EasyChair Preprint

№ 14545

BFNet: a Bi-Frequency Fusion Semantic Segmentation Network for High-Resolution Remote Sensing Images

Chengkun Diao and Jinyu Shi

August 26, 2024

BFNet: A Bi-Frequency Fusion Semantic Segmentation Network for High-Resolution Remote Sensing Images

Chengkun Diao and Jinyu Shi✉

School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China
sjy1967@dlmu.edu.cn

Abstract. Semantic segmentation of remote sensing images plays a vital role in urban planning, traffic guidance and other fields. However, high-resolution remote sensing images typically include large and complex scenes and heterogeneous objects, leading to poor segmentation at the edges of objects, which in turn leads to undesirable segmentation of the whole image. Specifically, current semantic segmentation techniques highlight the superiority of CNNs in maintaining local ground details, but they still can't globally build when processing full-geomorphic images. To address these problems, we propose an effective bi-frequency fusion semantic segmentation network (BFNet) for high-resolution remote sensing images. BFNet uses a bi-branch structure, where the low-frequency branch captures low-frequency context information at different scales based on ESwin-Transformer; meanwhile, a pixel-attention mechanism is designed behind the low-frequency branch to select the optimal global context information; The high-frequency branch extracts high-frequency edge information based on stacked CNNs and transverse connections. In addition, to tackle the issue of detail loss caused by the direct fusion of high-frequency and low-frequency information, we designed a boundary fusion module for bi-frequency balancing to enable better segmentation. Our method achieves good performance on two recognized remote sensing datasets, Potsdam and LoveDA, with mIoU of 87.22% on Potsdam and 92.85% on F1. mIoU on LoveDA is 51.37%, which is a relatively good balance in inference speed and accuracy.

Keywords: Semantic Segmentation, Bi-Frequency Fusion, Boundary Fusion, Pixel Attention.

1 Introduction

The study of semantic segmentation in high-resolution remote sensing images is a major field of research and is essential for practical applications like urban planning and land use [1-6]. The aim of semantic segmentation is to assign a label of a specific category to each pixel in the image. However, remote sensing images are rich and varied in detail and the background is complex and variable, so how to accurately and efficiently complete the semantic segmentation of them remains a challenging problem. Recently,

deep learning methods have proven effective for segmenting remote sensing images. Owing to their inherent ability to extract local detail features, deep convolutional neural networks (DCNNs) have a natural advantage, researchers have studied many remote sensing segmentation models [7,10,11] with high performance on this basis. However, these methods still have two limitations:

- Inadequate modeling of global information for full geomorphic images. The receptive field is small, which makes it difficult to fully learn the global information and long-distance context information. For remote sensing image segmentation, global information, long-range spatial context information and edge detail information are particularly important, but most models usually ignore these information.
- Imbalanced fusion of high-frequency and low-frequency information. Fusion of localized low-frequency contextual information is crucial for segmentation results. Most of these models tend to directly fuse the captured low-frequency information (context information) and high-frequency information (edge information). This approach can overwhelm the detailed features leading to inadequate information fusion.

To address the above limitations, inspired by the boundary attention mechanism, we propose a Bi-Frequency Fusion Network (BFNet), as shown in Fig1.

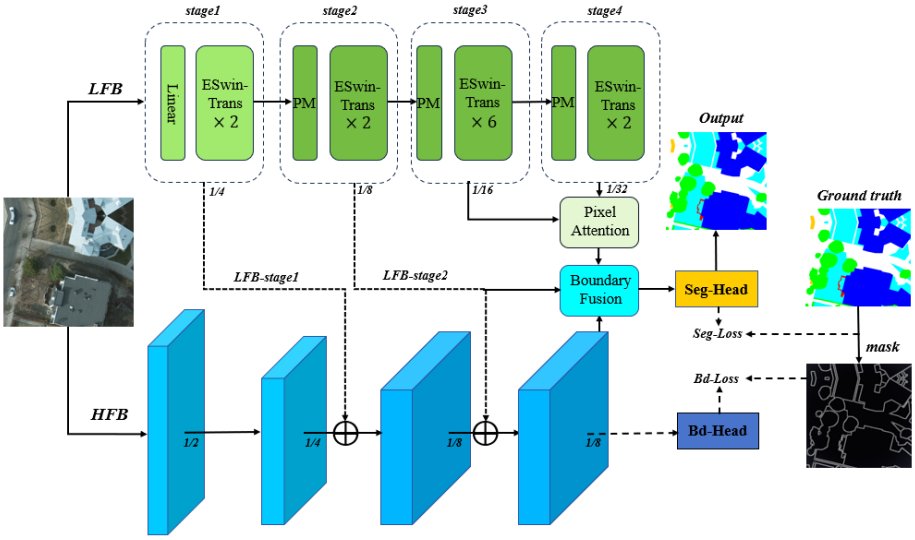

Fig. 1. The figure is a Model of Bi-Frequency Fusion Network.

The BFNet consists of two branches with complementary functions: A branch for obtaining low-frequency data and a branch for obtaining high-frequency data. The low-frequency branch extracts low-frequency information (contextual information) based on ESwin-Transformer blocks. The high-frequency branch uses stacked convolution and optimization loss to highlight the boundary information to extract high-frequency information (edge information). In order to address the issue of inadequate extraction

of global information from remote sensing images, a pixel attention module is designed to select the optimal context information and learn its representation; to address the problem of imbalance in the fusion of high-frequency and low-frequency information, a boundary fusion module is designed, which balances the high-frequency edge information and the low-frequency context information by using boundary attention to fuse the information from different domains. In conclusion, the key contributions of this work include:

1. BFNet is proposed to extract high-frequency information and low-frequency information through a novel two-branch structure composed of CNN and ESwin-Trans, which provides a new perspective to capture both contextual information and boundary information within a unified deep network.
2. A pixel attention module is designed, which uses pixel attention to horizontally select the detail information across feature maps of varying scales, which greatly improves the generalization ability of the model.
3. A boundary fusion module was designed. Boundary attention is utilized to balance the boundary and context information extracted from high-frequency and low-frequency branches to coordinate the fusion of high-frequency features and low-frequency features.

2 Related Works

2.1 Transformer-Based Segmentation Method

Most Transformer-based segmentation models have achieved good results. Transformer-based segmentation methods extract the local information of contextual information and fuse these information, these models combine Transformer and CNN with self-attention mechanism to fuse the local information of low-frequency contextual information. However, these segmentation methods use direct fusion, which ignores the adequacy of information fusion and leads to the loss of detailed information.

2.2 Bi-Frequency Fusion

The proposed bi-frequency fusion refers to the balanced fusion of high-frequency and low-frequency information from different domains to achieve better segmentation. The fusion methods in existing works usually simply sum or splice the information in the same domain. For example, FPN effectively fuses features of different scales by top-down and lateral connectivity. [24] proposed a TBN network architecture, which contains two different branches for obtaining contextual information and detail parsing. In order to fuse the features extracted from these two branches, the authors designed a feature fusion module, FFM, on which they proposed some subsequent work, which is used to improve its generalization ability. However, these methods are direct fusion of information at different scales, which usually leads to the detailed information being lost.

In general, existing works tend to neglect modeling global information and balancing high and low frequency information. Therefore, we propose a bi-frequency fusion network that focuses on sensing boundary information and utilizes pixel attention to balance high-frequency information with low-frequency information to obtain more accurate semantic segmentation results.

3 Method

3.1 Overview

The bi-frequency fusion semantic segmentation network uses a two-branch structure. The Low Frequency Branch (LFB) extracts the low-frequency information, i.e., contextual information, of the remote sensing images through four stages. Each stage consists of multiple consecutive ESwin-Transformer (ESTB) blocks, where stage 2, stage 3 and stage 4 are also down sample using Patch Merging (PM) operation. The different scales of context information generated by stage 3 and stage 4 in the LFB are then fed into the pixel attention module to selectively extract the more plausible low-frequency context information.

The High Frequency Branch (HFB) uses stacked convolution to capture high frequency information i.e. boundary information, every convolutional layer is accompanied by BN and ReLU. This branch fuses low frequency context information at the same scale in LFB at stage 3 and stage 4, and works with the boundary loss to capture the high frequency edge information of the image.

Then, the different domain outputs in LFB and HFB are fused using boundary attention to balance the high-frequency edge information and low-frequency context information. Finally the segmentation header module is added to transform the fused features into segmentation maps, achieving precise semantic segmentation.

3.2 Low-Frequency Branch

Low-frequency branch (LFB) utilises ESwin-Transformer to capture context information from the graph. We argue that only low-frequency context information of a single patch of an image can be extracted by a standard Transformer block, and local context information between neighboring patches cannot be captured. Inspired by Swin-Transformer[9], ESwin-Transformer block is designed in LFB.

The purpose of the Patch Merging layer in the ESwin-Transformer module is to down sample the feature map for hierarchical feature representation. This is shown in Fig.2.

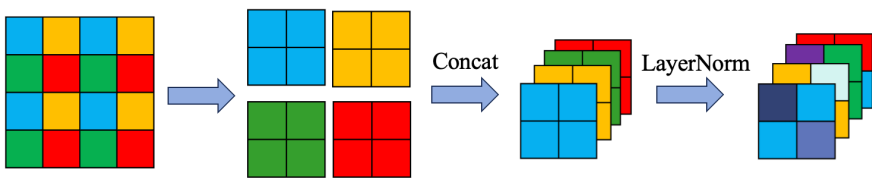

Fig. 2. The process of patch merging.

The Efficient Swin Transformer in the ESwin-Trans block is designed to extract multi-scale contextual information at different stages in the LFB. The block consists of a moving window based MSA module and the calculation process is shown below. Considering the input X, the context information after ESwin-Transformer can be represented as:

$$ESWTB(\boldsymbol{X}) = L(\boldsymbol{X}) + MLP\left(LN\left(L(\boldsymbol{X})\right)\right) \tag{1}$$

$$L(\boldsymbol{X}) = \boldsymbol{X} + EMSA(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \tag{2}$$

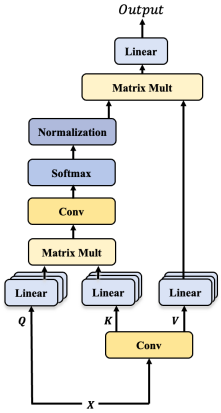

Fig. 3. ESwin-Trans calculation procedure

3.3 Pixel Attention Model

Inspired by[25], Pixel Attention Module uses a pixel attention mechanism to select the optimal low-frequency context information for learning. We believe that the direct fusion of different scale information of stage 3 and stage 4 will cause some small details to be lost. Therefore, we use the pixel attention mechanism to calculate the weights of different scales of information and select the optimal contextual information. The pixel attention selection process is shown in Fig.4.

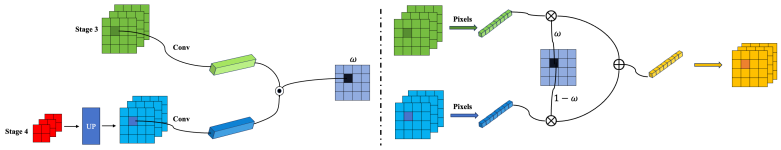

Fig. 4. Pixel Attention Selection Process

$\sigma(x)$ denotes Sigmoid; here all convolution kernels are of size 1*1.

First, the feature mapping of stage3 is bilinearly differenced and expanded to the same size as stage4, and then the corresponding element vectors in the feature mappings of stage 3 and stage 4 are linearly embedded respectively defined as $\vec{v}_{st3}$, $\vec{v}_{st4}$:

$$\vec{v}_{st3} = f\left(G^{st3}(x)\right) \tag{3}$$

$$\vec{v}_{st4} = f\left(G^{st4}(x)\right) \tag{4}$$

where f(·) is a linear function. The output of the Sigmoid function can be expressed as follows:

$$\omega = Sigmoid\left(f_t(\vec{v}_{st3}) \cdot f_i(\vec{v}_{st4})\right) \tag{5}$$

where ω denotes the probability that the two pixels belong to the same object. If ω is high, we trust $\vec{v}_{st3}$ more because the branch is semantically richer and more accurate, and vice versa. Therefore, the optimal low-frequency context information selected through the pixel attention mechanism is represented as follows: the output of pixel attention (PA) can be written as:

$$\vec{v}_{pa} = \omega\vec{v}_{st3} + (1-\omega)\vec{v}_{st4} \tag{6}$$

3.4 High-Frequency Branch

High-frequency information exists only at object boundaries [25], so we propose high-frequency branch (HFB) to highlight and extract the high-frequency semantic information to further predict the boundary region of the object.

HFB is a lightweight convolutional network consisting of stacked convolutions to build four different convolutional layers to capture high-frequency details. The branch takes the 1⁄2 feature map as input, passes through multiple layers of convolutional network, and fuses the low-frequency contextual information LFB details from the low-frequency information branch before stage 3 and stage 4, which is used to better predict the high-frequency detail features. Considering the input X, the output of the HFB can be formalized as:

$$\vec{v}_{hfb} = HFB(\mathbf{X}) = H_4\left(H_3\left(H_2\left(H_1(\mathbf{X})\right) + LF1\right) + LF2\right) \tag{7}$$

Here, H consists of a convolutional layer, BN layer and a ReLU.

3.5 Boundary Fusion

We designed a boundary fusion module as shown in Fig.5. The high-frequency region of the HFB is populated with pixel attention (PA) and stage2 detail features and context features. The context branch is semantically correct, but it lost too many details, especially boundary information and small objects. Due to the fact that HFB better preserves boundary information, we force the model to trust the HFB in boundary regions more and fill in other regions with contextual features from the LFB. Defining the vectors of pixel attention (PA), stage2, and the corresponding pixels of the output feature map of

the HFB as $\vec{v}_{pa}$, $\vec{v}_{st2}$, and $\vec{v}_{hfb}$, respectively, the output of the fused image is represented as:

$$\omega = Sigmoid\left(\vec{v}_{hfb}\right) \tag{8}$$

$$Output_{bfm} = f\left((1-\boldsymbol{\omega}) \otimes \vec{v}_{st2} + \vec{v}_{pa}\right) + f\left(\boldsymbol{\omega} \otimes \vec{v}_{pa} + \vec{v}_{st2}\right) \tag{9}$$

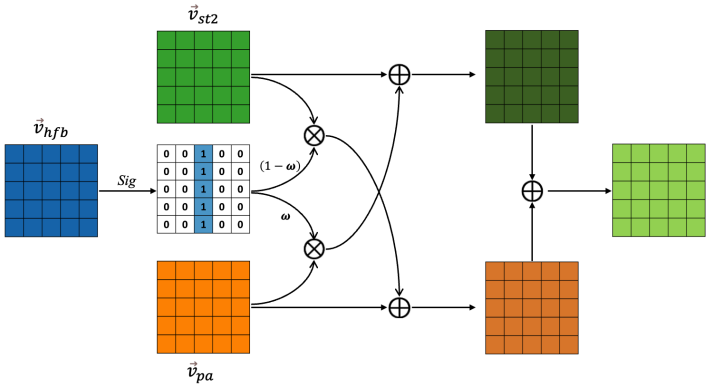

Fig. 5. Boundary fusion process

3.6 Losses Optimization

Bd-Loss denotes Weighted Binary Cross-Entropy Loss. It is employed to address the detection imbalance of HFB paths in the boundary highlighting process. In order to enhance the features of the small object boundaries, we use a coarse boundary to highlight the edge information. Seg-Loss denotes Cross-Entropy Loss, which uses the output of the segmentation head to optimize the semantic segmentation process and enhance the functionality of the BF module.

$$BdLoss = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \cdot \log\hat{y}_i + (1-y_i)\cdot\log(1-\hat{y}_i)\right] \tag{10}$$

where y_i denotes the segmentation ground-truth and $\hat{y}_i$ denotes the prediction result of the i-th pixel.

$$SegLoss = -\sum_{i,c}\{1: o_i > t\}\, y_i \log \hat{y}_i \tag{11}$$

where o_i denotes the output of the segmented head predicted to be of class c and t is the set threshold.

So, the total Loss of the model is defined as:

$$Loss = \lambda_0 BdLoss + \lambda_1 SegLoss \tag{12}$$

Based on the training experience, we set the final λ_0 = 30, λ_1 = 1, and t = 0.7.

4 Experiments

In this section, we train the BFNet and other works in Potsdam and LoveDA. And graphs are given comparing the results with other works performing on the Potsdam dataset.

4.1 Experimental Details

The experiments were trained using a single NVIDIA GTX 4090 GPU. We randomly cropped the training set images to 512 × 512 size. The data were enhanced by randomly flipping, resizing and rotating during the training process, with a batch size of 16. The three major metrics, F1, mIoU, and OA, were monitored during the training process, and if the three metrics did not grow within 20 epochs on the validation set, the training was stopped to prevent overfitting, and the maximum training epoch was set to 100.

4.2 Results

In this section, we compare BFNet's work with existing work[13,24,26,27,28] on semantic segmentation of remote sensing images, and the results are presented in tables below and we draw some conclusions.

Table 1. The results of the experiments conducted on the Potsdam dataset.

Network	Surf.	Building	Low Veg.	Tree	Car	Mean F1	OA	mIoU
BiSeNet	89.13	93.42	84.91	86.81	92.12	88.18	87.94	79.91
EANet	91.71	94.82	83.73	85.67	94.98	88.63	88.67	82.71
SwifNet	90.95	95.91	85.38	86.73	93.67	90.64	88.92	82.08
MANet	92.97	**96.88**	87.42	88.21	96.23	91.94	90.31	86.78
ShelfNet	91.93	95.48	85.89	86.89	94.09	91.31	89.87	83.68
BANet	93.34	96.51	87.18	88.92	95.87	92.50	91.03	86.88
Ours.	93.71	95.96	**88.71**	**89.88**	**96.74**	**92.85**	**91.64**	**87.22**

Table 2. The results of the experiments conducted on the LoveDA dataset. Agri. means Agriculture.

Network	Background	Building	Road	Water	Barren	Forest	Agri.	mIoU
PSPNet	43.87	51.91	53.37	76.32	8.97	43.93	57.43	47.9
DeepLabV3+	42.98	50.87	51.88	73.98	10.13	43.18	58.41	47.37
BANet	42.91	51.45	50.91	76.84	16.81	43.82	61.91	48.91
TransUnet	42.86	**55.82**	53.63	77.76	9.17	44.87	56.38	48.78
DC-Swin	40.87	54.32	**55.48**	77.92	14.42	**46.89**	62.29	50.29
Ours.	**44.13**	55.64	54.89	**78.63**	**19.12**	46.78	**62.53**	**51.37**

From the Table 1, our proposed BFNet exceeds the previous segmentation networks in many ways, achieving the highest F1 score of 92.85% on the Potsdam and mIoU is 87.22%. While on LoveDA dataset mIoU reaches 51.37% and all the fine categories also performed well .

The main difference between BFNet and BiSeNet is that we use ESwin-Transformer, and the experimental results in Table 1 show that BFNet outperforms BiSeNet in all the metrics. The training of the two models is based on the designed parameter script, which includes the image preprocessing format, training parameters and so on.

The main difference between BFNet and BANet is that we utilise boundary attention in the fusion of two-branch information, in addition to our low-frequency branches are appended at the end of the pixel attention module that extracts low-frequency context information. The experimental results indicate that BFNet surpasses BANet across all metrics.

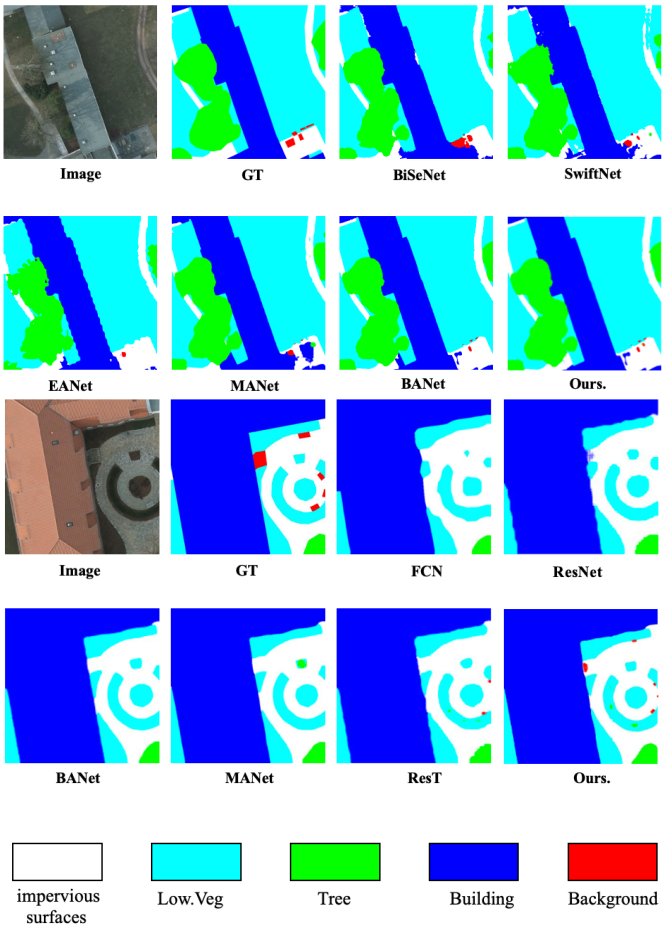

Fig. 6. The experimental results on the Potsdam dataset. GT represents Ground Truth.

To validate the model's effectiveness for image segmentation, we visualize the segmentation results of BFNet alongside other methods. As can be seen from Fig.6, existing methods tend to consider only local detail information, which leads to fragmentation of segmentation and classification confusion in some regions. BANet uses an attention mechanism and a two-branch network, and its segmentation effect has been improved.

4.3 Ablation Experiments

In this chapter, we employ a comprehensive set of ablation experiments on the dataset to substantiate the efficacy of each module. We performed the three experiments for each method and the final results are averaged results are shown in the Table3 and 4.

- LFB denotes that only low frequency branching is used to segment the image.
- $(PA)^-$ denotes that the pixel attention module is not used in the BFNet model.
- $(BA)^-$ denotes that the boundary fusion module is not used in the BFNet model.

We conduct experiments comparing the LFB model with the ResT model. ResT is the same as LFB and uses a single-branch structure. The difference between the two of them is that ResT uses the traditional Trans to extract features from the image, while LFB uses ESwin-Trans to extract features. From Table 3, it can be seen that LFB outperforms ResT in all the data of the metrics Mean F1, OA and mIoULFB while the computational power is improved. Thus, the effectiveness of our ESwin-Trans module can be verified.

Table 3. The results of the experiments on ResT and LFB.

Module	Surf.	Building	Low Veg.	Tree	Car	Mean F1	OA	mIoU
ResT	92.13	95.56	86.03	87.21	94.18	91.21	89.71	84.74
LFB	91.95	95.36	86.31	87.87	93.98	91.23	88.34	85.03
BFNet	93.58	96.01	88.38	89.18	96.73	92.58	91.58	86.73

In addition to verifying the effectiveness of the ESwin-Trans module, we also verified the efficiency of the pixel attention module (PA) and boundary fusion module (BF). The $(PA)^-$ lacks the pixel attention module compared to the BFNet model. The $(PA)^-$ does not optimize the different scales of low-frequency context information of the LFB branch, but directly sends the output of the final stage to the next module. A review of Table 4 reveals that the values of all the metrics of $(PA)^-$ are substantially lower than those of BFNet. Thus, it is clear that the pixel attention module has a crucial influence on the BFNet model.

The $(BA)^-$ lacks the boundary fusion module compared to the BFNet model. The $(BA)^-$ directly fuses the outputs of two branches without applying the boundary attention mechanism. Due to the fact that the low-frequency information and the high-frequency information are in different domains, direct summation or splicing is not the best approach. As can be observed in Table 4, the metrics of BFNet are optimal, and all the metrics of the $(BA)^-$ are lower than those of BFNet. The results demonstrate the efficacy of the boundary fusion module on the BFNet model.

Table 4. The results of the ablation study.

Module	Surf.	Building	Low Veg.	Tree	Car	Mean F1	OA	mIoU
$(PA)^-$	92.09	95.65	86.42	87.60	94.39	91.48	89.12	85.47
$(BA)^-$	92.38	95.87	86.44	88.14	94.78	91.69	89.65	85.77
BFNet	93.58	96.01	88.38	89.18	96.73	92.58	91.58	86.73

5 Conclusion

We propose a Bi-Frequency Fusion Network for high-resolution remote sensing images(BFNet). Specifically, the BFNet is a two-branch structure, where the low-frequency branch captures the low-frequency context information at different scales based on ESwin-Trans; the high-frequency branch extracts the high-frequency edge information and local context information based on stacked convolution and transverse connection. In particular, in order to better select the optimal context information, we design the pixel attention module to select the different scale context information. Meanwhile, In order to balance the features of different frequencies, we design the boundary fusion module, which applies boundary attention to the high-frequency data and low-frequency data from the unused domains for fusion, so as to realize accurate segmentation. A substantial number of experiments is conducted on both the Potsdam and LoveDA datasets have validated the efficacy of our proposed BFNet model. We hope that this paper can provide more researchers with ideas for semantic segmentation in the following years.

References

1. Zhang, Ce, et al. "Identifying and mapping individual plants in a highly diverse high-elevation ecosystem using UAV imagery and deep learning." *ISPRS Journal of Photogrammetry and Remote Sensing* 169 (2020): 280-291.
2. Zhang, Ce, et al. "Scale Sequence Joint Deep Learning (SS-JDL) for land use and land cover classification." *Remote Sensing of Environment* 237 (2020): 111593.
3. Li, Rui, et al. "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images." *IEEE Geoscience and Remote Sensing Letters* 19 (2021): 1-5.
4. Li, Rui, et al. "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images." *IEEE Geoscience and Remote Sensing Letters* 19 (2021): 1-5.
5. Wang, L., et al. "SaNet: Scale-aware neural network for semantic labelling of multiple spatial resolution aerial images." *arXiv preprint arXiv* 2103 (2021).
6. Huang, Zilong, et al. "Alignseg: Feature-aligned segmentation networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.1 (2021): 550-557.
7. Diakogiannis, Foivos I., et al. "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data." *ISPRS Journal of Photogrammetry and Remote Sensing* 162 (2020): 94-114.
8. Li, Rui, et al. "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery." *ISPRS journal of photogrammetry and remote sensing* 181 (2021): 84-98.

9. Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021
10. Liu, Rui, Li Mi, and Zhenzhong Chen. "AFNet: Adaptive fusion network for remote sensing image semantic segmentation." *IEEE Transactions on Geoscience and Remote Sensing* 59.9 (2020): 7871-7886.
11. Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
12. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer International Publishing, 2015.
13. Li, Rui, et al. "Multiattention network for semantic segmentation of fine-resolution remote sensing images." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021): 1-13.
14. Zhao, Qi, et al. "Semantic segmentation with attention mechanism for remote sensing images." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021): 1-13.
15. Zhang, Qinglong, and Yu-Bin Yang. "Rest: An efficient transformer for visual recognition." *Advances in neural information processing systems* 34 (2021): 15475-15485.
16. Wang, Libo, et al. "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 190 (2022): 196-214.
17. Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
18. Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." *arXiv preprint arXiv:1511.07122* (2015).
19. Zhao, Hengshuang, et al. "Pyramid scene parsing network." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
20. Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." *arXiv preprint arXiv:1706.05587* (2017).
21. Sun, Zhongyu, et al. "Multi-resolution transformer network for building and road segmentation of remote sensing image." *ISPRS International Journal of Geo-Information* 11.3 (2022): 165.
22. Gao, Liang, et al. "STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation." *IEEE journal of selected topics in applied earth observations and remote sensing* 14 (2021): 10990-11003.
23. Yu, Bing, Haoteng Yin, and Zhanxing Zhu. "St-unet: A spatio-temporal u-network for graph-structured time series modeling." *arXiv preprint arXiv:1903.05631* (2019).
24. Yu, Changqian, et al. "Bisenet: Bilateral segmentation network for real-time semantic segmentation." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
25. Xu, Jiacong, Zixiang Xiong, and Shankar P. Bhattacharyya. "PIDNet: A real-time semantic segmentation network inspired by PID controllers." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.
26. Wang, Haochen, et al. "Swiftnet: Real-time video object segmentation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
27. Zhuang, Juntang, et al. "Shelfnet for fast semantic segmentation." *Proceedings of the IEEE/CVF international conference on computer vision workshops*. 2019.
28. Wang, Libo, et al. "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images." *Remote Sensing* 13.16 (2021): 3065.