset and Massachusetts Road dataset are shown in Figures 10 and 11. The effectiveness of the different components in DRCNet analysis is given below.

1) EFFECTIVENESS OF DENSENET-121

We use DenseNet - 121 as the backbone to replace ResNet34 in LinkNet. The experimental results show that compared with the baseline network LinkNet, The IoU of DenseNet-121 on the DeepGlobe Road dataset and Massachusetts Road dataset is increased by 0.57% and 0.72%, respectively, and

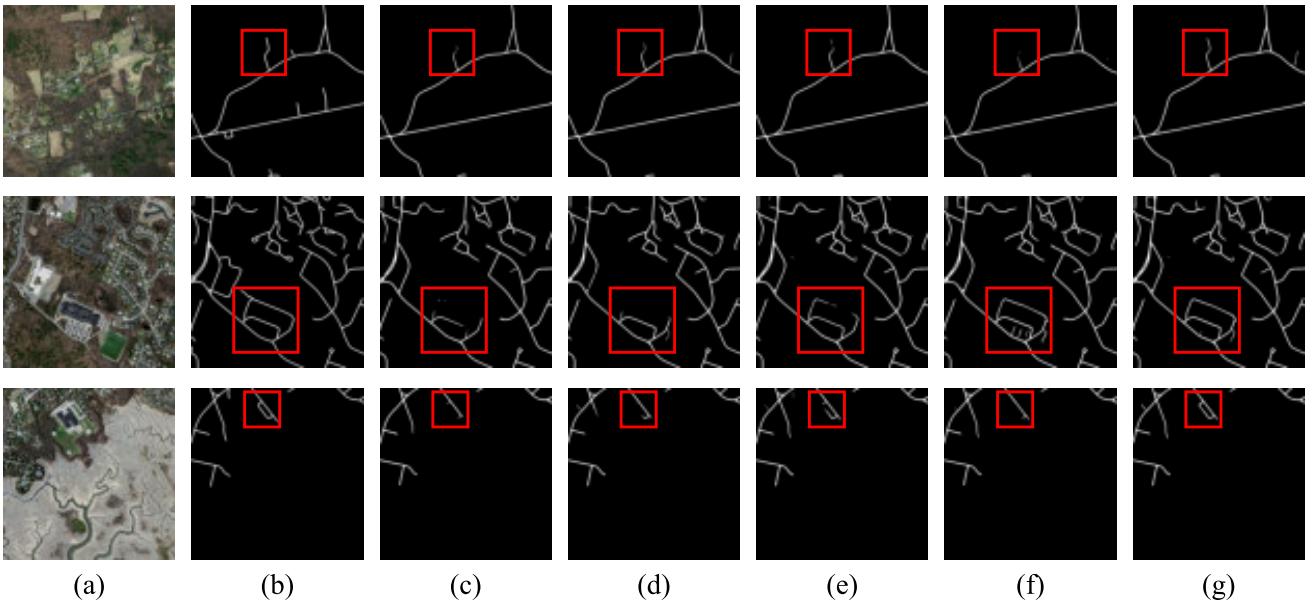


FIGURE 11. Visual comparison of ablation experiments on the Massachusetts Road dataset.(a) input image, (b) ground truth, (c) baseline(LinkNet), (d) DenseNet-121, (e) DenseNet-121 with CBAM, (f) DenseNet-121 with RCCA, (g) DRCNet. Red boxes highlight the hard regions to extract.

the F1-score is increased by 0.39% and 0.60%, respectively. We do this because DenseNet-121 has stronger feature extraction capability, better information transfer mechanism, and a structure more suitable for the road extraction task, which enables it to capture the complex features and shapes of roads more accurately, thus improving the accuracy and performance of road extraction.

2) EFFECTIVENESS OF CBAM

When we add CBAM on the basis of DenseNet-121. The experimental results on two datasets show that IoU and F1-scoore are improved, and the improvement on the DeepGlobe Road dataset is more obvious. Compared with DenseNet-121, DenseNet-121 with CBAM in ious and F1-score increased by 0.36% and 0.25% respectively. This suggests that the incorporation of the CBAM module contributes to the model's improved ability to discover and utilize feature correlations effectively, along with a more targeted emphasis on important channel and spatial information. This further strengthens the network model's performance in road extraction tasks, enabling it to more precisely capture and segment roads, thereby enhancing the quality and effectiveness of road extraction.

3) EFFECTIVENESS OF RCCA

When we join RCCA in DenseNet-121 on the basis of the module, the experimental results show that the DenseNet - 121 with RCCA on two data sets of all the indexes were improved significantly. Among them, the IoU and F1-score on the DeepGlobe Road dataset reach 71.38%and 83.30%respectively, which are 1.03%and 0.71%higher than those of DenseNet-121. The IoU and F1-score on the Massachusetts Road dataset reach 71.38% and 83.30% respectively, which are 0.19% and 0.14% higher than those of DenseNet-121. This shows that RCCA can help the network model predict the road area more accurately; so that the network shows better comprehensive performance.

VI. CONCLUSION

In this paper, we propose a DRCNet model for road extraction from complex high-resolution remote sensing images. Compared with the current popular road extraction models, the proposed model has better performance in road extraction. The network adopts an encoder-decoder structure to learn road features. Among them, the pre-trained DenseNet-121 acts as an encoder to solve the vanishing gradient problem during training. The recurrent cross-attention module is introduced into the encoder-decoder connection part to capture dense global context dependencies. The convolutional attention module is introduced into the skip connection part to highlight the road information while suppressing the background information, which effectively solves the problem of low road recognition accuracy. The experimental results on the DeepGlobe Road dataset and the Massachusetts Road dataset demonstrate that this model outperforms other comparative models in multiple metrics, including IoU and F1-score. This validates that the approach exhibits superior performance and extracts road structures more comprehensively. However, the model in this paper still needs to be further improved to improve its ability to extract roads from more complex background information in different scenarios; and to further optimize the model structure under the premise of ensuring the accuracy. In the next step, more cutting-edge and perfect deep learning methods can be combined to study.

In the future, the primary focus will be on expanding and optimizing the road extraction dataset to provide better support for neural network training and enhanced performance. Moreover, there will be ongoing efforts to refine neural network architectures in order to better align with the demands of road extraction. For instance, the utilization of encoders like ResNeSt [49] or Swin Transformer [50] within the road extraction network is anticipated to contribute to more accurate road information extraction. Furthermore, as technology continues to advance, there is an expectation that the exploration of additional innovative approaches, including the integration of multimodal data, such as the fusion of LiDAR and infrared images, will further enhance the performance of road extraction.

REFERENCES

[1] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

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

[3] D. Patil and S. Jadhav, "Road extraction techniques from remote sensing images: A review," in *Proc. Innov. Data Commun. Technol. Appl. (ICIDCA)*, 2021, pp. 663–677.

[4] P. P. Singh and R. D. Garg, "Automatic road extraction from high resolution satellite image using adaptive global thresholding and morphological operations," *J. Indian Soc. Remote Sens.*, vol. 41, no. 3, pp. 631–640, Sep. 2013.

[5] J. Xie, J. Cui, and F. Yu, "Semiautomatic extraction of belt-like roads from high spatial resolution remotely sensed imagery based on self-adaptively variable width of template matching metho," in *Proc. 3rd Int. Conf. Robot. Autom. Sci. (ICRAS)*, Jun. 2019, pp. 227–232.

[6] P. Lu, K. Du, W. Yu, R. Wang, Y. Deng, and T. Balz, "A new region growing-based method for road network extraction and its application on different resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 12, pp. 4772–4783, Dec. 2014.

[7] M. Darweesh, S. A. Mansoori, and H. AlAhmad, "Simple roads extraction algorithm based on edge detection using satellite images," in *Proc. IEEE 4th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2019, pp. 578–582.

[8] H. Zhou, X. Song, and G. Liu, "Automatic road extraction from high resolution remote sensing image by means of topological derivative and mathematical morphology," in *Proc. MIPPR Remote Sens. Image Process., Geographic Inf. Syst., Appl.*, vol. 10611, 2018, pp. 18–24.

[9] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517010.

[10] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, Jul. 2019.

[11] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513412.

[12] X. Wu, D. Hong, and J. Chanussot, "UIU-net: U-net in U-net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2023.

[13] X. Li, F. Xu, X. Lyu, H. Gao, Y. Tong, S. Cai, S. Li, and D. Liu, "Dual attention deep fusion semantic segmentation networks of large-scale satellite remote-sensing images," *Int. J. Remote Sens.*, vol. 42, no. 9, pp. 3583–3610, May 2021.

[14] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.

[15] X. Li, F. Xu, F. Liu, R. Xia, Y. Tong, L. Li, Z. Xu, and X. Lyu, "Hybridizing Euclidean and hyperbolic similarities for attentively refining representations in semantic segmentation of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[16] X. Li, F. Xu, R. Xia, T. Li, Z. Chen, X. Wang, Z. Xu, and X. Lyu, "Encoding contextual information by interlacing transformer and convolution for remote sensing imagery semantic segmentation," *Remote Sens.*, vol. 14, no. 16, p. 4065, Aug. 2022.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Munich, Germany: Springer, 2015, pp. 234–241.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

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

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

[21] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[22] 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 (CVPRW)*, Jun. 2018, pp. 192–1924.

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

[24] Y. Wang, Y. Peng, W. Li, G. C. Alexandropoulos, J. Yu, D. Ge, and W. Xiang, "DDU-net: Dual-decoder-U-net for road extraction using high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412612.

[25] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5792–5801.

[26] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jul. 2023, pp. 8115–8124.

[27] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 105–119, Jan. 2022.

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[29] J. Liu, J. Shang, R. Liu, and X. Fan, "Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5026–5040, Aug. 2022.

[30] J. Liu, G. Wu, J. Luan, Z. Jiang, R. Liu, and X. Fan, "HoLoCo: Holistic and local contrastive learning network for multi-exposure image fusion," *Inf. Fusion*, vol. 95, pp. 237–249, Jul. 2023.

[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

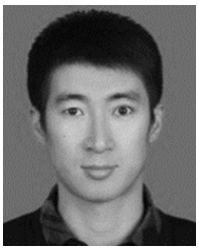
[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[33] L. Ding and L. Bruzzone, "DiResNet: Direction-aware residual network for road extraction in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10243–10254, Dec. 2021.

[34] Z. Ge, Y. Zhao, J. Wang, D. Wang, and Q. Si, "Deep feature-review transmit network of contour-enhanced road extraction from remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 3001805.

[35] Y. Jiang, C. Zhong, and B. Zhang, "AGD-linknet: A road semantic segmentation model for high resolution remote sensing images integrating attention mechanism, gated decoding block and dilated convolution," *IEEE Access*, vol. 11, pp. 22585–22595, 2023.

[36] 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, 2022.
[37] 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 Sens.*, vol. 61, 2023, Art. no. 5602213.
[38] V. Mnih, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[40] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.
[41] 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. (CVPR)*, Jun. 2019, pp. 3141–3149.
[42] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[43] X. Li, F. Xu, F. Liu, X. Lyu, Y. Tong, Z. Xu, and J. Zhou, "A synergistical attention model for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5400916.
[44] J. Wan, Z. Xie, Y. Xu, S. Chen, and Q. Qiu, "DA-RoadNet: A dual-attention network for road extraction from high resolution satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6302–6315, 2021.
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[46] 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. (ECCV)*, 2018, pp. 801–818.
[47] R. Li, C. Duan, S. Zheng, C. Zhang, and P. M. Atkinson, "MACU-net for semantic segmentation of fine-resolution remotely sensed images," 2020, *arXiv:2007.13083*.
[48] 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, 2022, Art. no. 8000405.
[49] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, and R. Manmatha, "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2022, pp. 2736–2746.
[50] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.



DEBIN WEI was born in 1978. He received the M.E. degree from Henan Normal University, Xinxiang, Henan, China, in 2004. He is currently pursuing the Ph.D. degree with the Nanjing University of Science and Technology, Nanjing, Jiangsu, China. He is currently an Associate Professor with Dalian University. His main research interests include network optimization, wireless communication, and image processing.

PINRU LI received the B.S. degree in digital media technology from Zhengzhou Normal University, Zhengzhou, China, in 2021. He is currently pursuing the master's degree with the Communication and Network Laboratory, Dalian University. His current research interests include computer vision and deep learning.

HONGJI XIE received the B.S. degree in food science and engineering from Shandong Agricultural University, Tai'an, China, in 2016. He is currently pursuing the master's degree with the Communication and Network Laboratory, Dalian University. His current research interests include computer vision and deep learning.

YONGQIANG XU received the B.S. degree in data science and big data technology from the Anhui University of Science and Technology, Huainan, China, in 2022. He is currently pursuing the master's degree with the Communication and Network Laboratory, Dalian University. His current research interests include computer vision and deep learning.