Sufficient Context for Real-Time Semantic Segmentation

Borui Fan¹, Wei Wu²

¹Department of Information Engineering, Wuhan University of Technology, Wuhan, Hubei Province, 430000, China
²Department of Information Engineering, Wuhan University of Technology, Wuhan, Hubei Province, 430000, China

*Corresponding author’s e-mail: wuwei@whut.edu.cn

Abstract. In recent years, the research on semantic segmentation has progressed rapidly, and many segmentation models with good effects have emerged, but these models often require a large amount of parameters. In actual application scenarios, the computing resources and memory of the device are limited, so the research on lightweight real-time semantic segmentation has very important practical significance. In this paper, we propose SDSNet, which is mainly composed of SDS module that includes some SDS-units. SDS unit is based on split-shuffle and dilated convolution, which can get sufficient context by increasing the interaction capability between channels and expanding the receptive field without increasing the amount of calculation. Our network combines the high feature generated by the SDS module with low-level spatial information, yielding a perfect segmentation results. Experiments on CityScapes prove that the proposed SDSNet achieved a comparable result to the state-of-the-art models.

1. Introduction

Semantic segmentation is a basic task of computer vision, which aims to assign corresponding labels to all pixels in an image. Semantic segmentation is widely used in many fields, including autonomous driving and video surveillance. In recent years, the segmentation accuracy of many studies has reached an ideal level. However, the price required to achieve such high accuracy is a large number of parameters and the accompanying large memory and computational capabilities. And in real-world applications, the computing performance and memory of the equipment are limited, but there are strict requirements on the inference speed. Therefore, how to maintain the efficient inference speed and high precision quality of high-resolution images is a key issue.

The mainstream model adopts a structure similar to FCN[1] and UNet[2]. This structure often cannot make full use of global information. At the same time, it requires more parameters so that the inference speed is very slow and cannot meet the requirements of real-time scenarios. Regarding the former, we use dilated convolution and pyramid pooling to enhance feature extraction capabilities; dilated convolution can obtain a larger receptive field without reducing the resolution, pyramid pooling can make full use of global information to produce different levels of features. Regarding the latter, many solutions have been proposed for real-time semantic segmentation. Part of the research can be summarized as model compression, reducing the input size or the number of channels to reduce the amount of model parameters such as Segnet[3]. However, this method will have a certain loss of accuracy due to the loss of spatial information. For more spatial information, some work adopt multi-branch which are respectively responsible for context and spatial information such as IcNet[4] and BiseNet[5]. However, the inference speed of the network will also be affected by the speed of slower
branch. In order to make the network lighter and faster, the factorized convolution is used to greatly reduce the model parameters without reducing the accuracy such as LedNet[6]. In particular, this method is also used in our network to reduce the computational cost. Based on the above observations, we propose a lightweight real-time semantic segmentation network named as SDSNet. Our main contributions are summarized as follows:

- We propose a SDSNet, which combine spatial information and semantic information by using dual branches and get comparable performance to the state-of-the-art real-time semantic segmentation models.
- We design a SDS module to get a larger receptive field and more sufficient context at a small cost.

2. Related Work

With the development of deep learning related research, real-time semantic segmentation is also constantly emerging new methods. In this chapter, we will give a brief introduction to the recent research progress of real-time semantic segmentation.

ENet[7] improves the network speed as much as possible by reducing the number of downsampling, but at the same time, the receptive field becomes smaller and the accuracy of the network is lost. EspnetV2[8] obtains a larger receptive field through using group point-wise and depth-wise dilated separable convolutions without increasing the amount of parameters. IcNet[4] constructs a cascade network and realizes real-time segmentation by forming multiple branches with different resolutions through multi-scale input. BiseNet[5] achieves effective and simple real-time segmentation by combining spatial branches and semantic context branches. At present, real-time segmentation has made great progress, but there is still a lot of room for improvement in the trade-off of accuracy and speed. The network proposed in this paper is a supplementary work to the existing real-time segmentation network.

2.1. Convolution factorization

There are many forms of convolution factorization now. A standard convolution can be taken apart in several steps to reduce the amount of calculation parameters while achieving the same result. Point-wise group convolution and channel shuffle are used in shufflenet[9] to promote information exchange between channels. In particular, shuffle operation is also used in our network. Xception[10] uses deep separable convolution to reduce the amount of parameters without reducing performance. Our network also uses factorized convolution to reduce the amount of parameters.

2.2. Spatial Information and Context Encoding

The difficulty of real-time semantic segmentation is that it can not effectively combines spatial information and contextual information. Downsampling can obtain high-level features, but at the same time, it will also reduce the resolution of the image and lose the spatial detail information and affect the segmentation accuracy. Our network uses dilated convolution to obtain a larger receptive field without reducing the resolution, and we proposed feature fusion module in the network, which combine the high level feature map with spation information from the skip connection.

3. Approach

In this chapter, we introduce our proposed SDSNet in detail, including SDS module, Pyramid Pooling Module (PPM). The entire network starts with an initial down-sampling which downsample the resolution to one-eighth of the original resolution by depth-wise separable convolution and then is divided into two branches: space and context. The core module of the context path is the SDS module. Each SDS module includes some SDS-unit with split-dilated-shuffle structure to obtain a larger receptive field at a smaller cost, and at the same time strengthen the information interaction between channels through the shuffle operation to enhance feature extraction capabilities. At the end of the branch is a simple pyramid pooling module to fuse context information from different regions. Spatial
branches are mainly realized through skip connections. The two branches are fused through a feature fusion module, and high-level semantic information guides low-level features to achieve better segmentation results. In order to achieve faster speed, we use native upsampling operation. The number of layers in the entire network is very small and lightweight, and the overall network architecture is shown in Figure 1.

Figure 1 Overview of our approach

3.1. SDS-Unit
With the development of deep learning, many efficient and lightweight networks have appeared in recent years, including the resnet[11] and inception series. These networks propose the residual layer and its related deformations including bottleneck (See Figure2(left)), and factorized convolutions (See Figure2(middle)). Inspired by these designs, we proposed the split-dilated-shuffle unit termed as SDS-unit (See Figure2(right)).

Figure 2. different types of residual modules. left:bottleneck;middle:non-bottleneck-1D;right:SDS-unit.“Conv” is a standard convolution, “DConv” is a depth-wise convolution, “DDConv” is a depth-wise dilated convolution.

The design purpose of bottleneck is to reduce the amount of calculation parameters. The implementation of bottleneck is to first reduce the dimensionality of the input tensor and then calculate, and in the end increase the dimensionality. In SDS-unit, the input tensor channel is divided into two parts into two branches respectively. The design ideas of the two branches here are inspired by Bisenet. One branch uses depth-wise convolution to be responsible for spatial information, and the other branch
uses depth-wise dilated Convolution to expand the receptive field to obtain more contextual information. Both branches use depth-wise convolution which is applied convolution factorization to reduce the amount of calculation, and finally a concat and shuffle operation to increase the information interaction and fusion between channels.

3.2. **Pyramid Pooling Module**
Pyramid pooling is a strategy often used in neural networks. It can extract and aggregate a feature map from different locations to capture accurate contextual information. Multi-scale feature extraction can also enhance the ability of capturing the small objects. In this work, we use pyramid pooling at the end of the context branch to obtain more effective context information.

3.3. **Feature Fusion Module**
In the task of semantic segmentation, the common encoder-decoder network mainly use multi-scale information to gradually restore spatial information, but it comes with huge computational loss and slow inference speed. The feature map after the pyramid is up-sampled and restored to the same size as the initial down-sampling, and then input into the feature fusion module together with the feature map on the spatial branch, concat them to calculate a weight vector, and use this weight vector to highlight Useful features, and finally the output feature map is subjected to a native upsampling to get the final result. The feature fusion module is shown in Figure 3.

4. **Experiments**
In this chapter, we use experiments to prove that our proposed SDSNet can achieve a good trade-off between speed and accuracy. We will introduce the data set, experimental details and result comparison in turn.

4.1. **datasets**
CityScapes is a public data set widely used in the field of driverless driving. It collects street scenes from 50 countries in different seasons and backgrounds to evaluate the performance of semantic segmentation models on street scenes. The data set mainly includes images_base and annotations_base, which correspond to the two folders leftImg8bit and gtFine respectively. The data set includes 5000 images, of which 2975 are used for training, 500 are used for verifcation, and the remaining 1525 are used for testing.

4.2. **Implement Details**
In this experiment, we test the performance of the proposed network on cityscapes. Our experiments are executed on the Ununtu 16.04 operating system with 2 Nvidia 2080 GPU, with CUDA 9.0 and CuDNN v7.

We use stochastic gradient decent (SGD) with momentum 0.9 and batch-size 15. We set learning rate with the base one as 0.045 and power as 0.9. On the cityscapes dataset, we train our model for 1000 epochs. Our experimental results are shown in Figure 4.
Figure 4. Example of segmentation in the test set of cityscapes. first column: input RGB images; second column: prediction of our experiments; third column: ground-truth

4.3. Comparison
We compared our method with the current mainstream methods on the cityscapes dataset. The comparison results are shown in Table 1. It can be seen that our method achieves a good trade-off between segmentation speed and accuracy on high-resolution images.

Table 1. Comparison with state-of-the-art methods on cityscapes dataset.

| Method     | InputSize | FPS  | MIou(%) | Parameters(M) |
|------------|-----------|------|---------|---------------|
| Enet[7]    | 512×1024  | 76.9 | 58.3    | 0.36          |
| ESPNet[8]  | 512×1024  | 112  | 60.3    | 0.36          |
| ERFNet[12] | 512×1024  | 41.7 | 68.0    | 2.1           |
| BiSeNet[5] | 768×1536  | 105.8| 65.4    | 5.8           |
| ICNet[4]   | 1024×2048 | 30.3 | 69.5    | 7.8           |
| ContextNet[13] | 1024×2048 | 18.3 | 66.1    | 0.85          |
| SQ[14]     | 1024×2048 | 16.7 | 59.8    | -             |
| Ours       | 1024×2048 | 28.6 | 69.8    | 1.34          |

The comparison results are shown in Table 1. We can see: 1) Our segmentation accuracy under high-resolution images is better than the current mainstream methods, thanks to the large receptive field of our network and more context information 2) our network parameters are comparable to other mainstream lightweight methods. This is due to our shallower network layer design and fewer downsampling times.

5. Conclusion
In this article, we propose a new real-time semantic segmentation network SDSNet based on the split-shuffle mechanism and factorized convolution, which increases the receptive field and feature extraction
capabilities of the network under the condition that the parameters are not changed or even reduced. Our network achieves an excellent trade-off between speed and accuracy.

Acknowledgments
This work was supported by the National Natural Science Foundation of China: Research on the visual perception enhancement method of unmanned surface ships in severe navigable environment.

References
[1] E. Shelhamer, J. Long, and T. Darrell, ‘Fully Convolutional Networks for Semantic Segmentation’, IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 640–651, 2017, doi: 10.1109/TPAMI.2016.2572683.
[2] O. Ronneberger, P. Fischer, and T. Brox, ‘U-net: Convolutional networks for biomedical image segmentation’, Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma., vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.
[3] V. Badrinarayanan, A. Kendall, and R. Cipolla, ‘SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation’, IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, 2017, doi: 10.1109/TPAMI.2016.2644615.
[4] H. Zhao, X. Qi, X. Shi, J. Shi, and J. Jia, ‘ICNet for Real-Time Semantic Segmentation on High-Resolution Images’, Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma., vol. 11207 LNCS, pp. 418–434, 2018, doi: 10.1007/978-3-030-01219-9_25.
[5] C. Yu, J. Wang, C. Gao, G. Yu, and N. Sang, ‘BiSeNet: Bilateral segmentation network for real-time semantic segmentation’, Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma., vol. 11217 LNCS, pp. 334–349, 2018, doi: 10.1007/978-3-030-01261-8_20.
[6] Y. Wang et al., ‘LedNet: A Lightweight Encoder-Decoder Network for Real-Time Semantic Segmentation’, Proc. - Int. Conf. Image Process. ICIP, vol. 2019-September, pp. 1860–1864, 2019, doi: 10.1109/ICIP.2019.8803154.
[7] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, ‘E Net: A Deep Neural Network Architecture for Real-Time Semantic Segmentation’, pp. 1–10, 2016.
[8] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, ‘ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network’, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 9182–9192, 2019, doi: 10.1109/CVPR.2019.00941.
[9] X. Zhang, X. Zhou, M. Lin, and J. Sun, ‘ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices’, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 6848–6856, 2018, doi: 10.1109/CVPR.2018.00716.
[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ‘Rethinking the Inception Architecture for Computer Vision’, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.
[11] K. He, X. Zhang, S. Ren, and J. Sun, ‘Deep residual learning for image recognition’, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.
[12] L. M. Bergasa, R. Arroyo, E. Romera, and M. Alvarez, ‘ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation’, IEEE Trans. Intell. Transp. Syst., pp. 1–10, 2018.
[13] R. P. K. Poudel, U. Bonde, S. Liwicki, and C. Zach, ‘ContextNet: Exploring context and detail for semantic segmentation in real-time’, arXiv, 2018.
[14] M. Treml et al., ‘Speeding up Semantic Segmentation for Autonomous Driving’, NIPS 2016 Workshop MLITS, no. Nips, pp. 1–7, 2016.