REAL-TIME SEMANTIC SEGMENTATION OF AERIAL VIDEOS BASED ON BILATERAL SEGMENTATION NETWORK

Yihao Zuo, Junli Yang, Zihao Zhu**, Ruizhe Li**, Yuhan Zhou, Yutong Zheng*

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

IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554952

ABSTRACT

In recent years, deep learning algorithms have been widely used in semantic segmentation of aerial images. However, most of the current research in this field focus on images but not videos. In this paper, we address the problem of real-time aerial video semantic segmentation with BiSeNet[1]. Since BiSeNet is originally proposed for semantic segmentation of natural city scene images, we need a corresponding dataset to ensure the effect of transfer learning when applying it to aerial video segmentation. Therefore, we build a UAV streetscape sequence dataset (USSD) to fill the vacancy of dataset in this field and facilitate our research. Evaluation on USSD shows that BiSeNet outperforms other state-of-the-art methods. It achieves 79.26% mIoU and 93.37% OA with speed of 148.7 FPS on NVIDIA Tesla V100 for a 1920×1080 frame size input aerial video, which satisfies the demand of aerial video semantic segmentation with a competitive balance of accuracy and speed. The aerial video semantic segmentation results are provided at Our Repository.

Index Terms— Real-time Aerial Video Semantic Segmentation, High-resolution Aerial Imagery, Deep Learning

1. INTRODUCTION

Semantic segmentation is one of the core tasks in computer vision which aims at assigning a class label to each pixel in an image (or a frame of a video) to achieve pixel-level understanding. In contrast to other tasks such as object detection, semantic segmentation of remote sensing images is able to provide much detailed semantic information, which means a broader application prospect including disaster damage assessment, urban planning and geographic mapping. And real-time aerial video semantic segmentation can bring more possibilities. However, this task is extremely challenging because: (1) Shooting angle is always vertical, which is very different from common videos; (2) More complicated texture, larger variation in object size, etc.; (3) Relevant datasets are scarce.

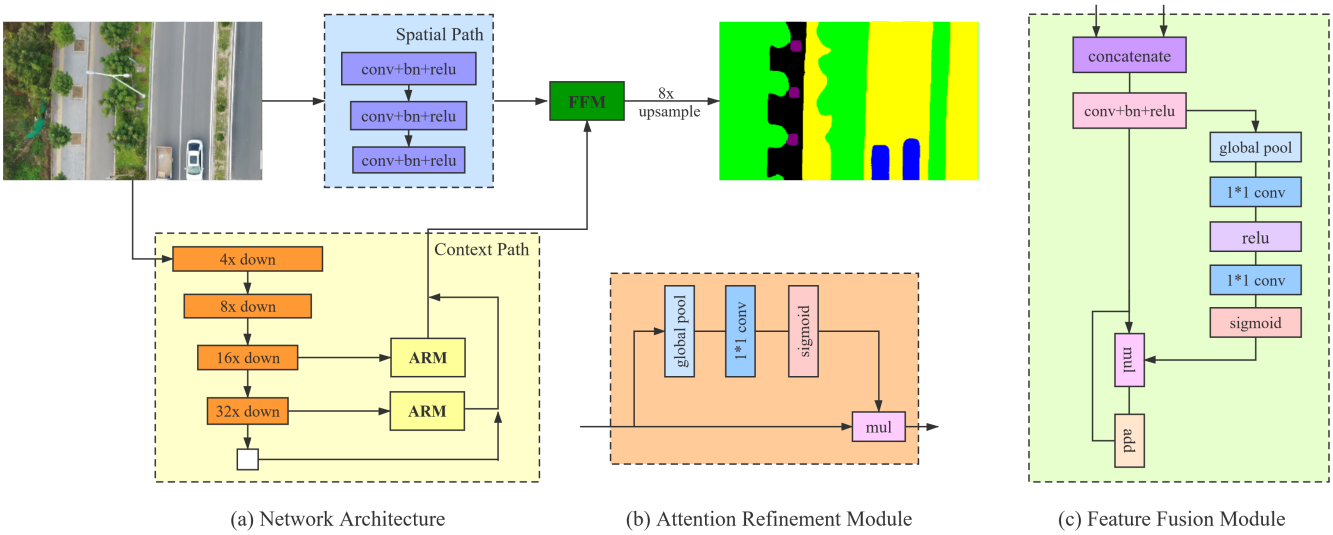
Convolutional Neural Networks (CNN) is a representative model of deep learning that has been extensively used in various networks for computer vision tasks. Fully Convolutional Networks (FCN)[2] is the first end-to-end semantic segmentation network based on CNN, which can accept inputs of any size and produce classification outputs of same size. Many subsequent outstanding semantic segmentation networks are based on the basic thought of FCN, such as SegNet[3], U-Net[4], etc.

Although these networks have achieved satisfactory performance on various benchmarks, the research on real-time semantic segmentation especially aerial videos segmentation is relatively insufficient. Networks like FCN are too computationally expensive to achieve real-time video inference. One solution is simply to take videos as a sequence of independent frames and input them to lightweight semantic networks. Some networks have attempted to reduce computation complexity and the number of parameters while ensuring satisfying accuracy. For example, LEDNet[5]adopts an asymmetric encoder-decoder structure and applies channel split and channel shuffle to reduce computation load. Other real-time semantic segmentation networks include DFANet[6], ESPNet v2[7], etc. BiSeNet [1] proposes a new bilateral structure composed of a spatial path (SP) and a context path (CP). SP stores spatial information and generates high-resolution features to cope with the loss of spatial information, while CP is used to prevent receptive fields shrinkage. A new feature fusion module (FFM) is designed to realize effective feature fusion of the SP and CP outputs.

This paper makes an attempt to introduce BiSeNet to carry out real-time aerial video semantic segmentation. We firstly evaluate it on high-resolution Potsdam dataset [8], and the performance demonstrates its applicability to remote sensing images. Then we conduct comparative experiments on USSD dataset built by ourselves. Experimental results show that BiSeNet outperforms other real-time semantic segmentation networks such as LEDNet [5]and ESPNet v2 [7]with a with higher accuracy and processing speed, and can output satisfying aerial video semantic segmentation results.

*Corresponding Author
** Equal Contribution

Authorized licensed use limited to: The Hong Kong University of Science and Technology (Guangzhou). Downloaded on February 19,2026 at 04:12:13 UTC from IEEE Xplore. Restrictions apply.



(a) Network Architecture (b) Attention Refinement Module (c) Feature Fusion Module

Fig. 1: An overview of BiSeNet. (a) Network Architecture. (b) ARM (c) FFM

2. APPROACH

In this section, we will briefly introduce the network we adopt: BiSeNet[1], and our work based on BiSeNet.

2.1. Overall Network Architecture

Semantic segmentation requires both rich spatial information and sizeable receptive field to generate an accurate prediction result. However, it is hard to meet the two requirements simultaneously, especially for real-time segmentation. The increase of one often leads to the loss of the other. As illustrated in Fig.1(a), BiSeNet proposes a bilateral structure consisting of a spatial path and a context path. These two paths are responsible for spatial information preservation and receptive field offering respectively. Moreover, Feature Fusion Module (FFM, Fig.1(c)) and Attention Refinement Module (ARM, Fig.1(b)) are designed to further improve the accuracy with acceptable cost. In this way, BiSeNet makes an appropriate balance between accuracy and speed.

2.2. Spatial Path and Context Path

Spatial Path (SP) is used to preserve the spatial size of the original input images and encode affluent spatial information. It contains only three identical convolution layers, which is composed of convolution, batch normalization and ReLU. Relatively large size of feature maps means more spatial information is preserved, leading to more accurate prediction especially around boundaries. And a small amount of convolution layers implies the calculation amount is acceptable.

Context Path (CP) is designed to provide sufficient receptive field with global context information. It contains a lightweight base model such as Xception and a global average pooling (GAP) layer used to capture global context information. This path utilizes the lightweight model to downsample the feature maps rapidly, which encodes high level semantic context information at a low cost.

An Attention Refinement Module (ARM) was proposed and used in CP as shown in Fig.1(b). Global pooling is used again to capture context information.The output vector can refine the features of each stage in CP by guiding the feature learning.

2.3. Feature Fusion Module

Since the features obtained by SP and CP are different in level of feature representation, they cannot be summed up directly. Therefore, a Feature Fusion Module (FFM) was proposed to integrate two levels of features. As shown in Fig.1(c), FFM concatenates the output features from SP and CP. After a series of calculations, a weight vector is obtained. This vector can re-weight the features, which is equivalent to feature selection and fusion. Then, the prediction result is obtained after upsampling.

2.4. Our Work

In this paper, in order to carry out real-time aerial video semantic segmentation task, we build a streetscape sequence dataset USSD and conduct experiments on it with several lightweight segmentation networks. After comparing and analyzing the experimental results, we finally adopt BiSeNet [1] with ResNet18 as the base model in CP. The performance metrics and visual effects indicate that BiSeNet satisfies the demand of aerial video semantic segmentation with an outstanding balance of accuracy and speed. The dataset and experiments will be demonstrated in the following section.

Authorized licensed use limited to: The Hong Kong University of Science and Technology (Guangzhou). Downloaded on February 19,2026 at 04:12:13 UTC from IEEE Xplore. Restrictions apply.

3. EXPERIMENTS

3.1. Dataset

In the experiments of this paper, we firstly evaluate BiSeNet on Potsdam[8]. Potsdam is a dataset consisting of 38 6000×6000 high resolution aerial urban images, published by International Society for Photogrammetry and Remote Sensing (ISPRS). USSD built by ourselves is a high-resolution aerial image dataset extracting sequentially from videos. These videos were captured by UAV in Changping District, Beijing, China. USSD consists of 614 images of 3840×2160 with corresponding fine labeled ground truth. Pixels in each image are labeled into 6 classes: background, road, vehicle, vegetation, ground and building. Besides, it contains a few videos used to evaluate visual effect of real-time segmentation. An example image-label pair is shown in Fig.2.

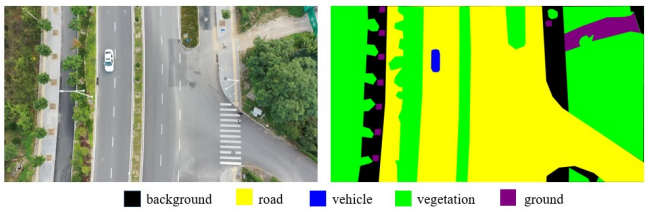


Fig. 2: An example of USSD

3.2. Implementation Details

3.2.1. Data Usage

Each image in Potsdam has five channels, we only use the RGB channels in our experiments. Each 6000×6000 image is split into 144 tiles of 512×512 with overlap for the last column and row while spliting. For USSD, each 3840×2160 image is split into 4 sub-images of 1920×1080. And the same data augmentation methods as BiSeNet[1] are employed. Besides, the videos in USSD are used to evaluate actual video processing effect.

3.2.2. Evaluation Metrics

In our experiments, we evaluate the network performance from three aspects: accuracy, speed and network complexity. Accuracy: (1)The standard metric of the mean intersection-over-union(mIoU), the mean intersection ratio of ground truth and predicted segmentation. (2) Overall accuracy (OA), the ratio of total correct prediction area to total area. (3) Kappa Coefficient, a metric to measure classification accuracy based on confusion matrix.

Speed: Frame per second (FPS). For segmentation networks, FPS is namely the number of images predicted per second. All of the experiments on speed are conducted using single NVIDIA Tesla V100.

Network complexity: (1)Floating point operations (FLOPs). FLOPs are the computation amount needed by the network; (2) The number of parameters. This indicates the memory needed by the network while running.

3.3. Results and Analysis

Table 1: Accuracy and speed evaluation of different networks

Model	Dataset	mIoU%	OA%	Kappa	FPS
BiSeNet[1]	Potsdam	72.01	87.18	0.8320	250.7
LEDNet[5]	USSD	68.53	90.28	0.8543	93.72
ESPNet v2[7]	USSD	64.09	87.55	0.8137	124.42
BiSeNet[1]	USSD	79.26	93.97	0.8998	148.7

Table 2: Complexity evaluation of different networks

Model	GFLOPs	Params (M)
LEDNet[5]	11.45	0.9140
ESPNet v2[7]	5.640	1.250
BiSeNet[1]	31.53	13.61

As presented by Table 1, BiSeNet achieves a quite high speed around 250FPS and satisfying accuracy on Potsdam [8]. This suggests that BiSeNet can be used for high-resolution remote sensing images. On USSD dataset, BiSeNet outperforms LEDNet and ESPNet v2 in terms of all accuracy metrics (mIoU, OA and Kappa). And its speed reaches 148.7 FPS, which is also the highest among all methods. These comparisons prove that BiSeNet achieves an outstanding balance of accuracy and speed.

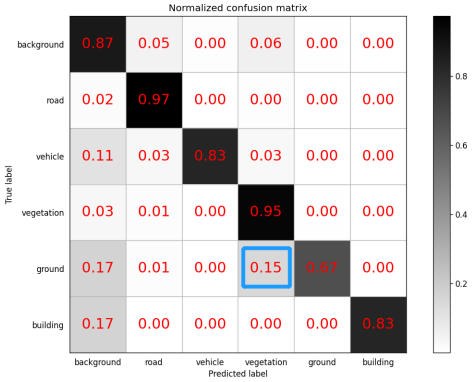


Fig. 3: Normalized confusion matrix of BiSeNet on USSD

Normalized confusion matrix of BiSeNet on USSD is shown in Fig.3. It can be found that BiSeNet has pixel classification precision more than 0.8 except for the "ground" (bare land). This is reasonable since this class is difficult to distinguish from land with sparse vegetation. At the intersection point in the confusion matrix(as shown in the

Authorized licensed use limited to: The Hong Kong University of Science and Technology (Guangzhou). Downloaded on February 19,2026 at 04:12:13 UTC from IEEE Xplore. Restrictions apply.

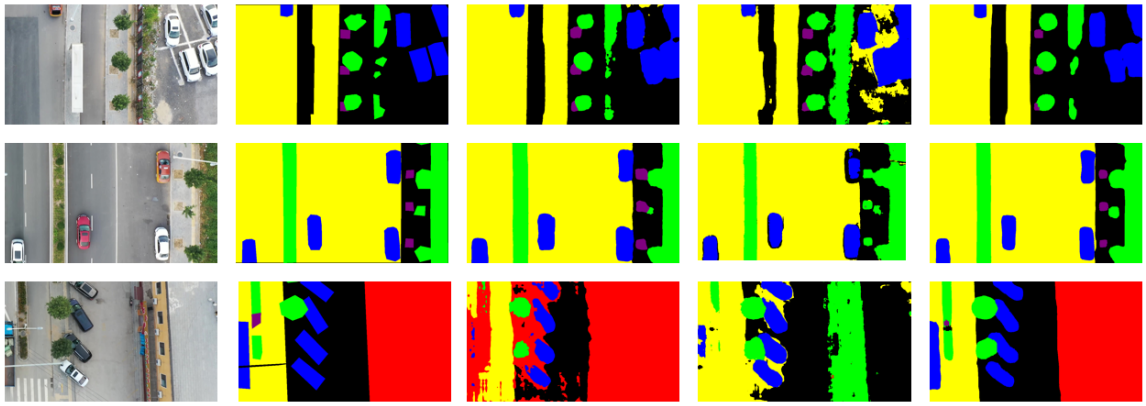


Fig. 4: Visualized comparison of different networks

blue square in Fig.3), we can see the probability the ground misclassified as vegetation is 15%, which is relatively high compared with other class pairs.

Table 2 shows the network complexity metrics of three networks. We can find that ESPNet v2 needs the least GFLOPs and LEDNet has the least parameter amount. BiSeNet needs the most GFLOPs and has the most parameters compared with the other two. This shows that the running cost of BiSeNet is the largest. Nevertheless, this level of overhead is acceptable in most cases. Take AI computing platform NVIDIA Jetson TX2 used for edge computing as an example, TX2's computation capability is 1.33 TFLOPs Jetson Modules, which is sufficient to run BiSeNet. Achieving significantly higher accuracy and speed with acceptable cost is a meaningful trade-off.

Fig.4 shows some visualization results generated by LEDNet, ESPNet v2 and BiSeNet with input images and ground truth. We can find that the result of BiSeNet is more smooth and continuous compared to that of LEDNet and ESPNet v2, which is consistent with the quantitative experimental results. This mainly benefits from the bilateral structure, which enables BiSeNet to acquire more spatial information and larger receptive field. Video segmentation results can be found at Our Repository.

4. CONCLUSION

In this paper, we apply BiSeNet[1] with well-designed bilateral structure composed of spatial path (SP) and context path (CP) to address the problem of aerial video semantic segmentation. And we build a dataset called USSD, which contains labeled sequential aerial images and videos captured by UAVs. Both quantitative and visual experiment results suggest that BiSeNet achieves an outstanding balance between accuracy and speed compared to other SOTA methods. Besides, the network complexity of BiSeNet is acceptable, which makes it possible to deploy on mobile devices and feasible for realistic application scenarios.

5. ACKNOWLEDGEMENT

This work is supported by the Research Innovation Fund for College Students (No.202011037) and Teaching Reform Project (2019JY-A05) of Beijing University of Posts and Telecommunications.

6. REFERENCES

[1] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[5] Yu Wang, Quan Zhou, Jia Liu, Jian Xiong, Guangwei Gao, Xiaofu Wu, and Longin Jan Latecki. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1860–1864. IEEE, 2019.

[6] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9522–9531, 2019.

[7] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9190–9200, 2019.

[8] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1*, 1(1):293–298, 2012.

Authorized licensed use limited to: The Hong Kong University of Science and Technology (Guangzhou). Downloaded on February 19,2026 at 04:12:13 UTC from IEEE Xplore. Restrictions apply.