ISSN: 2688-9323

Real-time Semantic Segmentation Network based on Regional Self-attention

Haoyu Cui[1], Hailong Bao[1], Jiebin Wen[2], Qian Ma[1], Shengcong Huang[1]

[1]School of Mechatronic Engineering, Southwest Petroleum University, Chengdu, Sichuan 610500, China

[2]School of Electrical Engineering and Information, Southwest Petroleum University, Chengdu, Sichuan 610500, China

Abstract

High precision semantic segmentation results often rely on rich spatial semantic information and detail information, but both of them require a large amount of computation. In order to solve this problem, a Real-time semantic segmentation network based on regional self attention is proposed by analyzing the similarity of local pixels. The network can calculate the regional correlation and channel attention information of feature information through a regional self attention module and a local interactive channel attention module respectively, and then obtain rich attention information with less computation. The experimental results on cityscapes dataset show that compared with the existing Real-time segmentation network, the segmentation accuracy of this network is higher and the speed is faster.

Keywords

Image Processing; Semantic Segmentation; Convolutional Neural Network; Attention Mechanism.

1. Introduction

With the rapid development of deep learning, image processing tasks such as infrared image and spectral image processing tasks in the optical field [1-2], computer vision tasks such as automatic driving with camera as the main application carrier, character recognition and remote sensing image segmentation [3-6] have developed rapidly. Semantic segmentation is a core technology of computer vision task. The purpose is to segment the image into several groups of regions with specific semantic categories, which belongs to pixel level dense classification problem [7]. Semantic segmentation can be used for infrared image segmentation to realize all-weather image analysis and understanding. It can also be applied to the segmentation of real street scenes and realize the environmental perception of automatic driving. For fast-moving scenes such as automatic driving, the segmentation rate and accuracy of the network are very important. In order to obtain high-precision segmentation results, the segmentation network must obtain enough semantic information and detail information [8-9]. However, both of them need to be realized by deepening the network parameters or improving the resolution of the input image, resulting in too much calculation of the network and too low segmentation efficiency [10-11].

Self attention (SA) mechanism [12-13] is a method used to obtain long-distance semantic information in the field of computer vision, which can greatly deepen the network's understanding of the whole feature map. However, this method needs to calculate the relationship between two feature points in the feature map to obtain the relationship weight of any feature point to the current feature point. The amount of calculation in this process is O (N^2C) (N= H× W, H and W are the length and width of the feature graph respectively, and C is

ISSN: 2688-9323

the number of channels of the feature graph), and the amount of calculation increases quadratic with the increase of the size of the feature graph (for example, the size of the feature graph increases from n to an, and the amount of calculation increases from N^2C to a^2N^2c), which is not suitable for the construction of Real-time network. Although the image resolution can be reduced and the amount of calculation can be reduced by down sampling such as pooling, a large amount of semantic information in the feature map will be lost, especially for high-level features with low resolution, which is not conducive to the improvement of network performance [14].

In practical application, the local pixel distribution of the image is similar. The same region or category generally has similar or even the same pixel values. The traditional SA mechanism [12-13] traverses and calculates the pairwise correlation of all feature points, which is redundant and unnecessary. Therefore, this paper proposes a lightweight area SA (RSA) module, which scales the feature map through the scaling factor R without losing feature information, and transforms the pixel level correlation calculation of the traditional SA mechanism into area level correlation calculation, so as to reduce the amount of calculation to $O(N^2C) / r^2$. Then, a lightweight local channel interactive attention (LCIA) module is proposed, which can improve the network performance without dimensionality reduction and channel information loss. Based on RSA and LCIA module, a Real-time segmentation network in the form of encoder decoder is built, and the image feature information in different stages is extracted by encoder; Then the RSA module is used to process the characteristics of each stage, to strengthen the global understanding of each layer of information; Finally, the information of each stage is effectively fused in the decoder combined with LCIA module to recover the size and detail information of the image in turn.

2. Design of Network Framework

2.1. Self Attention Mechanism

SA mechanism can obtain the pairwise correlation between all feature points and calculate the weighted influence of one feature point on all other feature points, to obtain more comprehensive semantic information, which can be expressed as

$$Y=f\left(Q,K^T\right)\cdot V \tag{1}$$

In the above formula, Y is the output of the self attention mechanism, f is the similarity calculation function, and t is the transpose operation of the matrix, Q、K and V are related characteristic diagrams which are through three different 1 × 1 convolution by feature original $X\in R(C\times H\times W)$. Among them, V contains the semantic information of the original pixels, Q and K calculate the correlation between two feature points by $f(Q,K^T)$. At the same time, the attention map is obtained by combining the Softmax function, which can be expressed as

$$Y_{j,i}=\frac{\exp\left(X_i\cdot X_j\right)}{\sum_{i=1}^{n}\exp\left(X_i\cdot X_j\right)} \tag{2}$$

In the above formula, X_i is the ith pixel in the feature map Q, X_j is the j-th pixel in the feature map K^T, n is the number of pixels of Q and K^T, $Y_{j,i}$ is the influence of pixel i on pixel j, he more similar the two, the greater the influence value [12]. In order to facilitate the calculation of

ISSN: 2688-9323

pixels in all spaces, the above three related feature images are obtained by matrix tiling: $X \in R^{C \times N}$, $f(Q, K^T)$ corresponding matrix row and column calculation formula is $(N, C) \bullet (C, N)$, the calculation amount is $O(N^2C)$. It can be seen that the amount of calculation is large, and it increases quadratic with the increase of the size of the characteristic graph.

2.2. Regional Attention Module

Figure 1 shows the actual processed image. It can be seen that the adjacent pixels in the local area are often of the same category and have similar or even the same pixel values. For these similar pixels, the global correlation should also be similar. Therefore, it is redundant and unnecessary to obtain attention information by traversing and calculating the correlation between all single feature points.

A lightweight RSA module can be designed by using the similarity of adjacent pixels in the local region to reduce the amount of calculation of SA mechanism. RSA module can reduce the amount of calculation without losing feature information and obtain the corresponding attention information. The structure is shown in Fig. 2 (a). RSA module includes two core operations: pixel shift (PS) and reverse pixel shift (R-PS), as shown in Fig. 2 (b).

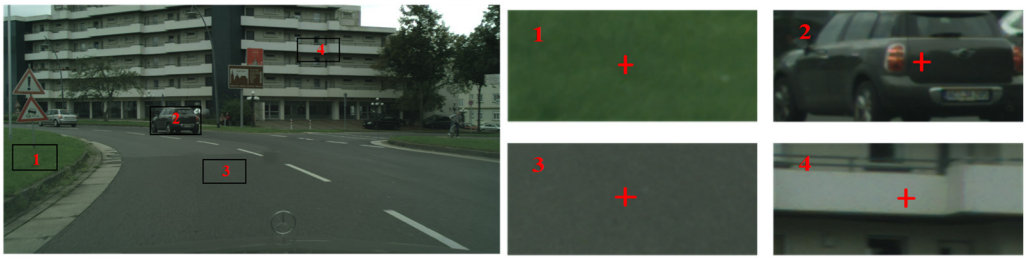
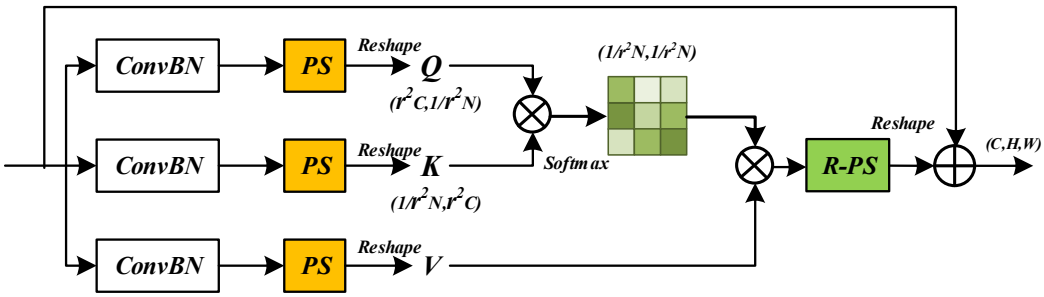
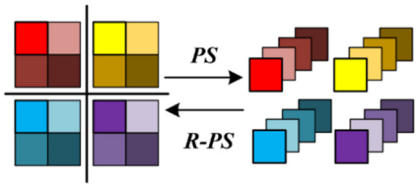


Figure 1. Pixel distribution in local area



(a) RSA Module



(b) PS and R-PS

Figure 2. Regional self attention module

First, use three different 1 × 1 convolution (Conv1×1) and batch standardization (BN) processing input characteristic diagram. Then, the PS module is used to control the scaling factor R to shift r^2-1 similar pixels around a pixel i to the adjacent position of the same channel as itself. In other words, the feature points of a region are arranged on the same channel, to reduce the image resolution without losing feature information。 R-PS is the reverse operation

ISSN: 2688-9323

of PS, which can restore the shifted r^2-1 pixels on the channel to the original spatial position around pixel i. Based on PS and RPS, equation (1) can be converted to

$$Y=f\left(PS(Q),PS\left(K^T\right)\right)\cdot PS(V) \tag{3}$$

Convert equation (2) to

$$Y_{j,i}=\frac{\exp\left(\sum_{k=1}^{r^2}X_{i,k}\cdot X_{j,k}\right)}{\sum_{i=1}^{n}\exp\left(\sum_{k=1}^{r^2}X_{i,k}\cdot X_{j,k}\right)} \tag{4}$$

Where X_{PS} is PS operation, $X_{u,k}$ and $X_{v,k}$ are the two scaling regions to be calculated， k is the k-th pixel point at the corresponding position of the two regions, and $Y_{v,u}$ is the influence of region $X_{u,k}$ on $X_{v,k}$.It can be seen that equation (4) obtains the correlation of regions $X_{u,k}$ and $X_{v,k}$ by calculating the correlation of pixels at corresponding positions.In terms of calculation amount, the characteristic diagram.In terms of calculation amount, characteristic graph $X\in R^{C\times H\times W}$ obtains $X\in R^{r^2C\times\frac{1}{r}H\times\frac{1}{r}W}$ through PS module with scaling factor R, and obtains $X\in R^{r^2C\times\frac{1}{r^2}N}$ after matrix tiling. The corresponding matrix row and column calculation formula is $\left(\frac{1}{r^2}N,r^2C\right)\cdot\left(r^2C,\frac{1}{r^2}N\right)$, and the calculation amount is reduced to $\frac{1}{r^2}O\left(N^2C\right)$. for example, when r = 4, the calculation amount is reduced to 1 / 16 of the original. Therefore, the amount of calculation of SA mechanism can be reduced by controlling the scaling factor r.

After completing the calculation of the global feature relationship at the regional level,first restore the feature map to a two-dimensional image in the spatial dimension by using the matrix reverse operation, and then restore the shifted pixels on channel C to the original spatial position by using the R-PS module, that is, $X\in R^{\left(r^2C\times\frac{1}{r}H\times\frac{1}{r}W\right)}$ changes back to $X\in R^{C\times H\times W}$ to restore the original dimension of the feature map. In order to ensure the integrity of detail information, the input and output are added and fused to form a residual connection.

2.3. Local Channel Interactive Attention Module

The channel attention mechanism can obtain the corresponding weight information for each channel of the feature map, and improve the expressive ability of the network. Existing channel attention mechanisms, such as SENet (Squeeze-andexcitation networks) [15] and Convolutional Block Attention Module (CBAM) [16], both use fully connected calculations to obtain weight information, and usually use channel dimensionality reduction operations (reduced to the original 1/16 of the image size) to reduce the computational burden of the full connection. Similar to spatial dimensionality reduction, channel dimensionality reduction will also lose a large amount of semantic information, and it is inefficient and unnecessary to capture the dependence between all channel information. Considering that CBAM can obtain spatial attention information through partial convolution, the LCIA module is designed, and the performance of the network is improved through a small number of parameter calculations. The structure of the LCIA module is shown in Figure 3.

ISSN: 2688-9323

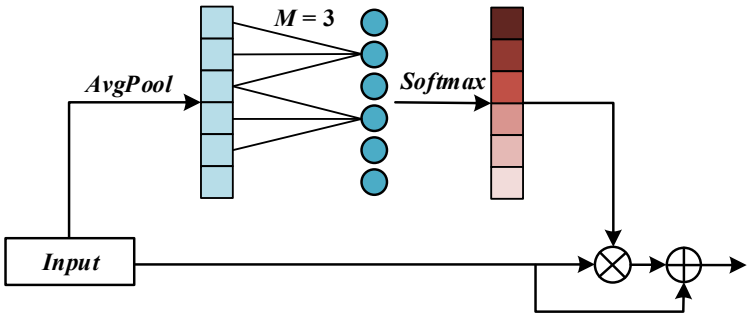


Figure 3. Structure of LCIA module

It can be seen from Figure 3 that after avgpooling(AvgPool), the ICLA module does not perform channel dimensionality reduction to ensure the integrity of information. Unlike full connection, ICLA module uses a one-dimensional convolution module with length m to extract only the current channel and its adjacent M-1 local channels to generate attention information, which can be expressed as

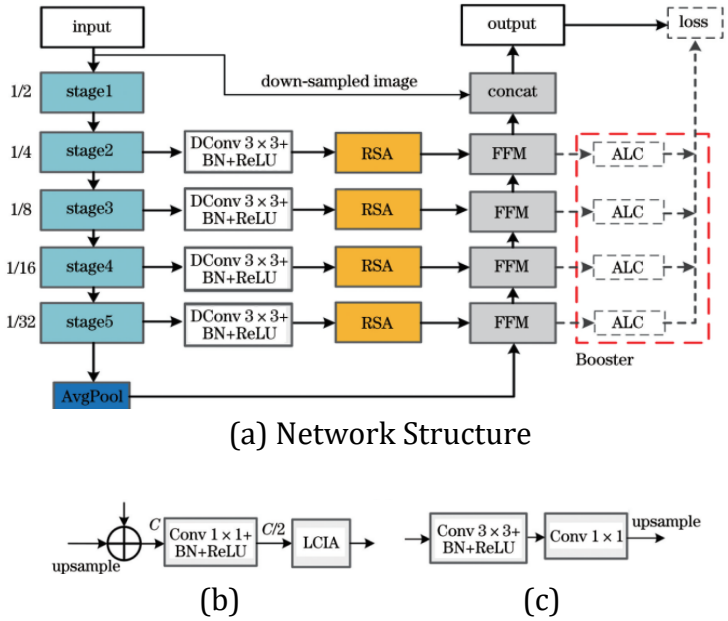
$$Y_o=\delta\left(\sum_{m=1}^{M} X_{o,m}\cdot W_m\right) \tag{5}$$

Where, L is the one-dimensional convolution kernel of size M, which can aggregate the values of M adjacent local channels, $X_{o,m}$ and m are the M adjacent channels of the O-th channel of the input characteristic, δ Is the Softmax activation function, and Y_o is the attention information of M local channels to the characteristics of the current channel. The weight sharing characteristic of convolution makes the whole LCIA module have only m parameters and the amount of calculation is O (MC), which ensures the efficiency of the network, and only using some adjacent channel information (e.g. M = 3) can also bring obvious performance gain.

2.4. Mixed Pooling Module Design

Combined with RSA and LCIA modules and using encoder decoder structure, a split network framework is built, as shown in Figure 4 (a). The decoder part uses the small residual network resnet-18 [17] combined with 18 layers as the skeleton network to obtain the characteristic information of the image,There are five stages in total, and the image is down sampled once in each stage. Finally, the feature image size output by the network is 1 / 32 of the original image size, and the features are processed in stages 2, 3, 4 and 5.For the feature information of each stage, first of all, using a size of 3 x 3 convolution module to process characteristics of local to local characteristic information fusion. Then, combined with hole convolution (dconv3) × 3) [18] improve the network receptive field, and each convolution module is followed by a BN and a modified linear unit (relu) activation function; Secondly, the RSA module is used to obtain the regional global correlation of feature information.Considering that the features of different stages have different resolutions and local similarity, the scaling rates of RSA modules in stages 2, 3, 4 and 5 are set to (4, 4, 2 and 1). In the decoder part, combined with the LCIA module, the resolution and detail information of the image are successively restored with the feature fusion module (FFM) shown in Fig. 4 (b).

ISSN: 2688-9323



(a) Network Structure

(b) (c)

Figure 4. This Network Structure

In order to further improve the segmentation effect, an intensive training booster [19] module is designed, that is, an auxiliary loss classifier (ALC) is set at each stage of the decoder to supervise and learn the initial segmentation results, as shown in Fig. 4 (c). The booster module can enhance the feature expression ability of the network in the training stage, and will not participate in the calculation when testing, so as to improve the segmentation accuracy of the network without affecting the efficiency of network segmentation.

3. Experimental Results

3.1. Experimental Setup

The effectiveness of this network is verified by the Cityscapes data set [20]. The Cityscapes data set includes street scene images in 50 different cities and a total of 5000 finely labeled images, of which 2975 are used for training and 500 are used for verification.1525 sheets were used for testing. Experiments are performed based on finely annotated image data, and images containing 19 types of objects are used for training and testing. Experimental environment: The software environment is the Pytorch deep learning framework, and the graphics card is 1080ti. During the experiment, the stochastic gradient descent (SGD) algorithm was used to optimize the convergence process; the poly learning rate decay strategy was adopted, the initial learning rate was e-2, the weight decay rate was e-4, and the momentum was 0.9. The loss function is a cross-entropy loss function, and the batch size is 10. In order to enhance the learning ability of the model, the data set is enhanced, including random mirroring, random size scaling, etc. The scaling range is {0.75, 1.0, 1.5, 1.75, 2.0}. The average interaction ratio (MIoU) is used to measure the segmentation accuracy of the network, and the number of frames per second (FPS) is used to measure the segmentation efficiency of the network.

3.2. Experimental Verification

3.2.1. Comparison Experiment of Zoom Ratio

The RSA module can obtain effective regional-level feature relevance, but the feature maps at different stages have different resolutions. The low-level feature maps have a larger resolution, which has rougher semantic information and broader similarities, while the high-level information is the opposite [21]. Therefore, different zoom ratios are set for different stages to conduct comparative experiments, and the results are shown in Table 1.

ISSN: 2688-9323

Table 1. Comparison experiment of zoom ratio

Serial number	Network	MIoU/%	FPS/frame
1	(1,1,1,1)	71.9	10
2	(4,2,2,1)	71.7	109
3	(4,4,2,1)	71.7	120
4	(8,4,2,1)	71.6	133

Among them, the scaling rate of the RSA module in the second, third, fourth, and fifth stages of the first group is (1, 1, 1, 1), which means that the feature map is not scaled, that is, the original SA mechanism [13], and its MioU is 71.9%, FPS is only 10frame, which cannot meet the needs of Real-time segmentation. After scaling, the MioU of the second, third, and fourth groups of parameters are 71.7%, 71.7%, and 71.6%, respectively, and the FPS are 109, 120, and 133 frames, respectively. It can be seen that the RSA module greatly improves the segmentation speed of the network without affecting the segmentation accuracy.

3.2.2. Ablation Experiment

In order to verify the improvement of the network expression ability of the RSA module, an ablation experiment was carried out, and the results are shown in Table 2. It can be found that when the RSA module is not used, the network's MioU is 68.4% and the FPS is 158frame; after the RSA module is used, the network's MioU is 71.7% and the FPS is 120frame. Compared with the direct fusion processing of feature information, the RSA module can help the network capture clearer feature correlation and long-distance information, and improve the expressive ability of the network.After adding the LCIA module, the MioU of the network is 72.3%, and the FPS is 115frame. Compared with CBAM, the LCIA module achieves a similar segmentation accuracy at a faster segmentation speed; compared with SENet, the LCIA module achieves a higher segmentation accuracy at a faster segmentation speed, which shows that the reduction of dimension of the channel will also affect the network performance, and only through the local channel interaction information can obtain effective attention information and improve the network performance. After combined with Booster enhanced training, the MioU of the network rises to 73.1%, and does not affect the segmentation speed, which shows that combined with auxiliary loss training can effectively enhance the expressive ability of the network.

Table 2. Results of ablation experiments

Network	MIoU/%	FPS/frame
Original network	68.4	158
RSA	71.7	120
RSA+LCIA	72.3	115
RS+CBAM	72.5	80
RSA+SENet	72.1	102
RSA+LCIA+Booster	73.1	115

In order to verify the ability of the RSA module to retain feature information, the average pooling and maximum pooling were selected for comparison experiments. The down sampling rate of the two is the same as that of the RSA module, and the original size of the image is restored through linear interpolation. The results are shown in Table 3. It can be found that compared to the maximum pooling and average pooling, after the RSA module is used, the network segmentation effect is better at similar speeds. This shows that the loss of information in the pooling process is harmful to the expressive ability of the network, and the RSA module

ISSN: 2688-9323

can more effectively retain the characteristic information, further explaining the importance of the integrity of the characteristic information to the network.

Table 3. Comparative experiment of downsampling method

Network	MIoU/%	FPS/frame
AvgPool	70.5	126
MaxPool	70.4	132
RSA	71.7	120

3.3. Comparative Experiment

Several common segmentation networks [22-28] are selected for comparison with the performance of this network. Among them, ENet, ESPNet, ERFNet and DABNet do not use backbone networks, while ICNet and DFANet use pre-training networks PSPNet50 and XceptionA as backbone networks, respectively. The resolution of the input image is 512 pixels × 1024pixel, and the results are shown in Table 4. It can be found that when the FPS of the lightweight network ResNet-18 is 115 and 126frame respectively, the MioU of this network on the test set is 72.1% and 71.8%, respectively, which is better than other Real-time segmentation in segmentation accuracy and segmentation rate.

Table 4. Experimental results of different networks

Network	Pretrain	MIoU/%	FPS/frame
ENet	No	58.3	76
ESPNet	No	60.3	112
ERFNet	No	68.0	41.7
ICNet	PSPNet50	69.5	30
DABNet	No	70.1	104
DFANet*	Xception-A	70.3	-
DFANet	Xception-A	71.3	100
Ours	ResNet-18	72.1/71.8	115/123

In order to show the superiority of this network more intuitively, some segmentation results are selected and visualized, and the results are visually compared with ERFNet. The result is shown in Figure 5. It is shown that this network can achieve a more refined segmentation effect in a local area, and can perform a more effective segmentation for small objects, and there are fewer intra-class inconsistencies and inter-class inconsistencies in the overall segmentation results.

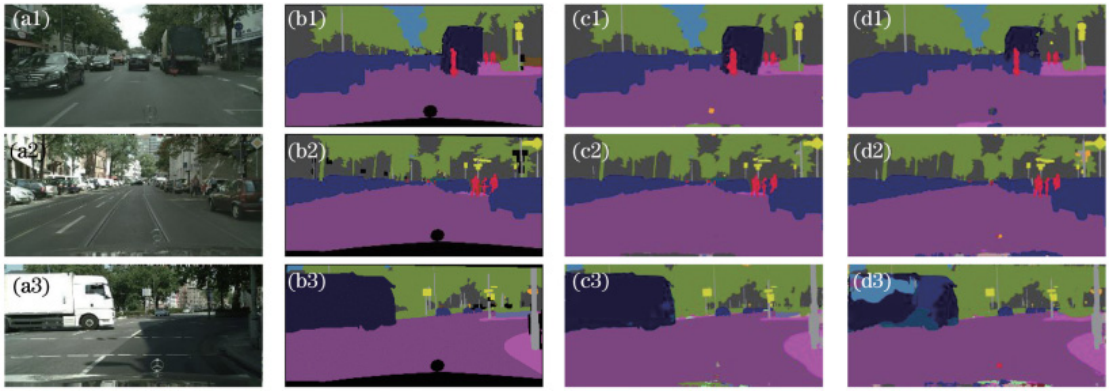


Figure 5. Visualize the results on Cityscapes.(a)Original Picture; (b)Real segmentation results; (c)Results of this network segmentation; (d) Results of ERFNet segmentation

ISSN: 2688-9323

4. Conclusion

Based on the similarity of local pixel distribution, a lightweight RSA module is designed, which can obtain the regional correlation of global information without losing feature information; It does not need to traverse and calculate the pairwise correlation of all feature points, which greatly reduces the amount of calculation of SA mechanism and improves the segmentation efficiency of the network. Then an LCIA module is proposed, which can obtain effective channel attention information only through adjacent local channels, avoid channel dimensionality reduction operation and retain the integrity of channel information. Based on RSA and LCIA module, a Real-time semantic segmentation network with encoder-decoder structure is built, and the regional correlation of features in each stage is extracted by RSA module to strengthen the expression ability of the network; In the decoder part, LCIA module is combined to improve the network performance. Experimental results show that compared with other networks, this network has better segmentation results and segmentation efficiency.

References

[1] Tang Chaoying, Pu Shiliang, Ye Pengzhao, et al. Fusion of Low-Illuminance Visible and Near-InfraredImages Based on Convolutional Neural Networks [J]. Acta Optica Sinica, 2020, 40(16): 161 0001.

[2] Kong Fanqiang, Zhou Yongbo, Shen Qiu, et al. End-to-eEnd Multispectral Image Compression Using Convolutional Neural Network[J]. Chinese Journal of Lasers, 2019, 46(10): 1009001.

[3] Liu Hui, Peng Li, Wen Jiwei. Multi-Scale Aware Pedestrain Detection Algorithm Based on Improved Full Convolutional Network[J]. Laser & Optoelectronics Progress, 2018, 55(9): 091504.

[4] He Y H, Wang H, Zhang B. Color-based road detection in urban traffic scenes[J]. IEEE Transactions on Intelligent Transportation Systems, 2004, 5(4): 309-318.

[5] Yao Lisha, Xu Guoming, Zhao Feng. Facial Expression Recognition Based on Local Feature Fusion of Convolutional Neural Network[J]. Laser & Optoelectronics Progress, 2020, 57(4): 041513.

[6] Zhang Zhehan, Fang Wei, Du Lili, et al. Semantic Segmentation of Remote Sensing Image Based on Encoder-Decoder Convolutional Neural Network[J]. Acta Optica Sinica, 2020, 40(3): 0310001.

[7] Zhang Xiangfu, Liu Jian, Shi Zhangsong, et al. Review of Deep Learning-Based Sematic Segmentation [J]. Laser & Optoelectronics Progress, 2019, 56(15): 150003.

[8] Lin G S, Milan A, Shen C H, etal. RefineNet multi-path refinement networksfor high-resolution semantic segmentation[C]‖2017 IEEE Conferenceon Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5168-5177.

[9] Peng C, Zhang X Y, Yu G, et al. Large kernel matters improve semantic segmentation by global convolutional network[C]‖2017 IEEE Conferenceon Computer Vision and Pattern Recognition (CVPR), July21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1743-1751.

[10] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]‖2017 IEEE Conference on ComputerVision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[11] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation [EB/OL]. [2019-12-09]. https:‖arxiv.org/abs/1706.05587.

[12] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]‖2018 IEEE/CVF Conferenceon Computer Vision and Pattern Recognition, June18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7794-7803.

[13] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]‖2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 3141-3149.

ISSN: 2688-9323

[14] Yu C Q, Wang J B, Peng C, et al. Learning a discriminative feature network for semantic segmentation[C]‖2018 IEEE/CVF Conferenceon Computer Vision and Pattern Recognition, June18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1857-1866.

[15] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks [EB/OL]. [2020-07-20]. https: ‖arxiv. org/abs/1709.01507.

[16] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module [EB/OL]. [2020-07-25]. https:‖ arxiv.org/abs/1807.06521.

[17] He K M, Zhang X Y, Ren S Q, et al. Deepresidual learning for image recognition[C]‖2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York, IEEE Press, 2016: 770-778.

[18] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets atrous convolution and fully connected CRFs[EB/OL]. [2020-07-21]. https:‖ arxiv. org/abs/1606.00915.

[19] Mehta S, Rastegari M, Shapiro L, et al. ESPNetv2: alight-weight, powerefficient and general purpose convolutional neural network[C]‖2019 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 9182-9192.

[20] Cordts M, Omran M, Ramos S, et al. Thecityscapes dataset for semantic urban scene understanding[C]// 2016 IEEE Conference on Computer Visionand Pattern Recognition (CVPR), June 27-30. 2016, Las Vegas, NV, USA. NewYork, IEEE Press, 2016: 3213-3223.

[21] Yu C Q, Wang J B, Peng C, et al. BiSeNet: bilateral segmentation network for Real-time semantic segmentation[M]‖Ferrari V, Hebert M, Sminchisescu C, et al. Computer Vision-ECCV 2018. Lecture Notesin Computer Science. Cham: Springer, 2018: 11217 334-349.

[22] Paszke A, Chaurasia A, Kim S, et al. ENet: a deepneural network architecture forReal-time semantic segmentation [EB/OL]. [2020-07-23]. https:‖arxiv.org/abs/1606.02147.

[23] Mehta S, Rastegari M, Caspi A, et al. ESPNet: efficient spatial pyramid of dilated convolutions for semanticsegmentation[M]// FerrariV, Hebert M, Sminchisescu C, et al. Computer Vision-ECCV2018. Lecture Notesin Computer Science. Cham: Springer, 2018, 11214: 561-580.

[24] Romera E, Álvarez J M, Bergasa L M, et al. ERFNet: efficient residual factorized ConvNet forReal-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(1): 263-272.

[25] Zhao H S, Qi X J, Shen X Y, etal. ICNet for Real-time semantic segmentation on high-resolution images[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer Vision-ECCV 2018. Lecture Notes in Computer Science. Cham Springer: 2018, 11207: 418-434.

[26] Wang Y, Zhou Q, Liu J, et al. Lednet: a lightweight encoder-decoder network for Real-time semantic segmentation[C]// 2019 IEEE International Conference on Image Processing (ICIP), Sept 22-25, 2019, Taipei, Taiwan, China. New York: IEEE Press, 2019: 1860-1864.

[27] Li G, Yun I, Kim J, et al. DABNet: depth-wiseasymmetric bottleneck for Real-time semantic segmentation [EB/OL]. [2020-07-22]. https://arxiv.org/abs/1907.11357.

[28] Li H C, Xiong P F, Fan H Q, et al. DFANet: deepfeature aggregation for Real-time semantic segmentation[C]// 2019IEEE/CVF Conferenceon Computer Visionand Pattern Recognition (CVPR), June15-20, 2019, Long Beach, CA, USA. NewYork: IEEEPress, 2019: 9514-9523.