Complex & Intelligent Systems
https://doi.org/10.1007/s40747-024-01364-9

ORIGINAL ARTICLE

AGF-Net: adaptive global feature fusion network for road extraction from remote-sensing images

Yajuan Zhang[1] · Lan Zhang[1] · Yunhe Wang[1] · Wenjia Xu[2]

Received: 28 September 2023 / Accepted: 18 January 2024

Abstract
Road extraction from remote-sensing images is of great significance for vehicle navigation and emergency insurance. However, the road information extracted in the remote-sensing image is discontinuous because the road in the image is often obscured by the shadows of trees or buildings. Moreover, due to the scale difference of roads in remote-sensing images, it remains a computational challenge to extract small-size roads from remote-sensing images. To address those problems, we propose a road extraction method based on adaptive global feature fusion (AGF-Net). First, a dilated convolution strip attention (DCSA) module is designed from the encoder–decoder structure. It consists of the dilated convolution and the strip attention module, which adaptively emphasizes relevant features in vertical and horizontal directions. Then, multiple global feature fusion modules (GFFM) in the skip connection are designed to supplement the decoder with road detail features, and we design a multi-scale strip convolution module (MSCM) to implement the GFFM module to obtain multi-scale road information. We compare AGF-Net to state-of-the-art methods and report their performance using standard evaluation metrics, including Intersection over Union (IoU), $F1$-score, precision, and recall. Our proposed AGF-Net achieves higher accuracy compared to other existing methods on the Massachusetts Road Dataset, DeepGlobe Road Dataset, CHN6-CUG Road Dataset, and BJRoad Dataset. The IoU obtained on these datasets are 0.679, 0.673, 0.567, and 0.637, respectively.

Keywords Deep learning · Road extraction · Remote-sensing images · Multi-scale feature · Attention mechanism

Introduction

Road information is an important part of mission systems such as intelligent transportation systems, urban infrastructure, and geographic information [1–5]. Automatic extraction of road information from remote-sensing images has important application value in life and industry. With the development of remote-sensing technology, more road information is clearly represented on high-resolution remote-sensing images, which provides the possibility for accurate extraction of roads. However, due to the complex background of remote-sensing images and the irregular distribution of roads, it brings challenges to road extraction. Therefore, a large number of road extraction methods have been proposed. Traditional methods for road extraction relied on manually designed features, such as texture and shape, which required laborious human intervention, resulting in time-consuming and error-prone processes [6]. However, it has been proven that traditional road extraction methods exhibit poor performance when confronted with issues such as multi-scale roads and complex textures [7, 8]. Consequently, automated road extraction from high-resolution remote-sensing images has garnered extensive attention. Deep learning methods exhibit powerful learning capabilities and excel at extracting deep features from high-resolution remote-sensing images. Within the realm of deep learning, road extraction is formulated as a binary semantic segmentation problem.

✉ Lan Zhang
93635626@qq.com

✉ Yunhe Wang
wangyh082@hebut.edu.cn

Yajuan Zhang
254973695@qq.com

Wenjia Xu
562153204@qq.com

[1] School of Artificial Intelligence, Hebei University of Technology, Xiping Road, Tianjin 300401, Asia, China

[2] Data Center, Hebei Prospecting Institute of Hydrogeology and Engineering Geological (Hebei Remote Sensing Center), Huaizhong Road, Shijiazhuang 050022, Asia, China

Published online: 05 March 2024

In recent years, approaches based on convolutional neural networks (CNNs) [9] have demonstrated efficacy in pixel classification. The classical encoder–decoder architectures have been widely applied to road extraction tasks. The Fully Convolutional Network (FCN) [10] replaces the fully connected layers in CNN with convolutional layers, enabling end-to-end training by upsampling the feature maps to their original resolution. U-Net [11] introduces skip connections to mitigate the loss of fine-grained details caused by downsampling. LinkNet [12] utilizes residual convolutional modules as encoders for semantic segmentation, increasing the network's depth to enhance accuracy. Due to the characteristics of roads in remote-sensing images, classical semantic segmentation models still encounter certain challenges that need to be addressed in the road extraction process.

(1) High-resolution remote-sensing images exhibit complex background information, and the pixels corresponding to road areas occupy a relatively small portion of the overall image. Roads are susceptible to occlusion by the vegetation and buildings on both sides of the road. The local nature of convolutional neural networks limits their ability to capture the broader contextual information, leading to the extraction of roads that may exhibit discontinuities.
(2) High-resolution remote-sensing images contain road information of different scales, and repeated downsampling is required in the process of feature extraction, resulting in the loss of detailed road information and making it difficult to extract small roads.

Establishing long-distance dependencies can effectively guarantee the connectivity of road extraction. By combining multiple expansion convolutions with different expansion rates [13, 14], the receptive field can be effectively expanded and long-distance dependencies can be established. However, irrelevant long-distance context information is learned at large dilation rates, leading to misclassification of pixels. In recent years, the effective incorporation of attention mechanisms has enabled the establishment of long-distance dependencies between elements [15–17]. Non-local [18] calculates the weight of each pixel in the input feature map and other position pixels, and establishes the long-distance dependence between pixel positions. The core part of Transformer [19], self-attention, obtains context information by capturing global information. Although these methods are effective in establishing long-distance dependencies, for road features, the weighted sum calculation of global information not only increases the computational load of the network, but also introduces irrelevant contextual information.

Multi-scale information fusion can alleviate the problem of missed detection of small roads. Multi-scale feature extraction techniques such as Spatial Pyramid Pooling [20], Atrous Spatial Pyramid Pooling [21], and dilated convolutions [22] have been developed. However, these methods extract contextual information from feature maps after multiple downsampling, which still leads to the loss of detailed information and results in a high rate of missing small roads [23, 24]. The DeepLab series [25–27] of methods utilize the ASPP (Atrous Spatial Pyramid Pooling) module for extracting features at multiple scales. However, dilated convolutions exhibit subpar performance in capturing small objects. Given that shallow-level information contains rich details, several methods have introduced improvements at the skip connection junctions [28–30]. These modifications aim to compensate for the loss of detailed information during the downsampling process and minimize interference caused by shallow-level information. The U-Net series [31, 32] incorporates skip connections to aggregate features from various hierarchical levels. However, these methods merely concatenate information from different levels, overlooking the semantic disparities between feature maps of different scales.

Establishing long-range dependencies and achieving multi-scale feature fusion remain challenging in road extraction methods. Motivated by the characteristics of roads, this paper proposes AGF-Net, a road extraction network based on adaptive global feature fusion. AGF-Net adopts a U-shaped encoder–decoder architecture. In the intermediate module, we combine the designed strip attention module with the dilated convolution module to simultaneously expand the receptive field and selectively emphasize relevant features along the road direction. This approach effectively establishes long-range dependencies among road pixels, enhancing the continuity of road extraction. Moreover, the multi-scale strip convolution module is employed to bridge the semantic gap between feature maps at different stages, facilitating the fusion of precise fine-grained details and rich semantic information. This integration helps mitigate the issue of missing small roads. The main contributions of this paper are as follows:

1. A dilated convolution strip attention (DCSA) module is developed in the middle stage to obtain global context information. In the DCSA module, we expand the receptive field using dilated convolutions, and utilize the strip attention module (SAM) to adaptively emphasize relevant features in different directions, suppressing the interference of irrelevant context information.
2. Multiple global feature fusion modules (GFFM) are designed at the skip connection to capture the multi-scale information of the road. In addition, a multi-scale strip convolution module (MSCM) is introduced to bridge the semantic gap between feature maps and effectively integrate information from different levels.
3. A road extraction method based on adaptive global feature fusion (AGF-Net) is proposed to alleviate the problems in

shadow occlusion and road scale. The proposed AGF-Net aims to capture shallow detail information, deep semantic information and global context information in the feature extraction process. The superior performance of AGF-Net can be observed from the extensive experiments on different road datasets. We conclude that AGF-Net is a promising method in road extraction.

Related work

Road segmentation

In the traditional road extraction methods, Gruen et al. [33] proposed a semi-automatic extraction scheme for remote-sensing image roads based on the dynamic programming model. Barzohar et al. [34] established a geometric probability model generated by road images based on the assumptions of width, length, curvature and pixel intensity of roads, and quickly found the main road from aerial images. Anil et al. [35] utilized the Snake model for road extraction. Ma et al. [36] automatically extracted roads using the Canny edge detection algorithm. However, traditional methods are not effective for road extraction in complex scenes.

With the development of deep learning, the accuracy of road extraction has been improved. Mnih and Hinton [37] pioneered the use of Convolutional Neural Networks (CNNs) for road extraction from high-resolution remote-sensing images. Zhang et al. [38] used the residual module [39] as the encoder of U-Net, and used the residual convolution module to extract road information more accurately. Xu et al. [40] utilized the powerful feature extraction capability of DenseNet [41] for road extraction tasks. Mei et al. [42] proposed CoA-Net, which uses a strip convolution module to capture long-range contextual information from different directions and avoid interference from irrelevant regions. DBRANet [43] employs a dual-branch structure in the encoder section to more effectively extract features at different scales and fuse feature maps from different scales. DiResNet [44] integrates context-aware, direction-aware, and structure-aware modules into the U-Net architecture. The direction-aware module aids the model in better learning the topological information of roads. However, these methods fail to establish long-distance dependence and make full use of the extracted multi-scale features, resulting in discontinuities in the road extraction process and the difficulty of extraction of small roads.

Long-range dependencies in networks

To establish long-distance dependence, alleviate the shadow occlusion of roadside trees or buildings. D-Linknet [13] and DGR-Net [14] expand the receptive field and establish long-distance dependencies using dilated convolution without increasing the amount of calculation and reducing the resolution. Attention mechanism has been widely used in road extraction. Wang et al. designed a network called NL-LinkNet [45], which introduces a nonlocal module into the encoder of LinkNet [12] to obtain long-distance dependence. Xie et al. [35] proposed HsgNet, which introduces bilinear pooling in the middle module of LinkNet to obtain the feature distribution of weighted spatial information and adaptive aggregation of long-distance context information. Lu et al. [46] proposed GAMSNet to solve the problem of road fragmentation by capturing relevant information between spaces and passages through global perception operations. Zhu et al. [47] proposed GCB-Net, which added the global context awareness module to the encoder–decoder structure, captured the relationship between any pixels through the global context awareness module, and established an effective global context information to solve the road extraction incomplete questions. Wan et al. [48] proposed DA-RoadNet, which uses the dual attention mechanism module to establish the relationship between features, as well as global dependencies, to obtain the dependencies between pixels and between channels. Zhang et al. [49] designed a U-Net-like network that combines the self-attention mechanism module in Transformer, using a dual-resolution Transformer module and a feature fusion section to obtain global context information. BDTNet [50] utilizes bi-direction transformer to capture contextual road information from different scale feature maps, improving the extraction capability of both global and local information. The attention mechanism helps to improve the performance of road extraction. However, considering the shape of the road, it is necessary to establish long-distance dependencies in different directions and adaptively emphasize the correlation of features in those directions.

Multi-scale feature fusion in networks

Some studies focus on the fusion of multi-scale features to improve the network's ability to extract multi-scale roads. FCN and U-Net use skip connections to fuse high-resolution features extracted from shallow convolutional layers to recover spatial detail information. In order to obtain multi-scale information, D-LinkNet [13] incorporates an expansion model in the middle. The dilated model contains dilated convolutions in cascade mode and parallel mode, each path has a different receptive field. Therefore, the network can capture multi-scale features. DGR-Net [14] utilizes the Atrous Spatial Pyramid Pooling Module to obtain multi-scale features, and expand the receptive field to obtain contextual information without reducing the resolution. GMR-Net [51] uses a gating mechanism to filter fuzzy features and irrelevant information in shallow information. MECA-Net [52] skip connection uses multi-scale encoding to fuse multi-scale road

Table 1 Related work

Study	Method	Datasets	IoU
GAMSNet: Globally aware road detection network with multi-scale residual learning	GAMSNet [46]	Boston image	0.630
		Birmingham image	0.567
		Shanghai image	0.452
BDTNet: Road Extraction by Bi-Direction Transformer From Remote Sensing Images	BDTNet [50]	Massachusetts dataset	0.660
		Google Earth dataset	0.880
MECA-Net: A MultiScale Feature Encoding and Long-Range Context-Aware Network for Road Extraction from Remote Sensing Images	MECANet [52]	DeepGlobe dataset	0.651
		Massachusetts dataset	0.658
DSMSA-Net: Deep Spatial and Multi-scale Attention Network for Road Extraction in High Spatial Resolution Satellite Images	DSMSANet [54]	DeepGlobe dataset	0.662
RADANet: Road Augmented Deformable Attention Network for Road Extraction From Complex High-Resolution Remote-Sensing Images	RADANet [55]	Massachusetts dataset	0.658
MSFANet: Multiscale Fusion Attention Network for Road Segmentation of Multispectral Remote Sensing Data	MSFANet [28]	Chongzhou dataset	0.787
		SpaceNet dataset	0.614

features to alleviate the problem of road scale differences. GA-Net [29] uses the feature fusion module to fully integrate multi-scale information and prevent data redundancy caused by direct fusion. DDU-Net [53] uses LinkNet as a baseline and designs a dual decoder structure to preserve details. MSP-Net [30] introduces strip pooling modules at skip connections to enhance road information. These studies demonstrate that fusing multi-scale features can enhance the performance of road extraction. DSMSANet [54] leverages a variety of convolutional kernels to extract multi-scale information from the feature maps of different residual blocks in the encoder. Moreover, it incorporates the Spatial Attention Unit to extract contextually significant information. RADANet [55] utilizes the Deformable Attention Module to fuse multi-scale information from both shallow and deep layers. MSFANet [28] utilizes HRNet [56] to perform multi-scale feature fusion on the feature maps generated by the encoder. However, these methods are not fully utilized for information extraction. The encoding process at different stages produces multi-scale features, and integrating the features from these different stages can enhance the network's capability to extract multi-scale roads (Table 1).

Proposed method

This section describes the details of our adaptive global feature fusion network (AGF-Net) including overall structure, DCSA module, GFFM and the loss function.

Overview of the network structure

In this paper, our model adopts the encoder–decoder structure as the basic framework, including four parts: encoder, decoder, intermediate module and skip connection, to address the challenges of disconnected road extraction and missed detection of small roads. The proposed method is shown in Fig. 1. Specifically, the model uses a pretrained ResNet-34 on the ImageNet dataset as the encoder. (1) In terms of the problem of extracting road discontinuity, a dilated convolutional strip attention module is developed in the intermediate module. This module can capture long-distance dependencies in vertical and horizontal directions across different receptive fields, thereby improving the connectivity of road extraction. (2) To address the issue of missing small roads, multiple global feature fusion modules have been designed in the skip connection part. These modules fuse features at different levels and utilize a multi-scale strip convolution module to enhance the extraction of multi-scale road features. The features extracted by the shallow network contain detailed information about the roads, but simple connections may introduce noise. Therefore, this module combines accurate spatial detail information and rich semantic information to improve the network's ability to extract roads of different scales.

Dilated convolutional strip attention module

In remote-sensing images, roads show the characteristics of connectivity and narrowness, and road extraction results often suffer from discontinuity, leading to fragmented road segments. To reduce the generation of fragmented roads, it

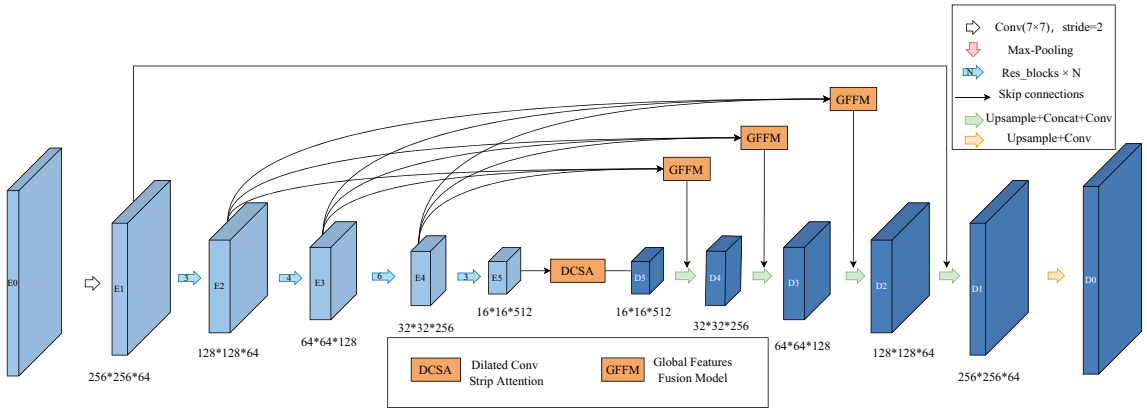

Fig. 1 Structure of the proposed AGF-Net method

is necessary to obtain remote context information during the road extraction process to improve result continuity. Therefore, the dilated convolution attention module is developed in the central module, as shown in Fig. 2. This module not only effectively extracts global road information through dilated convolutions but also utilizes the strip attention module to establish vertical and horizontal contextual relationships. It adaptively emphasizes relevant features in different directions and suppresses the interference of irrelevant contextual information.

The dilated convolution attention module is divided into several steps. First, the dilated convolutions with expansion rates of 1, 2, 4, and 8 are used to form five parallel branches with different receptive fields. The first four branches fuse feature maps of different receptive fields in series, while the last branch obtains global features through global average pooling. Second, a strip attention module is embedded in each parallel branch to capture long-distance dependencies in both vertical and horizontal directions. Lastly, these features are concatenated to obtain multi-scale feature-rich feature maps as output.

Roads in remote-sensing images often have large spans, narrow, and continuity. Ordinary attention mechanisms utilize N×N convolution kernels may introduce irrelevant contextual information. To extract more accurate road features, we should pay attention to the pixel distribution of rows and columns. Therefore, we designed the strip attention module (SAM), as shown in the Fig. 3. The SAM captures long-distance correlations of the road region along rows and columns, reducing the interference from surrounding pixels that are not correlated to the road area.

Given the input F, the strip attention mechanism uses pooling of $(H, 1)$ and $(1, W)$ to encode channels along rows and columns, respectively, and obtains two feature maps Z^h and Z^w:

$$Z^h = \frac{1}{W}\sum_{i=0}^{W} F(h, i) \tag{1}$$

$$Z^w = \frac{1}{H}\sum_{j=0}^{H} F(j, w) \tag{2}$$

We first concatenate them and then compress them using 1×1 convolutions to obtain $f \in \mathbb{R}^{C/r\times(H+W)}$.

$$f = \text{Conv}(1 \times 1)\left(\text{Concat}\left(Z^h, Z^w\right)\right) \tag{3}$$

We then spilt f into two separate tensors $f^h \in \mathbb{R}^{C/r\times H}$ and $f^w \in \mathbb{R}^{C/r\times W}$, and the feature map f^h and f^w are transformed into the same channel number as the input F using 1×1 convolution to obtain the horizontal attention weight g^h and the vertical attention weight g^w. Finally, the output of strip attention module can be expressed as F':

$$g^h = \sigma\left(\text{Conv}(1 \times 1)\left(f^h\right)\right) \tag{4}$$

$$g^w = \sigma\left(\text{Conv}(1 \times 1)\left(f^w\right)\right) \tag{5}$$

$$F' = F \times g^h \times g^w \tag{6}$$

where σ represents the sigmoid function and $\text{Conv}(1 \times 1)$ represents 1×1 convolution.

The feature maps obtained from dilated convolutions with different expansion rates have receptive fields of varying sizes. However, dilated convolutions suffer from grid effects, leading to which can lead to the loss of local information and degrade the detection of small-sized objects. In addition, irrelevant long-distance context information is learned under large dilation rates, resulting in misclassification of pixels. The strip attention module establishes long-distance dependence in both the horizontal and vertical directions, adaptively enhancing the importance of related features,

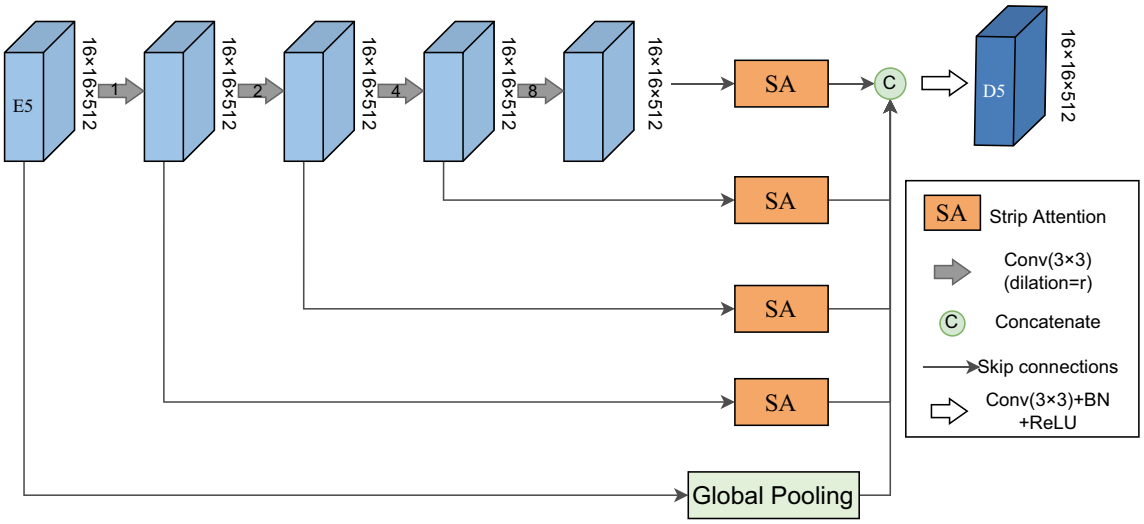

Fig. 2 Schematics of the DCSA module

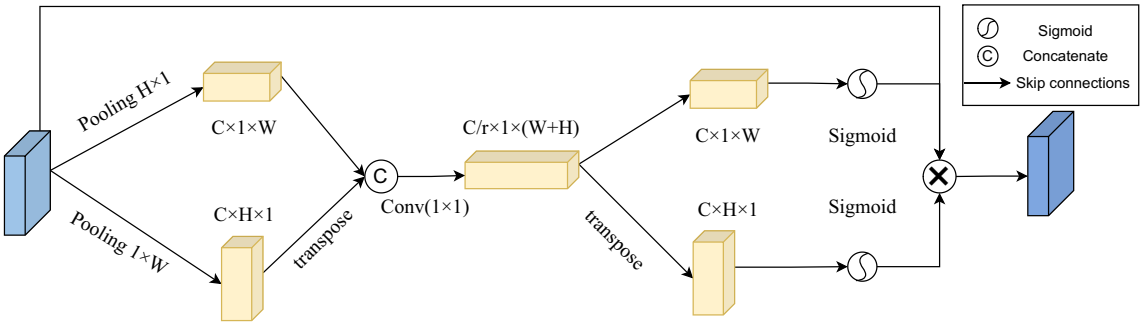

Fig. 3 Schematics of the SAM

reducing the influence of unrelated features, and obtaining more accurate road extraction results.

Global feature fusion module

The semantic segmentation model consists of multiple convolutional layers to capture multi-scale features. Deep features have large receptive fields and rich semantic information. In contrast, shallow features have smaller receptive fields, lack semantic information, but contain detailed information. During the road feature extraction process, the road extraction model loses the detailed information of the road due to continuous downsampling operations, and obtains a feature map with detailed semantic information. The fusion of deep and shallow features is essential for accurately extracting road details. The extraction and aggregation of multi-scale features at different stages of the backbone network can effectively provide road feature information of different scales for the decoder, and enhances the ability to extract roads of different scales. Considering the shape of the road, the multi-scale strip convolution module (MSCM) is designed within the global feature fusion module (GFFM) to extract accurate multi-scale road information.

The GFFM is shown in Fig. 4. To fully utilize the features generated at different stages during the encoding process, we adjust the deep feature map (E4) and the shallow feature map (E2) to the same scale as the E3 through convolution. Then, the three feature maps are fused using the concatenation operation. To conform to the distribution characteristics of roads in remote-sensing images, the merged feature map is inputted into the multi-scale strip convolution module (MSCM), which helps bridge the semantic gap between feature maps of different scales.

According to the characteristics of narrow and long roads, we have designed a multi-scale strip convolution module (MSCM), as shown in Fig. 5. This module utilizes three different scales of strip convolution to extract features. The strip convolution, with its long and narrow kernel, aligns with the characteristics of roads. The strip convolutions of different scales have different receptive fields, allowing them to extract road features of different scales and enhance road information. The module consists of three branches with convolutional kernels of sizes 5, 7, and 11, respectively. Each

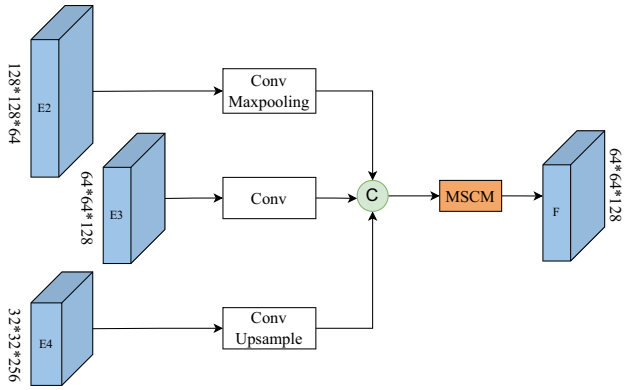

Fig. 4 Schematics of the GFFM

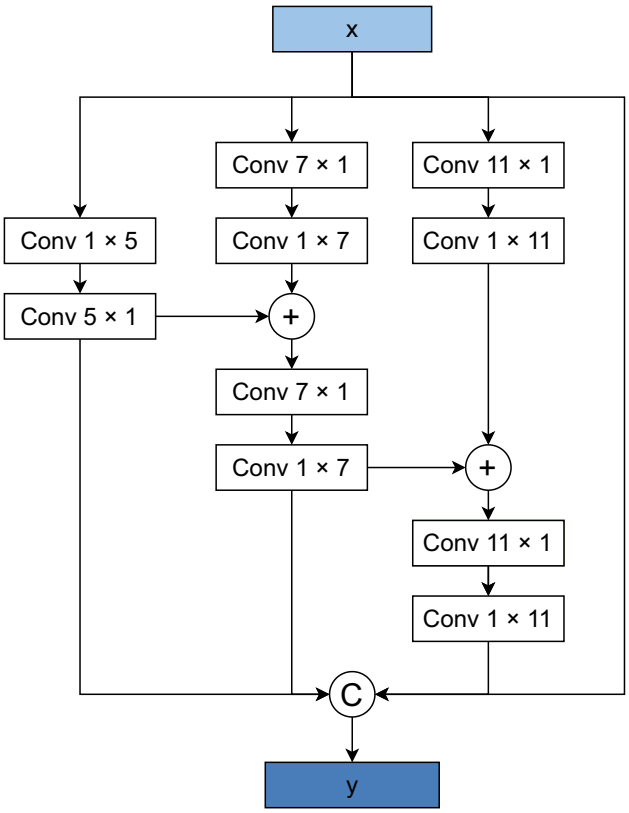

Fig. 5 Schematics of the MSCM

branch performs road feature extraction using a different convolutional kernel. To achieve effective fusion of features at different scales, the output features of the previous branch are fused with the current branch's output features and optimized using strip convolution operations:

$$x_i = x, i = 1, 2, 3 \tag{7}$$

$$y_1 = F_1(x_1) \tag{8}$$

$$y_i = F_i(F_i(x_i) + y_i - 1), i = 2, 3 \tag{9}$$

$$y = \text{Conv}(1 \times 1)\,(\text{Concat}\,(y_1, y_2, y_3)) + x \tag{10}$$

where x_i represents the input of the ith branch, y_i represents the output of the ith branch, F_i represents the operation of the ith branch and Conv(1×1) represents (1×1) convolution.

Loss function

Road extraction in remote-sensing images can be seen as a binary segmentation problem, assigning the background to 0 and the road to 1. Therefore, the binary cross-entropy (BEC) loss can be used as the loss function. However, the ratio of road to background in remote-sensing images is unbalanced, resulting in an imbalance of positive and negative samples. Therefore, we combine the two loss functions to alleviate the imbalance between positive and negative samples and improve the accuracy of road detection. The overall loss function is $L_{BCE+\,\text{Dice}}$:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}(g_i \times \log(p_i) + (1 - g_i) \times \log(1 - p_i)) \tag{11}$$

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N}(g_i \times p_i)}{\sum_{i=1}^{N} g_i^2 + \sum_{i=1}^{N} p_i^2} \tag{12}$$

$$L_{BCE+\,\text{Dice}} = L_{BCE} + L_{\text{Dice}} \tag{13}$$

Experimental result

Dataset

To prove the effectiveness and superiority of the model, we conducted experiments on the DeepGlobe Road dataset, Massachusetts Road dataset, CHN6-CUG dataset, and BJRoad dataset.

1. We performed experiments are performed on an open Massachusetts Road dataset to verify the effectiveness of the proposed network. The Massachusetts Road dataset covers over 2600 square kilometers of urban, suburban, and rural areas, including roads, buildings, vegetation, vehicles, and more. This dataset contains of 1171 images with a road resolution of 1 m/pixel. We cropped the original dataset into 512×512 images and removed any problematic images. Finally, the dataset was divided according to a ratio of 7:2:1. The training set contains 7972 image blocks, the validation set contains 126 image blocks.
2. The images in the DeepGlobe road dataset primarily come from Thailand, Indonesia, and India, covering urban and suburban scenes. Each image has a spatial resolution of 0.5m/pixel and a size of 1024 × 1024. In this study, we conducted experiments using 6226 labeled images from the dataset. To reduce memory consumption, the images

were cropped to 512×512. Finally, the dataset consists of 15,024 training images, 3760 validation images, and 1530 test images.

3. The CHN6-CUG Road dataset contains images of new large-scale representative cities in China. To enrich the dataset, it includes cities with different geomorphological features, city scales, development statuses, structural characteristics, and historical backgrounds. The dataset contains of 4511 images with a resolution of 512×512, divided into 3158 for model training, 902 for model validation and 450 for testing and result evaluation, with a resolution of 0.5 m/pixel.
4. The BJRoad dataset was captured in Beijing, China and consists of 350 high-resolution aerial images covering an area of approximately 100 square kilometers. The aerial images have a spatial resolution of 0.5 ms per pixel and a size of 1024×1024 pixels. Due to the limited number of samples in the dataset, overfitting can be a concern. To address this, we employed data augmentation techniques, specifically rotation, to increase the number of images in the dataset. This resulted in a total of 1750 aerial images. To reduce memory consumption, the images were resized to 512×512 pixels, and subsequently split into training, validation, and testing sets in a ratio of 7:2:1.

Training details

All experiments were conducted on a Windows 64-bit operating system with the following hardware configuration: an AMD Ryzen 7 5800H CPU with Radeon Graphics and an NVIDIA GeForce RTX 3060 Laptop GPU. The proposed model and other network architectures were implemented using the PyTorch framework. We used ResNet-34 as the backbone network and pretrained the model on ImageNet to help improve the convergence of the model. During the training process, we employed the Adam optimizer with an initial learning rate of 2e−3. The training batch size was set to 8. To prevent overfitting and ensure proper learning, we implemented a learning rate decay strategy. If the IoU did not improve for five epochs, the learning rate was multiplied by 0.2 until it reached 0, and the training process was terminated. The Params and FLOPs of AGF-Net are 37.22 (M) and 50.32 (G). The training process of AGF-Net took about 18.5, 33.5, 6.3, and 9.5 h for the Massachusetts Road dataset, the DeepGlobe Road dataset, the CHN6-CUG Road dataset, and the BJRoad dataset, respectively. To ensure that the test data remains completely separate from model development and parameter tuning, we divided the dataset into three distinct subsets: training, validation, and testing. During the model development and parameter tuning stages, we exclusively utilized the training and validation subsets. Subsequently, we evaluated the final performance of the model using a completely independent testing subset that had not been previously accessed, thus mitigating any potential overfitting issues.

Performance metrics

This paper uses four evaluation indicators: IoU [57], Precision, Recall, and $F1$-score. IoU represents the ratio of the intersection and union between the predicted result and the label. mIoU (mean Intersection over Union) represents the average IoU value for each class. In the road extraction task, usually accuracy (Precision) indicates the proportion of the model's prediction results on the image, which is the proportion of the pixel blocks of the road that are correctly predicted. Recall indicates the percentage of the correctly predicted portion of all road pixels. In order to balance the impact between precision and recall, $F1$-score performance indicators are used to comprehensively reflect the average of precision and recall:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{14}$$

$$mIoU = \frac{1}{K+1}\sum_{i=0}^{K} IoU_i \tag{15}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

$$F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

Results

To validate the proposed AGF-Net, we conducted comparative experiments on the Massachusetts Road dataset, DeepGlobe Road dataset, CHN6-CUG Road dataset, and BJRoad dataset. We compared the performance of our method with other classical semantic segmentation methods such as U-Net [11] and Linknet [12], as well as recent models such as DlinkNet [13], DDU-Net [53], and DICN [58]. The objective was to demonstrate the effectiveness of the proposed dilated convolutional attention module and global feature fusion module in road extraction. The performances of the five approaches are listed in Tables 2, 3, 4 and 5, and our method outperforms the other four methods in most of the metrics. Figures 6, 7, 8 and 9 showcase the detection results of our method compared to the other methods. The proposed AGF-Net demonstrates superior performance in extracting fine roads and ensuring road continuity compared to the other four methods.

Accurately extracting road information from remote-sensing images has broad applications in various fields,

Table 2 Experiment on Massachusetts Roads dataset

Models	Precision	Recall	*F*1-score	IoU	mIoU
U-Net	0.703	0.693	0.644	0.608	0.790
LinkNet	0.792	0.775	0.783	0.634	0.796
D-LinkNet	0.822	0.743	0.780	0.641	0.805
DDU-Net	0.815	0.755	0.784	0.644	0.807
Ours	0.837	0.783	0.809	0.679	0.827

Table 3 Experiment on DeepGlobe Roads dataset

Models	Precision	Recall	*F*1-score	IoU	mIoU
U-Net	0.725	0.741	0.733	0.579	0.779
LinkNet	0.775	0.761	0.769	0.625	0.804
D-LinkNet	0.786	0.777	0.781	0.641	0.812
DDU-Net	0.806	0.768	0.796	0.661	0.819
Ours	0.812	0.796	0.809	0.673	0.827

Table 4 Experiment on CHN6-CUG Roads dataset

Models	Precision	Recall	*F*1-score	IoU	mIoU
U-Net	0.643	0.575	0.644	0.536	0.742
LinkNet	0.726	0.682	0.703	0.542	0.742
D-LinkNet	0.722	0.689	0.705	0.545	0.743
DDU-Net	0.746	0.679	0.711	0.551	0.749
Ours	0.732	0.715	0.724	0.567	0.757

including autonomous driving, urban planning, traffic management, and disaster response [59–61]. The performance of road extraction models directly relates to road safety. Accurately extracting road areas can help drivers and traffic management systems better understand the road environment, thereby providing more accurate navigation, traffic control, and driving decisions. For advanced driver assistance systems, road extraction models can serve as the foundation for functions such as lane-keeping assistance and adaptive cruise control. In addition, road extraction models can be used in traffic monitoring and control systems to assist traffic management authorities in real-time traffic monitoring, formulation of traffic strategies, and adjustment of signal timings. In addition, accurately extracting road areas can assist in constructing precise geographic datasets that support urban planning, navigation services, map updates, and other related tasks.

Comparative experiments U-Net [11] is a commonly used classical semantic segmentation model that employs an encoder–decoder architecture and fuses detailed and semantic features through skip connections. LinkNet [12] replaces the encoder of U-Net with residual convolutional modules, increasing the depth of the network. D-Linknet [13], the winner of the 2018 DeepGlobe Road Extraction Challenge, utilizes the LinkNet architecture while incorporating dilated convolutions in the central part to capture contextual information. DDU-Net [53], based on ResNet as a backbone, adopts a dual decoder structure to preserve details and improve the extraction of small roads. LinkNet, D-Linknet, and DDU-Net all utilize a ResNet pretrained on ImageNet as their backbone. DICN [58] utilizes two encoding networks to extract relative information from two input sources separately. It then applies an attention mechanism to fuse them, effectively capturing complementary correlations. All methods employ the same input and undergo identical preprocessing of the input data, ensuring a fair comparison between algorithms.

Table 5 Experiment on BJRoad dataset

Models	Precision	Recall	*F*1-score	IoU	mIoU
U-Net	0.703	0.672	0.708	0.548	0.741
LinkNet	0.707	0.743	0.745	0.594	0.766
D-LinkNet	0.763	0.742	0.762	0.616	0.778
DICN	0.778	0.766	0.772	0.629	0.785
Ours	0.805	0.756	0.780	0.639	0.789

Experiment using the Massachusetts Roads dataset The partial testing results of our model and other models on the Massachusetts Road dataset are presented Fig. 6. The red frames highlight the discontinuous parts caused by obstructing tree shadows, and the yellow frames highlight the easily missed small size roads duo to large-scale differences in roads. In the first and second rows, U-Net, LinkNet, D-LinkNet, and DDU-Net missed the detection of small roads in road extraction due to the significant difference in road size. However, AGF-Net exhibited higher accuracy in extracting small roads. In the third, fourth, and fifth rows, U-Net and D-LinkNet extracted roads are discontinuous due to the obstruction of tree shadows on both sides of the road, while the results of AGF-Net extraction are more continuous. Therefore, the model based on global feature fusion and directional adaptation in this article can obtain richer contextual information and spatial relationships of roads on the input image, making it an effective road extraction model.

Table 2 shows the results of AGF-Net compared to other models. The data in the table demonstrate that our model outperforms other models across various metrics. The proposed AGF-Net performs much better in terms of recall, *F*1-score and IoU. Specifically, compared to Unet, our method improved the IoU and *F*1-score by 7.1% and 16.5%, respectively. The proposed model outperforms the best-performing model by 3.5% and 2.5%. The roads in the Massachusetts dataset are mostly rural roads, and the roads in the image are obscured by the shadows of vegetation. Despite this, our AGF-Net still achieves good results.

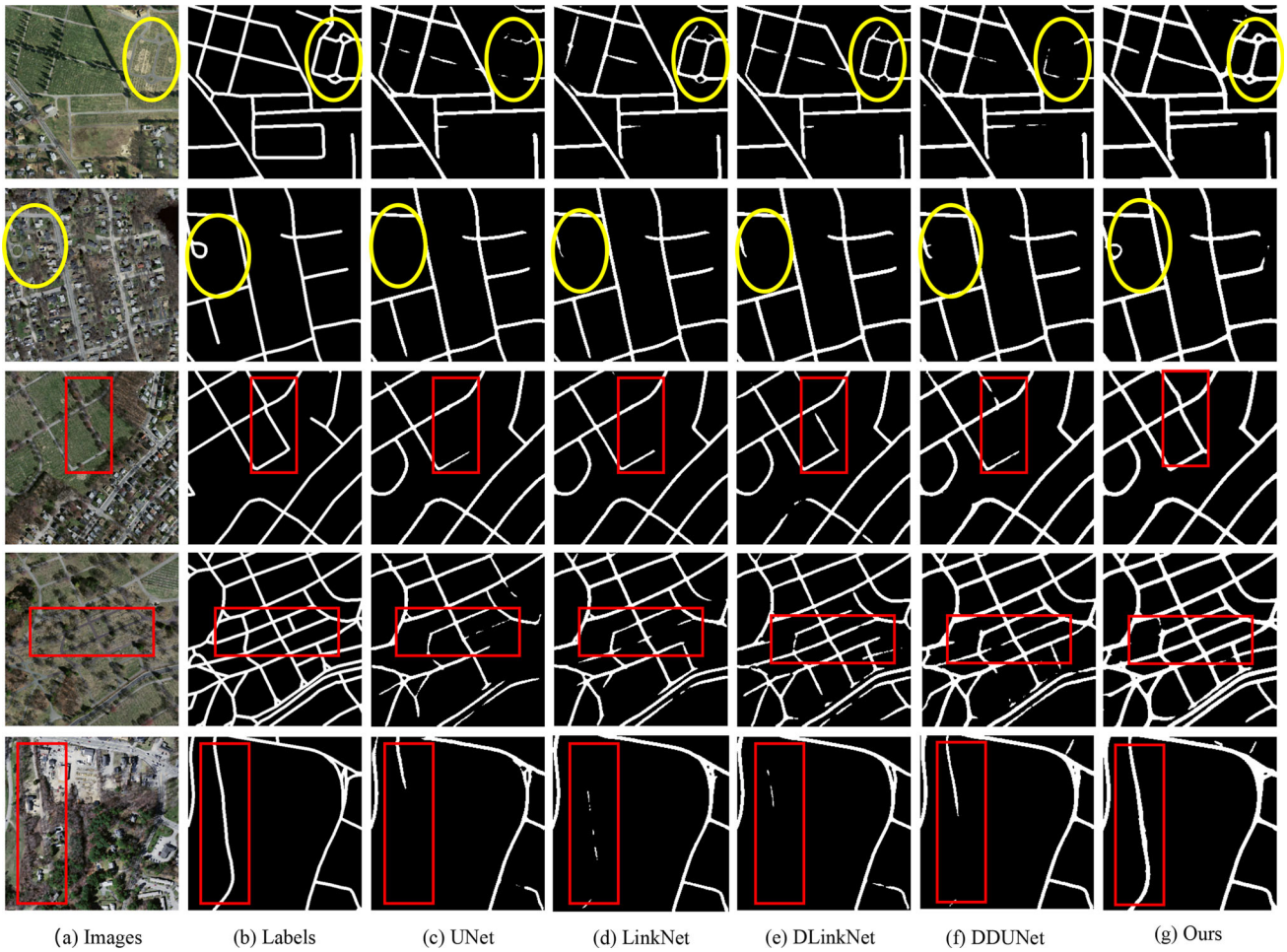

Fig. 6 Visual experimental results on the Massachusetts test set. **a** Images, **b** Labels, **c** U-Net, **d** LinkNet, **e** D-Linknet, **f** DDU-Net and **g** Ours. The yellow highlights the easily missed small-size roads duo to large scale differences in roads. The red highlights the discontinuous parts caused by obstructing tree shadows

Experiment using the DeepGlobe Roads dataset As shown in Fig. 7, some visual experimental results in the first, second and third rows, where yellow frames highlight the easily missed parts due to their small size. The extraction results of other methods exhibit varying degrees of absence, while AGF-Net demonstrates better performance in extracting small roads. In the fourth, fifth, and sixth rows, the roads in the images are obscured by dense trees. Other methods fail to identify roads obscured by tree shadows, resulting in discontinuous road extraction. In contrast, AGF-Net ensures the continuity of extraction roads. In this case, AGF-Net addresses the problem of result discontinuity and lager scale differences in roads, and achieved better results by obtaining sufficient long-range context information and fusing multi-scale information.

The quantitative results on the DeepGlobe dataset are presented in Table 3, and our AGF-Net improves the extraction accuracy when compared to the other methods, with IoU and $F1$ of 67.3% and 80.9%, respectively. Compared to DDU-Net, the IoU and $F1$ of the proposed AGF-Net are improved by 1.2% and 1.3%. Compared to the DeepGlobe dataset, the CHN6-CUG dataset contains more types of roads, including urban and rural roads.

Experiment using the CHN6-CUG Road dataset For comparison with related methods, Fig. 8 shows the test results for some images in the CHN6-CUG Road dataset. The accuracy of AGF-Net in extracting small-sized roads is depicted in the first and second rows. The third, fourth, and fifth rows show the discontinuity of the road extraction results caused by the occlusion of building shadows, while AGF-Net produces more continuous extraction results. The visualization results show the superiority of AGF-Net in the road extraction process, but the boundary of the obtained road is not smooth enough. The quantification results are shown in Table 4. Our method outperforms other methods on the CHN6-CUG Road test dataset, achieving the best results across all metrics.

Experiment using the BJRoad dataset Figure 9 presents the test results on the BJRoad dataset. The roads in the

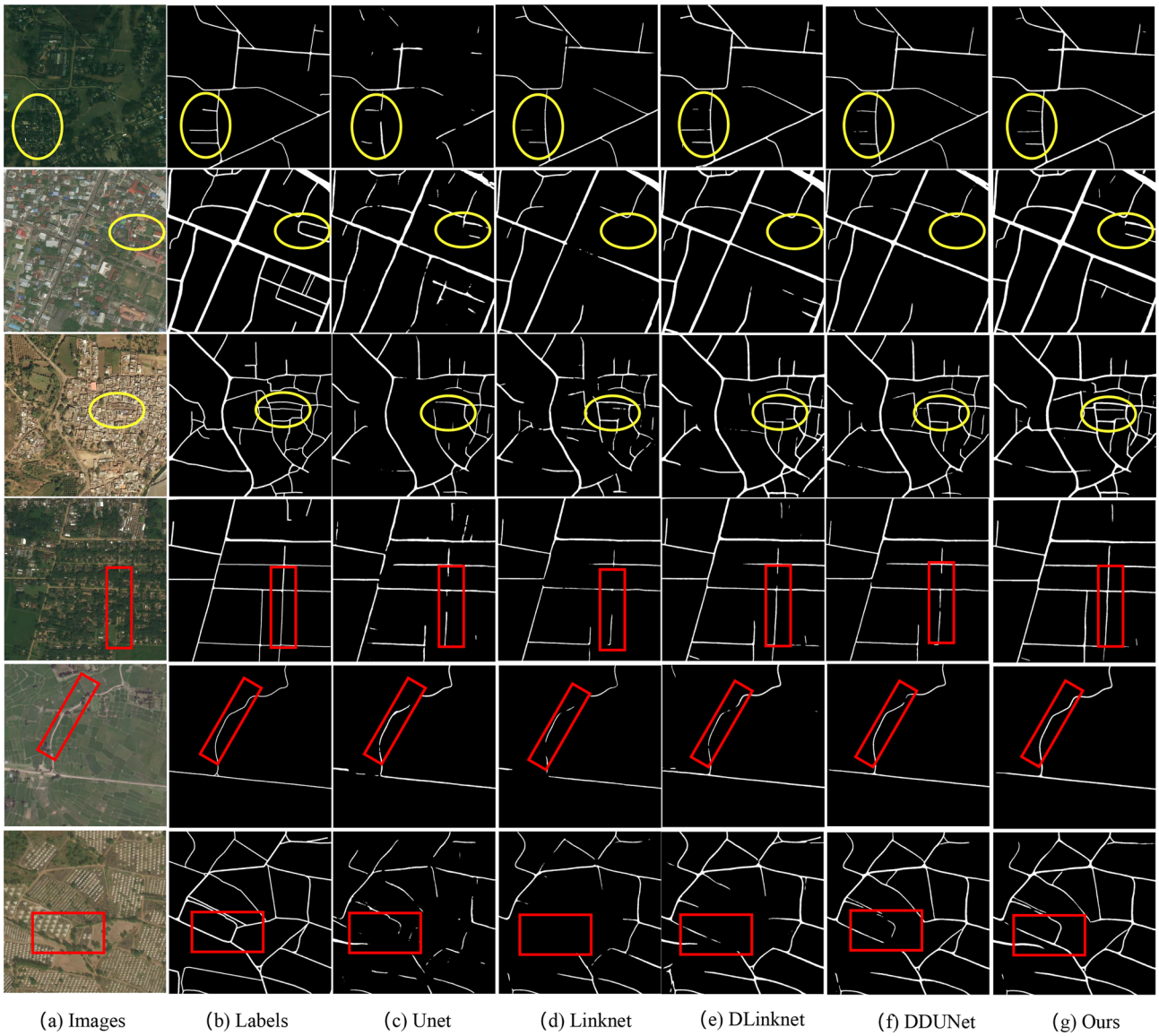

Fig. 7 Visual experimental results on the DeepGlobe test set. **a** Images, **b** Labels, **c** U-Net, **d** LinkNet, **e** D-Linknet, **f** DDU-Net and **g** ours. The yellow highlights the easily missed small-size roads duo to large-scale differences in roads. The red highlights the discontinuous parts caused by obstructing tree shadows

BJRoad dataset mainly consist of urban roads, and there are significant occlusions from trees and buildings along the roads. The visualizations demonstrate that AGF-Net performs better in extracting small roads and ensuring the continuity of road extraction. The quantitative results are shown in Table 5, where AGF-Net achieves the best performance in terms of IoU and $F1$ compared to other methods.

The test metrics and visualization results demonstrate that AGF-Net outperforms other comparison models in road extraction tasks, in complex environments where multi-scale roads coexist and roads are easily missed due to surrounding vegetation or building shadows. Since AGF-Net is capable of establishing long-distance dependencies and preserving more detailed information, it ensures better continuity of roads and more efficient extraction of small-size roads. Compared to mainstream road extraction models, AGF-Net has a lower false positive rate and a higher recall rate.

Analysis of the AGF-Net road extraction results in different environments

To validate the superiority of AGF-Net, we conducted a comparative analysis of road extraction in different environments using the Massachusetts dataset, comparing it with other methods. All methods were trained and tested on the same dataset. As shown in Fig. 10, the roads are located in forest

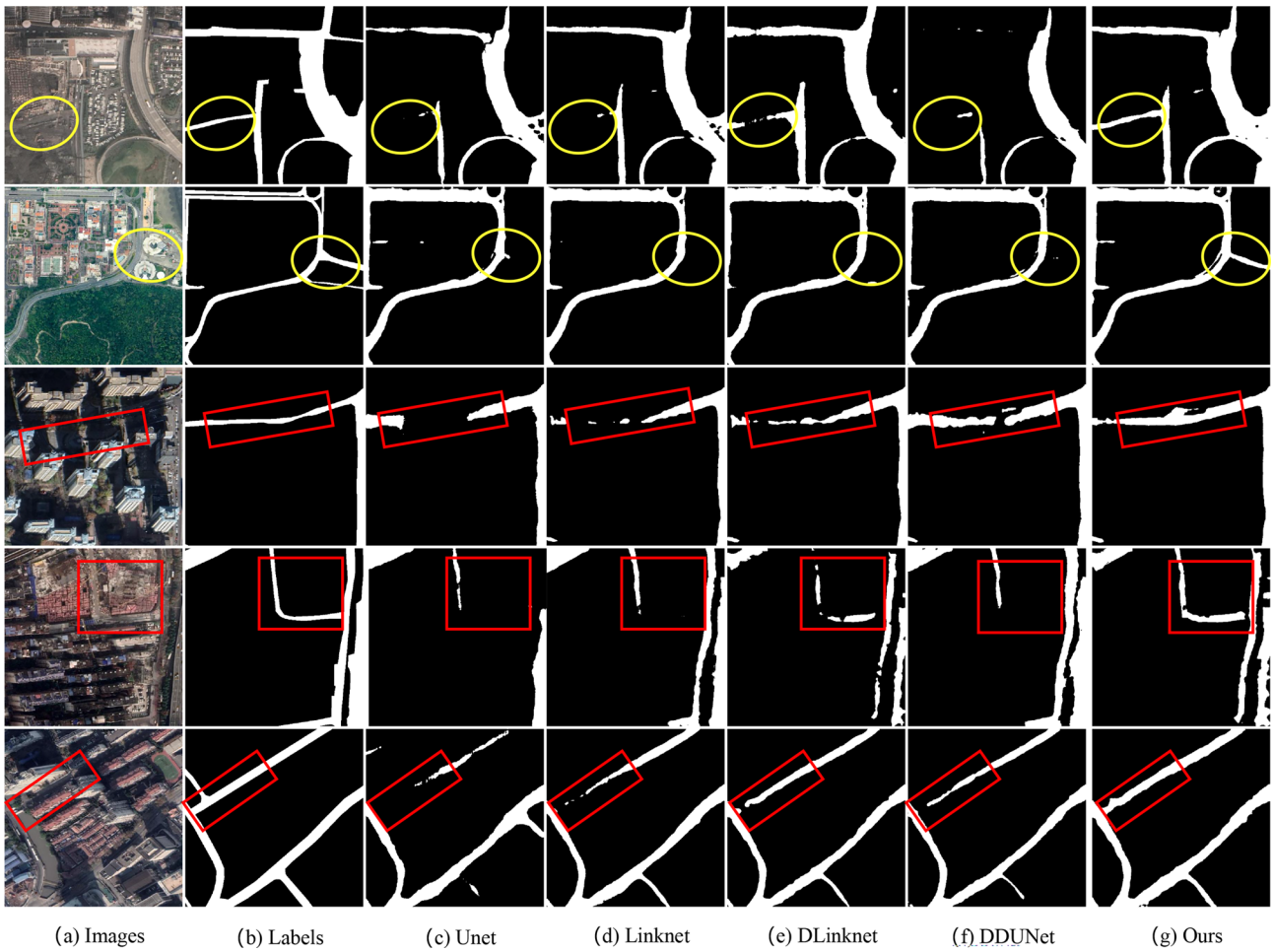

Fig. 8 Visual experimental results on the CHN6-CUG test set. **a** Images, **b** Labels, **c** U-Net, **d** LinkNet, **e** D-Linknet, **f** DDU-Net and **g** ours. The yellow highlights the easily missed small-size roads duo to large-scale differences in roads. The red highlights the discontinuous parts caused by obstructing tree shadows

areas. Road segmentation is primarily influenced by vegetation and tree shadows, which often leads to fragmented road instances. To evaluate the effectiveness of AGF-Net in such environments, we compared it with several commonly used methods. D-Linknet and DDU-Net can successfully extract a significant portion of obstructed roads. However, they struggle to accurately delineate heavily occluded areas. Considering that local information is easily lost when roads are obstructed by vegetation, shadows, and other factors, the proposed AGF-Net addresses this issue by employing dilated convolutions with different receptive fields to establish global contextual dependencies. In addition, the strip attention module (SAM) is utilized to enhance the relationships between road pixels. As a result, AGF-Net excels in extracting occluded roads while striving to maintain road continuity to the greatest extent possible.

As shown in Fig. 11, the roads belong to the suburbs, characterized by significant variations in road sizes within the images. This poses a challenge in accurately detecting small roads, often resulting in missed detections. The highlighted areas in red circles in the figure exhibit substantial differences compared to the main roads. If the details during the road extraction process are not preserved, there is a high risk of overlooking these differences. As depicted by U-Net and LinkNet, these networks demonstrate missed detections for small roads. Although DDU-Net shows some improvement in performance, the delineation results remain incomplete. In contrast, AGF-Net effectively integrates features from different levels while preserving precise details. Consequently, AGF-Net excels in extracting small roads and mitigates the issue of missed detections.

As shown in Fig. 12, the roads belong to urban areas. The presence of vehicles, vegetation, and urban structures on highways can interfere with road extraction. The segmentation maps indicate that U-Net and LinkNet are more susceptible to noise, resulting in poorer performance in urban road segmentation. These methods fail to accurately delineate urban roads under occlusion conditions. In contrast,

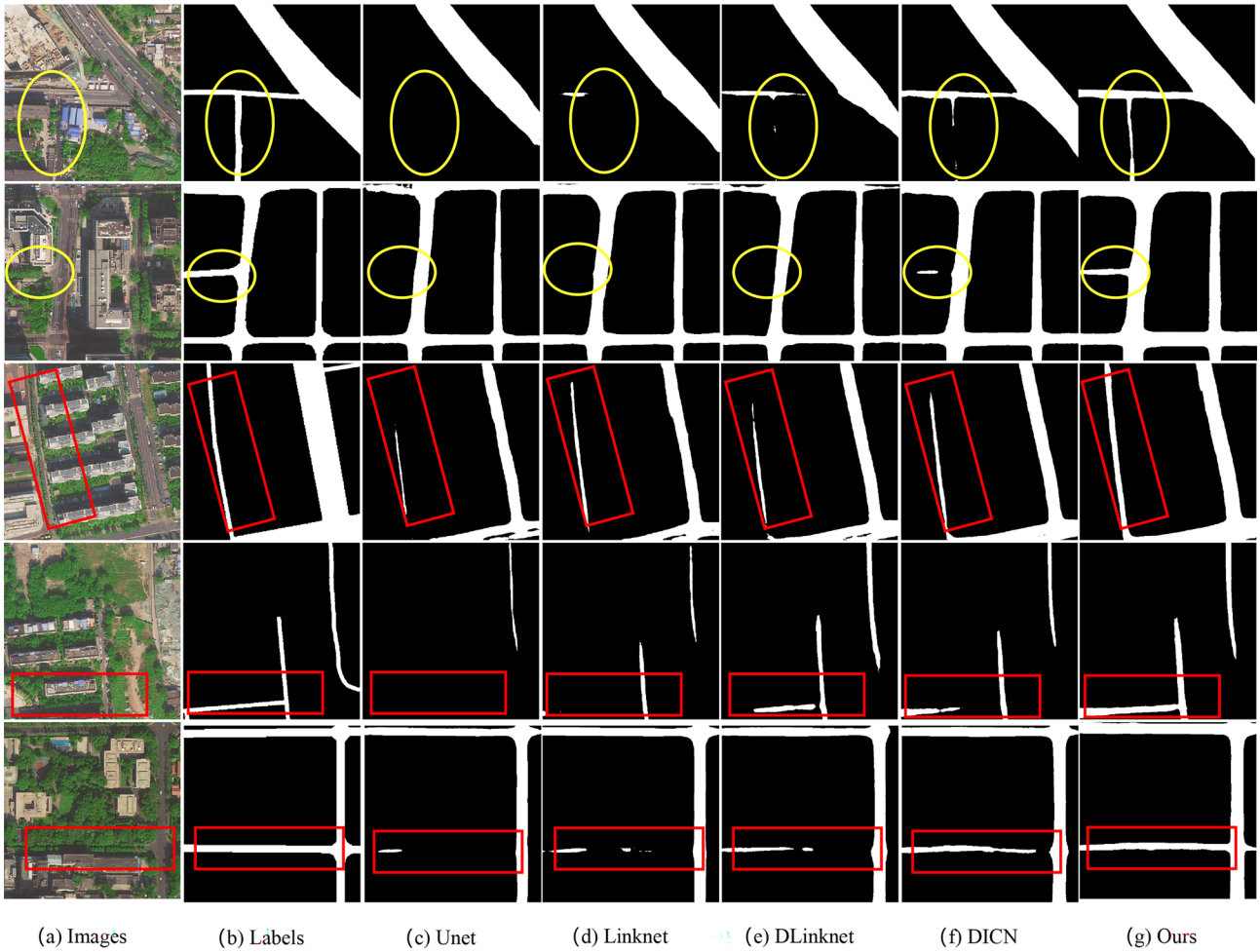

Fig. 9 Visual experimental results on the BJRoad test set. **a** Images, **b** Labels, **c** U-Net, **d** LinkNet, **e** D-Linknet, **f** DICN and **g** ours. The yellow highlights the easily missed small-size roads duo to large-scale differences in roads. The red highlights the discontinuous parts caused by obstructing tree shadows

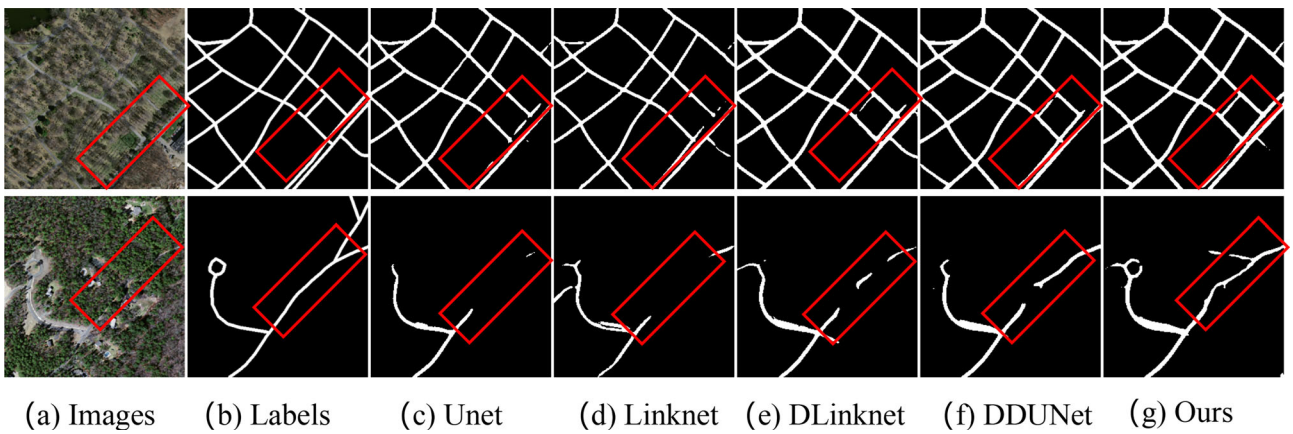

Fig. 10 Comparison of classification results of roads in forest areas. **a** Images, **b** Labels, **c** U-Net, **d** LinkNet, **e** D-Linknet, **f** DDU-Net and **g** ours

AGF-Net addresses this issue by focusing on global information in the images, enabling the extraction of more continuous roads in urban environments.

Ablation results

The proposed method incorporates the dilated convolution strip attention module and global feature fusion module.

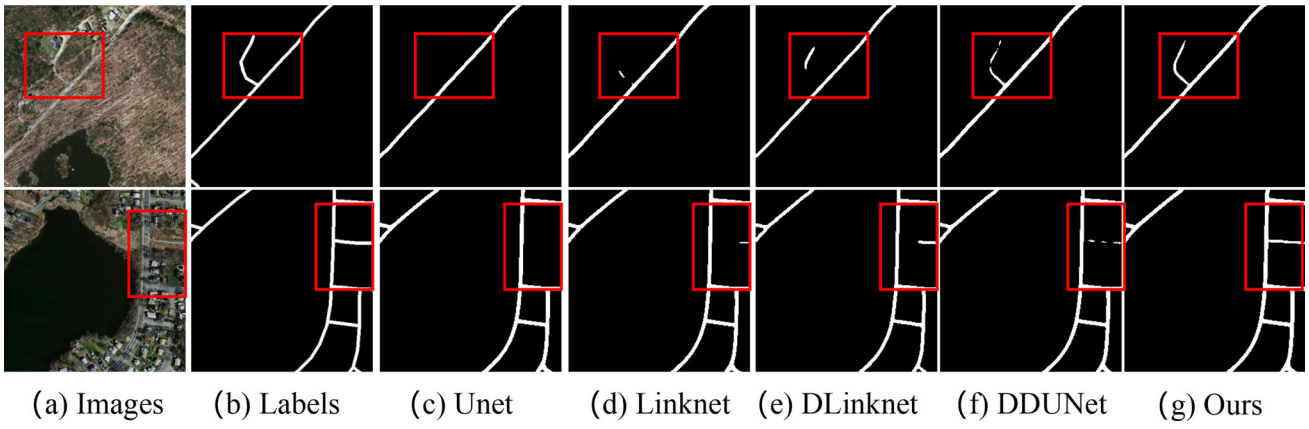

Fig. 11 Comparison of classification results of roads in suburbs. **a** Images, **b** Labels, **c** U-Net, **d** LinkNet, **e** D-Linknet, **f** DDU-Net and **g** ours

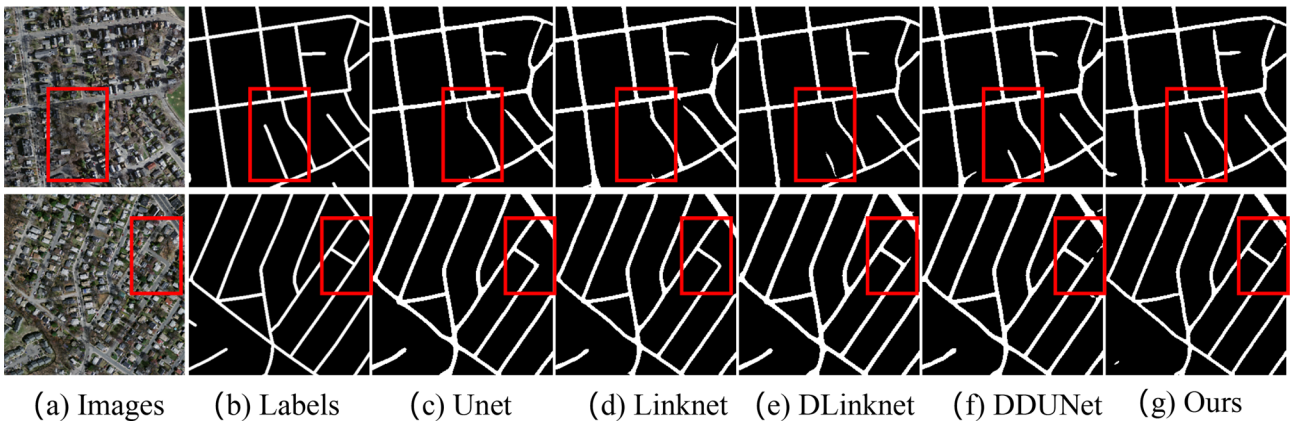

Fig. 12 Comparison of classification results of roads in urban areas. **a** Images, **b** Labels, **c** U-Net, **d** LinkNet, **e** D-Linknet, **f** DDU-Net and **g** ours

To demonstrate the effectiveness of these modules, we conducted ablation experiments on the Massachusetts dataset. In the following experiments, we employed the same configuration for consistency.

To discuss the effect of the number of feature maps generated by the encoder on the model, we conduct experiments on the Massachusetts dataset. E1–E5 represent feature maps of different scales generated by AGF-Net encoders. The fusion of E2–E4 is denoted as GFFM-3, the fusion of E1–E5 is denoted as GFFM-5, the fusion of E1–E4 is denoted as GFFM-4-1, and the fusion of E2–E5 is denoted as GFFM-4-2. GFFM-0 represents the fusion of shallow and deep features using simple skip connections. As shown in Table 6, the results are correlated with the number of feature maps generated by feature extraction. Compared to GFFM-0, GFFM-3 improves the IoU and mIoU by 2.7% and 2.1%. In addition, we verified the influence of MSCM on feature fusion. As shown in Table 7, when we performed experiments based on GFFM-3, the IoU of network without MSCM in the Massachusetts dataset decreases by 1.7%. Therefore, MSCM has an important impact on feature fusion.

Table 6 Ablation study for the proposed GFFM with fuse different numbers of feature maps

GFFM	Precision	Recall	*F*1-score	IoU	mIoU
GFFM-0	0.792	0.775	0.783	0.634	0.796
GFFM-3	0.810	0.757	0.783	0.661	0.817
GFFM-4-1	0.821	0.756	0.788	0.650	0.810
GFFM-4-2	0.841	0.733	0.783	0.645	0.808
GFFM-5	0.821	0.759	0.789	0.652	0.811

Table 7 Ablation study for the proposed GFFM with and without MSCM

	Precision	Recall	*F*1-score	IoU	mIoU
With MSCM	0.810	0.757	0.783	0.661	0.817
Without MSCM	0.829	0.742	0.783	0.644	0.807

To verify the effect of strip attention module on the model, we performed ablation experiments on DCSA with and without SAM, using D-Linknet as the baseline model. As shown in Table 8, the IoU of network without SAM in the Mas-

Table 8 Ablation study for the proposed DCSA with and without SAM

	Precision	Recall	*F*1-score	IoU	mIoU
With SAM	0.824	0.772	0.790	0.651	0.818
Without SAM	0.786	0.777	0.781	0.641	0.812

sachusetts datasets decreases by 1.0%. This indicates that SAM has an important impact on model accuracy.

Our method introduces the dilated convolutional strip attention (DCSA) module to establish long-distance dependence. In addition, the Global Feature Fusion module (GFFM) is utilized for skip connection to fuse features from multiple stages. Through ablation experiments conducted on the Massachusetts dataset, we demonstrate the effectiveness of the proposed DCSA and GFFM. As shown in Table 9, the experimental results indicate that the introduction of DCSA and GFFM significantly improves the accuracy of road extraction. This paper shows that the full utilization of shallow features and the establishment of long-distance dependence improve the *F*1-score and IoU performance of the road extraction model.

Efficiency comparison

We conducted extensive experiments on publicly available datasets in three different environments, comparing AGF-Net with other methods, and achieved significant improvements. As shown in Table 10, we discuss the trainable parameters and Floating-point Operations (FLOPs) of various methods on the Massachusetts Road dataset. These indicators represent the number of parameters and computational requirements of the model, and measure the complexity of the model. The Params and FLOPs of AGF-Net are 37.22 (M) and 50.32 (G), respectively. Compared to several other state-of-the-art methods, AGF-Net demonstrates outstanding performance in terms of IoU. However, the Params and FLOPs of our model are relatively large. In future work, we need to reduce the complexity of the model while ensuring accuracy.

Table 10 The accuracy and efficiency of the related researches

Models	Params(M)	FLOPs(G)	*F*1-score	IoU
U-Net	39.50	8.74	0.644	0.608
LinkNet	21.64	27.38	0.783	0.634
D-LinkNet	31.09	33.58	0.780	0.641
DDU-Net	33.76	40.61	0.784	0.644
Ours	37.22	50.32	0.809	0.679

Discussion

AGF-Net is an improvement upon the U-shaped encoder–decoder architecture, and its core lies in the utilization of the Dilated Convolution Strip Attention (DCSA) Module and Global Feature Fusion Module (GFFM) to extract more comprehensive road information. The Dilated Convolution Attention Module establishes long-range dependencies while reducing non-road information interference. The Global Feature Fusion Module fully utilizes the feature information generated during the encoding process. Compared to classical encoder–decoder models such as U-Net and LinkNet, AGF-Net leverages the GFFM to better utilize fine-grained details from shallow features. Compared to D-LinkNet, our proposed DCSA avoids introducing irrelevant information with large dilation rates while maintaining long-range dependencies. The analysis of IoU, *F*1-score, and prediction results indicates that AGF-Net achieves excellent performance in road extraction continuity and fine road extraction, attributed to the contributions of the Dilated Convolutional Strip Attention (DCSA) module and the Global Feature Fusion Module (GFFM).

The proposed AGF-Net exhibits strong performance in semantic segmentation of road information, effectively avoiding interference from non-road elements such as buildings and vegetation. In future work, we aim to explore the application of AGF-Net in road extraction by incorporating multi-source data fusion, including remote-sensing imagery from various data sources, to further enhance the accuracy of road extraction. Integrating the model into real-world scenarios would provide further validation of its practicality and feasibility. Furthermore, integrating the model with other relevant tasks or systems, such as traffic monitoring systems, navigation systems, or urban planning tools, can be consid-

Table 9 Ablation results

Models	Precision	Recall	*F*1-score	IoU	mIoU
Baseline	0.792	0.775	0.783	0.634	0.796
Baseline + DCSA	0.824	0.772	0.790	0.651	0.818
Baseline + GFFM	0.810	0.757	0.783	0.661	0.817
Baseline + DCSA + GFFM	0.837	0.783	0.809	0.679	0.827

ered. By combining the model with real-world applications, we can assess its performance in practical scenarios and evaluate its adaptability to factors such as complex backgrounds, occlusions, and lighting variations. Applying the road extraction model to real-world scenarios enables accurate road extraction and provides valuable data and support for fields such as urban planning, navigation services, and traffic monitoring.

Conclusion

In this paper, we propose AGF-Net for road extraction from remote-sensing images to tackle the challenges of road occlusion caused by tree shadows and the issue of small road detection. The model utilizes an encoder–decoder architecture to learn road features. Considering the characteristics of roads, we design the Dilated Convolutional Strip Attention (DCSA) module to establish long-range dependencies along the road direction. The Global Feature Fusion module (GFFM) is implemented based on the multi-scale strip convolution modules. This module can fuse features from different levels to obtain multi-scale road features. To more effectively integrate information from different scales, we designed the multi-scale strip convolution module to bridge the semantic gap of features at different scales. We conducted experiments on four public datasets (Massachusetts Road dataset, DeepGlobe Road dataset, CHN6-CUG dataset, and BJRoad dataset.), and the results showed that our AGF-Net outperformed the other models. In addition, we conducted ablation experiments under various conditions demonstrated that the modules we designed effectively enhance the accuracy of road extraction. The proposed AGF-Net has demonstrated significant advancements in enhancing the accuracy of road segmentation. However, it is noteworthy that the method exhibits a higher computational demand and slower training speed. In future work, there is a pressing need for further investigations to expedite the segmentation process while preserving the precision of road extraction.

Acknowledgements The work described in this paper was supported by the National Natural Science Foundation of China under Grant No. 62206086, the Natural Science Foundation of Hebei Province under Grant No. F2023202062, and also funded by Natural resources science and technology plan project of Hebei Province (454-0601-YBN-IBBM).

Data availability The data that support the findings of this study are available on request from the corresponding author, Zhang, upon reasonable request.

References

1. Kaack LH, Chen GH, Morgan MG (2019) Truck traffic monitoring with satellite images. In: Proceedings of the 2nd ACM SIGCAS conference on computing and sustainable societies, pp 155–164
2. Javed AR, Hassan MA, Shahzad F, Ahmed W, Singh S, Baker T, Gadekallu TR (2022) Integration of blockchain technology and federated learning in vehicular (iot) networks: a comprehensive survey. Sensors 22(12):4394
3. Xu Y, Xie Z, Feng Y, Chen Z (2018) Road extraction from high-resolution remote sensing imagery using deep learning. Remote Sens 10(9):1461
4. Li Y, Guo L, Rao J, Xu L, Jin S (2018) Road segmentation based on hybrid convolutional network for high-resolution visible remote sensing image. IEEE Geosci Remote Sens Lett 16(4):613–617
5. Hassan MA, Javed R, Granelli F, Gen X, Rizwan M, Ali SH, Junaid H, Ullah S et al (2023) Intelligent transportation systems in smart city: a systematic survey. In: 2023 international conference on robotics and automation in industry (ICRAI). IEEE, pp 1–9
6. Liu B, Wu H, Wang Y, Liu W (2015) Main road extraction from zy-3 grayscale imagery based on directional mathematical morphology and vgi prior knowledge in urban areas. PLoS ONE 10(9):0138071
7. Abdollahi A, Pradhan B, Shukla N, Chakraborty S, Alamri A (2020) Deep learning approaches applied to remote sensing datasets for road extraction: a state-of-the-art review. Remote Sens 12(9):1444
8. Lian R, Wang W, Mustafa N, Huang L (2020) Road extraction methods in high-resolution remote sensing images: a comprehensive review. IEEE J Sel Top Appl Earth Obs Remote Sens 13:5489–5507
9. Lin M, Chen Q, Yan S (2013) Network in network. arXiv preprint arXiv:1312.4400
10. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
11. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. Springer, pp 234–241
12. Chaurasia A, Culurciello E (2017) Linknet: exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE visual communications and image processing (VCIP). IEEE, pp 1–4
13. Zhou L, Zhang C, Wu M (2018) D-linknet: linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 182–186
14. Wu Q, Luo F, Wu P, Wang B, Yang H, Wu Y (2020) Automatic road extraction from high-resolution remote sensing images using a method based on densely connected spatial feature-enhanced pyramid. IEEE J Sel Top Appl Earth Obs Remote Sens 14:3–17

15. Xie Y, Miao F, Zhou K, Peng J (2019) Hsgnet: a road extraction network based on global perception of high-order spatial information. ISPRS Int J Geo Inf 8(12):571
16. Liu X, Wang Z, Wan J, Zhang J, Xi Y, Liu R, Miao Q (2023) Roadformer: road extraction using a swin transformer combined with a spatial and channel separable convolution. Remote Sens 15(4):1049
17. Tao J, Chen Z, Sun Z, Guo H, Leng B, Yu Z, Wang Y, He Z, Lei X, Yang J (2023) Seg-road: a segmentation network for road extraction based on transformer and cnn with connectivity structures. Remote Sens 15(6):1602
18. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, vol 30
20. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125
21. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
22. Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122
23. Lan M, Zhang Y, Zhang L, Du B (2020) Global context based automatic road segmentation via dilated convolutional neural network. Inf Sci 535:156–171
24. Gao L, Song W, Dai J, Chen Y (2019) Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network. Remote Sens 11(5):552
25. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2014) Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062
26. Chen L-C, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587
27. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder–decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 801–818
28. Tong Z, Li Y, Zhang J, He L, Gong Y (2023) Msfanet: multiscale fusion attention network for road segmentation of multispectral remote sensing data. Remote Sens 15(8):1978
29. Chen X, Sun Q, Guo W, Qiu C, Yu A (2022) Ga-net: a geometry prior assisted neural network for road extraction. Int J Appl Earth Obs Geoinf 114:103004
30. Qu S, Zhou H, Zhang B, Liang S (2022) Mspnet: multi-scale strip pooling network for road extraction from remote sensing images. Appl Sci 12(8):4068
31. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J (2018) Unet++: a nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer, pp 3–11
32. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen Y-W, Wu J (2020) Unet 3+: a full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1055–1059
33. Gruen A, Li H (1995) Road extraction from aerial and satellite images by dynamic programming. ISPRS J Photogramm Remote Sens 50(4):11–20
34. Barzohar M, Cooper DB (1996) Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation. IEEE Trans Pattern Anal Mach Intell 18(7):707–721
35. Anil P, Natarajan S (2010) A novel approach using active contour model for semi-automatic road extraction from high resolution satellite imagery. In: 2010 second international conference on machine learning and computing. IEEE, pp 263–266
36. Ronggui M, Weixing W, Sheng L (2012) Extracting roads based on retinex and improved canny operator with shape criteria in vague and unevenly illuminated aerial images. J Appl Remote Sens 6(1):063610
37. Mnih V, Hinton GE (2010) Learning to detect roads in high-resolution aerial images. In: Computer vision—ECCV 2010: 11th European conference on computer vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part VI 11. Springer, pp 210–223
38. Zhang Z, Liu Q, Wang Y (2018) Road extraction by deep residual u-net. IEEE Geosci Remote Sens Lett 15(5):749–753
39. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
40. Xu Y, Xie Z, Feng Y, Chen Z (2018) Road extraction from high-resolution remote sensing imagery using deep learning. Remote Sens 10(9):1461
41. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
42. Mei J, Li R-J, Gao W, Cheng M-M (2021) Coanet: connectivity attention network for road extraction from satellite imagery. IEEE Trans Image Process 30:8540–8552
43. Chen S-B, Ji Y-X, Tang J, Luo B, Wang W-Q, Lv K (2021) Dbranet: road extraction by dual-branch encoder and regional attention decoder. IEEE Geosci Remote Sens Lett 19:1–5
44. Ding L, Bruzzone L (2020) Diresnet: direction-aware residual network for road extraction in VHR remote sensing images. IEEE Trans Geosci Remote Sens 59(12):10243–10254
45. Wang Y, Seo J, Jeon T (2021) Nl-linknet: toward lighter but more accurate road extraction with nonlocal operations. IEEE Geosci Remote Sens Lett 19:1–5
46. Lu X, Zhong Y, Zheng Z, Zhang L (2021) Gamsnet: globally aware road detection network with multi-scale residual learning. ISPRS J Photogramm Remote Sens 175:340–352
47. Zhu Q, Zhang Y, Wang L, Zhong Y, Guan Q, Lu X, Zhang L, Li D (2021) A global context-aware and batch-independent network for road extraction from VHR satellite imagery. ISPRS J Photogramm Remote Sens 175:353–365
48. Wan J, Xie Z, Xu Y, Chen S, Qiu Q (2021) Da-roadnet: a dual-attention network for road extraction from high resolution satellite imagery. IEEE J Sel Top Appl Earth Obs Remote Sens 14:6302–6315
49. Zhang Z, Miao C, Liu C, Tian Q (2022) Dcs-transupernet: road segmentation network based on cswin transformer with dual resolution. Appl Sci 12(7):3511
50. Luo L, Wang J-X, Chen S-B, Tang J, Luo B (2022) Bdtnet: road extraction by bi-direction transformer from remote sensing images. IEEE Geosci Remote Sens Lett 19:1–5
51. Zhang Z, Sun X, Liu Y (2022) Gmr-net: road-extraction network based on fusion of local and global information. Remote Sens 14(21):5476
52. Jie Y, He H, Xing K, Yue A, Tan W, Yue C, Jiang C, Chen X (2022) Meca-net: a multiscale feature encoding and long-range context-

aware network for road extraction from remote sensing images. Remote Sens 14(21):5342
53. Wang Y, Peng Y, Li W, Alexandropoulos GC, Yu J, Ge D, Xiang W (2022) Ddu-net: dual-decoder-u-net for road extraction using high-resolution remote sensing images. IEEE Trans Geosci Remote Sens 60:1–12
54. Khan SD, Alarabi L, Basalamah S (2023) Dsmsa-net: deep spatial and multi-scale attention network for road extraction in high spatial resolution satellite images. Arab J Sci Eng 48(2):1907–1920
55. Dai L, Zhang G, Zhang R (2023) Radanet: road augmented deformable attention network for road extraction from complex high-resolution remote-sensing images. IEEE Trans Geosci Remote Sens 61:1–13
56. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder–decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 801–818
57. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J (2017) A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857
58. Yuan H, Wang S, Bao Z, Wang S (2023) Automatic road extraction with multi-source data revisited: completeness, smoothness and discrimination. Proc VLDB Endow 16(11):3004–3017
59. Zhang Y, Hsueh Y-L, Lee W-C, Jhang Y-H (2015) Efficient cache-supported path planning on roads. IEEE Trans Knowl Data Eng 28(4):951–964
60. Wang T, Zhao Y, Wang J, Somani AK, Sun C (2020) Attention-based road registration for gps-denied uas navigation. IEEE Trans Neural Netw Learn Syst 32(4):1788–1800
61. Wang Q, Han T, Qin Z, Gao J, Li X (2020) Multitask attention network for lane detection and fitting. IEEE Trans Neural Netw Learn Syst 33(3):1066–1078