Chen et al. EURASIP Journal on Wireless Communications and Networking (2019) 2019:146
https://doi.org/10.1186/s13638-019-1445-x

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

AtICNet: semantic segmentation with atrous spatial pyramid pooling in image cascade network

Jin Chen[1,2], Chuanya Wang[1,2] and Ying Tong[1,2*]

Abstract

This paper describes a new type of image segmentation method based on deep convolutional neural networks (DCNN) in the actual autonomous driving scene. The spatial pyramid pooling model is used to identify and segment the actual scene to complete the machine-aware task. In order to improve the information aggregation of the whole image, we use atrous convolution for multi-scale feature extraction based on the pyramid structure of image cascade network (ICNet), removing a residual module in the fifth stage of the network, in order to reduce the scale of the convolutional layer. The feature map is preprocessed by padding and atrous convolution before the four-level spatial pyramid model. Then, conventional convolutions are introduced to compose the atrous spatial pyramid pooling (ASPP) structure. Finally, the four feature maps in the pyramid are merged with the feature maps before input into the pyramid. This paper analyzes the spatial pyramid model, receptive field, and dilation convolution in detail and propose atrous image cascade network (AtICNet). Experiment results in the cityscape dataset have shown that AtICNet has some improvements over ICNet, by improving the accuracy of the segmentation.

Keywords: Semantic segmentation, Dilated convolution, Spatial pyramid pooling, Atrous image cascade network

1 Introduction

Deep convolutional neural network has shown strong capabilities in computer vision and machine perception in recent years, including image classification [1], semantic segmentation [2], object detection [3], and other recognition tasks. The emergence of fully convolutional neural network lays the foundation for the current semantic segmentation based on pixel method, and most models are based on this network. ICNet [4] and pyramid scene parsing network (PSPNet) [5] both adopt residual neural network (ResNet)'s basic architecture [6], and use the spatial pyramid pooling model to replace the pooling operation of the last layer. Long et al. [7] describe the internal tension between semantics and location faced by semantic segmentation: what global information solves and where to solve local information. Multi-scale feature extraction of images through local-to-global pyramid aggregates global information and local information.

In the initial network, a deep convolution neural network is applied to semantic segmentation: (1) replacing the original full-connected layer by a series of convolution layers and (2) enlarging the receptive field by dilated convolution to increase the feature pixels. In this paper, we have improved ICNet to reduce the number of residual modules in the fifth stage, which reduces the size of convolution layer and greatly reduces the amount of computation. Atrous convolution is used in the pyramid model, so the context information of the image can be aggregated effectively after multi-scale feature extraction, thus the experimental results with better accuracy than ICNet can be obtained. In fact, each label on the scene analysis contains a strong space correlation. Semantic segmentation needs to understand different categories of spatial information [7–9] to identify similar things.

* Correspondence: tongying2334@163.com
[1]Tianjin Key laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
[2]College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

2 Related work

This paper focuses on the current deep network for semantic segmentation and object detection. In the traditional network structure, the size of the input feature map of fully connected layers is fixed, and the appearance of spatial pyramid pooling (SPP) [10] network has changed this situation. SPP replaces the last pooling layer with a space pyramid pooling, which can generate some fixed space areas and transform them into fixed length vectors. These vectors will be transmitted to the fully connected layer so that it can receive feature maps of any size.

To further aggregate the information of multi-scale feature maps, pyramid scene parsing network (PSPNet) [5] uses several grid scales of spatial pooling to pool feature maps into fixed area blocks and extracts one feature in each area block. These features are connected in series with the original feature map before entering the pyramid pooling model to form a cascade feature graph, thereby aggregating global information.

PSPNet achieves good segmentation results by using the pyramid pooling model, but its segmentation speed is affected by its deep network. Based on this situation, ICNet proposes an image cascade network structure to compress the depth of PSPNet network on the basis of PSPNet. It uses three branches with different resolutions, which can simplify the network structure of PSPNet, improve the running speed, and not excessively reduce the segmentation accuracy. In ICNet, low-resolution and medium-resolution images are separately sent to the first and the second branches. The low-resolution feature maps are obtained by downsampling the medium-resolution feature maps, and their resolutions are 1/4 and 1/2 of the original image resolution, respectively. The network uses an image cascade structure to cascade the feature maps from the first branch and the second branch, so as to achieve the purpose of sharing parameters between the first branch to the second. The third branch is responsible for receiving high-resolution images, and then high-resolution feature maps will be cascaded into the space pyramid together with the feature map obtained by cascading the first two branches.

ICNet uses the pyramid model to average pool the received cascade feature map and fuses the result with the feature map before entering the pyramid. In SPP and PSPNet, the above two are connected in series to form a fixed vector; therefore, the choice of fusion or series connection will have a great impact on the performance of the network.

The new type of method proposed in the paper is named as atrous image cascade network (AtICNet). We add atrous convolution to the four-layer pyramid pooling model of ICnet, which can expand the range of receptive field and obtain more image information by setting different dilated rates to correlate information of different distances. Compared with ICNet's pyramidal pooling model, this method can obtain more global information, so that the final fused feature map contains more spatial pixel information.

3 Method

3.1 Atrous convolution and receptive field

In image segmentation, the steps of using CNN network to segment image are convolution first and then pooling, which reduces the size of image and enlarges the receptive field. Since image segmentation prediction is a pixel-wise output, it is necessary to upsample the pooled image to the size of the original image for preprocessing. It can be seen that there are two key points in this traditional processing method: one is to reduce the size of the image by pooling, and the other is to restore the image to its original size by upsampling. In the above two processes, a lot of useful information will be lost and the ideal segmentation effect will not be achieved.

Long et al. [7] show that atrous convolution can systematically aggregate multi-scale context information without losing resolution.

Atrous convolution is applied to one-dimensional or two-dimensional information input data $x[i]$. After filtering $w[k]$, the output $y[i]$ is obtained as follows.

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

In (1), i is the location of the pixels, r is the dilated rate of the atrous convolution, and k is the size of the convolution kernel. Standard convolution is a special atrous convolution with a dilated rate of 1. Different dilated rates can be set to adjust the range of the receptive field. The smaller the rate, the more detailed the segmentation of the rough feature map, but more time will be spent in training.

For standard $k \times k$ convolution operations, stride is S, which can be divided into three cases:

(1) $S > 1$, which means downsampling while doing convolution, the size of the feature map obtained by convolution will decrease;

(2) $S = 1$, representing the convolution of the normal step size of 1;

(3) $0 < S < 1$, representing the fractionally strided convolution, which is equivalent to upsampling the image. The size of the feature map obtained by convolution will increase. For example, $S = 0.5$ means padding a blank pixel behind each pixel of the image, and the size of the resulting feature map is twice as large as that of the convolution of $S = 1$ under the same conditions.

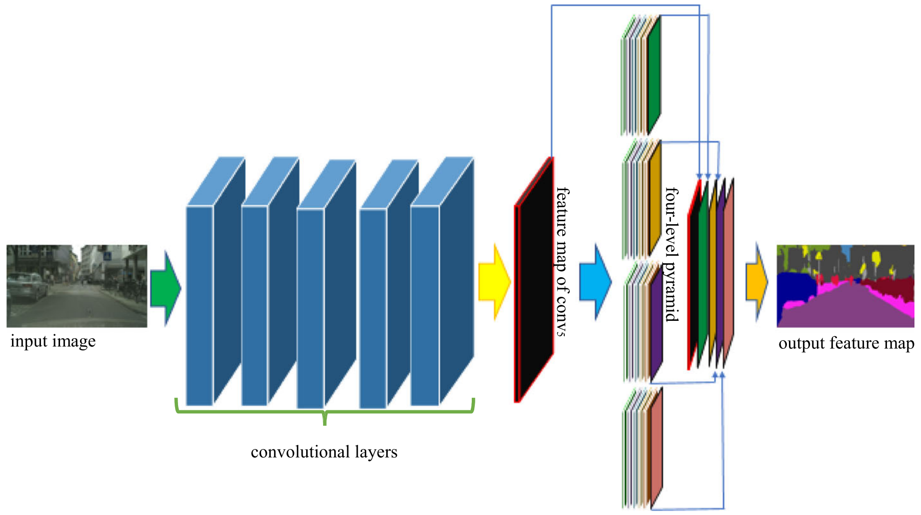

Fig. 1 Spatial pyramid pooling model. Four layers of pyramids are paralleled and feature extraction after the atrous convolution, and finally the information of the aggregated context is cascaded and outputted

Atrous convolution does not pad the blank pixels between the pixels, but skips some pixels on the existing pixels, or keeps the input unchanged, adding some weights of 0 to the parameters of the convolution kernel, so as to expand the receptive field. Of course, the convolution with $S > 1$ can achieve the same effect, but it will do downsampling at the same time of convolution, which will reduce the size of the feature map and is not suitable for use.

If the void rate of a void convolution is r and the size of the convolution nucleus is k, then the size of the receptive field F obtained is:

$$F = (r-1)(k-1) + k \tag{2}$$

We use the parallel atrous convolution layers with different dilated rates in the pyramid model to capture

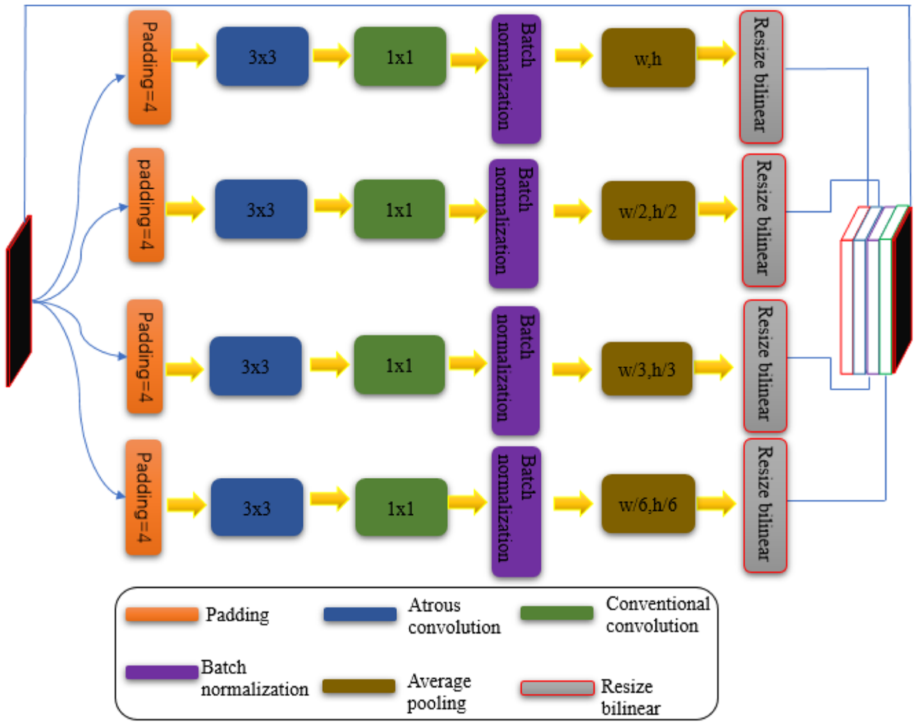

Fig. 2 The internal structure and specific operation of the four-level pyramid

multi-scale information. The smaller rate correlates the nearest pixels, while the larger rate correlates the long-range pixels. Because of the image boundary effect, it cannot capture the remote boundary information accurately in some cases. This design is different from the average pooling output of the feature map input directly into the pyramid model by ICNet. Although the parameters and training time are increased, more detailed information can be obtained to improve the segmentation accuracy.

3.2 Spatial pyramid and multi-scale feature extraction

Input image will get a semantic feature map in ICNet. The method described in [4] is as follows: firstly, a rough prediction map is obtained from a low-resolution image through a complete semantic perception network [11], and then a cascade fusion unit is used to introduce medium-resolution and high-resolution image features, and then the coarse semantic map is gradually improved. As shown in Fig. 1, the feature map of the last residual module of the deep neural network is imported into a pyramid pooling module [10]. Note that these pyramids are parallel and independent of each other. Chen et al. [12] mentioned that the atrous spatial pyramid pooling extracts multiscale features by using multiple parallel filters with different rates. Feature maps can extract four layers of feature information through different sub-regions and at the same time connect the context information of all feature images that are effective for aggregated images. From the previous description, we can see that the SPP and the PSPNet output the four-layer feature map information in series as the final output, while this network adopts a fusion approach, which are two completely different network structures.

At present, many networks have adopted global pooling; for example, Chen et al. [13] mentioned the use of global maximization pooling for target detection under weak supervision, and using global average pooling can not only reduce the size of the model, but also avoid over-fitting [12].

The four-layer space pyramid pooling the feature map is the global average. Max pooling is also a popular pooling method, but it only takes a single maximum, which is not suitable for this network. At each level, we divide the feature map with size of (w, h)(w width, h height) equally. For example, if we divide it into 4 blocks, the size of each block will be $(w/2, h/2)$; if we divide it into 9 blocks, the size of each block will be $(w/3, h/3)$; if we divide it into 36 blocks, the size of each block will be $(w/6, h/6)$. In this paper, we divide the four-layer feature map into 1, 4, 9, and 36 blocks, so that we can get 50 sub-regions and extract 50 features. These features are combined with the feature map of the fifth residual module to form a new feature map for output.

As shown in Fig. 2, we use a four-tier model structure to feed the feature map from the fifth stage into the pyramid model. In the fifth stage, we used two residual modules, one less than ICNet. Low-resolution and medium-resolution feature maps are processed by the first and second branches respectively, and the two branches can share parameters, while high-resolution feature maps are processed by the third branch. In this process, the first and second branches have stored most of the information of the image, so the third branch can use fewer convolution layers to process high-resolution feature maps, thus reducing the computational complexity.

Each layer in Fig. 2 performs atrous convolution, in which the size of the convolution kernel is set to 3 × 3, and the dilated rate of the four-layer atrous convolution is 2, 4, 8, and 12 in turn. Table 1 shows the parameters of the four-tier pyramid.

Atrous convolution with a larger dilated rate ignores the information of small objects and affects the accuracy of network. Therefore, we use atrous convolution with a smaller dilated rate to segment small objects. So the advantage of dilation is that it can increase the receptive field without losing the pooling information, so that the output of each convolution can contain a larger range of information.

Atrous convolution in each layer of the network is associated with multi-scale feature extraction. Different dilated rates are set to correspond to different scale feature maps. Among them, a small dilated rate is used to correlate short-range information, while a large dilated rate is used to correlate long-range information. Each branch is independent of each other. In the final stage of

Table 1 Hyperparameter settings for operation within a four-level pyramid

Level	padding	Atrous convolution						Conventional convolution					Batch normalization	Average pooling	
		k	c	d	p	r	s	k	c	s	b	r	r	k	s
One	4	3	1024	2	Same	Relu	1	1	1024	1	/	/	Relu	w, h	w, h
Two	4	3	1024	4	Same	Relu	1	1	1024	1	/	/	Relu	$w/2,h/2$	$w/2,h/2$
Three	4	3	1024	8	Same	Relu	1	1	1024	1	/	/	Relu	$w/3,h/3$	$w/3,h/3$
Four	4	3	1024	12	Same	Relu	1	1	1024	1	/	/	Relu	$w/6,h/6$	$w/6,h/6$

k kernel, *c* channel, *d* dilated rate, *p* padding, *r* relu, *s* stride, *b* biased the network, the feature maps of different scales obtained from the pyramid model will be fused with the feature map before entering the space pyramid.

4 Experiments results

Our experimental and training learning platform is the TensorFlow deep learning framework. The basic working platform of the experiment is the GPU graphics card of Tesla pc100, which contains 16G memory and is equipped with CUDA8.0 and cudnn6.0. All training, learning, and testing evaluation are carried out on this GPU graphics card.

AtICNet is modified on the basis of ICNet: (1) it reduces a residual module in the fifth stage of ICNet and (2) it adds filling, atrous convolution, 1 × 1 conventional convolution and batch normalization layers before the average pooling of the four-layer pyramid model. This is to enable each layer to capture different range of image information, so as to better aggregate global information.

4.1 Datasets and evaluation metrics

Cityscapes is an image segmentation data set driven by Mercedes-Benz. It is mainly used to evaluate the performance of visual algorithms in urban scene semantic understanding. This data set can provide 1024 × 2048 high-resolution images, including street scenes of 50 cities in different scenes, backgrounds and seasons. It can be divided into 5000 fine-labeled images, 20,000 rough-labeled images and 30 types of labeled objects. Of the 5000 fine-labeled images provided by the Cityscapes dataset, 2975 were used for training, 500 for evaluation, and the remaining 1525 for testing. The data set provides 30 types of data labels, but we only use 19 of them for training and evaluation. The standard of evaluation is mean intersection over unit (mIoU).

Table 2 Without loading the pre-training model, our model mIoU is slightly higher than ICNet

	ICNet	AtICNet
mIoU (%)	24.9	25.1
Training memory (M)	4553	8649
Evaluation memory (M)	7111	7655
Evaluation time (S)	127	136

4.2 Experimental detail and evaluation

In the experiment, we adopted different training strategies and set different hyper-parameters: the size of the input training picture was set to 720 × 720; the mini-batch size was set to 10; the basic learning rate was set to 0.001; the power was set to 0.9; and the momentum and weight attenuation were set to 0.9 and 0.0001, respectively. For a better comparison, the number of iteration steps is set to 5 K. ICNet uses the same hyper-parameter settings as AtICNet in training. In the first training, the pre-training model was not loaded, and the training was carried out from the beginning.

The mIoU in Table 2 is 0.2% higher than that in ICNet when the pre-training model is not loaded for the first time. However, when the iteration is set to 5 K, the training time is longer.

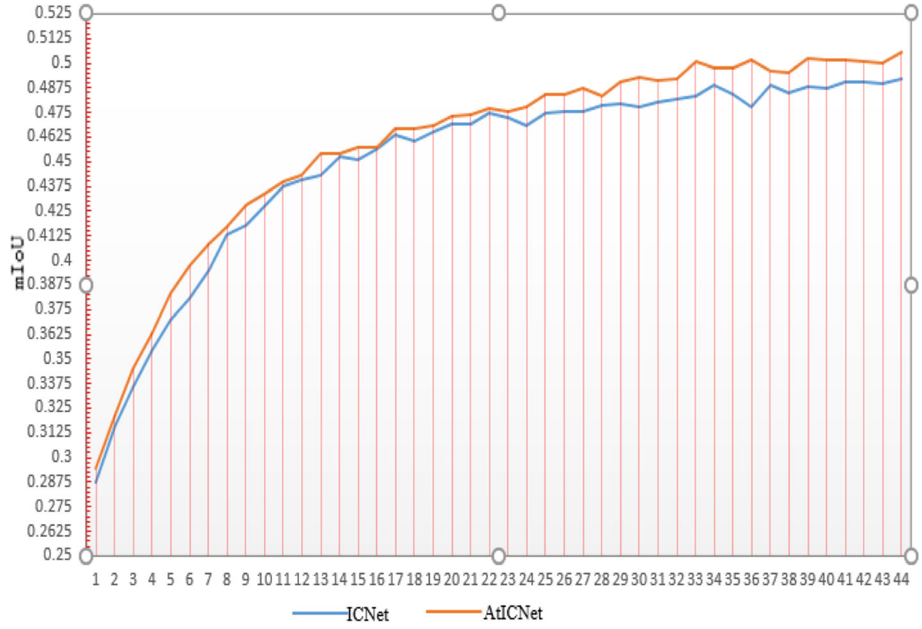

Fig. 3 AtICNet and ICnet network training mIoU. It can be seen that Atrous ICNet's MIoU is higher than ICNet

In the next training, we will take the weight obtained when the pre-training model is not loaded as the weight parameters of the pre-training model, and record it in Fig. 3. Each subsequent training will take the weight obtained from the previous training as the initial weight for training. The same method is used in training ICNet. From Fig. 3, we can see that the mIoU of this network is higher than ICNet in every training, and it is the first network to make the mIoU reach 50%.

We use the model with the last iteration number of 5 K as the training pre-training model, in which the size of the training image is set to 720 × 720, the batch size is set to 16, the iteration number is set to 60 K, and the other hyperparameters remain unchanged. As can be seen from Table 3, the mIoU of AtICNet is higher than that of ICNet in the training process, which indicates that the performance of the improved model is better than that of ICNet.

Table 3 The last training result AtICNet's mIoU is higher than ICNet. The evaluation time is only 11 s longer than that of ICNet

	ICNet	AtICNet
mIoU (%)	48.67	50.19
Training memory (M)	4553	8649
Evaluation memory (M)	7111	7655
Evaluation time (S)	124	135

Figure 4 is a comparison of the segmentation effects of different networks. As can be seen from the figure, AtICNet has some improvements in the details of segmentation compared with ICNet. For example, in the part II, the road and the pedestrians on both sides of the road are clearer in the segmentation results obtained by AtICNet, and for the greening on both sides of the road in part IV, AtICNet is more detailed than ICNet, especially the lawn under the trees.

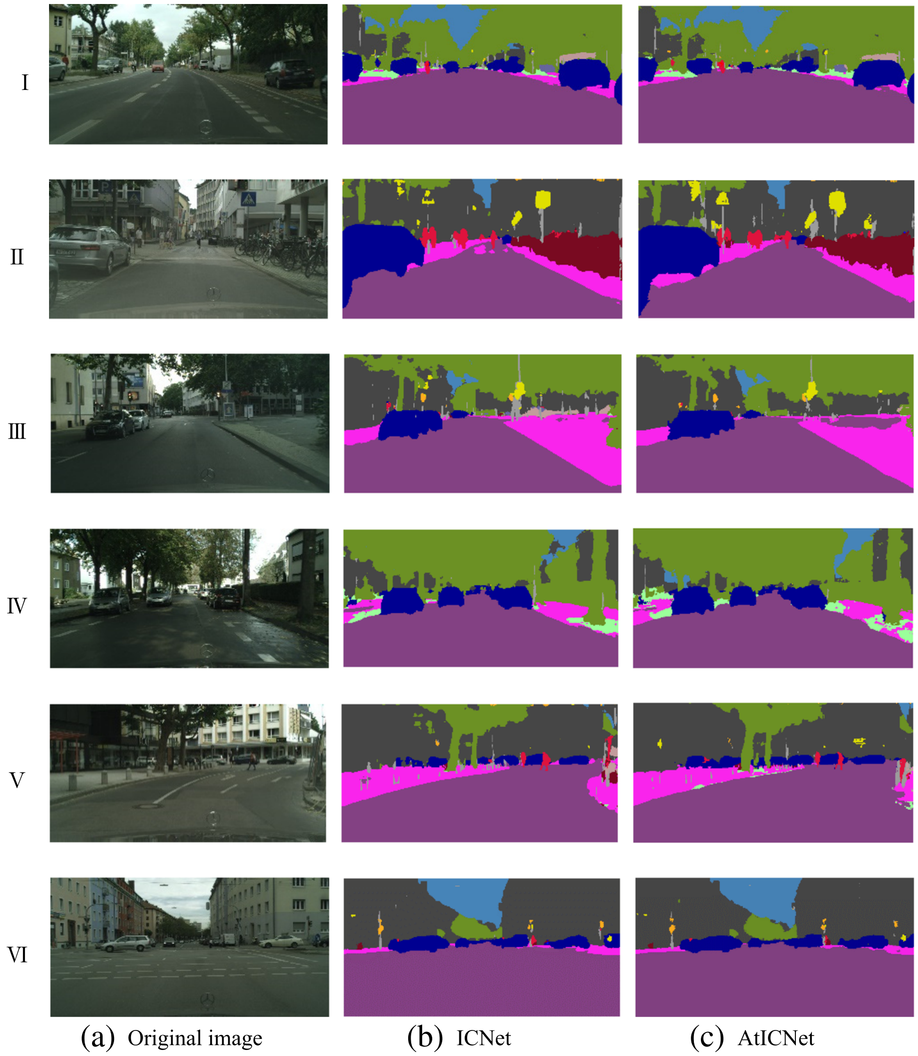

Fig. 4 Comparison of ICNet and AtICNet results. **a** Original image. **b** ICNet. **c** AtICNet

5 Discussion

This paper mainly improves the spatial pyramid pool structure of ICNet. Each layer of the pyramid is reset to (1 × 1), (2 × 2), (3 × 3), and (6 × 6) sub-regions respectively, from which 50 sub-regions can be obtained. A 50-dimensional feature can be obtained by extracting an eigenvalue from each sub-region. Each layer of the improved pyramid model includes padding, atrous convolution, conventional convolution, batch normalization, and average pooling. Finally, the feature maps of different scales obtained by the model are cascaded to realize the aggregation of different kinds of spatial relations. As can be seen from Table 2 and Fig. 3, the mIoU of AtICNet is significantly higher than that of ICNet, reaching 1.53% in Table 3, which indicates that the accuracy of image segmentation has been improved, because more global information can be obtained through a pyramid model. Through training, it is found that the recognition accuracy of the network for small objects has also been greatly improved, which provides good support for automatic recognition and machine perception, so it can be well applied in the field of automatic driving.

Abbreviations

ASPP: Atrous spatial pyramid pooling; AtICNet: Atrous image cascade network; DCNN: Deep convolutional neural networks; FCN: Fully convolutional networks; ICNet: Image cascade network; PSPNet: Pyramid scene parsing network; ResNet: Residual neural network

Acknowledgements

The authors would like to thank prof. F.J. Duan and associate prof. J.J. Jiang from Tianjin University in China for having started the collaboration which lead to this work.

Funding

This work has been partly supported by National Natural Science Foundation of China (Grant No: 61701344), Tianjin Normal University Application Development Foundation (52XK1601), Tianjin Normal University Doctoral Foundation (52XB1603, 52XB1713), and Tianjin Higher Education Creative Team Funds Program in China.

Availability of data and materials

The datasets used and analyzed during the current study are publicly available online.

Authors' contribution

JC gives the overall research direction and ideas, carried out the improved CNN studies, and helped to draft the manuscript. CW read the relevant literature and books and drafts the article and makes the corresponding experimental simulation. YT also gives the original ideas and research direction and makes the corresponding experimental simulation. All authors read and approved the final manuscript.

Authors' information

JC was born in Wuhu, China, in 1976. He received the M.S. degree from Tianjin Normal University and the Ph.D. degree from Tianjin University, in 2005 and 2013 respectively. Since 2005, he has been working at Tianjin Normal University in China. He is an associate professor of Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission. His research interests include image and acoustic signal acquisition and processing, broadband sensor array signal processing, and artificial intelligence.

CW was born in Shandong, China, in 1992. He received the B.S. degree from Qingdao Agricultural University of Haidu College in 2016. He is currently working toward the M.S. degree of Tianjin Normal University. His research interests include image processing and artificial intelligence.

YT was born in Tianjin, China, in 1982. She received the B.S. and M.S. degree from Tianjin Normal University, the Ph. D degree from Tianjin University in 2004, 2007 and 2015 respectively. Since 2007, she has been working at Tianjin Normal University in China. She is a lecturer of Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission. Her research interests include computer vision and digital signal processing.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 26 January 2019 Accepted: 18 April 2019
Published online: 03 June 2019

References

1. A. Krizhevsky, I. Sutskever, G.E. Hinton, in *ImageNet Classification with Deep Convolutional Neural Networks (Harrahs and Harveys,Lake Tahoe, 2012)*. Neural Information Processing Systems (NIPS) (2012)
2. R. Girshick, J. Donahue, T. Darrell, J. Malik, in *Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation (IEEEColumbus, 2014). IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 580–587
3. S.Q. Ren, K.M. He, R. Girshick, J. Sun, *Neural Information Processing Systems (NIPS). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks* (Palais des Congrès de Montréal, Montreal, 2015), p. 2015
4. H.S, Zhao, X.J. Qi, X.Y. Shen, J.P. Shi, J.Y. Jia, *in 2018 European Conference on Computer Vision (ECCV). ICNet for Real-Time Semantic Segmentation on High-Resolution Images* (GASTEIG Cultural Center, Munich, 2018), pp. 405–420
5. H.S. Zhao, J.P. Shi, X.J. Qi, X.G. Wang, J.Y. Jia, in *Pyramid scene parsing network (IEEEHawaii, 2017). IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 2881–2890
6. K.M. He, X.Y. Zhang, S.Q. Ren, J. Sun, in *Deep Residual Learning for Image Recognition (IEEELas Vegas 2016). IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778
7. J. Long, E. Shelhamer, T. Darrell, in *Fully convolutional networks for semantic segmentation (IEEEBoston, 2015). IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3431–3340
8. C. Szegedy, A. Toshev, D. Erhan, in *Deep neural networks for object detection (Harrahs and Harveys,Lake Tahoe, 2013)*. Neural Information Processing Systems (NIPS) (2013)
9. F. Yu, V. Koltun, in *Multi-scale context aggregation by dilated convolutions (Caribe Hilton, San Juan, Puerto Rico, 2016)*. International Conference of Learning Representation (ICLR) (2016)
10. K.M. He, X.Y. Zhang, S.Q. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (Sept. 2015)
11. V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (Dec. 2017)
12. L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (April. 2018)
13. L.C. Chen, Y. Yang, J. Wang, W. Xu, A.L. Yuille. *in* 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Attention to scale: Scale-aware semantic image segmentation (IEEELas Vegas, 2016), pp. 3640–3649