Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **6**, Number **4**, November **2021**

Semantic segmentation based on Improved Pyramid Scene Parsing Network

Ren-Jie Song

Department of Computer Science and Technology
Northeast Electric Power University
169 Changchun Road, Chuanying District, Jilin City, Jilin Province
1939811347@qq.com

Fan Zhang*

Department of Computer Science and Technology
Northeast Electric Power University
169 Changchun Road, Chuanying District, Jilin City, Jilin Province
Corresponding author: Zhangf_08@163.com

Kwang Ho Park

Department of Electrical and Computer Engineering
Chungbuk National University
Cheongju 28644, Republic of Korea
khblack@dblab.chungbuk.ac.kr

Received April 2021; revised July 2021

Abstract. *To address the problems of semantic segmentation in practical applications such as the difficulty in achieving precise segmentation of small target objects and the significant performance degradation when facing data that is quite different from the training set, we proposes a semantic segmentation method based on improved Pyramid Scene Parsing Network (PSPNet). First, a style transfer module is added before semantic segmentation to achieve style transfer from the training set to the current task, and the generated composite image is used as the input of semantic segmentation, so as to improve the generalization ability of semantic segmentation network for different data. Then, the semantic segmentation part is improved, and the SE structure is integrated for weighted processing of each channel in the pooling result in PSPNet, which can increase useful features, so as to effectively improve the segmentation precision for small targets. We conducted comparative experiments on the dataset SYNTHIA, and the experimental results proved the effectiveness of our improved method.*
Keywords: semantic segmentation, style transfer, pyramid pooling, SE structure

1. **Introduction.** The semantic segmentation of image is an important branch in the field of computer vision. It refers to classifying and labeling each pixel according to the semantic information contained in image, so as to segment the entire image into various regions with different semantic meaning, which has broad applications in realistic scenarios such as medical treatment [1], autonomous driving [2] and indoor navigation.Traditional image segmentation is generally achieved by setting thresholds [3-6] or edge detection [7-10]. Although these methods are relatively simple, they need to design image features manually, and the segmentation results have no semantic annotation. With the fast development of deep learning in recent years, the semantic segmentation of image has

become a hot issue. Scholars at home and abroad have done a lot of research on how to apply convolutional neural networks [11] used in prediction [12] and other fields to image semantic segmentation.Based on the improved CNN, Fully Convolutional Networks [13] (FCN) realizes pixel-level classification for the first time, which is the pioneering work of semantic segmentation. FCN replaces the last fully connected layer in the traditional convolutional neural network with convolutional layer, and then upsamples the convolutional feature map to achieve deconvolution. Finally, it is classified through the SoftMax layer and the segmentation results are output. Many researchers have made improvements based on FCN, and many advanced semantic segmentation networks have been proposed since then.To address the problem that FCN usually extracts local features of images for pixel classification and lacks the use of global features of images, Liu et al. [14] proposed ParseNet in 2016. The network first obtains global features of images through global average pooling, and then fuses global features with local features for classifier learning, so as to the global context information is used effectively to improve the accuracy of semantic segmentation. But for images with many kinds of objects, global average pooling may make the image lose its spatial correlation and cause blur.To solve this problem, Zhao et al. [15] proposed the Pyramid Scene Parsing Network (PSPNet). In this network, the global average pooling of image is changed to the pooling of four different scales, so as to obtain the feature maps at four different levels, and various feature maps are fused after up-sampling. By fusing features of different scales and combining the local and global context information, better segmentation results can be achieved. In the structure pre-trained with dataset MS-COCO [16] and built with ResNet101 [17], the mean intersection-over-union (MIoU) could reach 85.4%, which is at an advanced level. Even though can provide good segmentation results in general, its performance will decline significantly when applied to data remarkably different from the training set. In addition, because PSPNet does not conduct fusion based on the importance of information when combining local and global information, it may lead to loss of partial feature information, so it cannot provide a precision during segmentation of small targets.

To address the problems described above, we proposes a semantic segmentation method based on improved PSPNet. Our main contributions are summarized below: (1) First of all, before semantic segmentation, the AdaIN [18] method is used to achieve style transfer between the training set and current task under different environment, and the composite image after style transfer is used as the input of segmentation network, so that the segmentation network can maintain its original high performance when processing data that significantly different from the training set. (2) In view of the problem that it is easy to lose feature information when combining local information and global information in PSPNet, we improve it to achieve more accurate semantic segmentation of small target objects: We use the same pyramid pooling structure as pspnet to pool the input features at different scales, but we use different processing methods from pspnet when fusing the pooling results at different scales. PSPNet only simply fuses the pooled results of different scales, but we use SE structure to weight the different channels of the fusion results, so that the features can be promoted or inhibited according to their importance, which can improve the segmentation precision for small targets. (3) We use Miou as the standard to experient on the SYNTHIA [19] subset, and the results verify the effectiveness of our semantic segmentation method compared with the original pspnet.

2. **Design idea of the semantic segmentation method based on improved PSPNet.**

The network structure we proposed is as shown in Figure 1, which mainly consists of two parts: the image style transfer network and the image segmentation network. After the image is input into the network, it will go through style transfer first, which will make

the input image transfer to the style of the training set of this network, but it will also be able to maintain its original contents; then, the improved PSPNet is used to conduct semantic segmentation of the image after style transfer. Because the composite image may lose some details of the original image, which may cause performance degradation of segmentation network, the result of the segmentation needs to be processed. Semantic segmentation is also conducted to the original image and the KL (Kullback–Leibler) divergence is used to restrain the segmentation results of composite image and original image, which can reduce the segmentation error caused by composite image.

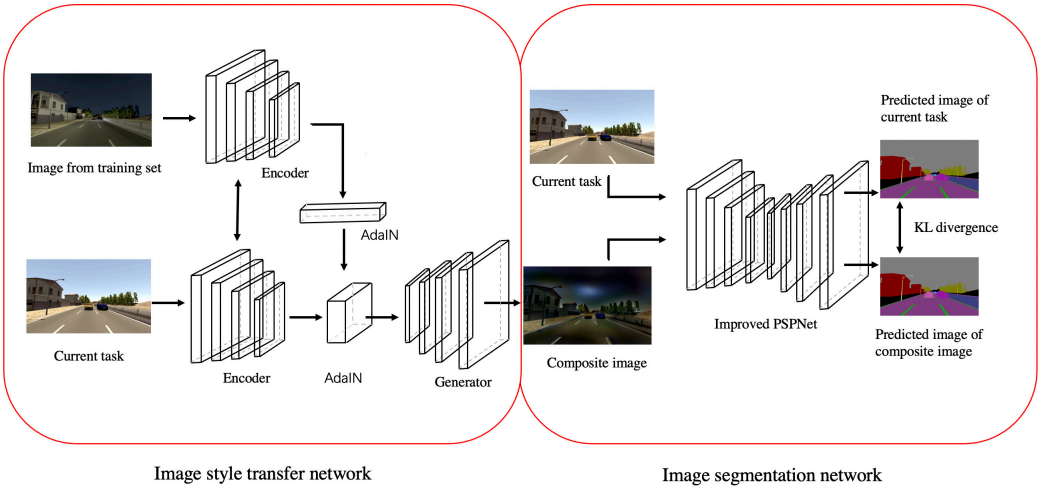

FIGURE 1. Overall structure of the network designed in our idea

3. **Image style transfer network.** The image style transfer network aims to realize style transfer between the training set and current task, which consists of the three parts of encoder, AdaIN module and generator. First, the pre-trained RESNet50 network is used as the encoder to extract feature maps $Z0$ and Zi of current task $X0$ and image Xi from the training set, respectively; then, the AdaIN module is used to change the information of feature map $Z0$ of current task to $\widehat{Z0}$, and to normalize the content of feature map.$\widehat{Z0}$ represented as:

$$\widehat{Z0} = AdaIN(Z0, Zi) = \alpha(Zi)\frac{Z0 - \mu(Z0)}{\alpha(Z0)} + \mu(Zi) \tag{1}$$

where, the α function calculates the mean value of each channel, and the μ function calculates the variance of each channel. Finally, a network structure symmetrical to encoder is used as the generator, and feature map $\widehat{Z0}$ after modification is transformed to composite image $\widetilde{Z}$ after style transfer.The weighting parameters in the generator are trained to minimize the loss function, so that feature map $Z0$ of current task can maintain its own content as much as possible, but its image style can be close to the style of feature map Zi of images from the training set. The loss function is represented as:

$$f = \| \widetilde{Z} - (\widehat{Z0}) \|_2 + \| \mu(\widetilde{Z}) - \mu(Zi) \|_2 + \| \alpha(\widetilde{Z}) - \alpha(Zi) \|_2 \tag{2}$$

The first term of the function is to calculate the difference between the composite image feature of current task and the feature after style transfer, and the purpose is to maintain the content of current task. The other two terms aim to calculate the differences between the composite image and the training set image in terms of mean value and variance, force

the style of composite image to be close to that of the training set image, and complete style transfer from the training set image to current task.

4. **Improved PSPNet.**

4.1. **PSPNet.** Because the common pooling method adopted in traditional semantic segmentation involves a fixed receptive field, it can't well utilize the context information during segmentation, and the pixel classification can only be conducted based on the local features of image, so that it may miss the target during segmentation. The proposal of PSPNet has solved the problem mentioned above.The network structure is shown in Figure 2.This network can extract the context information of different scales through different scales of pooling, and then fuses feature maps of different scales. This network can well utilize the global and local features to improve the segmentation precision for targets of different sizes.

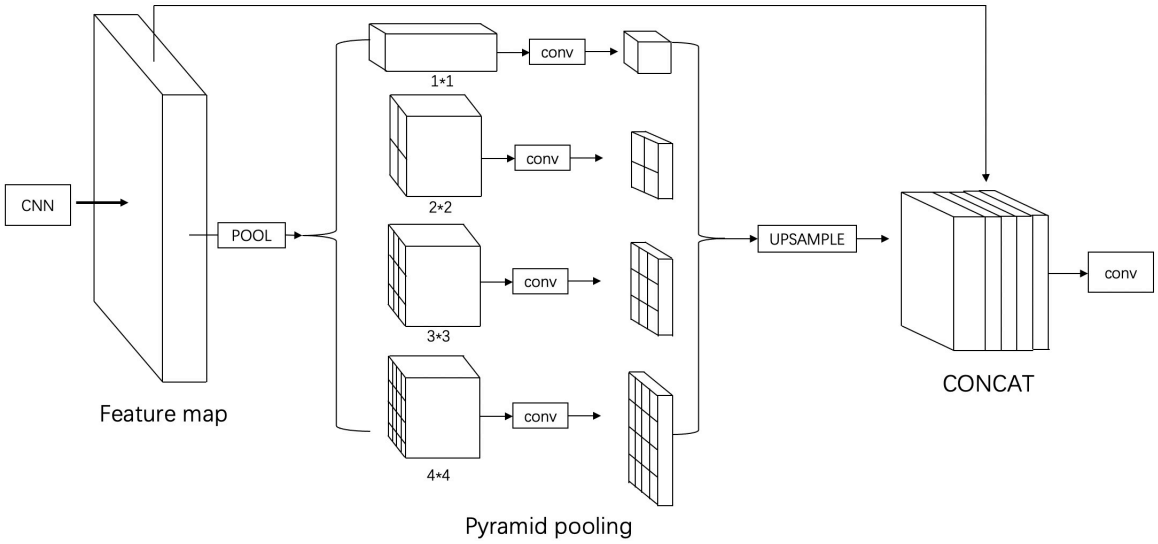

FIGURE 2. Structure of PSPNet

As shown in Figure2, after the image is input, the feature map is extracted from the input image first, and then pooling of four scales is used on the feature map. On the top of pyramid, the global average pooling is used to extract the global features of feature map; for the rest three layers, pooling of scales 2×2, 3×3 and 4×4 is used successively to extract the features of different sub-regions on the feature map.Convolution, batch normalization and relu are performed for different sizes of features in turn to learn local feature information and reduce the dimension of the feature map. The learned feature maps of different scales are up-sampled to the size of the input feature map, and the pyramid feature map is superimposed with the original feature map channel. Convolution and up-sampling are performed on the result to obtain the final segmentation map.Pyramid pooling can effectively fuse the global features and multi-scale local features of image, which can become the global priori information and effectively reduce the segmentation error.

4.2. **Improved PSPNet.** Integrating attention mechanism into convolutional neural network has increasingly become an important method for semantic segmentation to strengthen feature learning, which was first applied to the field of machine translation

and achieved remarkable results [20]. Since then, networks integrating attention mechanism have been proposed, such as Senet [21], Ocnet [22], Danet [23]. We consider that PSPNet can't fuse information according to the importance of local and global information, the SE (Squeeze and Excitation) structure is integrated to improve the PSPNet. The SE structure was proposed to address the problem that the traditional convolutional network can only conduct feature fusion based on the spatial dimension, while ignoring the relation among feature channels. The SE structure mainly includes the two parts of Squeeze and Excitation. By re-calibrating the channel features, the features of important channels are promoted and the insignificant features are suppressed. See Figure 3 for the structure of SE:

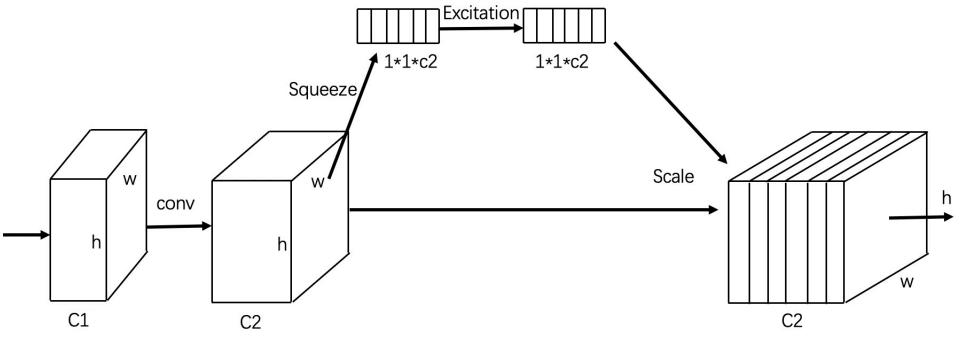

FIGURE 3. Structure of SE

Input feature map X, $X \in R^{H\times W\times C}$, in which, H, W and C refer to the height, width and channel number of feature map, respectively. First of all, conduct convolution operation of input, and the input channel number changes from $C1$ to $C2$; then, conduct Squeeze operation to the result after convolution, i.e., global average pooling, and generate $y \in R^{C\times 1}$,in which, y_m is the m_{th} element of y, and X_m is the m_{th} feature map of m.

$$y_m = F_s(X_m) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_m(i,j) \tag{3}$$

The global descriptive features are obtained through the Squeeze operation. Then, grasp the relation among channels via the Excitation operation. The Excitation operation mainly consists of two fully-connected layers with parameters W_1 and W_2 and two activation functions, which is represented as:

$$\widetilde{y} = F_e(y, W) = \alpha(g(y, W)) = \alpha(W_2\delta(W_1 y)) \tag{4}$$

where, α represents the sigmoid activation function, while δ represents the ReLU [24] activation function. Obtain a number within the range of $(0, 1)$ through calculation with sigmoid function, which is the corresponding weight of each channel; finally, conduct reweight operation to multiply the weights obtained from the Excitation operation with previous features channel by channel, so as to complete the recalibration work in the channel domain and generate the recalibrated feature map $\widetilde{X}$:

$$\widetilde{X_k} = F_{scale}(X_m, y_m) = X_m \cdot y_m \tag{5}$$

where, $F_{scale}(X_m, y_m)$ refers to the channel-by-channel multiplication, while $\widetilde{X_k}$ is the k_{th} feature map of $\widetilde{X}$.

In this paper, the segmentation network obtained after improvement of PSPNet by integrating the SE module is as shown in Figure 4:

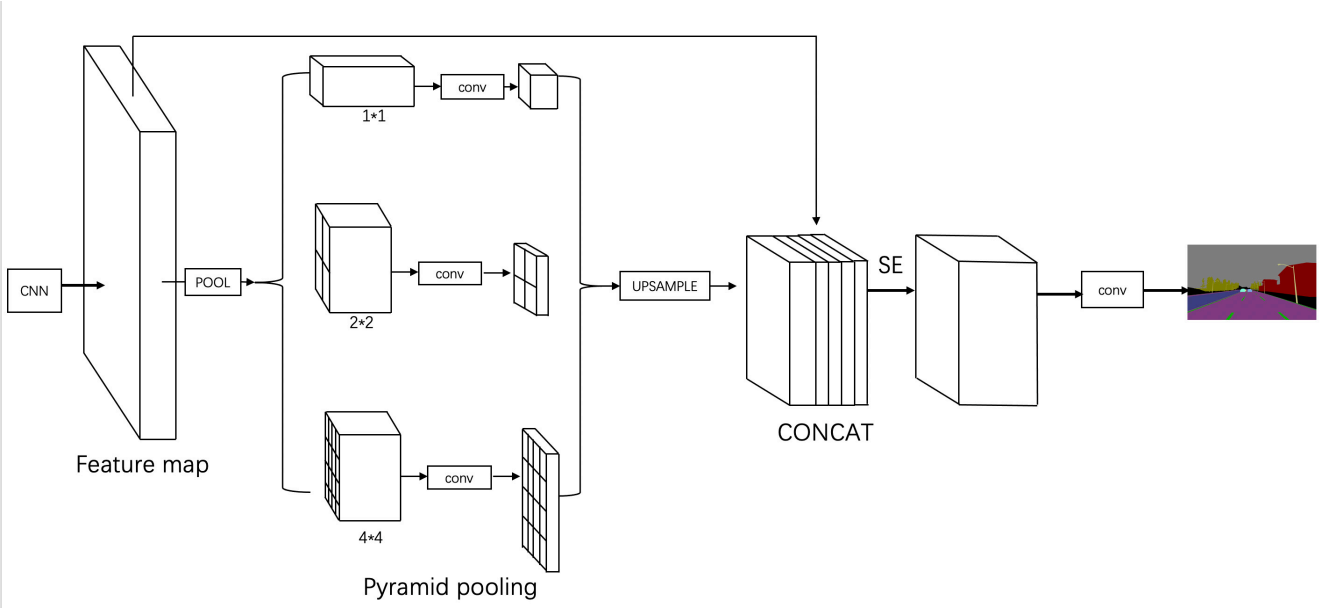

FIGURE 4. Improved segmentation network model

In the coding part of this network, the ResNet with dilated convolution is used to extract features. In the decoding part, features of different scales are obtained through pyramid pooling first, and these features are fused; then, the features after fusion learn the importance of each channel via the SE model, so as to promote important feature while ignoring ineffective features; at last, the final results are obtained through convolution, so as to achieve semantic segmentation of image.

5. **Experiment and analysis.**

5.1. **Experimental dataset.** The purpose of our network improvement is to improve the segmentation precision for small targets and the generalization ability for dataset significantly different from the training set. Therefore, dataset SYNTHIA was used in our experiment. It contains many small targets and scenes of different cities, weather and light. This dataset includes the subsets of the environment of different cities, and the subsets of each city also consists of the subsequences of different lighting and weather conditions, such as spring, summer, autumn, winter, rain, sunset, fog, night and dawn. In the experiment, the images of highway and New York were selected from the subsets as the subjects, which are also divided into the training set and verification set. First of all, the labeled images in subsequence Dawn were used to train the segmentation network; then, style transfer from the labeled images in Dawn to other subsequence images to be segmented at present was completed via the image style transfer network, and the composite images were generated for segmentation training; next, the segmentation results were further refined.

5.2. **Settings of experimental parameters.** Our experiment was based on the Pytorch. The momentum parameter was set at 0.99, the initial learning rate was 0.001, and the total iterations of experiment were 1000. Each training includes one forward propagation process and one back propagation process. The results were inferred and predicted during forward propagation, and the cross entropy loss was generated by comparing with the truth lable, the model parameters were updated during back propagation.The specific parameters of our semantic segmentation network are shown in Table 1 and Table2, Res, AtrousRes, and ConvInterp respectively represent residual block, residual block with dilated convolution, and linear interpolation block.

Table 1. Encoder parameters

layer name	kernel size	kernel number	next layer
Input	3*3	540	Conv1
Conv1	3*3	540	Conv2
Conv2	3*3	1080	Conv3
Conv3	3*3	1080	MaxPooling
MaxPooling	3*3	-	Res1
Res1	1*1	64	Atrous2Res2
	3*3	64	
	1*1	256	
Atrous2Res2	1*1	128	Atrous4Res3
	3*3	128	
	1*1	512	
Atrous4Res3	1*1	256	Atrous8Res4
	3*3	256	
	1*1	1024	
Atrous8Res4	1*1	512	AvePooling
	3*3	512	
	1*1	2048	

5.3. **Experimental results and analysis.** In order to evaluate the segmentation effects of the network proposed in this paper on dataset SYNTHIA, the MIoU of each task on the verification set and the average MIoU of all tasks were used as the criteria of experiment performance. Assuming there are k different classes, let n_{ij} denote the number of pixels actually of class i but predicted as class j, t_i represents the total number of pixels in class i, and p_i denotes the total number of pixels with prediction result of i. The computational formula of MIoU is as follows:

$$MIoU = \frac{1}{k} \times \sum_i \frac{n_{ii}}{t_i + p_i - n_{ii}} \tag{6}$$

As shown in Table 3 and Table 4, experiment was carried to compare the performances of the original PSPNet and the improved network in our work:

According to Table 3 and Table 4, the MIoUs obtained with our network are all higher than the MIoUs obtained with the original PSPNet .On the highway dataset with relatively simple scenarios, Mean mIOU of PSPNet is 64.6 and Ours can achieve 69.2. On the New York City dataset with complicated scenarios, Mean mIoU of PSPNet is 52.0 and Ours can achieve 56.4. The Mean mIoU of our proposed networks is increased by $4 \sim 5$, which shows that the segmentation accuracy of our network is improved to a certain extent, furthermore, for the scenarios in which the dataset environment is significantly from

Table 2. Decoder parameters

layer name	kernel size	kernel number	next layer
AvePooling1	16*16	-	Conv4
AvePooling2	8*8	-	Conv5
AvePooling3	4*4	-	Conv6
AvePooling4	2*2	-	Conv7
Conv4	1*1	512	Conv8Interp16
Conv5	1*1	512	Conv8Interp8
Conv6	1*1	512	Conv8Interp4
Conv7	1*1	512	Conv8Interp2
Atrous8Res4	-	-	Conv9
Conv8Interp16			
Conv8Interp8			
Conv8Interp4			
Conv8Interp2			
Conv9	16*16	-	GlobalAvelPoolConv10
Atrous8Res4	-	-	Conv11
GlobalAvelPoolConv10			
Conv11	3*3	512	Conv12
Conv12	1*1	3	Conv13
Conv13	-	-	Conv13Interp16

Table 3. Comparison results of Highway

Scene	PSPNet	Ours
Dawn	67.3	71.4
Fall	65.6	69.6
Fog	65.8	69.2
Night	59.4	66.9
Spring	69.7	72.4
Summer	67.9	70.9
Sunset	68.4	71.2
Winter Night	57.2	64.7
Winter	60.6	67.2

Table 4. Comparison results of New York

Scene	PSPNet	Ours
Dawn	60.6	62.5
Fall	52.7	57.2
Fog	54.4	57.9
Night	49.2	56.7
Spring	55.8	58.1
Summer	56.7	58.9
Sunset	54.2	56.7
Winter Night	39.2	47.4
Winter	45.8	52.3

the training set environment (e.g.: Night, Winter, etc.), the network proposed in this paper has better generalization ability. Figure 5 shows the test image, PSPNet prediction image, the prediction image obtained with our network and the actually labeled image:

(a) Test image (b) PSPNet (c) Ours (d) Labeled image

FIGURE 5. Segmentation effect comparison on dataset SYNTHIA

According to Figure 5, during image segmentation, the original PSPNet was only able to roughly segment road, vehicles and mountain and to obtain a rough profile, while it could not well segment small targets, such as zebra crossing. By contrast, the network proposed in this paper could make more accurate segmentation of vehicle and trees faraway, and it could also effectively segment small targets like zebra crossing. The experimental results show that the improved network proposed in this paper can provide better segmentation results.

6. **Conclusions.** To address the problems of poor segmentation effects of current semantic segmentation networks during segmentation of small targets and significant decline of performance when processing data with marked difference from the training set, we improved PSPNet then propose a new semantic segmentation method. The results of experiment carried out on dataset SYNTHIA show that the improved network structure can more accurately segment small targets, the generalization ability of network on different datasets is also increased, and the segmentation results are effectively improved.

REFERENCES

[1] E. K. Wang, C. M. Chen, M. M. Hassan, A. Almogren, A deep learning based medical image segmentation technique in Internet-of-Medical-Things domain, *Future Generation Computer Systems*, vol. 108, pp. 135-144, 2020.

[2] K. K. Tseng, J. Lin, C. M. Chen, M. M. Hassan, A fast instance segmentation with one-stage multi-task deep neural network for autonomous driving, *Computers and Electrical Engineering*, vol. 93, no. 12, 107194, 2021.

[3] N. Otsu, A threshold selection method from gray-level histograms, *IEEE Transactions on Systems Man and Cybernetics*, vol. 9, pp. 62-66, 2007.

[4] T. Pun, A new method for gray-level picture thresholding using the entropy of the histogram, *Signal Processing*, vol. 2, no. 3, pp. 223-237, 1980.

[5] J. C. Yen, F. J. Chang, S. Chang, A new criterion for automatic multilevel thresholding, *IEEE Transactions on Image Processing*, vol. 4, no. 3, pp. 370-378, 1995.

[6] L. Han, J. Wang, Research on global threshold segmentation method of color image based on two-dimensional histogram, *Journal of Northeast Electric Power University*, vol. 40, pp. 76-83, 2020.

[7] A. Rosenfeld, The max roberts operator is a hueckel type edge detector, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-3, no. 1, pp. 101-103, 1981.

[8] W. Gao, X. Zhang, L. Yang, H. Liu, An improved sobel edge detection, *IEEE International Conference on Computer Science and Information Technology*, pp. 67-71, 2010.

[9] L. Yang, X. Wu, D. Zhao, H. Li, J. Zhai, An improved prewitt algorithm for edge detection based on noised image, *2011 4th International Congress on Image and Signal Processing*, pp. 1197-1200, 2011.

[10] E. S. Li, S. L. Zhu, B. S. Zhu, Y. Zhao, C. G. Xia, L. H. Song, An adaptive edge-detection method based on the canny operator, *2009 International Conference on Environmental Science and Information Application Technology*, pp. 465-469, 2009.

[11] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, 2012, https://doi.org/10.1145/3065386
[12] J. M. T. Wu, Z. Li, N. Herencsar, B. Vo, J. C. W. Lin, A graph-based CNN-LSTM stock price prediction algorithm with leading indicators, *Multimedia Systems*, 2021, https://doi.org/10.1007/s00530-021-00758-w
[13] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, 2015.
[14] W. Liu, A. Rabinovich, A. C. Berg, Parsenet: Looking wider to see better, *arXiv:1506.04579*, 2015, https://arxiv.org/abs/1506.04579v1
[15] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230-6239, 2017.
[16] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, *European conference on computer vision*, pp. 740-755, 2014.
[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.
[18] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1510-1519, 2017.
[19] None, Synthia: Synthetic collection of imagery and annotations [ITS Research Lab], *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 4, pp. 138-140, 2017.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *arXiv: 1706.03762*, 2017, https://arxiv.org/abs/1706.03762v5
[21] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132-7141, 2018.
[22] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, J. Wang, Ocnet: Object context network for scene parsing, *arXiv:1809.00916*, 2018, https://arxiv.org/abs/1809.00916
[23] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3141-3149, 2019.
[24] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines vinod nair, *27th International Conference on International Conference on Machine Learning*, pp. 807-814, 2010.