BiSeNet with Depthwise Attention Spatial Path for Semantic Segmentation

Seongmin Kim
School of Electrical Eng.
University of Ulsan, Korea
dailysmile3347@gmail.com

Kanghyun Jo
School of Electrical Eng.
University of Ulsan, Korea
acejo@ulsan.ac.kr

2022 International Workshop on Intelligent Systems (IWIS) | 979-8-3503-9893-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWIS56333.2022.9920717

***Abstract*—This paper proposes a new structure to obtain similar results while reducing the computational amount of BiSeNet for Real-Time Semantic Segmentation. Among the Spatial Path and Context Path of BiSeNet, the study was conducted focusing on the large size kernel of the Spatial Path. Spatial Path has rich spatial information by creating a feature map 1/8 times the size of the original image through three convolution operations. The convolution operation used at this time is performed in the order of 7x7, 3x3, and 3x3. When a general convolution is used for a kernel of such a large size, the calculated cost increases due to a large number of parameters. To solve this problem, this paper uses Depthwise Separable Convolution. At this time, in Depthwise Separable Convolution, loss occurs in Spatial Information. To solve this information loss, an attention mechanism [1] was applied by elementwise summing between the input and output feature maps of depthwise separable convolution. To solve the dimensional difference between input and output, PPM: Pooling Pointwise Module is used. PPM uses Maxpooling to change the Spatial Dimension of input features and Channel Dimension through Pointwise Convolution (1x1 Convolution) [2]. This paper propose to use Depthwise Attention Spatial Path for BiSeNet using these methods. Through our proposed methods, mIoU in SS, SSC, MSF, and MSCF were 72.7%, 74.1%, 74.3%, and 76.1%. Proposed network can segment the part that the original one can't when using our Depthwise Attention Spatial Path.**

***Index Terms*—Semantic Segmentation, Depthwise Separable Convolution, Attention, BiSeNet, Spatial Path**

I. Introduction

Currently, in the computer vision field, many studies progress to solve the semantic segmentation problem, which uses deep learning to classify objects by pixel. To be used in fields that require real-time segmentation, like autonomous driving, high accuracy, and high processing speed are required. However, Real-Time Semantic Segmentation models had to trade-off between accuracy and inference speed. BiSeNet [3] is the model designed to solve the trade-off problem. BiSeNet [3] uses a two-branch structure called Spatial Path and Context Path. This structure solves spatial information loss and increases accuracy. The Context Path in Fig. 1 uses Xception [4] or ResNet18 [5] as a backbone for downsampling and obtains context information using a large Receptive Field. Spatial Path obtains spatial information through 1/8 size feature map of the input image through three convolution, batch normalization [6], and ReLU [7] operations. However, the Spatial Path uses a 7x7 Kernel in the first convolution operation to obtain rich spatial information. If a large size kernel is used, the number of parameters and calculate cost are increased. To solve this problem, this network use Depthwise Separable Convolution [8] (here after D.S.C) instead of the standard convolution operation to reduce the amount of computation of the Spatial Path. D.S.C [8] separates the operation of spatial axis and channel axis, loss occurs in spatial information. To solve this loss, after each D.S.C [8] operation (in this case, the operation includes convolution, BN [6], and ReLU [7]), the feature map input before the operation and the operation result are elementwise sum. This paper proposes the Depthwise Attention Spatial Path through the following method. BiSeNet [3], using our proposed new Spatial Path showed 72.7% mIoU on single scale evaluation, 74.1% mIoU on single scale crop evaluation, 74.6% mIoU on multi scale evaluation with flip augment and 76.1% mIoU on multi scale crop evaluation with flip augment.

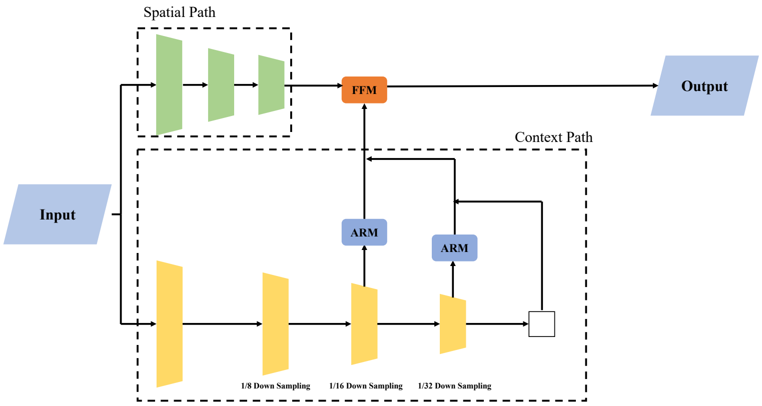


Fig. 1: Architecture of BiSeNet. Denote that ARM is Attention Refinement Module and FFM is Feature Fusion Module.

II. Proposed Method

A. Depthwise Separable Convolution

D.S.C [8] refers to an operation that sequentially proceeds with depthwise convolution that operates only on the spatial axis and pointwise Convolution (1x1 Convolution) [2] that operates on only the channel axis. When input tensor has C channels, kernel of K size has N, and output tensor has resolution of W×H, the number of general convolution and

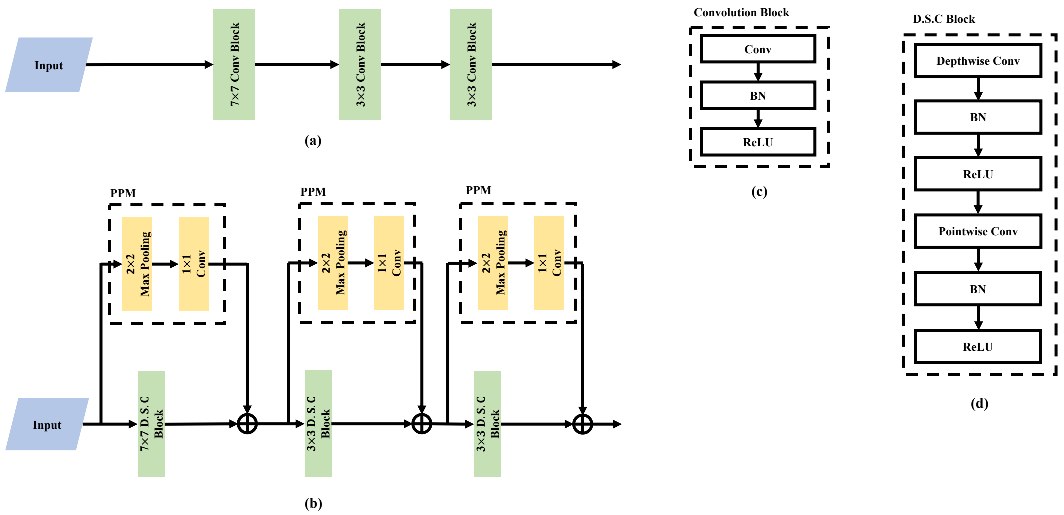


Fig. 2: Structure of Spatial Path (a) Original Spatial Path of BiSeNet. (b) Depthwise Attention Spatial Path. (c) Covolution Block. (d) Depthwise Separable Convolution Block.

D.S.C parameters follows Eq.(1) and (2) below.

$$W_{conv} = K^2 \times C \times N \tag{1}$$

$$W_{D.S.C} = K^2 \times C + C \times N \tag{2}$$

And the product calculate cost of convolution and D.S.C follows Eq.(3) and (4).

$$C_{conv} = W_{conv} \times W \times H \tag{3}$$

$$C_{D.S.C} = W_{D.S.C} \times W \times H \tag{4}$$

The following Table I compares the number of parameters and the amount of computation between D.S.C [8] and standard convolution according to Eq.(1), (2), (3), and (4). Since the convolution used in the original Spatial Path has a stride of 2 and Zero Padding is 3, 1, 1, assuming that the tensors input and output in each operation are T_i, $T_{7\times7}$, $T^1_{3\times3}$, and $T^2_{3\times3}$, the dimensions of each are as follows.

$$T_i \in \mathbb{R}^{3\times1024\times1024} \; T_{7\times7} \in \mathbb{R}^{64\times512\times512}$$

$$T^1_{3\times3} \in \mathbb{R}^{64\times256\times256} \; T^2_{3\times3} \in \mathbb{R}^{128\times128\times128}$$

TABLE I: Comparison of Number of Parameters and Calculated Cost for General Convolution and D.S.C

		7x7	3×3^1	3×3^2	Total
Convolution	Parameter	9,408	36,864	73,728	120,000
	Cost	2.47G	2.42G	1.21G	6.1G
D.S.C	Parameter	339	4,672	8,768	13,779
	Cost	88.9M	306.1M	143.7M	538.7M

M and G each denote Mega 10^6 and Giga 10^9. In this Table I, neglecting BN [6] and ReLU [7] when calculating cost and the number of parameters. Based on these results, this paper propose to use D.S.C [8] instead of general convolution in Spatial Path.

B. Depthwise Attention Spatial Path

The advantage and disadvantage of D.S.C [8] are that it separates operations for Spatial and Channel Axis. Therefore, loss of spatial information occurs when the feature map is output. To solve the problem, this paper proposes the Depthwise Attention Spatial Path using the Pooling Pointwise Module. Figure 2(a) shows the structure of Depthwise Attention Spatial Path. To compensate for the Spatial Information Loss that occurs during D.S.C [8], the Attention Mechanism [1] is taken by performing elementwise summation between the input feature map and the output feature map. Where F_i and F_o are the input and output of the current D.S.C [8] and F_{i+1} is the input feature map of the next D.S.C [8]. The feature map input to the next D.S.C [8] through the Attention Mechanism [1] becomes as in (5).

$$F_{i+1} = \mathcal{P}(F_i) + F_o \tag{5}$$

where $\mathcal{P}$ is Pooling Pointwise Module.

C. PPM: Pooling Pointwise Module

Pooling Pointwise Module is designed because of the output dimension that varies for each D.S.C [8] operation. The structure of the PPM is detailed in figure 3. MaxPooling was used to adjust the Spatial Dimension of the input required for attention according to the resolution reduced by 1/2 after the D.S.C [8] operation, and a 2x2 size pooling layer was used. Pointwise Convolution [2] was used to adjust the Channel Dimension of the Input required for attention according to the changing Channel after the D.S.C [8] operation. With a simple example, this paper will explain the process of applying attention after adjusting the dimension. Let F_i's dimension is

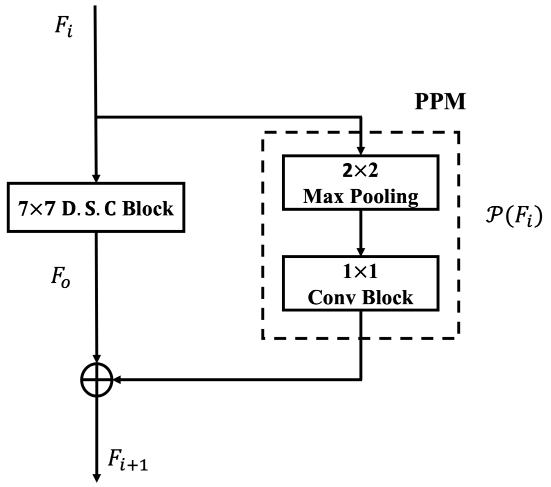


Fig. 3: D.S.C with PPM

$F_i \in \mathbb{R}^{3\times1024\times1024}$ and after D.S.C operation F_o's dimension is $F_o \in \mathbb{R}^{64\times512\times512}$. For attention, F_i goes through PPM to match the dimension with F_o. F_i is Maxpooled through a pooling layer of 2×2 size, and F_i spatial dimension becomes $F_i \in \mathbb{R}^{3\times512\times512}$. Then, through 1×1 Convolution [2], it is adjusted to the same Channel Dimension $F_i \in \mathbb{R}^{64\times512\times512}$ as F_o. F_i whose dimension is transformed through PPM is called $\mathcal{P}(F_i)$, and attention is taken through the elementwise sum. The feature map with attention taken in this way is used as the next input, F_{i+1}.

III. Experiment

A. Dataset

In this paper, an experiment was performed using the Cityscapes dataset. Cityscape is an image of a city while driving a car. It consists of 2975 training data, 500 validation data, and 1525 test data. The resolution of the image is 2048x1024, and it consists of a total of 30 classes. This paper use 19 classes on whole classes because there are some unclear classes in the dataset. The description of each class this paper used is as follows.

- Flat: road, sidewalk
- Construction: building, wall, fence
- Object: pole, traffic light, traffic sign
- Nature: vegetation, terrain
- Sky: sky
- Human: person, rider
- Vehicle: car, truck, bus, train, motorcycle, bicycle

B. Evaluation Metric

This paper measured IoU and mIoU for each class in the following cases through Cityscapes dataset.

- SS: single scale evaluation
- SSC: single scale crop evaluation
- MSF: multi-scale evaluation with flip augment
- MSCF: multi-scale crop evaluation with flip evaluation

C. Implementation Setup

In training and testing, Intel Core i9-10900X CPU and 4 RTX 2080Ti GPUs were used. In train, 8 batches are given to each GPU, so the total batch size is 32. And the total iteration was 160,000.

D. Result

First, this paper evaluated our model and get each class's IoU(%) and mIoU(%).

TABLE II: IoU and mIoU of BiSeNet with Depthwise Attention Spatial Path

		SS	SSC	MSF	MSCF
IoU	Road	97.7	97.8	98	98
	Sidewalk	81.9	82.3	83.6	83.9
	Building	91.3	91.8	92	92.3
	Wall	42	52.9	42.3	50.6
	Fence	50.1	52.7	53.2	56.4
	Pole	63.5	64	67.3	67.6
	Traffic Light	70.3	70.5	73.2	73.7
	Traffic Sign	77.6	78.1	80.6	81.1
	Vegetation	91.8	92	92.4	92.6
	Terrain	59	59.8	61.6	62.3
	Sky	94.8	94.9	95.1	95.2
	Pearson	80.2	80.5	82.3	83.0
	Rider	58.1	58.8	60.9	62.2
	Car	94.1	94.6	94.5	95.1
	Truck	59.8	64.7	60.1	63.6
	Bus	77	76.6	79	79.5
	Train	64.4	66.1	65.9	70.2
	Motorcycle	52.9	54.3	56.4	60
	Bicycle	75.3	75.6	78	78.6
mIoU	-	72.7	74.1	74.5	76.1

Table II shows very high accuracy for Car, Road, and Sky, but low mIoU for Wall and Motorcycle.

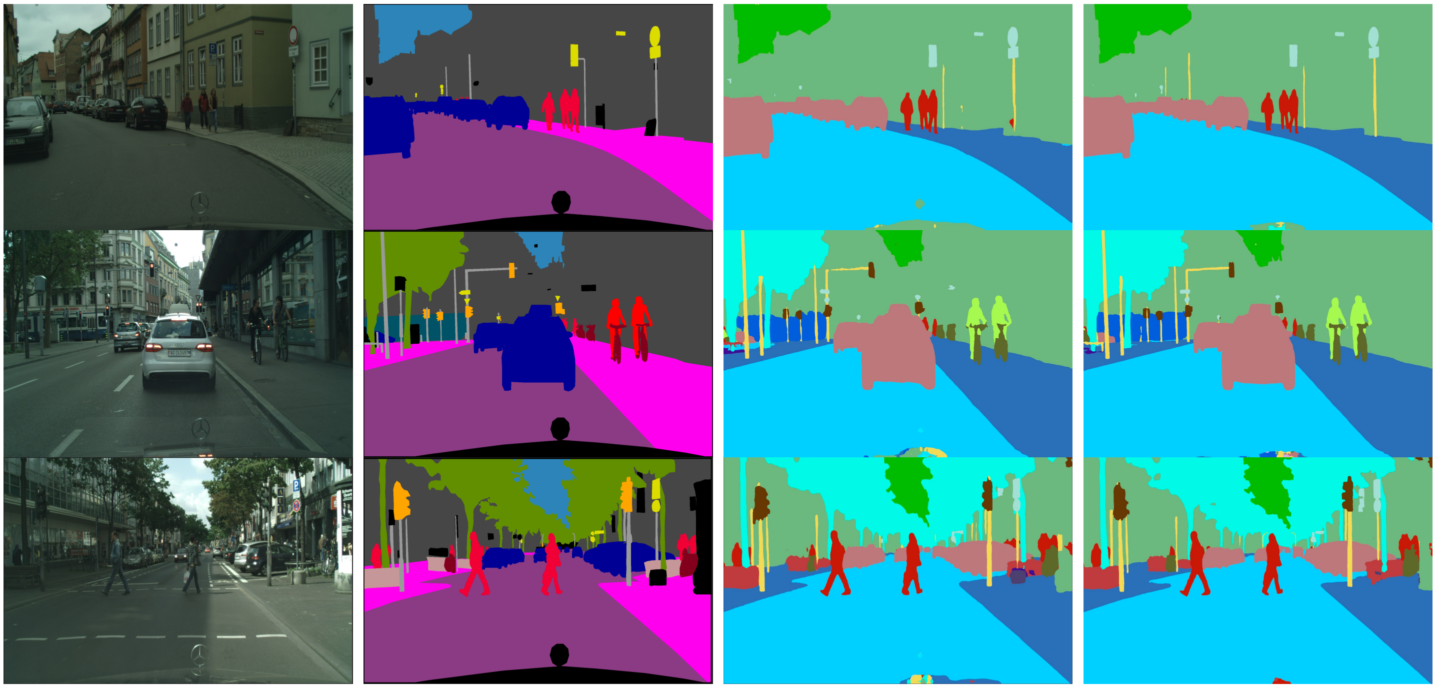
Based on Table II, compared the mIoU(%) between the original BiSeNet [3] and BiSeNet [3] with Depthwise Attention Spatial Path. The dataset and equipment used are the same as those used for training. When evaluating the original BiSeNet [3], weights pre-trained with the cityscapes dataset were used.

TABLE III: Accuracy analysis on each evaluation metric between original BiSeNet and our proposed method.

	SS	SSC	MSF	MSCF
BiSeNet	75.4	76.9	77.4	78.9
Ours	72.7	74.1	74.3	76.1

From Table III, it can be confirmed that the accuracy of our proposed method is close to that of the original BiSeNet [3]. Our proposed method greatly reduced amount of parameters. So our alternative BiSeNet [3] has a little bit low mIoU than the original one.

Next, checked the result of the original BiSeNet and ours. Fig. 4 are example results of the original BiSeNet [3] and our proposed method. Fig. 4 (c) and (d), our Spatial Path can segment objects that the original one can't. Because the Depthwise Attention Spatial Path rewards spatial information loss using attention mechanism [1] but the original spatial path not do.



(a) Image (b) GT (c) BiSeNet (d) Ours

Fig. 4: Example result of the output about the original BiSeNet and adding Depthwise Attention Spatial Path. Denote GT is ground truth..

IV. Conclusion

This paper proposed using Depthwise Separable Convolution [8] to reduce calculation costs on the spatial path of BiSeNet. And, using Attention Mechanism [1] on each D.S.C's input with Pooling Pointwise module to reward spatial information loss after D.S.C [8] operation. When training and evaluating our model, this paper uses cityscapes datasets that have 19 classes. The mIoU of our proposed structure is 72.7%, 74.1%, 74.3%, and 76.1% in the four evaluation metrics of SS, SSC, MSF, and MSCF. When using the Depthwise Attention Spatial Path, it is possible to segment the part that was not possible when using the original Spatial Path [3]. In future research, study a new type of attention mechanism [1] to improve the accuracy of semantic segmentation with low calculate cost.

References

[1] J. Choi and K. Jo, "Attention based object classification for drone imagery," in *IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society*, 2021, pp. 1–4.

[2] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[3] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[4] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[7] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.