REAL-TIME SEMANTIC SEGMENTATION OF VHR REMOTE SENSING IMAGES BASED ON BILATERAL SEGMENTATION NETWORK

Yu Lei , Youguang Lin , Junli Yang, Yifan Sheng*

Beijing University of Posts and Telecommunications
yangjunli@bupt.edu.cn

ABSTRACT

Deep Convolutional Neural Networks have been successfully used in semantic segmentation of remote sensing images(RSI). However, most methods compromise either spatial resolution or receptive field size, which are especially important for RSI with large image size and abundant spatial details, to achieve real-time prediction speed. In this paper, we address the problem of real-time semantic segmentation of very high resolution(VHR) remote sensing images with a bilateral segmentation network named BiSeNet. BiSeNet is a two-sided structure composed of spatial path (SP) and context path (CP) to solve the problem of loss of spatial information and shrinkage of receptive field. It is very efficient to store spatial information with small steps to generate high-resolution feature maps and to use context paths with rapid downsampling strategies to obtain sufficient receptive field. Experimental results on the public available Postdam dataset show that BiSeNet achieves 88.1% pixelAcc and 80.78% mIoU with speed of 97.1 FPS on NVIDIA GTX 1080 for a 512*512 input, which is competitive on balance of accuracy and speed compared to other state-of-the-art methods.

Index Terms— Real-time semantic segmentation, very high resolution remote sensing images, deep convolutional neural networks, spatial information

1. INTRODUCTION

Semantic segmentation is one of the basic tasks in computer vision which aims at assigning a class label to each pixel in the image. As a core problem of computer vision, the importance of scene understanding is becoming more and more prominent in some applications, because these application scenarios in reality need to deduce relevant knowledge or semantics from images, such as automatic driving, medical imaging, image search engine, augmented reality and so on.

Convolutional neural network (CNN) is a neural network specially designed for image recognition. It imitates the multi-layer process of human image recognition by filtering out the contours between adjacent pixels. On the basis of CNN, the full convolutional neural network replaces the full connection layer with the convolutional layer to output the spatial domain mapping (deconvolution) instead of the probability of output category, thus transforming the image segmentation problem into an end-to-end image processing problem.

Semantic segmentation needs not only rich spatial information but also a large receptive field. However, modern methods often sacrifice spatial resolution for real time reasoning speed, resulting in poor performance. BiSeNet proposed a new two-sided structure composed of spatial path (SP) and context path (CP) to cope with the loss of spatial information and the shrinkage of receptive field. BiSeNet stores spatial information and generates high-resolution features through a spatial path with small steps. At the same time, the context path of the fast downsampling strategy is used to obtain sufficient acceptance fields. On this basis, a new feature fusion module is proposed to realize effective feature fusion.

This paper is the first attempt to introduce BiSeNet to address the semantic segmentation problem of high-resolution remote sensing images. We evaluate our method on Postdam dataset and experimental results show that BiSeNet outperforms other state-of-the-art methods and achieves a competitive real-time processing speed on the basis of high accuracy.

2. METHODOLOGY

2.1. Overall Network Architecture

BiSeNet is a bidirectional model for real-time semantic segmentation which can achieve real-time performance and high accuracy at the same time. As illustrated in Fig. 1(a), it uses the pre-trained ResNet model as the backbone of the Context Path and three convolution layers with stride as the Spatial Path to address the problem that the speed reduction and spatial information loss caused by the complete U-shape structure. Finally, it fuses the output features of these two paths to make the final prediction.

*Corresponding author.

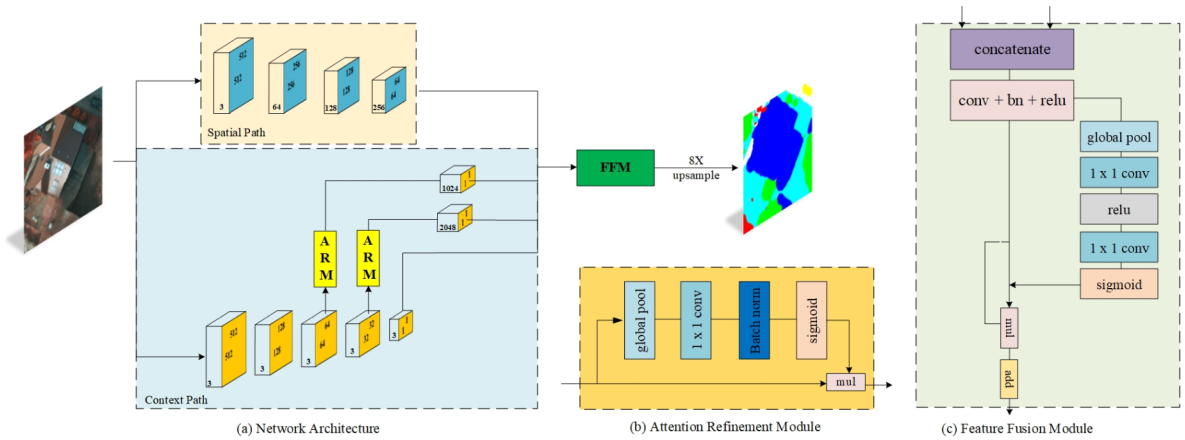

Fig. 1: An overview of the Bilateral Segmentation Network. (a) Network Architecture. The blue layers in spatial path: convolutional blocks with successive Conv, BN, ReLU operations. (b) Components of the Attention Renement Module (ARM). (c) Components of the Feature Fusion Module (FFM).

2.2. Method

2.2.1. Attention Refinement Module

Global Average Pooling (GAP) has proven to be more robust on spatial transformation to pool than FC since GAP uses feature map as a unit instead of window to perform averaging, which reduces the amounts of calculated parameters while summing up spatial information.

Attention Refinement Module is a specific module employs GAP to capture global context and compute a vector of importance weights to guide the feature learning. The advantage of this design is that you can more clearly distinguish the output characteristics of each stage in the context path while easily obtaining global information without any upsampling. Hence, its calculation cost is negligible. (see Fig. 1(b))

2.2.2. Feature Fusion Module

Feature Fusion Module is a feature integration method proposed to address the problem that the features of the two paths can't be simply sum up due to the different level feature representation. Given the different level of the features, it first concatenates the output features of Spatial Path and Context Path. Then, it uses batch normalization [1] to balance the scale of the features. Finally, it sets the connected features into a feature vector and calculates its weight vector, like SENet [2]. The role of this weight vector is to select and combine features. Fig. 1(c) shows the details of this design.

3. EXPERIMENTS

In this section, we evaluate our method on Potsdam Dataset [3] and present the results obtained by BiSeNet compared with FastFCN and other state-of-the-art models.

3.1. Dataset description

Potsdam is a famous benchmark dataset for 2D semantic labelling which is provided by Commission II of ISPRS [3]. It consists of 38 high-resolution of 6000 × 6000 pixels areial images, whose ground sample distance is 5cm, and each aerial image is captured in five channels(nDSM, NIR, R, G, and blue(B)). In addition, all the pixels are categorized into six semantic classes, including the Impervious Surfaces, Building, Low Vegetation, Tree, Car and Clutter.

3.2. Implementation Details

3.2.1. Data usage

In our work, we only use three channels(IR, R, G), where 24 images are used for training, and the rest 14 images are for testing. We split the each 6000 × 6000 image into 144 tiles of 512 × 512 with overlap, padding pixels that overlap with the previous picture if needed. For all of the methods evaluated, the usage of data is identical.

3.2.2. Evaluation metrics

For evaluating the performance of BiSeNet, we choose the overall accuracy(OA) [9], mean pixel intersection-over-union(mIoU) as our metrics, which are the most frequently used evaluation metrics for semantic segmentation. Overall accuracy is a metric that considers the global aspects of the classfication, it takes into account all correctly classified pixels indistinctly. On the other hand, average accuracy reports the average (per-class) ratio of correctly classified samples, it outputs an average of the accuracy of each class. Intersection over union is a metric that calculates the intersection ratio of two sets, which are ground truth and predicted segmentation.

3.2.3. *Processing*

We train and evaluate BiSeNet on two NVIDIA GTX 1080 GPU with 8 GB memory. To optimize the dice loss function, we use the Adadelta with rho of 0.9 and decay of 10^{-4}. The initial learning rate is 0.025 and decays as the Poly scheduler with power of 0.9. To avoid overfitting, we use the method of bilinear interpolation to randomly select the scale to change the size of the image and the horizontal flip.

3.3. Results and Analysis

We plot the miou curve, oa curve and loss curve of BiSeNet with different encoder.

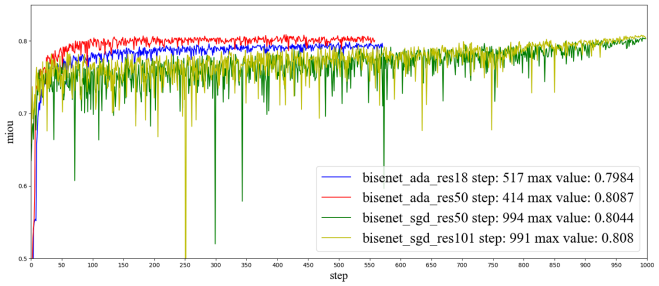

Fig. 2: miou curve

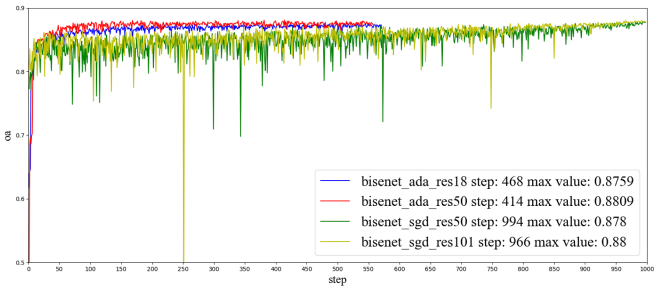

Fig. 3: oa curve

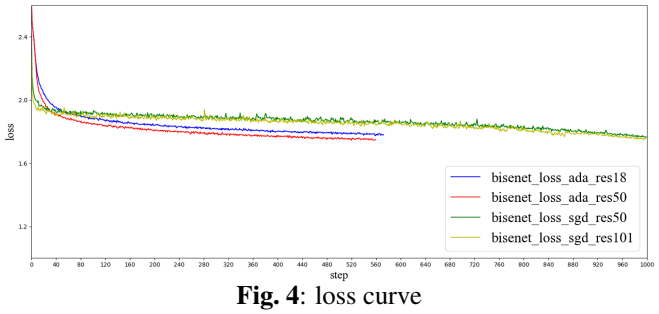

Fig. 4: loss curve

Normalized confusion matrix generated by results of BiSeNet is shown in Fig. 5. From the diagonal line, we can find that BiSeNet can classify pixels of all classes with a relatively high precision. Particularly, building class is the main source of well-classfication, which is resonable since this class actually contains large scale objects and BiSeNet performs better than other models at this field such as the close shot of cars and houses. As for the Clutter/Background, we hold that because of the existence of multiple diverse classes such as construction site, fence, closure and open field not covered by vegetation, BiSeNet gets mis-classification at this class.

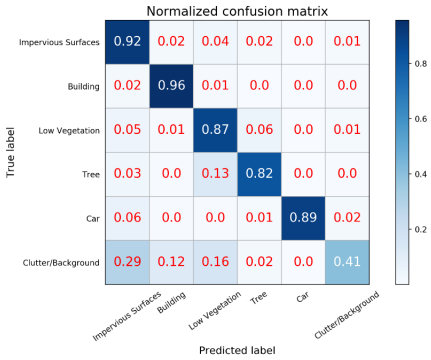

Fig. 5: Visualization of classification performance of BiSeNet

Table 1 presents the quantitative results obtained, using the compared approaches. It can observed that, BiSeNet outperforms other models which achieves a relatively high accuracy and fast test frame per second. From the Table2 we can find that BiSeNet with different backbones or optimizer gets a different performance. Compared to resnet50, though resnet101 improves accuracy (0.2% higher OA and 0.36% mIoU), it lost nearly half of its FPS. And from different optimizer, as is clearly shown in Fig. 3, adadelta converges faster than sgd and gets a higher OA than sgd with same encoder. Therefore, it is worth noting that we chose an intermediate result (resnet50+adadelta) where BiSeNet has both a high precision and a fast frame-per-second.

Table 1: Quantitative Performance Evaluation on the ISPRS Potsdam dataset.

Model	Encoder	OA%	mIoU%
FCN[4]	VGG-16	82.75	61.71
SegNet[5]	VGG-16	82.93	63.42
SHG[6]	Hourglass-104	85.38	67.26
EncNet[7]	ResNet-101	86.52	69.45
CxtHGNet[8]	Hourglass-104	87.15	70.28
BiSeNet	ResNet-50	88.1	80.87

Table 2: Performance Evaluation of different encoder.

Encoder	Optimizer	OA%	mIoU%	F1	Fps
ResNet-50	sgd	87.8	80.44	88.56	97.1
ResNet-101	sgd	88	80.8	89	56
ResNet-18	adadelta	87.6	79.84	87.56	183.3
ResNet-50	adadelta	88.1	80.87	89.3	97.1

Visualization of results generated by BiSeNet, with original image, ground truth and difference image of each, are shown in Fig. 6.

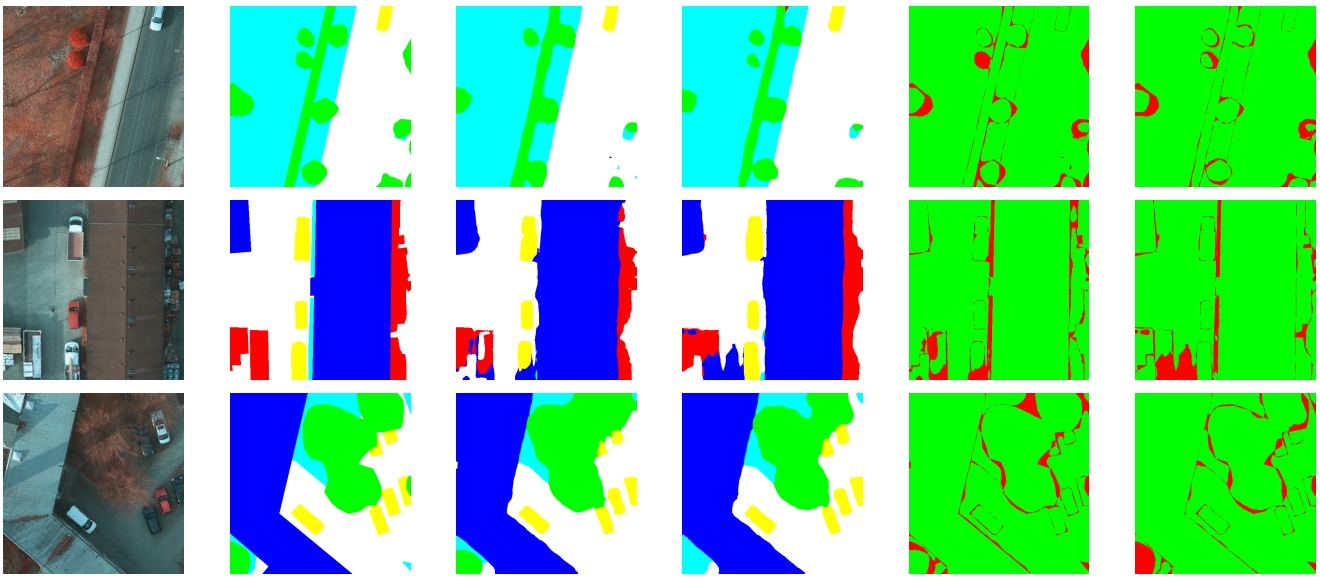

Fig. 6: Examples of segmentation results on the Potsdam dataset. Where (a) and (b) is difference image of resnet18 and resnet50. Red color stands for the misclassified category while green color represents the category that are properly classfied.

4. CONCLUSION

In this paper, we apply a method named BiSeNet that composed of spatial path (SP) and context path (CP), to address the problem of poor performance under compromise spatial resolution to achieve real-time inference speed. In context path, Attention Refinement Modul is adopted to obtain sizeable receptive field rapidly and pre-trained classic network to extract high level features while the spatial path encodes affluent spatial information. Experimental results on the ISRPS Postdam dataset show that BiSeNet achieves the best performance compared with other state-of-the-art methods in terms of overall accuracy and average intersection ratio, and is worthwhile to mention that with BiSeNet it can achieve a balance between accuracy and efficiency, which makes it suitable for time-sensitive applications.

Acknowledgements

This work is funded by College Student Innovation Project of Google, Student Innovation Training Programme(No.201911025) and Teaching Reform Project(No.2019JYA05) of Beijing University of Posts and Telecommunications .

5. REFERENCES

[1] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv (2017) 4, 7

[2] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456 (2015) 5, 7, 9, 10

[3] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf,"The isprs benchmark on urban object classification and 3D building reconstruction," ISPRS Ann. Photogramm, Remote Sens. Spat. Inf.Sci, vol. 1, no. 3, pp. 293-298

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in CVPR, 2015.

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "SegNet: A deep convolutional encoderdecoder architecture for image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481–2495, Dec 2017.

[6] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in ECCV, 2016.

[7] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal, "Context encoding for semantic segmentation," in CVPR, 2018.

[8] Panfeng Li, Youzuo Lin, and Emily Schultz-Fellenz, "Contextual hourglass network for semantic segmentation high resolution of aerial imagery," in arXiv preprint arXiv:1810.12813, 2018.

[9] X. Liang and Z. Fu, "MHNet: Multiscale Hierarchical Network for 3D Point Cloud Semantic Segmentation," in IEEE Access, vol. 7, pp. 173999-174012, 2019.