J Inf Process Syst, Vol.19, No.1, pp.98~108, February 2023
https://doi.org/10.3745/JIPS.02.0191

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Deep Learning-Based Image Semantic Segmentation Algorithm

Chaoqun Shen and Zhongliang Sun*

Abstract

This paper is an attempt to design segmentation method based on fully convolutional networks (FCN) and attention mechanism. The first five layers of the Visual Geometry Group (VGG) 16 network serve as the coding part in the semantic segmentation network structure with the convolutional layer used to replace pooling to reduce loss of image feature extraction information. The up-sampling and deconvolution unit of the FCN is then used as the decoding part in the semantic segmentation network. In the deconvolution process, the skip structure is used to fuse different levels of information and the attention mechanism is incorporated to reduce accuracy loss. Finally, the segmentation results are obtained through pixel layer classification. The results show that our method outperforms the comparison methods in mean pixel accuracy (MPA) and mean intersection over union (MIOU).

Keywords

Attention Mechanism, FCN, Image Semantic Segmentation, Skip Structure, VGG16

1. Introduction

Since it was proposed [1], image segmentation has sparked strong scholarly interest at home and abroad for its wide requirements for application. With continuous research and improvement, image segmentation has become an important content in the field of computer vision [2]. In order to meet the needs of various tasks and obtain better segmentation results, a large number of image segmentation algorithms have been proposed so as to achieve better segmentation results and meet the needs of various tasks [3]. The complexity of the scenes and the uncertainty of the image data structure, however, make it difficult to design a general image segmentation method applicable to various tasks [4]. With the deepening of image segmentation awareness, image segmentation algorithms have formed their own systems in different application directions. At present, image segmentation is mainly divided into traditional and deep learning-based segmentation methods [5,6].

Traditional image segmentation algorithms start from image features and use different features between categories in the image to distinguish image categories [7]. Grayscale, color, texture, etc., are some of the commonly used features in traditional image segmentation. The authors of [8] proposed a hierarchical statistical model of image segmentation by combining independent feature extraction. The joint image segmentation and labeling model is further derived based on the segmentation layer, which can accurately

Manuscript received March 10, 2022; first revision May 26, 2022; second revision July 26, 2022; accepted August 4, 2022.
* **Corresponding Author:** Zhongliang Sun (scqbs@163.com)
Henan Mechanical and Electrical Vocational College, Zhengzhou, Henan, China (41240212@qq.com)

approximate the partition function on the block and label. In [9], the authors estimated the statistical information of various segmentation map attributes from a large number of images, and established a Markov random segmentation map model containing statistical data. The segmentation effect, however, is not ideal [10] when the training process is absent. The geometry-based method uses 2D images to infer the spatial layout of 3D images. The effective structure assumption in physics is used to find the best fitting model of the line segment and convert it into a complete 3D model, which solves the problem of target occlusion in semantic segmentation. For example, Di Mauro et al. [11] emphasized that when the geometric shape is better than the method based on object detection, the method based on image segmentation is preferable. This method, however, has certain limitations in image recognition of complex backgrounds. Xie et al. [12] proposed a semantic segmentation method for multi-scale features and context information. A multiple adjacency tree model is proposed to generate adjacency graphs to achieve accurate segmentation of image semantics. However, the extraction of unobvious features needs further research. Convolutional neural networks (CNN) have become the frontier of research [13] thanks to the advances in deep learning technology and robust demand for image semantic segmentation tasks. Image semantic segmentation methods are mainly based on fully supervised learning methods. The training samples provide huge amounts of local features and information, which help to improve segmentation effect. Fully convolutional network (FCN) is an excellent paper of CVPR2015, and is a pioneering work of semantic segmentation technology based on deep learning [14]. FCN replaces the fully connected layers in the CNN network with fully convolutional layers, and constructs an end-to-end, pixel-to-pixel semantic segmentation network. In 2017, PSPNet improved the segmentation problem in the FCN network and proposed a spatial pyramid module [15]. The contextual information and multi-scale information of the image will be extracted, which reduces the probability of mis-segmenting the image category. Le et al. [16] proposed the Graph-FCN method, which uses the graph node classification method in the field of image semantic segmentation. This method expands the receptive field of each node while ensuring local location information, thereby improving the accuracy of the algorithm. Wang et al. [17] designed a multi-scale feature fusion semantic segmentation network Res-Seg based on the residual network, which can improve the segmentation accuracy of the target edge. Li et al. [18] proposed to introduce disparity map into the network to improve accuracy. However, the efficiency of algorithm recognition is still too low. Wang et al. [19] designed a ship target detection method based on CNN. But the speed of segmentation recognition needs to be optimized. The CNN-based semantic segmentation method can automatically learn the features, which is end-to-end [20,21]. These algorithms occupy a large proportion in image segmentation algorithms because they use neural networks to perform image segmentation, and they have high precision and strong anti-noise ability.

To reduce the loss of image feature extraction information and improve the accuracy of semantic segmentation, we propose an image segmentation method based on FCN and attention mechanism. This method uses convolution instead of pooling to deepen the network depth while retaining the feature space dimensionality reduction function. At the same time, the skip structure is used to fuse different levels of feature information during the deconvolution process. Our method proves effective in enhancing the weight of beneficial features.

2. Method

2.1 FCN

As the first image semantic segmentation algorithm the FCN algorithm uses convolution to replace the

fully connected layer. The final feature map reaches the same size as input through upsampling, which can solve the pixel-level segmentation problem. At the same time, the design of this network structure allows FCN to accept pictures of any scale as input without changing the parameters in the network.

In addition to removing the fully connected layer, FCN is innovative in two aspects. First, the upsampling layer is used at the back end of the network to realize the enlargement of the feature map., To improve the receptive field, the pooling layer is used continuously for downsampling when extracting features from the network, which leads to a reduction in the feature map scale. In order to achieve end-to-end semantic segmentation, the FCN algorithm uses deconvolution and bilinear interpolation to upsample the feature map after feature extraction. Secondly, skip structure is adopted to obtain richer features. Although it is possible to get a feature map as large as input through stepwise deconvolution and upsampling, end-to-end semantic segmentation can be achieved. Image details, however, are unattainable for the presence of the pooling layer in feature extraction process. Therefore, directly using the final feature map for segmentation will lead to insufficient accuracy of the result. FCN adds a layer skip connection in the process of deconvolution and upsampling. The feature maps obtained by the last two pooling layers are added to the high-level features as supplementary information, which increases the spatial information to a certain extent and improves the accuracy of the algorithm in edge segmentation.

The FCN algorithm can achieve image semantic segmentation, its segmentation accuracy, however, remains relatively low because the multiple pooling layers make the spatial receptive field of the convolution kernel larger and at the same time cause the loss of spatial information. Even though the layer skip connection is added in the upsampling process, the upsampling process is not sophisticated enough and its ability to restore details is poor. Therefore, there is still a lot of room for improvement in segmentation accuracy.

2.2 Attention Module

Attention is actually universal but overlooked fact. For example, when a bird is spotted flying in the sky, one tends to focus his or her attention on the bird, and naturally the sky becomes background information in the visual system. The attention mechanism allows the system to ignore irrelevant information and focus on key information.

Based on the correlation between features and tasks attention mechanism assigns different weights, and extracts effective information that is beneficial to the task. Such an attention mechanism can expand the receptive field to obtain global information, and adaptively reflect the dependence between different positions according to different inputs so as to ignore useless information and extract key information.

The feature extraction of attention module is described in Fig. 1.

First, the global average pooling (GAP) is used to adjust the dense layer data output from $W\times H\times K$ to $1\times 1\times K$. The equation is as follows:

$$z_c = f_1(u_c)\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}u_c(i,j) \tag{1}$$

where f_1 represents GAP function.

The second step is to continue two FC processes. C is the dimensionality reduction coefficient. Performance is best when C=16. The calculation is as follows:

$$s_c = f_2(z, W) = \sigma\big(g(z, W)\big) = \sigma\big(W_2\delta(W_1 z)\big)$$
$$W_1, W_2 \in R^{\frac{K}{C}\times K} \tag{2}$$

where σ and δ are the activation functions of sigmoid and ReLU, respectively.

Then scale and adjust the data to $W\times H\times K$.

$$X_c = f_3(u_c, s_c) = s_c \cdot u_c \tag{3}$$

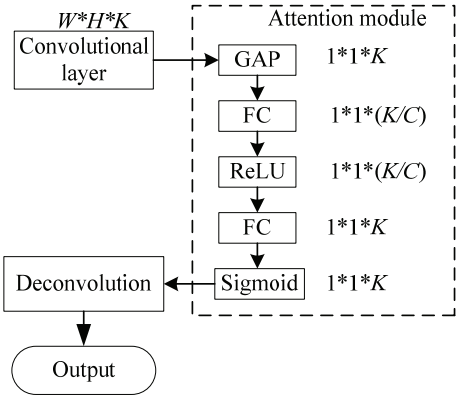


Fig. 1. Feature extraction part.

2.3 Network Structure Design

The current CNN and FCN use pooling operations when processing image tasks, which leads to the loss of the image feature information during the processing. To remove the pooling layer and retain its function of reducing the dimensionality of the feature space, we propose a semantic segmentation network using a common convolutional layer with a larger step size instead of pooling layer. Since the VGG16 image is reduced by a factor of 32 after five poolings, the encoder is based on the first five layers of VGG16. Deconvolutional units are used for upsampling of the decoder network. And the attention module is integrated into the decoding network. Finally, the pixel-level classification layer is integrated into the decoding network. The structure is described in Fig. 2.

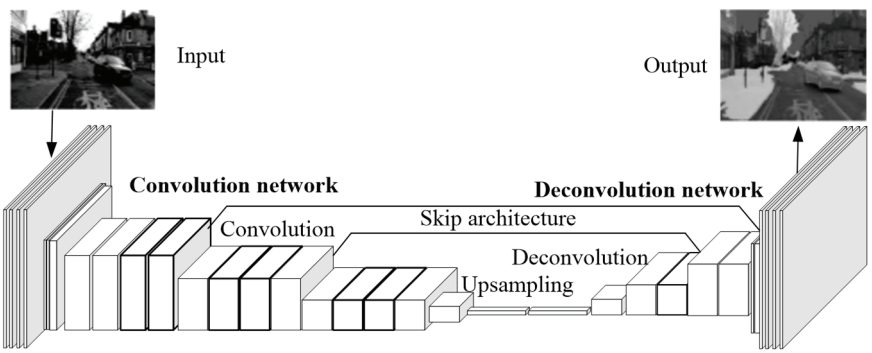


Fig. 2. Structure of our segmentation network.

Before the network is added to attention module, the batch size is 256, the input image size is 224×224, and one forward propagation process time is 42 ms. After passing the attention module, the time increases to 47 ms. After we reduce and upgrade the dimension of the network and reduce the complexity, there is still possibility for the time to increase, but compared with performance improvement, it can be ignored [22,23].

Also, each layer is linked by a skip connection, which differs from DenesNet. One difference is that DenesNet only has connections in downsampling, while the proposed method has connections in all layers. The other lie in pooling, and the proposed network uses Atrous spatial pyramid pooling (ASPP). Image features are combined with GAP to increase global context. The schematic diagram of network skip connection is described in Fig. 3.

To verify the complexity of our model, the experiment compares the parameters of the network without jump-layer connection with the improved network in this paper. The parameter volume of the network without using skip layer connection is 1.58M, and the parameter volume of the improved network is 1.63M. The improved network does not significantly increase the complexity of the network.

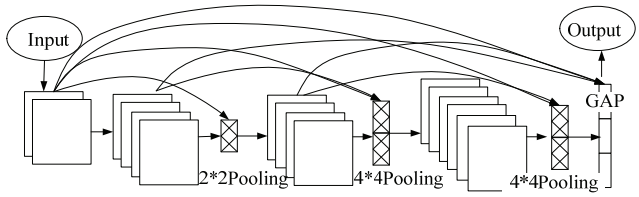


Fig. 3. Structure of the network skip connection.

3. Experiment and Analysis

The image semantic segmentation network is constructed through TesorFlow deep learning framework released by Google. The main experimental environment is listed in Table 1.

Table 1. Environment parameters

Parameter	Configuration
System	Ubuntu 18.04
GPU	RTX2080Ti
CPU	i7-8700k
RAM	12G
Python	3.6

3.1 Network Parameter

The batch size is 256. The learning rate is initial 1×10^{-3} multiplied by $(1-current\_iter/max\_iter)^{power}$ reduction strategy. Where, $power=0.95$ and $current\_iter$ are current iteration times, max_iter is the maximum iteration times.

3.2 Evaluation Index

m_{ii} represents the correct number of divisions. m_{ij} is the number of pixels that at first belonged to class i, but became class j. m_{ji} is the number of pixels that at first belonged to class j but became class

i. A total of $k+1$ classes.

Pixel accuracy (PA) is proportion of correctly marked pixels to the total pixels:

$$PA = \frac{\sum_{i=0}^{k} m_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} m_{ij}} \tag{4}$$

Mean pixel accuracy (MPA) is an improvement of PA.

$$MPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{m_{ii}}{\sum_{j=0}^{k} m_{ij}} \tag{5}$$

Mean intersection over union (MIOU) sums and averages the ratio of the intersection of each type of predicted result and the true value to the union.

$$MIOU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{m_{ii}}{\sum_{j=0}^{k} m_{ij} + \sum_{j=0}^{k} m_{ji} - m_{ii}} \tag{6}$$

3.3 Cityscapes Dataset

Cityscapes is a dataset of semantic understanding images, which has 5,000 high-quality pixel-level annotated pictures of driving scenes (2,975 for train, 500 for val, 1,525 for test, 19 classes in total). Also, it has 20,000 coarsely annotated pictures (gt coarse). The annotation map of images is described in Fig. 4.

Fig. 4. Some annotation images of the Cityscapes dataset.

Fig. 5 describes the changes in evaluation index of validation set during training process. As can be seen from Fig. 5, the evaluation index shows very fast convergence, and achieves better results.

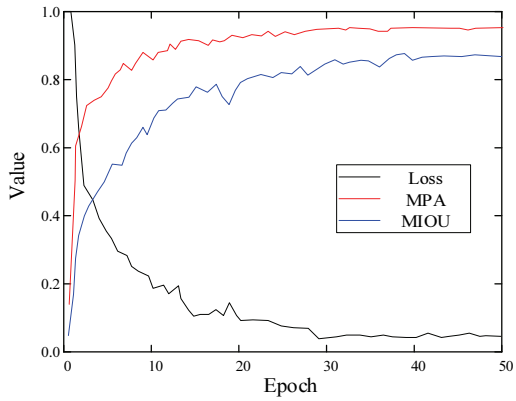
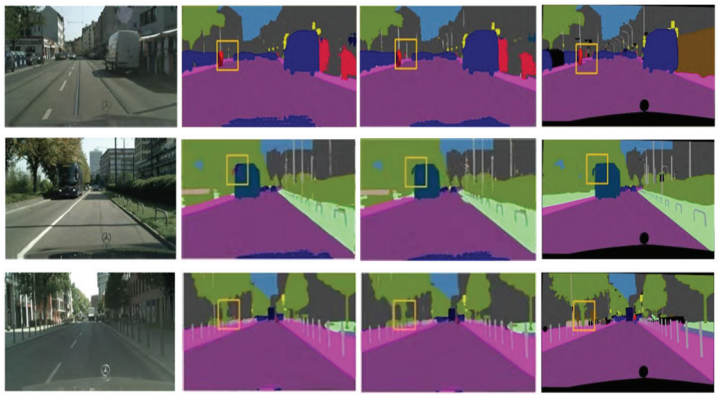


Fig. 5. Changes in indicators during training.

Fig. 6 shows a comparison of segmentation result between the improved model in this paper and the original FCN.



(a) (b) (c) (d)

Fig. 6. Comparison of results between the improved model and the original FCN in Cityscapes dataset: (a) original image, (b) FCN, (c) improved FCN, and (d) manual making.

Fig. 6 shows that the original FCN can hardly segment objects such as pedestrians, rearview mirrors and other small targets. And the segmentation results of the original FCN contain many interruptions and discontinuities. Our methodcan segment objects such as pedestrians or rearview mirrors more accurately than the original FCN, and is, therefore, more accurate and continuous for small target segmentation. However, a small number of segmentation errors may occur to small objects, such as discontinuous segmentation of grass. Relatively speaking, the improved segmentation network has achieved satisfactory results. The performance of our method is compared to several studies [8,12,15,17,18], which is shown in Table 2.

Table 2. Results on Cityscapes dataset

Study	MPA	MIOU
Ion et al. [8]	0.793	0.756
Xie et al. [12]	0.859	0.795
Zhao et al. [15]	0.846	0.781
Wang et al. [17]	0.891	0.821
Li et al. [18]	0.881	0.812
Proposed study	0.907	0.829

As can be seen from Table 2, our algorithm achieves the largest value in both MPA and MIOU, reaching 0.907 and 0.829, respectively. Ion et al. [8] further derives the joint image segmentation and labeling by combining the image segmentation statistical model of independent feature extraction to achieve image semantic segmentation. But this method is more traditional and simple. Xie et al. [12] uses multi-scale features and context information to achieve semantic image segmentation, but the extraction of unobvious features needs further optimization. Other CNN comparison methods cause loss of image feature information to varying degrees during the training process and fail to achieve optimal results.

3.4 PASCAL VOC2012 Dataset

PASCAL stands for Pattern Analysis, Statistical Modelling and Computational Learning. The full name of VOC is Visual Object Classes. The PASCAL VOC competition is a world-class computer vision challenge, namely, to provide pictures and corresponding labels, and use these data to achieve three tasks: image classification, target detection and recognition, and image segmentation. There are a total of four classes and 20 small classes of objects in these images. In the image segmentation competition, each pixel of the image needs to be classified according to the corresponding 20 categories, otherwise it is classified as "background." One of the labels is shown in Fig. 7.

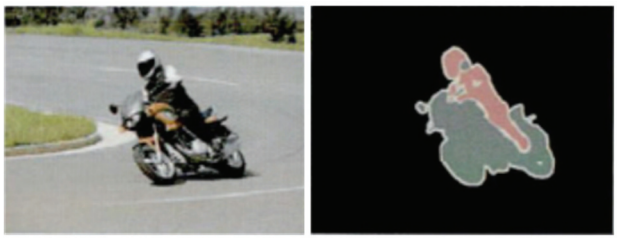


Fig. 7. Partial image labels of PASCAL VOC2012 dataset.

In the experiment, training ceased after 50 epochs. The changes in evaluation index during the training process are described in Fig. 8. As can be seen from Fig. 8, the evaluation index shows very fast convergence, and achieves better results.

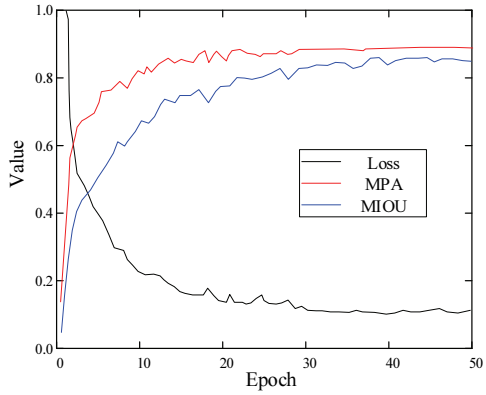


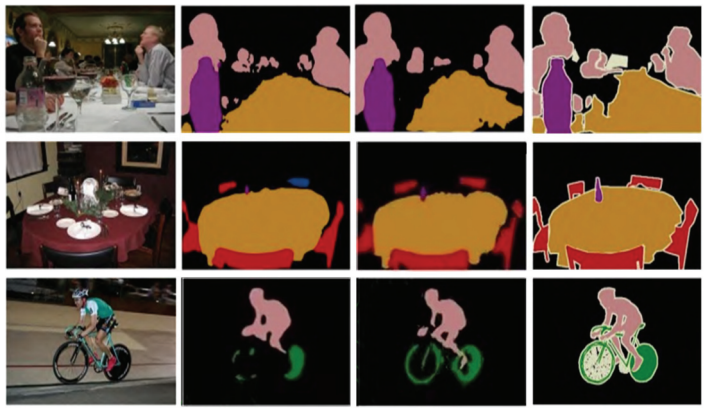
Fig. 8. Changes in indicators during training.

Fig. 9 compares our improved model and the original FCN segmentation results, which shows our algorithm is effective in segmentation results.

It can be found that the original FCN segmented part of the chair incorrectly, and the human legs and bicycle wheels were segmented only partially. The segmentation of small objects and image details is far from satisfactory with many interruptions and discontinuities in the original FCN segmentation results. While the improved segmentation network is more accurate and continuous in segmentation of small targets with all the chairs divided correctly and the edges of the bicycle tires and human legs well divided. This clearly shows that the network in this article improves the segmentation effect.

Our method is compared with the studies [8,12,15,17,18] in segmentation performance. The results on PASCAL VOC2012 dataset are described in Table 3.

Ion et al. [8] further derives the combined image segmentation and labeling by combining the image segmentation statistical model of independent feature extraction to achieve image semantic segmentation. However, this method has a simple structure and poor segmentation effect, and MIOU is only 0.751. Xie et al. [12] uses multi-scale features and context information to achieve semantic image segmentation, but the extraction of unobvious features needs further optimization. Therefore, the MIOU is lower than 0.800. Other CNN comparison methods cause loss of image feature information to varying degrees during the training process and fail to achieve optimal results. The proposed method improves both precision and performance.



(a) (b) (c) (d)

Fig. 9. Comparison of the segmentation results between improved model and original FCN: (a) original image, (b) FCN, (c) improved FCN, and (d) manual marking.

Table 3. Results on PASCAL VOC2012 dataset

Study	MPA	MIOU
Ion et al. [8]	0.808	0.751
Xie et al. [12]	0.865	0.794
Zhao et al. [15]	0.871	0.814
Wang et al. [17]	0.895	0.821
Li et al. [18]	0.887	0.819
Proposed study	0.902	0.827

4. Conclusion

Given the fact that small-scale target image features are hard to extract and are prone to mis-segmentation as well,, we use a simple encoder and decoder structure to perform segmentation on target image based on the in-depth study of VGGNet and FCN network models. Among them, the first five layers of VGG16 are used as encoding part of the segmentation structure, and the up-sampling and deconvolution unit of the FCN is used as the decoding part of the segmentation network. And the attention mechanism is integrated to reduce accuracy loss. The image feature information obtained through coding

part analysis is processed by pixel layer classification to generate the segmentation results. Results show that our method achieves higher segmentation results, and outperforms other comparison methods in MPA and MIOU on Cityscapes and PASCAL VOC2012 datasets.

Although our method can achieve better segmentation effects than fully convolutional neural network, it has defects in the semantic segmentation of smaller objects and does not demonstrate whether the processing speed meets demand. Therefore, how to improve its detection effect on small objects segmentation and the speed of network operation is the focus of our follow-up research.

Acknowledgement

This work is supported by 2019 training plan for young backbone teachers in Henan Higher Vocational School (No. 2019GZGG100).

References

[1] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834-848, 2018.

[2] H. Yu, Z. Yang, L. Tan, Y. Wang, W. Sun, M. Sun, and Y. Tang, "Methods and datasets on semantic segmentation: a review," *Neurocomputing*, vol. 304, pp. 82-103, 2018.

[3] S. Cui, C. Liu, M. Chen, and S. Xiong, "Brain tumor semantic segmentation from MRI image using deep generative adversarial segmentation network," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 9, pp. 1913-1919, 2019.

[4] M. Shahzad, A. I. Umar, M. A. Khan, S. H. Shirazi, Z. Khan, and W. Yousaf, "Robust method for semantic segmentation of whole-slide blood cell microscopic images," *Computational and Mathematical Methods in Medicine*, vol. 2020, article no. 4015323, 2020. https://doi.org/10.1155/2020/4015323

[5] S. Ghosh, A. Pal, S. Jaiswal, K. C. Santosh, N. Das, and M. Nasipuri, "SegFast-V2: semantic image segmentation with less parameters in deep learning for autonomous driving," *International Journal of Machine Learning and Cybernetics*, vol. 10, pp. 3145-3154, 2019.

[6] Q. Ning, J. Zhu, and C. Chen, "Very fast semantic image segmentation using hierarchical dilation and feature refining," *Cognitive Computation*, vol. 10, pp. 62-72, 2018.

[7] Z. Lu and D. Chen, "Weakly supervised and semi-supervised semantic segmentation for optic disc of fundus image," *Symmetry*, vol. 12, no. 1, article no. 145, 2020. https://doi.org/10.3390/sym12010145

[8] A. Ion, J. Carreira, and C. Sminchisescu, "Probabilistic joint image segmentation and labeling by figure-ground composition," *International Journal of Computer Vision*, vol. 107, pp. 40-57, 2014.

[9] E. Akbas and N. Ahuja, "Low-level hierarchical multiscale segmentation statistics of natural images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1900-1906, 2014.

[10] C. Menart, J. W. Davis, M. N. Akbar, and R. Ilin, "Scene-based priors for bayesian semantic image segmentation," *International Journal of Smart Security Technologies (IJSST)*, vol. 6, no. 1, pp. 1-14, 2019.

[11] D. Di Mauro, A. Furnari, G. Patane, S. Battiato, and G. M. Farinella, "Estimating the occupancy status of parking areas by counting cars and non-empty stalls," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 234-244, 2019.

[12] J. Xie, L. Yu, L. Zhu, and X. Chen, "Semantic image segmentation method with multiple adjacency trees and multiscale features," *Cognitive Computation*, vol. 9, no. 2, pp. 168-179, 2017.

[13] X. Xia, Q. Lu, and X. Gu, "Exploring an easy way for imbalanced data sets in semantic image segmentation," *Journal of Physics: Conference Series*, vol. 1213, no. 2, article no. 022003, 2019. https://doi.org/10.1088/1742-6596/1213/2/022003

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, 2017.

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," 2016 [Online]. Available from: https://arxiv.org/abs/1612.01105.

[16] Y. Lu, Y. Chen, D. Zhao, and J. Chen, "Graph-FCN for image semantic segmentation," in *Advances in Neural Networks – ISNN 2019*. Cham, Switzerland: Springer, 2019, pp. 97-105.

[17] Y. Wang, J. Han, J. Lu, L. Bai, and Z. Zhao, "TIG stainless steel molten pool contour detection and weld width prediction based on Res-Seg," *Metals*, vol. 10, no. 11, article no. 1495, 2020. https://doi.org/10.3390/met10111495

[18] L. H. Li, B. Qian, J. Lian, W. N. Zheng, and Y. F. Zhou, "Study on semantic image segmentation based on convolutional neural network. *Journal of Intelligent & Fuzzy Systems*, vol. 33, no. 6, pp. 3397-3404, 2017.

[19] W. Wang, Y. Fu, F. Dong, and F. Li, "Semantic segmentation of remote sensing ship image via a convolutional neural networks model," *IET Image Processing*, vol. 13, no. 6, pp. 1016-1022, 2019.

[20] V. Romanuke, "A generator of a toy dataset of multi-polygon monochrome images for rapidly testing and prototyping semantic image segmentation networks," *The Scientific Journal of Riga Technical University-Electrical, Control and Communication Engineering*, vol. 15, no. 2, pp. 54-61, 2019.

[21] R. Jayadevan and V. S. Sheeba, "A semantic image retrieval technique through concept co-occurrence based database organization and DeepLab segmentation," *Journal of Computer Science*, vol. 16, no. 1, pp. 56-71, 2020.

[22] H. Zhou, A. Han, H. Yang, and J. Zhang, "Edge gradient feature and long distance dependency for image semantic segmentation," *IET Computer Vision*, vol. 13, no. 1, pp. 53-60, 2019.

[23] F. Jiang, A. Grigorev, S. Rho, Z. Tian, Y. Fu, W. Jifara, K. Adil, and S. Liu, "Medical image semantic segmentation based on deep learning," *Neural Computing and Applications*, vol. 29, pp. 1257-1265, 2018.

Chaoqun Shen **https://orcid.org/0000-0002-6955-2271**

She was born in May 1982 with a master's degree. She graduated from Harbin University of Science and Technology in 2008 and is currently an associate professor. Her main research fields are image processing and intelligent control.

Zhongliang Sun **https://orcid.org/0000-0002-6324-928X**

He was born in September 1965 with the highest degree of doctor. He graduated from Xi'an Jiaotong University in 2009 and currently serves as a senior professor, engineer/professor. His main research fields are network manufacturing and intelligent manufacturing.