MBNET: A MULTI-RESOLUTION BRANCH NETWORK FOR SEMANTIC SEGMENTATION OF ULTRA-HIGH RESOLUTION IMAGES

Lianlei Shan [1], *Weiqiang Wang* [1,*]

[1] University of Chinese Academy of Sciences, CAS, Beijing, China
Computer Vision and Multimedia Technology Laboratory
shanlianlei18@mails.ucas.edu.cn, wqwang@ucas.ac.cn

ABSTRACT

Semantic segmentation of ultra-high resolution images is more challenging than ordinary images since high-resolution images need to be cropped into patches in training due to GPU memory limitation. To solve this problem, we design a multi-branch structure to deal with multi-resolution inputs, called **M**ulti-resolution **B**ranch Network (MBNet). MBNet takes patches of various instead of only one resolution as inputs, so it can make the extracted features pure and different from each other so as to cover the complex scenes with tremendous variation. Moreover, to make full use of the multi-branch structure, we design a zoom module. Zoom module abandons the previous L2-norm before feature concatenation but combines different features according to the learned attention, which fully releases the advantages of multi-resolution. Results on two benchmark datasets show that our method improves significantly over the previous state-of-the-art methods.

***Index Terms*—** Ultra-high resolution images, semantic segmentation, multi-resolution, zoom module, attention

1. INTRODUCTION

An image with more than 4 million pixels is regarded as the ultra-high resolution image [1]. With the advancement of photography and sensor technologies, more and more ultra-high resolution images can be accessible, which leads to a growing need for an efficient and effective process of ultra-high-resolution images.

The deep learning-based approaches have made great progress and achieve satisfactory results in the semantic segmentation of normal resolution images. However, there is still a wide gap in accuracy between ultra-high resolution and normal size images. The critical difference is that ultra-high resolution images can not be directly put into GPUs for training but need to be cropped into patches. Patch-based training lacks contexts and neighborhood dependency information, making it difficult to distinguish targets with similar textures or colors and thus causes errors.

To solve this problem, people proposed amounts of approaches. GLNet [2] extracts global features from downsampled images and local features from cropped patches and then enforces global and local streams to exchange with each other to integrate multi-resolution information. [3] uses a similar idea with GLNet, but trains global and local feature extracted networks respectively. GLNet and [3] achieve the state-of-art performance on ultra-high resolution image segmentation. However, when the images with original size are downsampled to the small patches, such an amount of valuable information is lost. Moreover, compared with the global information, the surrounding information around the patch is more valuable for the segmentation of this patch. It is wise to design a multi-branch network to deal with local and global information respectively, but the works mentioned above only execute rough fusion and pay equal attention to different branches without considering that the different complex parts need different resolution information. These rough operations significantly weaken the effectiveness of the multi-branch structure and thus causes the improvement far from enough. The contribution of our work is to take full advantage of the multi-branch network structure to eliminate the accuracy difference between high-resolution images and ordinary size images.

In general, previous works fuse global and local information, while we fuse local and the more valuable surrounding dependent information. Moreover, to cooperate with the multi-branch network structure, we design the zoom module. Zoom module learns to pay different attention to various resolution branches through the complexity of image parts.

To summary, the main contributions of our work are as follows,

- A multi-branch network is designed to deal with multi-resolution patches. Employing different branches to extract features of various resolutions can make the extracted features different from each other so as to cover scenes of multiple complexities.
- To fully exploit the advantages of multi-branch structure, we abandon the previous L2-norm before concate-

* Corresponding author.
This work is supported by NSFC Key Projects of International (Regional) Cooperation and Exchanges under Grant 61860206004, and NSFC projects under Grant 61976201.

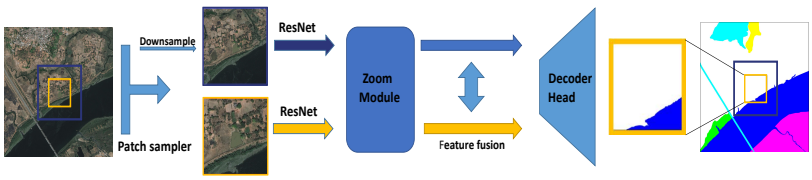

Fig. 1. Overview of the whole network architecture. The entire network consists of a patch sampler, ResNet [4], zoom module, feature fusion module, and decoder head. For brevity, the figure only shows two branches, and there could be more than two. The input of the ResNet represented by the yellow arrow is the yellow patch, and the blue arrow is the blue patch.

nation but combine different features according to the learned attention, which vastly boosts the performance.

2. RELATED WORK

2.1. Aggregation of Multi-resolution Features

Integrating multi-level features can capture patterns of different granularity, and thus it is a key element to achieving state-of-the-art performance on segmentation. RefineNet [5], Feature pyramid Network (FPN) [6], and High-Resolution Network (HRNet) [7] are typical representatives of this idea. In HRNet, a variety of the multi-resolution aggregated strategies, including from high resolution to low, from low to high, and the same level, have been proposed and achieve enormous success. We carry forward the idea to multi-resolution features and multi-resolution inputs to deal with the segmentation of ultra-high-resolution images.

2.2. Attention Mechanism

After non-local [8], the attention mechanism in segmentation has been developed rapidly, including channel attention and spatial attention. SENet [9] and SKNet [10] propose and develop the channel attention. DANet [11] puts the spatial attention and the channel attention together into one network and achieves good results in panoptic segmentation. However, all of the above works focus on features from one patch or one image with the same fixed resolution. For our multi-branch network structure with multi-resolution, only this attention level is not enough. Therefore, under the precondition of retaining the original attention, we combine the features from different patches and learn to pay attention to different resolutions.

3. METHOD

3.1. Overview of Architecture

The overall network structure is shown in Fig. 1, which consists of five parts, patch sampler, backbone, zoom module, feature fusion, and decoder head. Patch sampler obtains sets of patches, and patches of one set have different sizes. These patches go through ResNet [4], zoom module, and feature fusion in turn. Zoom module is cooperated with the following fusion module to combine various resolution features, and the fusion is the most used concatenation operation. Finally, the output features are sent to the decoder head to get the prediction. The decoder head contains several convolutions to keep the channel numbers of final features consistent with the class number.

Multi-resolution patch sampler is introduced in 3.2, and zoom module in 3.3.

3.2. Multi-resolution Patch Sampling

GLNet fuses local patch information and global information, which not only increases the receptive field but also introduces a lot of irrelevant information. Intuitively, objects that are farther apart are more likely to be independent. Thus, the information around the patch is the most meaningful for pixel segmentation within this patch. On the contrary, features that are too far away are apt to be interfering noise. Patch sampling is used to obtain information about this patch and its surroundings.

Patch **S**ampler is $PS(\mathbf{I}, i) = \mathbf{P}_{\mathbf{I}}^{(1)}(i), \ldots, \mathbf{P}_{\mathbf{I}}^{(l)}(i)$, which extracts l different patches $\mathbf{P}$ of varying resolutions from the original ultra-high resolution images $\mathbf{I}$ centered at pixel i. We set l to 2 in this article if without other special statements. $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$. Pixel i is selected randomly if its largest patch does not exceed the image boundary. All the cropped patches are downsampled to the same size as the smallest patch so that all patches have the same number of pixels but different resolutions. As shown in Fig. 1, the yellow box is the cropped patch, and the blue box contains surrounding information.

3.3. Zoom Module

The direct fusion of features with different resolutions will make the network challenging to train and perform poorly [2, 12, 13]. ParseNet [13] and GLNet carry out L2-norm for each feature map before fusion, while the proposed zoom module is a more flexible and effective processing mode. The zoom module concatenates two resolution inputs and then learns to give different weights to the different resolutions. The operation makes it possible to focus on high resolution for complex parts and low resolution for parts requiring large receptive fields, which behavior is similar to the zoom function of eyes, so we call it zoom module.

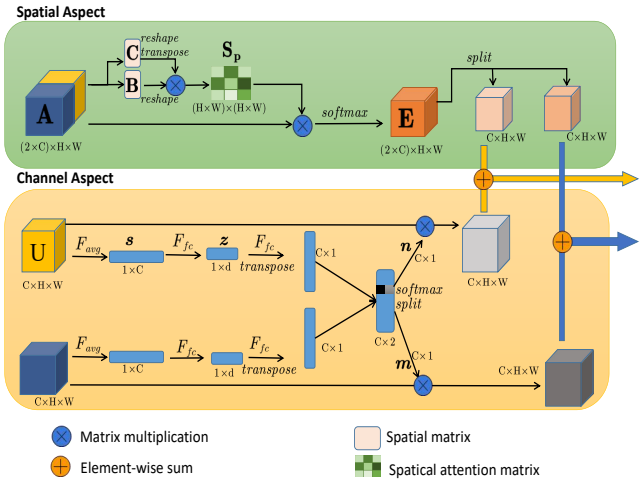

Fig. 2. Overview of the zoom module. Blue and yellow features are from different branches. The same color represents the same features, and the different represent different features. The color is consistent with Fig. 1.

The overview of the zoom process is shown in Fig. 2. The input features come from the last layer of ResNet, and blue and yellow in the figure mean features from two different resolution branches. Zoom module is divided into two parts: spatial aspect and channel aspect, and we introduce respectively in the following.

Spatial aspect. The spatial aspect is shown in the green part of Fig. 2. Before computing, features go through one convolution to change the number of channels to $\frac{1}{8}$ of the original. Then, we combine the features of various branches together to calculate the attention.

Firstly, the features from different resolution branches are concatenated in the channel dimension to get $\mathbf{A}$. Given the combined feature $\mathbf{A} \in \mathbb{R}^{(2\times C)\times H\times W}$, we then feed it into a convolution layers to generate two new feature maps $\mathbf{B}$ and $\mathbf{C}$, respectively, where $\{\mathbf{B}, \mathbf{C}\} \in \mathbb{R}^{(2\times C)\times H\times W}$. Further, we reshape them to $\mathbb{R}^{(2\times C)\times M}$, where $M = H \times W$ is the number of pixels. After that, we perform a matrix multiplication between the transpose of $\mathbf{C}$ and $\mathbf{B}$, and apply a softmax layer to calculate the spatial attention map $\mathbf{S}_p \in \mathbb{R}^{M\times M}$:

$$s_p^{ji} = \frac{\exp\left(B_i \cdot C_j\right)}{\sum_{i=1}^{M} \exp\left(B_i \cdot C_j\right)} \tag{1}$$

where s^{ji} measures the i^{th} position's impact on j^{th} position. The more similar feature representations of the two positions contribute to a more significant correlation between them. We reshape $\mathbf{A}$ to $\mathbb{R}^{(2\times C)\times M}$. Then we perform a matrix multiplication between $\mathbf{A}$ and the transpose of $\mathbf{S}_p$ and reshape the result to $\mathbb{R}^{(2\times C)\times H\times W}$. Finally, we obtain the final output $\mathbf{E} \in \mathbb{R}^{C\times H\times W}$ as follows:

$$\mathbf{E} = \mathbf{S}_p\mathbf{A} \tag{2}$$

Finally, $\mathbf{E}$ is split in the channel dimension. Besides, totally different from SENet and DANet, the function of the zoom module is to better integrate features of different resolutions rather than enhance features. Therefore, the attention mask and feature map are multiplied as output directly, and the add operation is discarded, which also reduces the memory occupation.

Channel aspect. Channel aspect is shown in the orange part of the figure, and are calculated by global average pooling, full connected (fc), and softmax in turn, introduced as follows,

$$s_c = \mathcal{F}_{avg}\left(\mathbf{U}\right) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{U}(i,j), \tag{3}$$

$$\mathbf{z} = \mathcal{F}_{fc}(\mathbf{s}) = \delta(\mathcal{B}(\mathbf{s}\mathbf{W})), \tag{4}$$

$$m_c = \frac{e^{\mathbf{z}\mathbf{M}_c}}{e^{\mathbf{z}\mathbf{M}_c} + e^{\mathbf{z}\mathbf{N}_c}}, n_c = \frac{e^{\mathbf{z}\mathbf{N}_c}}{e^{\mathbf{z}\mathbf{M}_c} + e^{\mathbf{z}\mathbf{N}_c}}, \tag{5}$$

where δ is the ReLU function [14], $\mathcal{B}$ denotes the batch normalization [15]. $\mathbf{W} \in \mathbb{R}^{C\times d}$ and $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{d\times C}$, which represent the matrix parameters of the first and second full connected operation. $\mathbf{M}$ and $\mathbf{N}$ come from different branches, and $\mathbf{M}_c \in \mathbb{R}^{1\times d}$ is the c-th row of $\mathbf{M}$. d=32 is a typical setting in our experiments. Eq. 3 and 4 need to be calculated in each branch, and we only represent them in one branch for brevity. The whole procedure is in line with yellow part in Fig. 2. It can be observed that the channel aspect compares the features from different resolutions and then gives greater weight to more discriminating features through the softmax operation. Compared with the previous fixed norm method (like L2-norm), ours is data-driven and thus more flexible and effective.

4. EXPERIMENTS

4.1. Implement Details

Our work is based on MMSegmentation codebase [16], and the super parameters are all consistent with the original document. Two scales of patch sizes are 500×500 and 700×700 respectively, and the latter becomes 500×500 after downsampling. All experiments are performed on a workstation with 2 NVIDIA GTX 2080 Ti 11G GPU cards. The comparative experiment of GLNet has the same parameters as its original paper. The training set, validation set, and test set in Deepglobe are divided in line with GLNet. We train all models with the Adam [17] optimizer with the learning rate set to 4e-4, and we decay the learning rate with cosine [18] annealing to the minimum value of 1e-6 in the last epoch (we do not perform any warm restarts). We use the command-line tool 'gpustat' to measure the GPU memory usage of a model with the minibatch size of 1, and avoid calculating any gradients. All experimental results are obtained on single-scale tests.

Dataset: Deepglobe [19] is a high-quality satellite dataset focusing on rural areas, which provides 803 images in 7 classes with 2448 × 2448 pixels. ISIC [20] is an ultra-resolution medical dataset for pigmented skin lesions. The

average resolution of ISIC is up to 9 M. The largest image is up to the size of 6748×4499.

4.2. Experimental Analysis

We do sufficient experiments and show comprehensive performance of models from accuracy, memory, speed, and stability.

Table 1. Performance on Deepglobe.

Model	mIOU (%)	Memory(MB)	Time (s)
PSPNet [21]	53.3	6289	127.96
ICNet [22]	40.2	2557	23.79
DeepLabv3 [23]	63.6	3199	84.55
BiSeNet [24]	47.0	**1801**	**10.19**
DANet [11]	**63.4**	6914	60.94
GLNet [2]	71.6	1865	235.39
MBNet (ours)	**72.6**	**1549**	**162.86**

Table 2. Performance on ISIC.

Model	mIOU (%)	Memory(MB)	Time (s)
PSPNet [21]	**70.1**	3679	127.42
ICNet [22]	35.8	**1617**	23.87
DeepLabv3 [23]	66.7	2033	85.81
BiSeNet [24]	50.7	1975	**13.74**
DANet [11]	64.4	3888	67.88
GLNet [2]	75.2	1912	638.85
MBNet (ours)	**76.0**	**1604**	**436.54**

Accuracy, memory and speed. We present our experimental results in terms of accuracy, inference memory, and inference time, as shown in Table 1 and 2. The upper part in the table is classical segmentation networks, and the lower part is segmentation networks specifically for high-resolution images. Compared with the classical segmentation network, we have an incomparable advantage in accuracy. And compared to the previous SOTA method GLNet, our network also has a significant improvement in accuracy without GPU memory and inference time increase. Our method gets a 1% improvement in Deepglobe and 0.8% in ISIC and keeps the superiority in memory and speed.

Table 3. Results with different patch sizes.

Method	patch size			
	500	400	256	128
GLNet [2]	71.6	70.6	66.4	64.6
MBNet (ours)	72.6	72.2	71.1	69.1

Stability to patch size. Table 3 shows the relationship between segmentation accuracy and patch size. It can be seen that the segmentation accuracy of our network is less affected by the change of patch size. This stability enables the network to be trained with smaller patches, thus reducing the requirement for hardware.

4.3. Ablation Study

In this section, we introduce the respective functions of spatial and channel aspect in zoom module, as well as their comparison with SENet and DANet. Then we present the results with different numbers of branches.

Table 4. Ablation study for zoom module:performance on Deepglobe.

Model	Result (mIOU/%)
L2-norm	66.8
only spatial aspect	68.1
only channel aspect	69.3
dual aspects calculated separately	69.9
channel attention in SENet [9]	70.1
zoom module without spatial aspect	71.7
dual attention in DANet [11]	71.8
zoom module	72.6

The Effect of zoom module. The results are shown in Table 4. L2-norm can be considered as the baseline for comparison. The addition of channel aspect improves by 2.5%, and spatial aspect improves 1.3%. The most critical factor is to combine the features together for calculation. The separate calculation is only 69.9%, but the combined calculation (zoom module in table) reaches 72.6%, which really gives full play to the advantages of multi-branch structure. The result of SENet and DANet is better than our any single aspect, but the two aspects cooperate with each other to get the best final result. Compared with DANet, we have a 0.8% improvement. In a nutshell, the remarkable improvements prove the effectiveness of the zoom module.

Table 5. Results with different numbers of branch.

Branch number	mIOU (%)	Memory(MB)	Time (s)
2	72.6	1549	162.86
3	71.0	2046	193.66
5	66.5	3135	202.45

The result with different number of branches. The results are shown in Table 5, branch of 3 adds 600×600, and 5 adds 800×800 and 1000×1000 on the basis. According to the results, more branch number is not always the better, and two branches are the best, which indicates that feature fusion will introduce a large amount of irrelevant noise. Our method works much better than GLNet because we remove most of the extraneous information by replacing the global feature of the image with the surrounding feature. Similarly, the zoom module adaptively selects the most reasonable features to remove interference accordingly.

5. CONCLUSION

In this paper, we design a multi-branch network to process multi-resolution patches for the segmentation of ultra-high-resolution images. Meanwhile, the zoom module is designed to take full advantage of the multi-branch network structure. Our method can be widely used in a variety of scenarios using high-resolution images.

6. REFERENCES

[1] Paul Lilly, "Samsung launches insanely wide 32:9 aspect ratio monitor with hdr and freesync 2," 2017.

[2] Wuyang Chen et al., "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8924–8933.

[3] Lei Ding et al., "Semantic segmentation of large-size vhr remote sensing images using a two-stage multiscale training architecture," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[4] Kaiming He et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] Guosheng Lin et al., "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[6] Tsung-Yi Lin et al., "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[7] Ke Sun et al., "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[8] Xiaolong Wang et al., "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[9] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[10] Xiang Li et al., "Selective kernel networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[11] Jun Fu et al., "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[12] Lianlei Shan, Minglong Li, Xiaobin Li, Yang Bai, Ke Lv, Bin Luo, Si-Bao Chen, and Weiqiang Wang, "Uhrsnet: A semantic segmentation network specifically for ultra-high-resolution images," in *2020 25th International Conference on Pattern Recognition*. IEEE, 2021, pp. 1460–1466.

[13] Wei Liu, Andrew Rabinovich, and Alexander C Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[14] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[15] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[16] Kai Chen et al., "MMSegmentation: Open mmlab segmentation toolbox and benchmark," 2019.

[17] Diederik P Kingma et al., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] Ilya Loshchilov et al., "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[19] Ilke Demir et al., "Deepglobe 2018: A challenge to parse the earth through satellite images," in *2018 CVPRW*. IEEE, 2018, pp. 172–17209.

[20] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[21] Hengshuang Zhao et al., "Pyramid scene parsing network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[22] Hengshuang Zhao et al., "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 405–420.

[23] Liang-Chieh Chen et al., "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[24] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, pp. 1–18, 2021.