DSNET: ACCELERATE INDOOR SCENE SEMANTIC SEGMENTATION

Feng Jiang[1], *Feng Guo*[1], *Rongrong Ji*[1⋆]

[1]Fujian Key Laboratory of Sensing and Computing for Smart City, Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, 361005, China
edjiang.xmu@gmail.com, betop@xmu.edu.cn, rrji@xmu.edu.cn

ABSTRACT

In this paper, we address the problem of real-time image semantic segmentation for indoor scene. As for scene parsing, both accuracy and speed are equally important. However, most of existing methods mainly focus on improving accuracy rather than speed. How to find a balance between accuracy and speed is crucial for real-time semantic segmentation tasks. To tackle this problem, we propose a lightweight framework with *depthwise dilation residual module* and *multi-scale information integration module*. A single DSNet yields the performance of mIoU accuracy 32.12% on SUN RGB-D dataset and accuracy 26.32% on ADE20K dataset, which is the most challenge scene parsing dataset. Besides, our system yields real-time inference on a single NVIDIA GPU.

Index Terms— Semantic Segmentation, Indoor Scene, Real-time, Deep Neural Network

1. INTRODUCTION

Recent years have witness the growth with autonomous driving which has created great need for real-time semantic segmentation. Although *Convolution Neural Networks* (CNNs) have achieved high accuracy in image semantic segmentation task, but very deep network structure results in slow inference speed, which is not allowed when ported to mobile devices.

In order to make deep neural network less time consuming, lots of efficient network structures have been proposed, such as MobileNet [2], MobileNetV2 [3], ShuffleNet [4], ShuffleNetV2 [5] and Xception [6], etc. MobileNet proposed a brand new form of convolution called *depthwise convolution*, which can effectively reduce the number of parameters without reducing the accuracy of the network. As the upgraded version, MobileNetV2 uses a modified version of linear bottleneck proposed by ResNet [7] called inverted residual with linear bottleneck. ShuffleNet proposed a strategy to shuffle channels for group convolution which also reduces unnecessary parameters effectively. These networks have shown excellent performance in image classification task. However, they perform unsatisfactory for semantic segmentation task.

To solve above problems, we propose a new framework termed as Dilated Speed Network (DSNet). The proposed method uses *depthwise dilation residual module* and *multi-scale information integration module* to achieve both extremely fast inference speed and high segmentation performance. Besides, we apply attention mechanism in our network structure, which improves performance of our system.

Our main contributions are summarized as follows:

1)We proposed a new lightweight network structure keeping a balance between speed and accuracy, which achieves better results compared to other real-time semantic segmentation system.

2)The proposed *depthwise dilation residual module* and *multi-scale information integration module* further improve the receptive filed without increasing parameters.

3)Our proposed method achieves impressive results on datasets of SUN RGB-D [8], ADE20K [9]. More specifically, we obtain the result of mIoU 32.12% on the SUN RGB-D test dataset with speed of 45 FPS on single NVIDIA GPU.

2. RELATED WORK

In the following, we review recent advances in semantic segmentation tasks.

High Quality Semantic Segmentation: Compared to graph-based methods, lots of effective approaches based on CNNs have achieved state-of-the-art performance on different benchmarks of semantic segmentation tasks, such as *fully connected nerual networks* (FCNs) [10] based methods. Most of these methods are aiming at improving parsing accuracy, ignoring speed which is as important as accuracy.

Receptive filed of convolution neural network is very important in scene parsing. To enlarge the receptive filed, most

*Corresponding Author.

This work is supported by the National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Nature Science Foundation of China (No.U1705262, No.61772443, and No.61572410), Post Doctoral Innovative Talent Support Program under Grant BX201600094, China Post-Doctoral Science Foundation under Grant 2017M612134, Scientific Research Project of National Language Committee of China (Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No. 2017J01125 and No. 2018J01106).

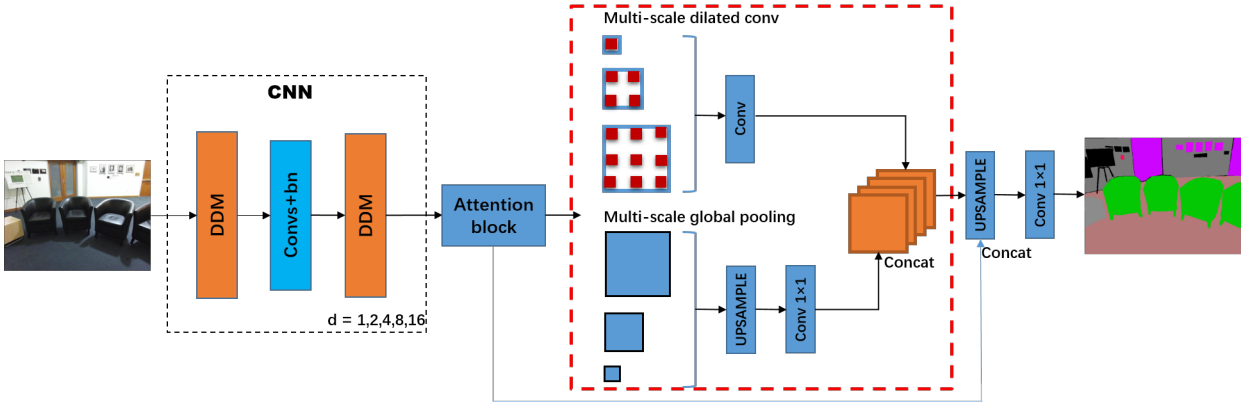

Fig. 1. Architecture of our network. We stack convs and DDMs as encoder network and use multi-scale diltaed convolutons and global poolings to obtain sufficient global context information.

of the recent approaches use very deep neural network or employ *dilated convolution* in convolution layers to capture more context information. Inspired by the image pyramid, multi-scale feature ensembling is often used in semantic segmentation network structure. *Atrous Spatial Pyramid Pooling* (ASPP) proposed by [11] contains different rates of dilation convolution to capture different scales of receptive field. Similar with "ASPP", PSPNet [12] proposed a *Pyramid Pooling Module* (ppm) which is composed of several scales of pooling operation to capture different scales of feature. Zhang *et.al.* [13] proposed a method called scale adaptive convolution layer to make the neural network learn adaptive receptive field and obtain more efficient context information.

Attention Mechanism: Attention mechanism can guide the network to focus on the most important information we want, it is used to recalibrate the feature map to emphasize on useful channels. Recently, attention module becomes an important part in deep neural networks, the *"Squeeze-and-Excitation"*(SE) block proposed by [14] can be seamlessly integrated with any CNN structure. In [15], they learn a global context as attention to recalibrate the channel importance.

Real-time Semantic Segmentation: In order to generate high-quality prediction in real-time, SegNet [12] used a small network structure by abandoning some layers to reduce parameters. ICNet [16] used cascade network structure to speed up semantic segmentation. However, when porting the network structure to mobile devices, limited by inference framework running on mobile devices, it may not support these cascade structures and slow down the parsing speed. So in our proposed method, we design a lightweight network with sufficient receptive filed, while its shallow structure reduces time complexity and provides adequate spatial information.

3. THE PROPOSED FRAMEWORK

In this section, we start with our analysis of recent FCN-based methods. Then, we elaborate on the details of our proposed *Dilated Speed Network*, which uses *depthwise dilation residual module* and *multi-scale information integration module* to address real-time semantic segmentation task.

3.1. Network Architecture

Similar to image classification tasks, high quality features extracted by deep convolution neural networks are very important to semantic segmentation task. Deep neural network can provide sufficient receptive filed and good abstraction of features, which is better suited to semantic segmentation. However, as network going deeper, computation complexity grows quickly. At the same time, we will loss lots of spatial details which is hard to recover.

Based on this two observations, we propose an efficient network as encoder network to get $1/8$ feature map. With this design, the network can keep most of the spatial information. Our proposed network is consisted of a shallow but receptive filed sufficient structure and an efficient multi-scale information integration module as shown in Fig 1.

Depthwise Dilation Residual Module: To get large receptive filed, we need larger convolution kernels or strides. However, large kernels will bring more number of parameters and large stride will down-sample the feature map quickly which results in loss of spatial information. To solve this problem, more efficient convolution operations are needed. *Dilated convolution* is first proposed in [11] to increase receptive filed of network without down sampling the feature maps. The dilated convolution can be defined as follows:

$$y[i] = \sum_{k=1}^{K} x[i + d \cdot k] \cdot w[k] \tag{1}$$

where $x[i]$ denotes input feature map, $y[i]$ denotes output, d is the dilation rate, K is the size of kernel, $w[k]$ is k-th parameter of kernel. Dilated convolution is equivalent to standard convolution when $d = 1$. When $d > 1$, dilated convolution

will insert $d-1$ zeros between kernel values to get a larger receptive filed. We notice that, to get large receptive filed in

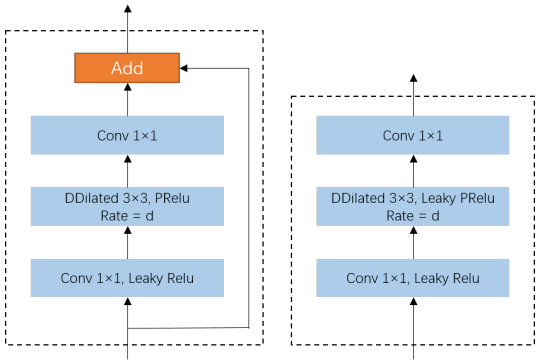

Fig. 2. Illustration of Depthwise Dilation Residual Module.

a shallow network, dilated convolution is a good choice. In our method, we propose a module called *Depthwise Dilation Residual Module* (DDM), shown in Fig 2.

DDM contains a 1×1 convolution that expands the dimensionality, a special form of depthwise convolution, we call it depthwise dilated convolution which can effectively reduce the computation cost, and a 1×1 convolution to reduce the dimensionality. We place Batch Normalization [17] and PRelu [18] between convolutions. Besides, skip connection is also used in this module to capture more context information. Through this way, we can achieve larger receptive in shallow network and keep most of the spatial information at the same time.

Attention mechanism: At the end of the encoder network, we design an attention block similar to SE block proposed by SENet [14] to refine the output feature maps. Attention block will compute an attention vector to guide the network learning more important features.

Multi-scale Information Integration Module: One convolution can only catch fixed-size receptive filed which has little influence on image classification task for objects in this task are usually in the center of image. However, fixed-size receptive filed matters a lot when it comes to indoor scene parsing. Objects in these scenes often have different sizes and shapes. If the object is too large for receptive filed to capture, then fixed-size receptive filed can only catch part of the object. If the object is too small for receptive filed to capture, it can be ignored and seen as background.

To solve this problem, we propose a *Multi-scale Information Integration Module*, as illustrated in the red dashed box in Fig 1. This module contains multi-scale rates of dilated convolution, yielding different sizes of receptive filed. Apart from that, we use different scales of global pooling that can handle various size of global context information, as mentioned in [12].

3.2. The Loss Function

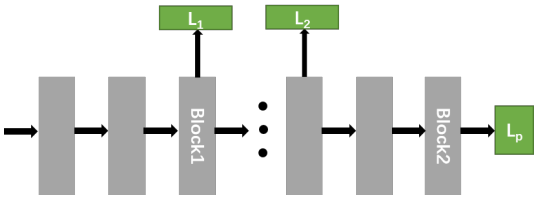

Fig. 3. Illustration of auxiliary loss.

To achieve better result, we utilize the auxiliary loss function to supervise the training of our proposed method. We add additional two auxiliary loss functions L_1, L_2 to help supervising the output of the network at the feature maps with $1/8$ of the original resolution. All loss functions are cross-entropy loss on the corresponding downsampled score maps , as illustrated in Fig 3.

We we have a set of balance factors $\{\lambda_1, \lambda_2\}$ to represent the importance between these loss functions, as Equation 2 shows.

$$L = L_p + \lambda_1 \cdot L_1 + \lambda_2 \cdot L_2 \tag{2}$$

where L_p is the principal loss of the final output.

Our experiments show that, auxiliary loss functions help increase the performance of the network. We only use auxiliary loss during training time.

3.3. Boundary Refine

In semantic segmentation task, pixels belonging to boundary are often hard to predict precisely, because the spatial information around boundary will lose after several times of down-sample operation. The prediction around boundary region may have blur details and incorrect prediction. To overcome this problem, we add additional balance factor to the loss function. During training, we extract pixels belonging to boundary region, and pay more attention to them by increasing the loss weights of prediction. Particularly, we obtain the region around boundary as region B. Finally, the object function can be modified as follow:

$$L_f = \alpha_1 \sum L(p|p \in B) + \alpha_2 \sum L(p|p \notin B) \tag{3}$$

In the above equation, α_1 denotes the loss weight where pixel p belongs to boundary region and α_2 is the opposite. The network is trained to minimize the above loss function, as Equation 3 shows.

4. EXPERIMENTS

In this section, we will introduce the details of our experiments on two challenge indoor scene parsing datasets, including SUN RGB-D and ADE20K. We tabulate mIoU performance and inference time of our method and other methods.

Table 1. Performance comparison on SUN RGB-D.

Method	Mean IoU(%)	Pixel Acc.(%)
FCN-8s [10]	24.05%	67.31%
Segnet [21]	22.52%	70.73%
ENet [1]	19.7%	59.5%
DeepLab-LargeFOV [22]	30.67%	70.70%
Ours	**32.12%**	**75.60%**

Table 2. Performance comparison on ADE20K.

Method	Mean IoU(%)	Pixel Acc.(%)
FCN-8s [10]	24.83%	64.77%
Segnet [21]	21.64%	71.00%
DilatedNet [23]	25.92%	65.42%
Ours	**26.32%**	**71.23%**

Table 3. Speed comparison on GTX 1080Ti with different resolutions, the higher is better. Image size is $W \times H$.

Method	640×360 (fps)	1280×720 (fps)	1920×1080 (fps)
Segnet [21]	14.6	3.5	1.6
ENet [1]	135.4	46.8	21.6
ICNet [16]	28	12.3	6.3
Ours	**48.2**	**32.4**	**25.2**

We use MobileNetV2 [3] as our baseline network structure and modify it to our network. We use SGD optimization algorithm [19] with initial learning rate 0.02, momentum 0.9 and weight decay to train our network. We set $\lambda_1 = 0.4, \lambda_2 = 0.6, \alpha_1 = 2, \alpha_2 = 1$ empirically. Inspired by [11], we apply "poly" learning rate adjust strategy during training. We train our network on GTX 1080Ti GPUs for about $100,000$ iterators. We use random flip, random scale and other ways to augment the dataset during training.

SUN RGB-D: The SUN RGB-D [8] is a very challenge indoor scene parsing datasets, it contains $5,285$ training images and $5,050$ testing images. It contains 37 object classes including wall, table, ceiling, window, chair, etc. This task is hard for indoor objects due to various shapes, sizes, poses and overlapping with each other. We do not use any depth information of the dataset during training.

Table 1 shows our results compared to other methods on SUN RGB-D dataset. As we can see, our proposed method achieves better or similar performance, but faster speed. The visualization results are shown in Fig 4.

ADE20K: The ADE20K [9] dataset is the most challenging for it contains 150 classes and diverse scenes with total 25K images, including 20K for training, 2K for validation and 3K for testing. Table 2 shows our results compared to other methods on ADE20K dataset. The visualization results are shown in Fig 5.

Speed: We test our proposed method on a single GTX 1080Ti GPU with different resolution images. As we can see from Table 3, our method achieves the fastest speed in high resolution compared to other popular real-time segmentation networks.

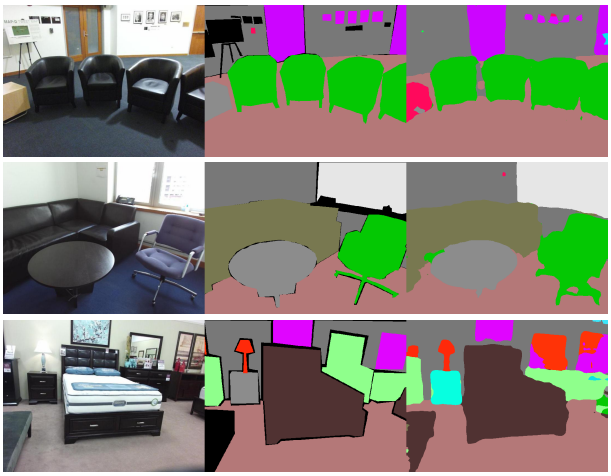

Fig. 4. Visualization results on the SUN RGB-D dataset when employing our best model.

Fig. 5. Visualization results on the ADE20K dataset when employing our best model.

5. CONCLUSIONS

We propose an efficient network for semantic segmentation. It achieves both high performance and fast speed in popular datasets. The major contribution of our work is figuring out that it is possible to find a balance between deeper network which means higher segmentation accuracy and shallow network which helps obtain faster speed. We believe that, speed and accuracy are equally important, our research shows a good balance between them.

6. REFERENCES

[1] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[2] Andrew G Howard and Menglong Zhu, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[3] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018.

[4] Michael G Hluchyj and Mark J Karol, "Shuffle net: An application of generalized perfect shuffles to multihop lightwave networks," *Journal of Lightwave Technology*, 1991.

[5] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," *arXiv preprint arXiv:1807.11164*, 2018.

[6] François Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint*, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[8] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *CVPR*, 2015.

[9] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017.

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

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," in *PAMI*, 2018.

[12] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *CVPR*, 2017.

[13] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan, "Scale-adaptive convolutions for scene parsing," in *ICCV*, 2017.

[14] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2017.

[15] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang, "Learning a discriminative feature network for semantic segmentation," in *CVPR*, 2018.

[16] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia, "Icnet for real-time semantic segmentation on high-resolution images," *arXiv preprint arXiv:1704.08545*, 2017.

[17] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *CVPR*, 2015.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[20] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun, "Megdet: A large mini-batch object detector," in *CVPR*, 2018.

[21] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," in *PAMI*, 2017.

[22] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *ICLR*, 2015.

[23] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.