ResSCNN: A semantic segmentation method for fast processing of large-scale input

Haoyang Yu¹, Siyuan Xu¹, Wenliang Su¹, Zihan Xiao¹, Jie Yang¹, ² and Jiong Mu¹, ²,*

¹College of Information Engineering, Sichuan Agricultural University, Ya’an, Sichuan, China
²Key Laboratory of Agricultural Information Engineering of Sichuan Province, Ya’an, Sichuan, China

*Corresponding author e-mail: jmu@sicau.edu.cn

Abstract. With the development of driver-assistance systems and driverless cars, vehicles are becoming more and more intelligent. However, today's intelligent vehicles rely more on large sensors to sense the environment, so it is becoming more and more important for vehicles to be able to understand road information, and semantic segmentation has developed greatly. We find that the current mainstream semantic segmentation model is slow and inaccurate in the case of large scale input images. In this paper, we combine the large-scale input Fast-SCNN with the residual module to construct the residual semantic segmentation convolutional neural network (ResSCNN), after the experimental test, when the input scale is 1024×2048 px, yielding an accuracy of 68.4% mean intersection over union at 123.1 frames per second on Cityscapes dataset. Our network has achieved better results than the previous mainstream methods.

1. Introduction

With the development of science and technology, people's living standards continue to improve, cars have become a necessity of people's life, but with the increasing number of cars, the number of traffic accidents is increasing. The number of traffic accidents in China has been on the rise since 1997, reaching a peak in 2002 of about 770,000. In recent years the number of traffic accidents has fallen to about 20,000 a year, but that is still a lot [1]. However, up to 93% of accidents are caused by improper driving or human error. Therefore, the effective prevention of traffic accidents has become a very concerned research.

Deep learning has been widely used in many fields and environments and has proved to be a very efficient method. Of course, deep learning also plays an important role in intelligent driving. Semantic image segmentation, also known as pixel-level classification, is the clustering of parts belonging to the same object class in the image [2, 3]. Semantic segmentation technology is to segment pixels according to the different semantic meanings expressed in the image. By extracting semantic information from the image, the content of the image can be understood, and then recognition and classification can be realized [4]. By segmenting the road information from the car's front-facing...
camera, we can allow the computer to effectively distinguish the road from other vehicles, pedestrians and buildings, so that the car can safely drive on the road.

In this paper, inspired by the paper [5], we use the semantic segmentation model to apply the residual module to the Fast-SCNN network structure and compare it with other classical semantic segmentation models such as DeepLab v2 [6], PSPNet [7], SegNet [8], ERFNet [9], Fast-SCNN [10]. Experiments show that the accuracy of our network model in cityscapes dataset is better than other models. Specific experiments are discussed in Section 2.

2. Materials and methods

2.1. cityscapes dataset

Cityscapes has 5,000 images of driving scenes in an urban environment, including 2,975 training sets, 500 verification sets, and 1,525 test sets. It has 19 categories of dense pixel annotations, eight of which have instance-level segmentation. The Cityscapes data set, or cityscape data set, is a new large-scale data set that contains a different set of stereo video sequences that record street scenes in 50 different cities. [13]

Cityscapes dataset has a huge amount of data and extremely fine labelling fineness. From the perspectives of occurrence frequency and practicability, 30 object categories are defined, which are mainly divided into eight categories: flat, construction, nature, vehicle, sky, object, human and void. Given the utility of all annotations and their compatibility with other data sets, the Cityscapes dataset uses 19 common categories for benchmarking.

Table 1. Structure of Fast-SCNN network.

| Type      | Convolutional | Num_channel | Stride parameter | Output          |
|-----------|---------------|-------------|------------------|-----------------|
| Conv2D    |               | 32          | 2                | 1024×2048×3     |
| DSConv    |               | 48          | 2                | 512×1024×32     |
| DSConv    |               | 64          | 2                | 256×512×48      |
| bottleneck|               | 64          | 2                | 128×256×64      |
| bottleneck|               | 96          | 2                | 64×128×64       |
| bottleneck|               | 128         | 1                | 32×64×96        |
| PPM       |               | 128         | -                | 32×64×128       |
| FFM       |               | 128         | -                | 32×64×128       |
| DSConv    |               | 128         | 1                | 128×256×128     |
| Conv2D    |               | 19          | 1                | 128×256×128     |

2.2. Fast-SCNN

In the Fast-SCNN [10] proposed in 2019, the author proposed a Fast Semantic Segmentation convolutional neural network (Fast-SCNN), a semantic segmentation model that can also achieve real-time effects on high-resolution images (1024×2048px) and is suitable for running on embedded devices with low memory consumption. This network combines high-resolution spatial details with low-resolution depth features, with a speed of 123.5 frames per second and a mean intersection over union of 68% on the Cityscapes dataset.

The main contributions of this article are as follows:

- It is proposed that Fast-SCNN can be used for real-time semantic segmentation of high resolution (1024×2048px) images with a frame rate of 123.5 and an accuracy rate of 68%;
- Using the popular skip connection method in offline DCNN, this paper proposes a shallow learning lower sampling module for fast and efficient multi-branch low-level feature extraction;
- A low-capacity Fast-SCNN was specifically designed, and it was empirically verified that running more iterations of training on this network architecture was as successful as using ImageNet pretraining or using additional refinement training dataset;
If the input data is sampled twice by using Fast-SCNN, the performance of SOT can be achieved without the need to redesign the network.

### Table 2. Structure of the ResSCNN network.

| Type       | Convolutional | Num channel | Stride parameter | Output       |
|------------|---------------|-------------|------------------|--------------|
| Conv2D     |               | 32          | 2                | 1024×2048×3  |
| DSConv     |               | 48          | 2                | 512×1024×32  |
| DSConv     |               | 64          | 2                | 256×512×48   |
| Residual   |               | -           | -                | 256×512×48   |
| bottleneck |               | 64          | 2                | 128×256×64   |
| bottleneck |               | 96          | 2                | 64×128×64    |
| bottleneck |               | 128         | 1                | 32×64×96     |
| PPM        |               | 128         | -                | 32×64×128    |
| Residual   |               | -           | -                | 32×64×128    |
| FFM        |               | 128         | -                | 32×64×128    |
| Residual   |               | -           | -                | 32×64×128    |
| DSConv     |               | 128         | 1                | 128×256×128  |
| Conv2D     |               | 19          | 1                | 128×256×128  |
| Residual   |               | -           | -                | 128×256×128  |

2.3. ResNet

ResNet [11] was proposed by four Chinese, including Kaiming He from Microsoft Research, who successfully trained a 152-layer neural network by using the ResNet Unit, which solved the vanishing gradient problem that perplexed previous people. ResNet's authors proposed a residual structure, which also became an important representative of ResNet's differences from other network structures:

![Figure 1. Unit structure of the ResNet network.](image)

The output of the residual block can be defined as a \((l-1)\) th, based on the output of the previous Layer X\(l-1\), where \(\mathcal{F}(Xl-1)\) is the output after various operations. The final output x\(l\) of the residual unit can define the following equation [12]:

\[
Xl = \mathcal{F}(Xl-1) + Xl-1
\]

2.4. An improved network ResSCNN based on Fast-SCNN

In this paper, our network model integrates the residual-module of Fast-SCNN and ResNet. And we call this network model as ResSCNN. Inspired by the idea of the paper, the two ResNets were merged to form a network with a Shared weight F. Data was input into the network in pairs. Finally, a relative Loss function was used to guide the training of the network the network structure is shown in table 2.
3. Results and Analysis
We tested ResSCNN on Cityscapes dataset to verify the effectiveness of residual module. During the exercise, we use random horizontal flips were used, random rotation between -10 and 10 degrees, and scaling between 0.5 and 2 degrees. None of the models adopted the Cityscapes data set with coarse annotation. The learning rate is based on the learning rate multiplied by an attenuation factor, and the calculation formula is as follow, the basic learning rate $l_r$base is set to 0.01 and power is set to 0.9:

$$\text{lr} = \text{lr}_{base} \times (1 - \frac{\text{epoch}}{\text{max\_epoch}}) \quad \text{power (2)}$$

For the accuracy discrimination of various classification results, we use the IOU discrimination method, whose specific formula is as follow:

$$\text{IOU} = \frac{\text{area(ROI\_GT)}}{\text{area(ROI\_GT) + area(ROI\_Pred) - area(ROI\_I\_GT \cap \_Pred)}}$$

In semantic segmentation, the commonly used accuracy discriminant is mIOU, which formula is as follow:

$$\text{mIOU} = \frac{1}{K+1} \sum_{l=0}^{K} \frac{\text{area(ROI\_GT)}}{\text{area(ROI\_GT) + area(ROI\_Pred) - area(ROI\_I\_GT \cap \_Pred)}}$$

Part of the experimental results are shown in figure 2. The comparison between ResSCNN and other networks is shown in table 3. The server platform was configured as an Intel® Core (R) CPU I7-9700 K @ with a 3.60 GHz processor, 16 GB running memory, 2 TB hard drive capacity, 8 GB NVIDIA GeForce RTX 2080 GPU, and the operating system was Windows 10.

![Figure 2. The Cityscapes data set verified the results, from top to bottom, as follows: Input image, Ground truth and Prediction of ResSCNN.](image)

| model    | DeepLab V2 | PSPNet | SegNet | ERFNet | Fast-SCNN | ResSCNN (Ours) |
|----------|------------|--------|--------|--------|-----------|----------------|
| InputSize| 512×1024   | 713×713| 360×640| 512×1024| 1024×2048 | 1024×2048      |
| FPS      | <1         | <1     | 14.6   | 41.7   | 123.5     | 123.1          |
| Parameters (M) | 44       | 65.7   | 29.5   | 2.1    | 1.11      | 1.11           |
| mIOU (%) | 70.4       | 78.4   | 56.1   | 68.0   | 68.0      | 68.4           |
4. Conclusion
In this paper, inspired by ResNet [11], we combined the double-branched Fast-SCNN with ResNet's residual module and construct the ResSCNN network structure. A large number of experiments were designed to verify the effectiveness of residual module in the SCNN network, and the network experiment results of ResSCNN were compared with those of other networks on Cityscapes dataset.

Acknowledgments
This paper was supported by Key Laboratory of Agricultural Information Engineering of Sichuan Province, This study was supported by Sichuan Agricultural University Innovation and Entrepreneurship Training Program (Grant No. 201910626069).

References
[1] Wang, X. and Z. Jin, Research on the tendency of road traffic accidents in China. Technology & Economy in Areas of Communications. 2019. 21.
[2] Liu, X., Z. Deng, and Y. Yang, Recent progress in semantic image segmentation. Artificial Intelligence Review, 2018. 52(2): p. 1089-1106.
[3] Lateef, F. and Y. Ruichek, Survey on semantic segmentation using deep learning techniques. Neurocomputing, 2019. 338: p. 321-348.
[4] Kuang, H. and J. Wu, Survey of Image Semantic Segmentation Based on Deep Learning. 2019. 55.
[5] Zagoruyko, S. and N. Komodakis, Learning to Compare Image Patches via Convolutional Neural Networks, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015.
[6] Chen, L.-C., et al., DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.
[7] Zhao, H., et al., Pyramid Scene Parsing Network, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.
[8] Badrinarayanan, V., A. Kendall, and R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015.
[9] Romera, E., et al., ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. IEEE Transactions on Intelligent Transportation Systems, 2018. 19(1): p. 263-272.
[10] Poudel, R.P. and S. Liwicki, Fast-SCNN: Fast Semantic Segmentation Network, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019.
[11] He, K., et al., Deep Residual Learning for Image Recognition, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.
[12] Alom, M.Z., et al., A State-of-the-Art Survey on Deep Learning Theory and Architectures. Electronics, 2019. 8(3).
[13] Cityscapes Dataset[Online available]. https://www.cityscapes-dataset.com/