DOI: 10.19884/j.1672-5220.202401004

A Dense Feature Iterative Fusion Network for Extracting Building Contours from Remote Sensing Imagery

WU Jiangyan[1,2], WANG Tong[1,2*]
1. College of Information Science and Technology, Donghua University, Shanghai 201620, China
2. Engineering Research Center of Digitized Textile & Apparel Technology, Ministry of Education, Donghua University, Shanghai 201620, China

Abstract: Extracting building contours from aerial images is a fundamental task in remote sensing. Current building extraction methods cannot accurately extract building contour information and have errors in extracting small-scale buildings. This paper introduces a novel dense feature iterative (DFI) fusion network, denoted as DFINet, for extracting building contours. The network uses a DFI decoder to fuse semantic information at different scales and learns the building contour knowledge, producing the last features through iterative fusion. The dense feature fusion (DFF) module combines features at multiple scales. We employ the contour reconstruction (CR) module to access the final predictions. Extensive experiments validate the effectiveness of the DFINet on two different remote sensing datasets, INRIA aerial image dataset and Wuhan University (WHU) building dataset. On the INRIA aerial image dataset, our method achieves the highest intersection over union (IoU), overall accuracy (OA) and *F*1 scores compared to other state-of-the-art methods.

Key words: remote sensing image; building contour extraction; feature iteration
CLC number: TP751.1 **Document code**: A
Article ID: 1672-5220(2024)06-0654-08

Open Science Identity (OSID)

0 Introduction

Building contour extraction has a variety of applications in remote sensing, such as urban planning[1], disaster monitoring[2] and population estimation[3-4]. In recent years, the quality of high-resolution aerial imagery has continuously improved with the development of drone detection technology. As a result, the extraction of building contours from these images has received increasing attention. However, accurately extracting complete building contours from complex backgrounds is still challenging. Different types of buildings vary greatly in appearance and size. Vehicles and vegetation also disturb the accuracy of extracting buildings. At the same time, buildings bear a certain degree of similarity to certain objects, such as parking lots, roads and swimming pools, which further increases the difficulty of this study.

Traditional methods for extracting building features involve extracting features from active contours, line detectors, corner detectors and shadow context. These methods heavily rely on prior knowledge of the image and are only applicable to specific scenarios, while they cannot achieve satisfactory accuracy. Recently, researchers have made significant progress in developing methods for extracting building contours. This progress is because of the advancement of deep learning techniques such as convolutional neural networks (CNNs). Generally, CNN-based methods obtain semantic feature information through an encoder-decoder structure and then extract features from building contours to separate building boundaries. Fully convolutional networks (FCNs) derived from CNNs can adeptly accomplish pixel-by-pixel semantic segmentation.

Many studies focus on improving the receptive field of the building extraction network structures. In this way, the network captures the full semantic information of buildings. Chen et al.[5] proposed a Deeplab v3+ model with an encoder-decoder architecture. This model enhanced the receptive field of the model to capture global semantic information. Liu et al.[6] innovated a spatial residual module to boost the network's feature extraction capability. Cai et al.[7] presented a groundbreaking approach that merged mixed depthwise convolution and dense upsampling convolution. This approach effectively extracted features and enhanced the model's capability to capture details of small-scale buildings. Yuan et al.[8] introduced the Swin Transformer as the encoder to extract features and employed a scale-adaptive decoder to

Received date: 2024-01-31
Foundation items: National Natural Science Foundation of China (No. 61903078); Fundamental Research Funds for the Central Universities, China (No. 2232021A-10); Shanghai Sailing Program, China (No. 22YF1401300); Natural Science Foundation of Shanghai, China (No. 20ZR1400400)
* Correspondence should be addressed to WANG Tong, email: wangtong@dhu.edu.cn
Citation: WU J Y, WANG T. A dense feature iterative fusion network for extracting building contours from remote sensing imagery [J]. *Journal of Donghua University (English Edition)*, 2024, 41(6): 654-661.

generate prediction results. The Transformer-based approach can adequately capture global information but overlooks contextual details. These methods aim to improve the accuracy of building extraction by increasing the feature receptive field. However, they do not effectively fuse multiscale features by utilizing both global and local information. As a result, feature loss is inevitable during the upsampling stage of the decoder.

In this research, we propose a novel dense feature iterative (DFI) fusion network, denoted as DFINet for building contour extraction. The network uses a DFI decoder to extract features at different scales, and then gradually iterates to fuse them. It achieves this by merging high-scale features with multi-scale features through a feature refinement module. Additionally, the network avoids feature loss caused by downsampling by adding a dense feature fusion (DFF) module. As a result, the final prediction effectively captures valid features at various scales. Finally, the model uses a contour reconstruction (CR) module to generate the ultimate prediction result. To validate the effectiveness of DFINet, we conduct comprehensive experiments on two different remote sensing datasets, INRIA aerial image dataset and Wuhan University (WHU) building dataset. The main contributions of this paper can be listed as follows.

1) A DFINet is proposed for building contour extraction from aerial imagery. This method iteratively fuses dense features, gradually outputs the prediction results, and refines them step by step. Furthermore, it effectively improves the prediction accuracy by enhancing the boundary information and retaining both the semantic and spatial information.

2) A DFI decoder is designed to enhance feature fusion across different scales. The next-scale feature map is generated through a feature refinement module. The decoder generates a prediction result for each scale, followed by a step-by-step refinement process. Ultimately, the stepwise refinement process produces the final prediction.

3) A DFF module is introduced to compensate for the loss of feature information caused by downsampling. The effective fusion of features at different scales enhances both semantic and spatial information. It prevents the neglect of high-scale semantic information and loss of low-scale detail information.

The remaining parts of this paper are organized as follows. Section 1 introduces the structure and composition of the DFINet. Section 2 showcases experimental results and analyses. Conclusions are given in Section 3.

1 Methodology

Researchers in the field of remote sensing building contour extraction have recognized that feature maps at low scales are rich in local spatial information. This means that at smaller scales, the network obtains information about the local details of buildings. Meanwhile, high-scale feature maps contain rich global semantic information about the overall layout and structure of the buildings. By utilizing both low-scale and high-scale feature maps, the network acquires more contextual information, leading to more accurate and comprehensive building contour extraction results. However, during the downsampling process, inevitably, high-scale feature maps will lose some feature information, resulting in suboptimal predicted results. Based on the above issues, we introduce a novel DFINet model for building contour extraction. Figure 1 illustrates the architecture of DFINet. M_i represents multi-scale features, E_i denotes encoder features, D_i indicates decoder features, and RRM is short for residual refinement module. This section will sequentially introduce the DFI decoder, DFF module, CR module and loss functions used for training DFINet.

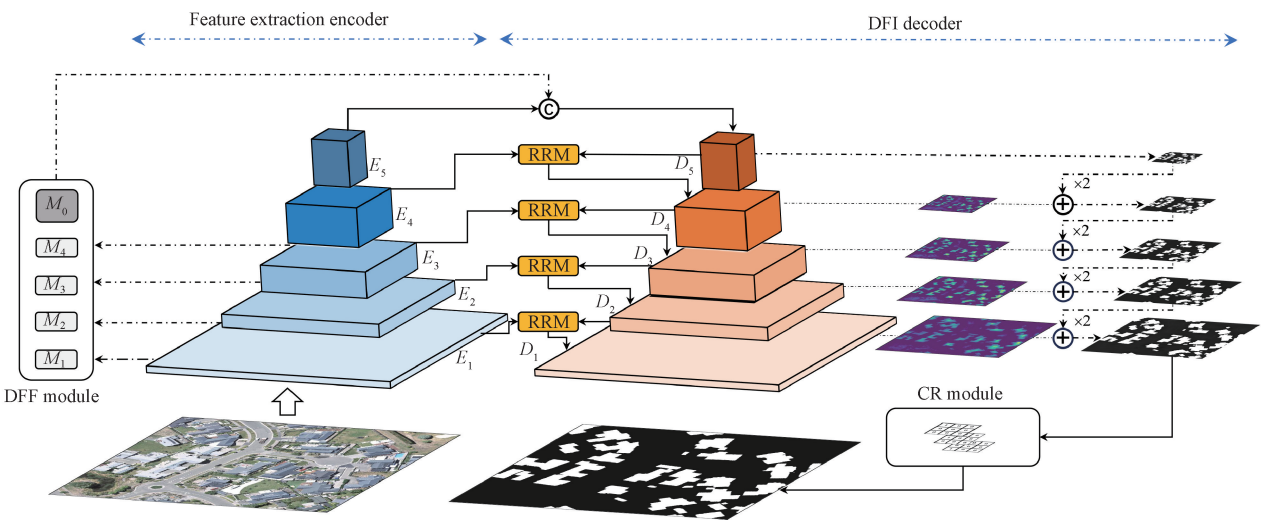

Fig. 1 Architecture of DFINet

1.1 DFI decoder

The DFINet is built upon the U-Net encoder-decoder architecture. The network uses the basic VGG-16 as an encoder, and a novel DFI decoder in the decoder module, effectively enhancing the network's receptive field. By iteratively generating the next stage feature maps through the combination of low and high-scale features, the network fully utilizes global semantic information and local spatial information. Additionally, to enhance the network's understanding of boundary features, it introduces a building boundary predictor. This predictor is incorporated into each decoding layer and is responsible for predicting the building boundaries. Our network model becomes capable of learning an extensive range of boundary features.

The DFI decoder is a decoder that combines the encoding features with the decoding features through iterative fusion to generate decoding features at different scales. At each decoding layer, it generates building prediction results and building boundary information at that scale. By iteratively fusing the features and continuously refining the building prediction results, the decoder generates the final feature map. Then, using the CR module, the network generates the ultimate prediction result. The decoder achieves the generation of decoding features at each layer by utilizing the RRM. The input for the RRM is the output features from the dense feature decoder at the previous scale and the output features from the encoder at the current scale. Figure 2 demonstrates the structure of the RRM.

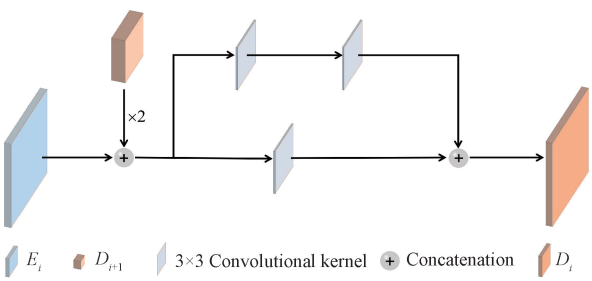

Fig. 2 Structure of RRM

To optimize feature integration, the RRM merges the feature extracted from encoder with the decoder feature at the previous scale. Firstly, the features from the decoder at the previous scale are upsampled to the same resolution as the features from the encoder at this scale. Then, a residual block connection combines the features from the encoder and the features from the decoder to generate the decoding features at this scale using two 3×3 convolutional kernels. The decoder generates the building prediction results at each scale by combining the building prediction results from the previous scale with the decoding features at the current scale. In this way, the low-resolution prediction results with global semantic information can be gradually refined by the high-resolution features with local spatial details. Finally, the model generates the final prediction by inputting the decoding features from the last layer and the prediction result at this scale into the CR module.

1.2 DFF module

To extract building outline features, the network must downsample the information through the encoder. Unfortunately, this downsampling process inevitably leads to a loss of feature information. To address feature loss during downsampling, we introduce a DFF module. This module extracts encoding features at different scales and then merges and accumulates them onto the first decoding layer, effectively solving the problem of feature loss during the downsampling process. The building outline prediction result generated in the first decoding layer also contains richer details, laying a better foundation for the subsequent gradual refinement process.

In general, networks can extract features directly for prediction by utilizing convolutional layers. However, many studies have shown that different scales of features play an indispensable role in understanding the overall scene information of images. The DFF process effectively enhances the information representation in deeper layers by seamlessly integrating the DFI decoder with multi-scale features. This approach enhances the overall performance of the system. The DFF module utilizes downsampling and channel concatenation to merge encoder features of four different scales. The module downsamples the shape size of the encoder layers to the size of the first decoder layer, setting the default number of channels to 128. A 5×5 convolutional layer combines these downsampled features into multi-scale features. After that, the module adds the resulting multi-scale features to the last encoder layer to generate the first layer of decoder features. Based on this, the network generates the building extraction prediction result of the first decoder layer and then gradually refines it through the DFI decoder.

1.3 CR module

To further address blurry building contours, we introduce the CR module. By inputting high-resolution building features into the CR module, the module can fix the boundary pixels to make the boundary information more prominent. This module refers to the design concept of Guo et al.[9] and introduces orientation prediction in the process of predicting building boundaries. The CR module takes 360° and converts it into 8 directions, with each direction spaced 45° apart. It then uses a direction classifier and the softmax function to predict the refinement of each pixel at the building boundary corresponding to a fixed direction.

The module takes the output features of the last dense feature iteration decoder as input. It combines the refined boundary feature maps with the building prediction maps generated from the previous layer of decoder features to produce the final building prediction results.

1.4 Loss functions

In this paper, the loss function $\mathcal{L}$ consists of three parts: the dense feature supervision loss l_d, the building precision loss l_p, and the self-supervision loss l_s. The composition of the loss function is defined as

$$\mathcal{L} = l_s + l_p + \alpha \cdot l_d, \tag{1}$$

where α is the weight coefficient of the dense feature supervision loss, and this paper uses $\alpha = 0.25$ in all experiments.

The dense feature supervision loss is calculated using the cross-entropy loss between the predicted values and the downsampled ground truth values. The dense feature supervision loss is defined as

$$\begin{aligned} l_d = & \sum_{k=1}^{3} \left[-\frac{1}{N} \sum_{i=1}^{N} 0.85 \times \mathrm{I}(E) \times \log(\hat{E}_k) + \right. \\ & \left. 0.15 \times (1 - \mathrm{I}(E)) \times \log(1 - \hat{E}_k) \right] + \\ & \sum_{k=1}^{4} \left[-\frac{1}{N} \sum_{i=1}^{N} \mathrm{I}(B) \times \log(\hat{B}_k) + \right. \\ & \left. (1 - \mathrm{I}(B)) \times \log(1 - \hat{B}_k) \right], \end{aligned} \tag{2}$$

where B is the real truth of the building contour and E is the edge truth of the building; $\hat{E}$ and $\hat{B}$ are the prediction results of B and E respectively from the kth DFI decoder layer; k stands for the number of decoder layers; I is a bilinear interpolation function to perform downsampling; N represents the quantity of model prediction object groups.

The building precision loss l_p can be expressed as

$$l_p = \mathrm{CE}(B, \hat{B}) + \mathrm{CE}(E, \hat{E}) + \mathrm{CE}(D, \hat{D}) + \mathrm{CE}(B, \hat{B}_R), \tag{3}$$

where CE refers to the cross-entropy loss function; D represents the direction of the ground truth; $\hat{D}$ and $\hat{B}_R$ represent the direction prediction and the network-refined building prediction results, respectively.

The self-supervision loss l_s can be concluded as

$$\begin{aligned} l_s = & -\frac{1}{N} \sum_{i=1}^{N} \frac{1}{1 + \mathrm{e}^{-\hat{B}_R}} \times \log(\hat{B}) + \\ & \frac{\mathrm{e}^{-\hat{B}_R}}{1 + \mathrm{e}^{-\hat{B}_R}} \times \log(1 - \hat{B}). \end{aligned} \tag{4}$$

2 Experiments and Analyses

2.1 Dataset

2.1.1 INRIA aerial image dataset

This dataset contains 360 high-resolution remote sensing images, each with a resolution of 5 000 pixel × 5 000 pixel. Since the ground truth results for the test set are not publicly available, we only conduct experiments on the training set. To conduct the experiments, we divide each 5 000 pixel × 5 000 pixel remote-sensing image into small blocks with a resolution of 512 pixel × 512 pixel. The cut images are then divided into the training set, validation set and test set with the ratio of 6∶2∶2.

2.1.2 WHU building dataset

This dataset comprises aerial image and satellite image datasets. In our experiment, we only use aerial images from the dataset. These aerial images cover an area of over 450 km^2, with approximately 22 000 buildings. The dataset consists of 8 188 remote sensing satellite images, each sized at 512 pixel × 512 pixel. In our experiments, we follow the official settings.

2.2 Evaluation metrics

To assess the accuracy of our model's predictions, we use intersection over union (IoU), overall accuracy (OA) and $F1$ score. These evaluation metrics are widely used in building contour extraction. I_{IoU} represents the ratio of the intersection area between the building prediction and the actual situation in the union area. O_{OA} represents the proportion of correctly classified samples to all samples. P and R represent the precision and recall, respectively. The computation of the $F1$ score involves using the harmonic mean of precision and recall. As these metrics increase, the accuracy of the prediction results gets better.

The evaluation metrics can be represented through the following formulas:

$$I_{IoU} = T_{TP} / (T_{TP} + F_{FP} + F_{FN}), \tag{5}$$

$$O_{OA} = (T_{TP} + T_{TN}) / (T_{TP} + T_{TN} + F_{FP} + F_{FN}), \tag{6}$$

$$F_1 = (2 \times P \times R) / (P + R), \tag{7}$$

$$P = T_{TP} / (T_{TP} + F_{FP}), \tag{8}$$

$$R = T_{TP} / (T_{TP} + F_{FN}), \tag{9}$$

where T_{TP} represents true positive samples; F_{FP} stands for false positive samples; T_{TN} denotes true negative samples; F_{FN} signifies false negative samples.

2.3 Implementation details

During the training process, we used the PyTorch framework to implement our method and train DFINet on four NVIDIA 2080TI GPUs. DFINet network is trained using the Adam optimizer with 150 iterations. The initial learning rate is set to 0.001, and the weight decay rate is set to 1×10^{-5}. Additionally, during training, we set the batch size to 12. To prevent overfitting, we also perform data augmentation, including horizontal and vertical flips, and random rotation angles between 0°–90°. The probability of horizontal and vertical flips is set to 0.5.

2.4 Experimental results

We validate the model by conducting comprehensive experiments. In these experiments, DFINet is compared with other state-of-the-art building extraction techniques and well-known image labeling approaches, including UNet[10], SegNet[11], DeepLabV3[12], DANet[13], SETR[14], ESFNet[15], MAP-Net[16], BRRNet[17], SRI-Net[6], DAN-Net[18], STT[19], CBR-Net[9] and BCT-Net[20].

2.4.1 Quantitative comparisons

Table 1 showcases the outcomes of different building extraction methods on the INRIA aerial image dataset, whereas Table 2 presents the corresponding results on the WHU building dataset. The best results are in bold and the second best are underlined. The arrows in the table indicate that the higher the value of metrics the better the performance of methods. The experimental results clearly show that our method outperforms others in all indicators on the INRIA aerial image dataset. Furthermore, our method achieves a superior OA result on the WHU building dataset, while also exhibits IoU and $F1$ scores that closely approach the state-of-the-art performance. These data strongly demonstrate the accuracy and effectiveness of DFINet.

Table 1 Results on the INRIA aerial image dataset

Methods	IoU ↑	OA ↑	$F1$ ↑
SETR[14]	70.34	94.87	82.06
DeepLabV3[12]	73.90	95.54	84.39
DANet[13]	76.02	95.94	85.80
SegNet[11]	76.32	96.10	85.83
UNet[10]	77.29	96.25	86.56
DAN-Net[18]	76.73	96.08	86.17
MAP-Net[16]	76.91	96.13	86.34
SRI-Net[6]	76.84	96.12	86.32
BRRNet[17]	77.05	96.47	86.61
STT[19]	79.42	96.59	87.99
BCT-Net[20]	—	—	—
CBR-Net[9]	<u>81.10</u>	<u>96.63</u>	<u>89.56</u>
DFINet (ours)	**81.21**	**96.68**	**89.60**

Table 2 Results on the WHU building dataset

Methods	IoU ↑	OA ↑	$F1$ ↑
SETR[14]	75.92	97.13	87.75
DeepLabV3[12]	80.69	97.76	89.24
DANet[13]	81.22	97.82	89.57
SegNet[11]	85.13	98.36	91.74
UNet[10]	87.52	98.65	93.11
DAN-Net[18]	87.69	98.80	92.81
MAP-Net[16]	88.99	98.82	94.12
SRI-Net[6]	88.84	98.79	93.98
BRRNet[17]	89.03	98.81	94.14
STT[19]	90.48	98.97	94.97
BCT-Net[20]	91.15	—	95.37
CBR-Net[9]	**91.40**	<u>98.98</u>	**95.51**
DFINet (ours)	<u>91.38</u>	**98.99**	<u>95.49</u>

2.4.2 Qualitative comparisons

Figures 3 and 4 showcase the visualization of extracted building contours on the INRIA aerial image dataset and WHU building dataset.

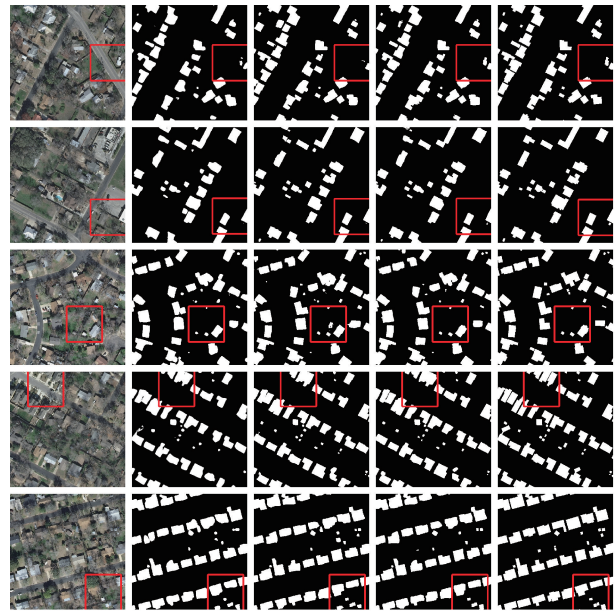

Fig. 3 Visualization of extracted building contours on the INRIA aerial image dataset: (a) aerial imagery; (b) STT; (c) CBR-Net; (d) DFINet (ours); (e) ground truth

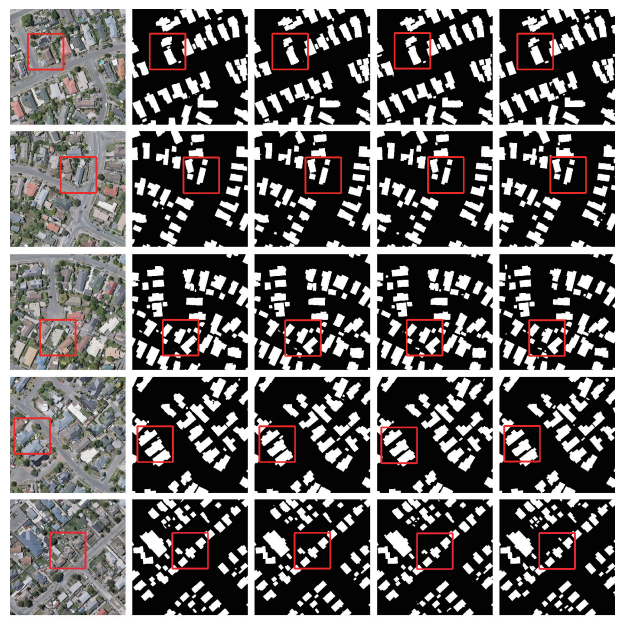

Fig. 4 Visualization of extracted building contours on the WHU building dataset: (a) aerial imagery; (b) STT; (c) CBR-Net; (d) DFINet (ours); (e) ground truth

These figures illustrate building extraction using a black-and-white binary map, where white indicates buildings and black represents non-building objects within the remote sensing image. The red boxes in the figures highlight specific sections of the buildings where the extraction effect is more prominent. By examining these framed building extraction prediction maps, it becomes

evident that the results produced by DFINet closely match the ground truth, capturing finer details and enhancing the edge profiles of the buildings. When faced with complex backgrounds, higher levels of noise or the distribution of numerous small-sized buildings, DFINet places a greater emphasis on the completeness of contour extraction and exhibits superior capability in distinguishing disturbances. This suggests that the DFI decoder and the DFF module contribute to predicting both the overall architectural spatial structure and fine boundaries effectively.

2.5 Ablation studies

We conduct ablation experiments on the INRIA aerial image dataset to evaluate the effectiveness of DFINet. The evaluation metrics obtained from the ablation experiments are presented in Table 3. In this context, "Baseline" refers to the direct combination of encoding features and decoding features through the encoding-decoding architecture of the UNet network to generate the output features, and then directly produce the final prediction results. The process does not entail the generation and decoding of features via the DFI decoder and the DFF module, nor does it encompass boundary optimization through the CR module. We conduct validation on four distinct architectural models, commencing with the Baseline model and augmenting it with the CR module, DFI decoder and DFF module sequentially. Table 3 presents data that demonstrate the increasingly precise results obtained through the sequential addition of each module ("√" denotes the module incorporated in the corresponding method). The most optimal performance of the experimental method is observed when all three modules are incorporated.

Table 3 Ablation experiments

Method	Module			Result		
	CR	DFI	DFF	IoU	OA	*F*1
Baseline				80.67	96.58	89.30
DFINet	√			80.83	96.63	89.40
DFINet	√	√		81.14	96.67	89.58
DFINet	√	√	√	**81.21**	**96.68**	**89.60**

To verify the superiority of the backbone, we conduct experiments on the WHU building dataset. Table 4 shows the experimental results for different backbones, and VGG-16 achieves the most superior performance.

Table 4 Backbone results

Backbone	IoU	OA	*F*1
ResNet-101	90.27	98.85	94.87
ResNet-50	90.34	98.87	94.91
VGG-16	**91.38**	**98.99**	**95.49**

The visual results of the ablation experiments are shown in Fig. 5. It can be observed that DFINet achieves the best building contour extraction results and boundary integrity. This fully demonstrates the excellent fusion effect of the DFI decoder on features of different scales, preserving both local spatial information and overall semantic features. The DFF module also effectively complements the semantic information loss caused by downsampling.

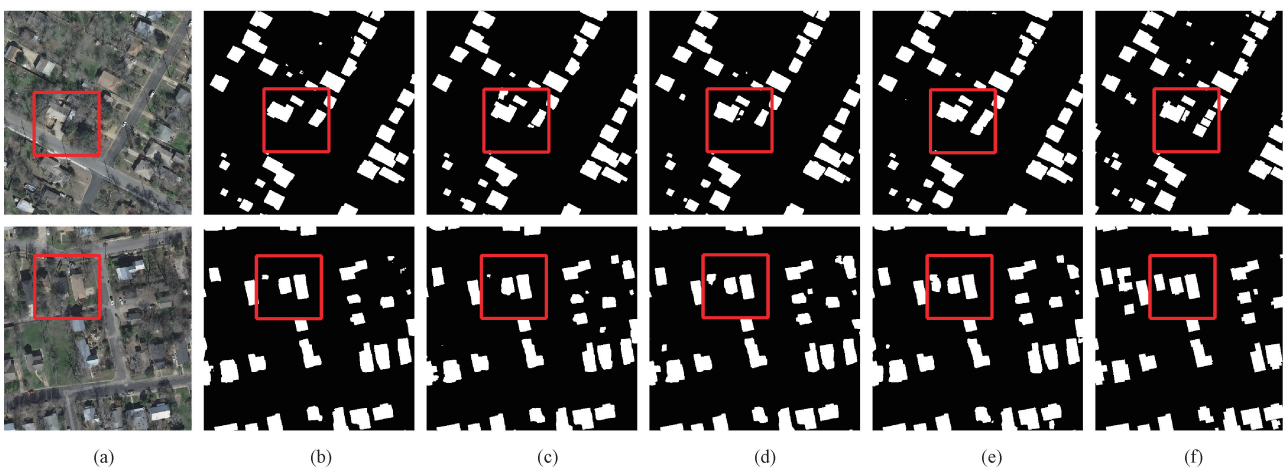

Fig. 5 Visualization of ablation experiments: (a) aerial imagery; (b) Baseline; (c) CR module; (d) CR module and DFI decoder; (e) DFINet; (f) ground truth

3 Conclusions

In this paper, an efficient DFINet model is introduced. It iteratively fuses encoding and decoding features of different scales and enhances the features through the DFF module to generate the final feature map. The CR module generates the final prediction result. The DFI decoder effectively combines features of different scales, giving the network a larger receptive

field. The DFF module effectively incorporates multi-scale features during the encoding and decoding stages, compensating for the feature loss caused by downsampling in the encoder. This method effectively addresses challenges related to blurred boundaries of intricate structures and complications in contour extraction of small-scale buildings. Our approach has been thoroughly validated through extensive experimentation. The validation is conducted on both the INRIA aerial image dataset and the WHU building dataset, producing compelling results.

Though DFINet is capable of extracting accurate building contours, it is difficult to effectively extract complete building boundary information in images with a lot of interference. To enhance the feature extraction capability of the network, it is necessary to further explore more applicable multi-scale feature fusion strategies.

References

[1] GUO M Q, LIU H, XU Y Y, et al. Building extraction based on U-net with an attention block and multiple losses[J]. *Remote Sensing*, 2020, 12(9): 1400.

[2] CHEN M, WU J J, LIU L Z, et al. DR-net: an improved network for building extraction from high resolution remote sensing image[J]. *Remote Sensing*, 2021, 13(2): 294.

[3] ZHOU D J, WANG G Z, HE G J, et al. Robust building extraction for high spatial resolution remote sensing images with self-attention network [J]. *Sensors*, 2020, 20(24): 7241.

[4] JI S P, WEI S Q, LU M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, 57(1): 574-586.

[5] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]//European Conference on Computer Vision. Cham: Springer, 2018: 833-851.

[6] LIU P H, LIU X P, LIU M X, et al. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network[J]. *Remote Sensing*, 2019, 11(7): 830.

[7] CAI J H, CHEN Y M. MHA-net: multipath hybrid attention network for building footprint extraction from high-resolution remote sensing imagery[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, 14: 5807-5817.

[8] YUAN W, XU W B. MSST-net: a multi-scale adaptive network for building extraction from remote sensing images based on swin transformer [J]. *Remote Sensing*, 2021, 13(23): 4743.

[9] GUO H N, DU B, ZHANG L P, et al. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022, 183: 240-252.

[10] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[M]//Lecture Notes in Computer Science. Cham: Springer, 2015: 234-241.

[11] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39 (12): 2481-2495.

[12] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation [EB/OL]. (2017-12-5) [2023-10-12]. https://arxiv.org/abs/1706.05587.

[13] FU J, LIU J, TIAN H J, et al. Dual attention network for scene segmentation [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitons: IEEE, 2019: 6877-6886.

[14] ZHENG S, LU J, ZHAO H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers [C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitons: IEEE, 2021: 6881-6890.

[15] LIN J B, JING W P, SONG H B, et al. ESFNet: efficient network for building extraction from high-resolution aerial images [J]. *IEEE Access*, 2019, 7: 54285-54294.

[16] ZHU Q, LIAO C, HU H, et al. MAP-net: multiple attending path neural network for building footprint extraction from remote sensed imagery[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59(7): 6169-6181.

[17] SHAO Z F, TANG P H, WANG Z Y, et al. BRRNet: a fully convolutional neural network for automatic building extraction from high-resolution remote sensing images [J]. *Remote Sensing*, 2020, 12(6): 1050.

[18] YANG H, WU P H, YAO X D, et al. Building extraction in very high resolution imagery by dense-attention networks [J]. *Remote Sensing*, 2018, 10(11): 1768.

[19] CHEN K Y, ZOU Z X, SHI Z W. Building extraction from remote sensing images with sparse token transformers[J]. *Remote Sensing*, 2021, 13(21): 4441.

[20] XU L L, LI Y, XU J Z, et al. BCTNet: Bi-branch cross-fusion transformer for building footprint extraction[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 4402014.

遥感图像中提取建筑物轮廓的密集特征迭代融合网络

吴江炎[1,2]，王　彤[1,2*]

1. 东华大学 信息科学与技术学院，上海 201620

2. 东华大学 数字化纺织服装技术教育部工程研究中心，上海 201620

摘　要：从航空图像中提取建筑物轮廓是遥感技术的一项基本任务。目前的建筑物提取方法无法准确提取建筑物轮廓信息，而且在提取小规模建筑物时存在误差。该文提出了 DFINet，一种用于提取建筑物轮廓的新型密集特征迭代融合网络。该网络使用密集特征迭代解码器融合不同尺度的语义信息，并学习建筑轮廓知识，通过迭代融合产生最后的特征。密集特征融合模块结合了多个尺度的特征，利用轮廓重构模块获取最终预测结果。大量实验验证了 DFINet 在两个不同遥感数据集（INRIA 航空图像数据集和武汉大学建筑数据集）上的有效性。在 INRIA 航空图像数据集上，与目前其他方法相比，DFINet 在重叠度（intersection over union，IoU）、总体准确率（overall accuracy，OA）和 *F*1 值等指标上取得了最高分。

关键词：遥感图像；建筑物轮廓提取；特征迭代