Semantic segmentation of urban street scene images based on improved U-Net network*

ZHU Fuzhen[1], CUI Jingyi[1], ZHU Bing[2], LI Huiling[1], and LIU Yan[1]**

1. College of Electronic Engineering, Heilongjiang University, Harbin 150080, China

2. Institute of Image Information Technology and Engineering, Harbin Institute of Technology, Harbin 150001, China

(Received 19 July 2022; Revised 18 November 2022)

To balance the speed and accuracy in semantic segmentation of the urban street images for autonomous driving, we proposed an improved U-Net network. Firstly, to improve the model representation capability, our improved U-Net network structure was designed as three parts, shallow layer, intermediate layer and deep layer. Different attention mechanisms were used according to their feature extraction characteristics. Specifically, a spatial attention module was used in the shallow network, a dual attention module was used in the intermediate layer network and a channel attention module was used in the deep network. At the same time, the traditional convolution was replaced by depthwise separable convolution in above three parts, which can largely reduce the number of network parameters, and improve the network operation speed greatly. The experimental results on three datasets show that our improved U-Net semantic segmentation model for street images can get better results in both segmentation accuracy and speed. The average mean intersection over union (*MIoU*) is 68.8%, which is increased by 9.2% and the computation speed is about 38 ms/frame. We can process 27 frames images for segmentation per second, which meets the real-time process and accuracy requirements for semantic segmentation of urban street images.

Document code: A **Article ID:** 1673-1905(2023)03-0179-7

DOI https://doi.org/10.1007/s11801-023-2128-8

Image semantic segmentation is a key technology in the task of environment perception[1,2], which is the core task of autonomous driving. Image semantic segmentation can classify common targets captured by in-vehicle cameras at pixel level[3], such as pedestrians, vehicles, traffic lights and railings in urban street images. It can assist the vehicle's decision planning module to make rational decisions and plans, and guarantee safe driving of autonomous vehicles. However, urban street images often exhibit complex and variable street targets. The characteristics of the same target in the image are diverse, and the characteristics of different types of targets are also similar. Moreover, the influence of illumination, shooting angle and occlusion will inevitably lead to the reduction of segmentation accuracy. Solving the imbalance problem of accuracy and speed in semantic segmentation of urban street images is critical and challenging for the application of autonomous driving technology.

Traditional image segmentation methods are mainly based on grey characteristics and can segment simple grey images. With the improvement of computer hardware and software performance, image semantic segmentation algorithms based on convolution neural networks are taking the lead in image segmentation field. The first end-to-end image semantic segmentation algorithm was implemented by the fully conventional network (FCN)[4]. In the FCN, 1×1 convolution layer is used to replace the fully-connected layers of the original network, which makes it possible to input images of arbitrary size into the network, and the semantic segmented images can be obtained directly through the network. A skip structure is used to connect the shallow network and the deep network of FCN network[5,6], which can fuse the low-level features of the shallow layer with high-level features of the deep layer, and get better results in the semantic segmentation task. After the FCN, coding and decoding network models appeared, such as SegNet[7], in which the coding network part is used to extract the features of the image and the decoding network part is used to extract the image features and restore the image dimensions. In order to solve the problem of limited perceptual field during downsampling, the concept of dilated convolution was proposed in Ref.[8], which increased the perceptual field without increasing the number

* This work has been supported by the National Natural Science Foundation China (No.61601174), the Postdoctoral Research Foundation of Heilongjiang Province (No.LBH-Q17150), the Science and Technology Innovative Research Team in Higher Educational Institutions of Heilongjiang Province (No.2012TD007).

** E-mail: zhufuzhen@hlju.edu.cn

of convolutional kernels. In 2014, Google team proposed the Deeplab series of network structures, deeplab_v1[9], which combined dilated convolution with conditional random field (CRF). Deeplab_v2 proposed the atrous space pyramid pooling (ASPP)[10] module, which set the null rate of the dilated convolution to different values to extract and integrate image features of different scales. Deeplab_v3+ used the multi-grid strategy[11] to add different rates of dilated convolution at the back end of the model, and added a normalization layer[12] to the ASPP module. Depthwise separable convolution is used in the ASPP module to reduce the computational effort. In addition to the above methods to improve the accuracy of semantic segmentation models by increasing the depth and width of the network, it was found that attention mechanisms can assist feature extraction networks in aggregating contextual information, which can also improve the accuracy of semantic segment models. In 2019, FU et al[13] proposed the dual attention network (DANet), after the ResNet network extracted the feature maps, the channel attention module and spatial attention module were used to process the feature maps in parallel, in order to build global contextual dependencies on local features and enhance feature representation. Above semantic segmentation models have achieved high metrics in accuracy for urban street images, but there are still shortcomings in segmentation speed. In the context of autonomous driving scene understanding, the task of semantic segmentation of urban street images requires a combination of accuracy and speed to meet practical application requirements. Therefore, to achieve real-time and accuracy of urban street images segmentation, we used a simple U-Net semantic segmentation model, combined with depth-separable convolution and attention mechanisms.

The U-Net[14] network is a typical coding and decoding model, which extended the hop structure of FCN, except that the number of hop structures usage is increased and the add connections in FCN are transformed into concat connections. The U-Net network combined coding features with decoding features, which can better extract the features of different categories of targets in complex urban street images. Since U-Net was originally proposed in the field of biomedical image segmentation, it is more suitable for datasets of small-scale images. For urban street images with complex features and large scales, it cannot provide enough features to support accurate semantic segmentation. Simply increasing the network depth will not only slow down the real-time processing speed of the model, but also bring problems such as gradient disappearance. Therefore, we improved the U-Net network structure, and combined the network feature extraction with attention mechanism to balance accuracy and speed. We designed shallow, intermediate and deep networks, and introduced matching attention mechanisms into them respectively. At the same time, for large-scale urban street view images, we replaced traditional full convolution with depthwise separable convolution to avoid the problem of reduced urban street segmentation accuracy caused by the traditional U-Net reducing the size of images.

The overall structure of the improved U-Net network model can be divided into three parts, shallow layer network, intermediate layer network and deep layer network. The spatial attention module, dual attention module and channel attention module were brought to optimize the network. The previous attention mechanism[15] is to directly fuse the spatial attention module and the channel attention module together, and then embed it into the backbone network to assist the extraction of features, which will lead to a surge in computation[16]. Therefore, we replaced the traditional convolution with depthwise separable convolution to reduce the number of network parameters and computation and improved real-time segmentation speed. The structure of the improved algorithm model is shown in Fig.1, where the three different types of attention modules are combined with the depth-separable convolutional blocks, indicated by short arrows of different colors. The short orange arrows represent the combination of a spatial attention module and a depth-separable convolutional block. The short grey arrows represent the combination of a dual attention module and a depth-separable convolutional block, and the short green arrows represent the combination of a channel attention module and a depth-separable convolutional block. Our improvements are described below.

Traditional U-Net network is suitable for small-sized images, while cannot guarantee the segmentation speed of the specific semantic segmentation for larger-sized urban street images. However, such small-sized training street scene images will lead to poor segmentation results. In order to ensure both accuracy and efficiency of the segmentation algorithm, we replaced the traditional convolution with depthwise separable convolution in the improved network to reduce the number of parameters and computation to achieve real-time segmentation.

Compared with the traditional convolution which keeps the same number of convolution kernels and channels, the depthwise separable convolution is composed of channel-by-channel convolution and point-by-point 1×1 convolution which can integrate channel dimension features, and reduce the dimension of the channel according to the output demand. The comparison of the traditional convolutional kernel and depthwise separable convolutional kernel is shown in Fig.2.

In the implement of depthwise separable convolution, the input feature maps are performed by convolution operations channel by channel, then the numbers of feature maps in input channels are the same to those of output. Then point-by-point convolution operation is used after the channel-by-channel convolution, to combine features between channels and to up-dimension or down-dimension the channels[17]. And the number of

point-by-point convolution kernels is equal to that of the final output feature maps. In this paper, 17 depthwise separable convolution operations are used in the modified U-Net network. In order to keep the consistence of image size throughout the convolution operation, the edges of the image are padded with 0 (Padding='same') before the depthwise separable convolution operation performed. The other depthwise separable convolutions have the same structure with that of the first depthwise separable convolution, with the only difference in the number of output feature maps by each convolution operation. We adjusted the image size of the dataset to 256×256. Taking the first depthwise separable convolution operation as an example, the computation of the normal convolution operation is $Q1$=256×256×3×3×3×3×64=113 246 208. The computation of the depthwise separable convolution is $Q2$=256×256×3×3×3+256×256×3×3×64=14 352 384.

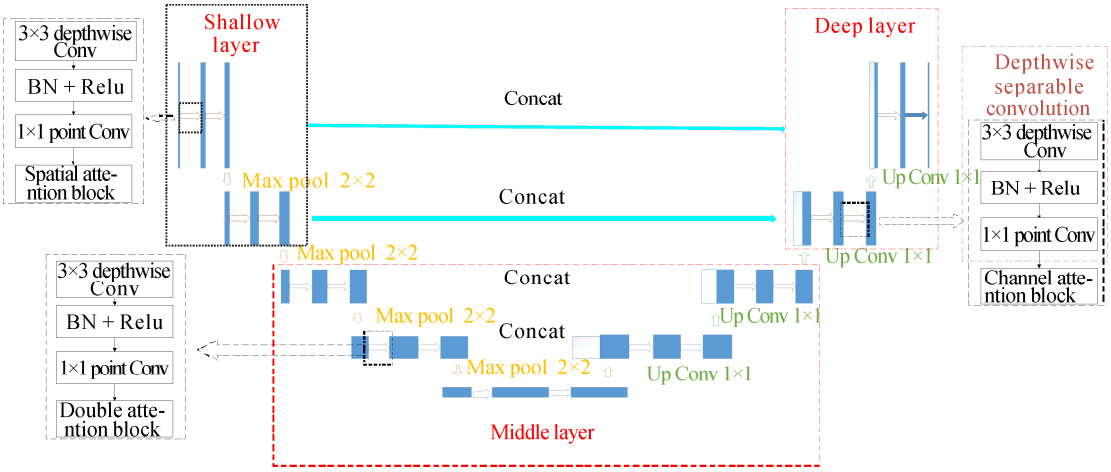


Fig.1 Improved U-Net model structure diagram

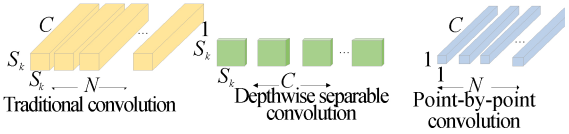


Fig.2 Convolution kernel structure comparison

We can see that the calculation in the first depthwise separable convolution operation is only about 0.13 time the computation of the ordinary convolution operation. When N feature maps are generated and the size of the convolution kernel is S×S, the ratio of depthwise separable convolution to normal convolution is $1/N+1/(S\times S)$. The number of parameters in the original U-Net algorithm model is 36M, while the number of parameters in our improved algorithm model is 13M. The number of parameters is reduced by about 2/3, which greatly improves the computational efficiency.

The semantic segmentation task needs to classify each pixel. In the process of extracting features by convolutional neural network, the low-level features are obtained from the shallow network, such as position, edge and contour, etc, among which position information is more crucial[18]. Therefore, we introduced a spatial attention module[19] into the shallow network, which combined with deep separable convolution to effectively get the overall space distribution of each channel and better targets location. Specifically, first, the feature map is obtained by deep separable convolution processing. Then, the spatial features of the input feature map are processed by global maximum pooling and global average pooling, respectively. The obtained features are fused into a new output feature, and a common convolution operation is performed on the new output feature. The convolution kernels number is set as 1 and its size is 7×7. The convolution layer is followed by the activation layer, and the Sigmoid function[20] is selected as the activation function. The activation output is normalized at the same time, which is more conducive to improve the calculation speed. And the output features are activated to obtain a spatial attention factor M_s, which is defined as Eq.(1). Finally, the input features are multiplied with the output spatial attention coefficients M_s to obtain the output features incorporating spatial attention.

$$M_s(F)=\sigma\{f^{7\times7}[\mathrm{Avg}(F);\mathrm{Max}(F)]\}, \quad (1)$$

where σ represents the Sigmoid activation operation, $f_{7\times7}$ represents the convolution operation with a 7×7 convolution kernel, Avg represents the global average pooling operation, and Max represents the global maximum pooling operation. The structure of the spatial attention module is shown in Fig.3.

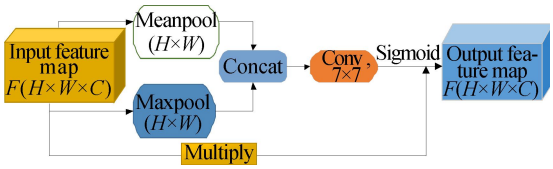


Fig.3 Spatial attention module

The features extracted by the intermediate layer network are mixed. In order to enable the intermediate layer

network to extract enough important features, we adopted a combination of channel attention and spatial attention, as shown in Fig.4. Firstly, the channel attention module is used to process the feature map, and the feature map fused with the channel attention coefficient is obtained. Then the spatial attention module is applied to obtain a feature map, in which the channel attention coefficients and the spatial attention coefficients[21] are fused. We called this new attention module as the dual attention module. To reduce the computations of the intermediate layer network, the depthwise separable convolution is applied to the intermediate layer network again, and its structure is the same as that of the shallow layer network.

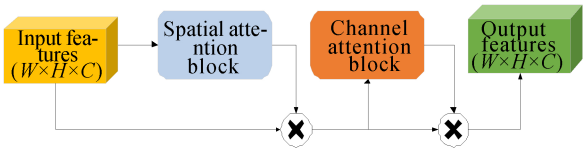


Fig.4 Dual attention module

The deep network is used to extract higher layer features which provide data support for the final classification prediction. We applied the channel attention module to the deep network. Channel-level feature information was extracted from the channel attention module, which helps the model to determine the target pixel category. The channel features of the input feature map are first processed by global average pooling and global maximum pooling to obtain two channel output features, and then fed into the fully connected layer. To avoid a surge in data volume, not all the input channel features were processed in this fully connected layer, and we selected 1/4 for processing. After the output channel features are processed by the first fully connected layer, they are fed into an activation layer for activation, where the linear rectification function (Relu)[22] is chosen as the activation function of the activation layer. The activated channel features are then fed into a second fully-connected layer, resulting in two new output channel features. The two new output channel features are added together to obtain a new output channel feature. Then it is sent to another activation layer for activation, and the Sigmoid function is chosen as the activation function of this activation layer to obtain a channel attention coefficient M_c, which is defined as shown in Eq.(2). Finally, the input features are multiplied with the channel attention coefficients M_c to obtain the output features incorporating the channel attention.

$$M_c(F)=\sigma\{W_0[\mathrm{Avg}(F)]+W_1[\mathrm{Max}(F)]\}, \tag{2}$$

where σ represents the Sigmoid activation operation, W_0 and W_1 represent the parameters in the two fully connected layers, Avg represents the global average pooling operation, and Max represents the global maximum pooling operation. To reduce the computational effort, we applied depthwise separable convolution in the deep network as well. The structural design of the deep separable convolution in the deep network remains the same as that of the shallow network. The structure of the channel attention module is shown in Fig.5.

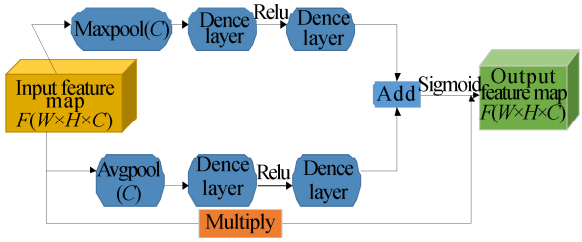


Fig.5 Channel attention module

Three publicly datasets for street scent image segmentation were used in our experiment, which were BDD100K, CamVid and Cityscapes, respectively. The Cityscapes dataset was used for our network training, the CamVid and Cityscapes datasets were used for test. Cityscapes dataset contains 25 000 coarsely and finely annotated images, 20 000 coarsely annotated images for pre-training and 5 000 finely annotated images for formal training. We used 2 975 as the training set, 1 525 as the test set and 500 as the validation set. This dataset has the problem of uneven number of target types. We eliminated the uncommon targets through mapping process, and finally the finely annotated data are mapped into 19 classes. In order to enrich the training data, the existing accurately annotated data are enhanced by fusing the original images with the corresponding annotated images and then randomly cropping them to 256×256 size. At the same time, the original images and their corresponding annotated images are randomly flipped left and right with a flip probability of 0.5, so that the final number of accurately annotated images involved in the training is about 4 500.

Pixel accuracy (*PA*) and mean intersection over union (*MIoU*) are respectively used as evaluation metrics for image semantic segmentation algorithms, defined as below.

$$PA=\frac{\sum_{i=0}^{k}P_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}P_{ij}}, \tag{3}$$

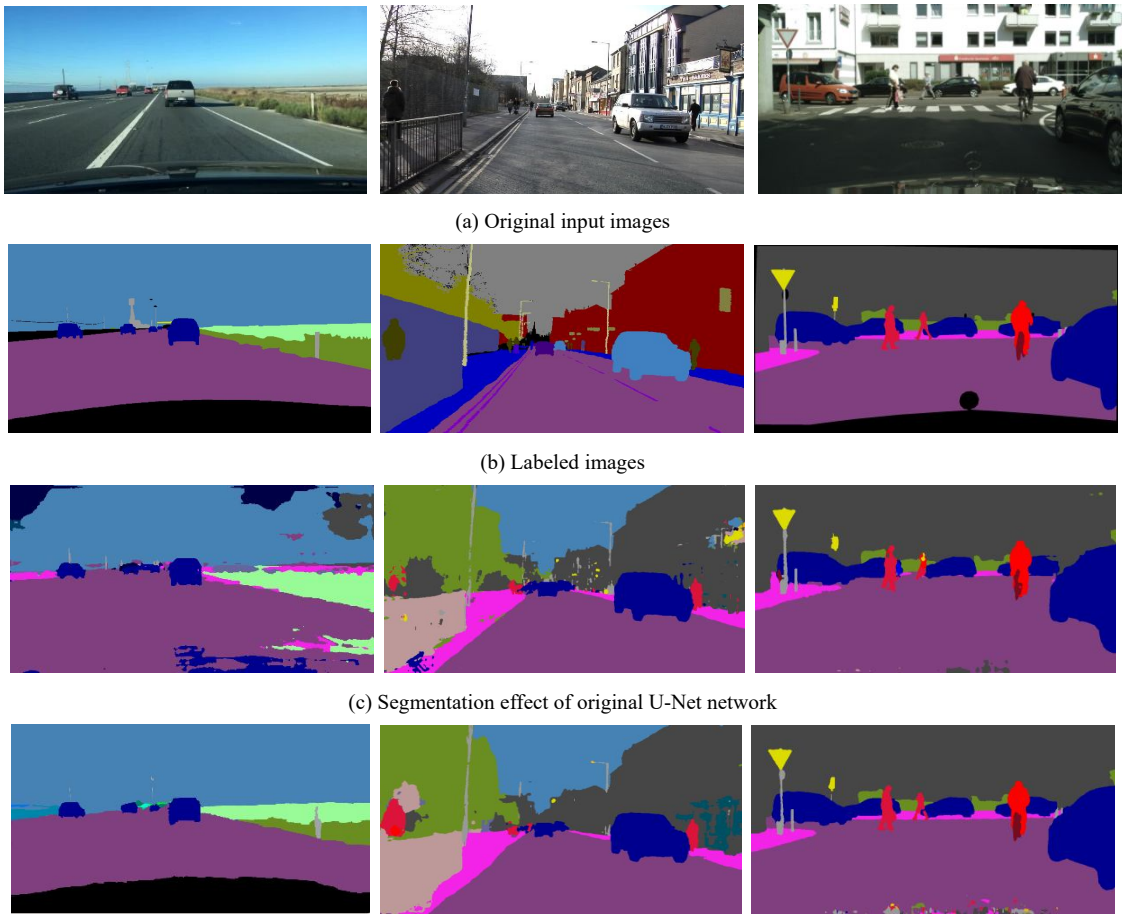
$$MIoU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}+\sum_{j=0}^{k}P_{ji}-P_{ii}}, \tag{4}$$

where the true positive pixels is represented as P_{ii}, while the number of false positive pixels as P_{ij} and the number of false negative pixels as P_{ji}. *PA* is calculated as the ratio of the number of pixels correctly predicted by the model to the total number of pixels, and the higher the *PA* value, the better the algorithm works. *MIoU* is the result of averaging the ratio of the intersection of the predicted results and the true values for each category.

The experimental environment was an Intel Xeon CPU E5-2640 v4 @2.40 GHz, 32 GB RAM, 64-bit Win10 system, dual RTX2080 Ti GPU accelerated training, and Pytorch as the deep learning framework. First, pre-training was performed on 20 000 roughly labeled data from the Cityscapes dataset to initialize the weights, and then continued on the accurately labelled dataset.

The cross-entropy loss function is selected, and the training weights were updated with the Adam optimization algorithm. The batchsize was set to 8, and the initial learning rate was 0.001, and after 50 epochs of training, the learning rate was adjusted to 0.005 for 30 epochs, and finally the learning rate was adjusted to 0.000 1 for 20 epochs to finish the net training.

A total of 13 hours was cost for our network training. The pixel accuracy was 90.3%, and an average cross-merge ratio (*MIoU*) was 68.8%. Then well trained network and weights were saved, 100 validation data were selected from BDD100K, CamVid and Cityscapes for verification, which took a total of 3.2 s. A comparison of the semantic segmentation results of the improved U-Net network urban street images is shown in Fig.6.



(a) Original input images

(b) Labeled images

(c) Segmentation effect of original U-Net network

(d) Segmentation effect of our improved U-Net network

Fig.6 Segmentation results comparison on datasets of BDD100K, CamVid and Cityscapes from left to right

As shown in Fig.6, subgraph (d) has smoother segmentation and more accurate classification of small targets than the original U-Net subgraph (c), and closer to the labeled subgraph (b) which is used for training. At the same time, our improved U-Net network segmentation algorithm is compared with the current leading deep learning image segmentation algorithm by the *IoU* value, as shown in Tab.1. It can be seen that the *IoU* value of our proposed network has remarkable advantages. Especially in the last two lines, our improved U-Net semantic segmentation algorithm has significantly improved the segmentation *IoU* of pedestrians, traffic lights, cyclists and other targets, with a maximum increase of 14.2% (cyclists). The *MIoU* value of the overall category reached 68.8%, which is 9.2% higher than that of the U-Net network algorithm before improvement.

Finally, ablation experiments were done for our improved optimization models, such as deep separable convolution, channel attention used in shallow network, double attention used in intermediate network, and spatial attention used in deep network, and they are respectively verified, which is shown in Tab.2. The ablation experiment results show that the accuracy is greatly improved by our optimization models.

Tab.1 *IoU* comparison on common categories with state-of-the-art image segmentation nets (%)

Categories	Road	Buildings	Traffic lights	Traffic signs	Pedestrians	Riders	Cars	Bus	ALL (*MIoU*)
SegNet	97.2	88.8	47.5	44.5	57.1	30.7	82.1	68.5	55.6
DeepLab v3+	93.3	81.3	41.6	52.1	48.7	54.7	82.8	77.4	63.5
HRNet	94.2	86.3	40.8	54.2	63.2	50.2	84.5	67.6	58.4
U-Net	92.1	86.0	53.5	63.5	65.9	40.6	89.7	63.5	59.6
Improved U-Net	96.7	86.1	65.0	66.0	78.4	54.8	89.9	68.7	68.8

Tab.2 Ablation experiments for our improved optimization models

Depthwise separable convolution	Channel attention module	Spatial attention module	Dual attention module	*MIoU* (%)
×	×	×	×	59.6
√	×	×	×	63.5
√	√	×	×	65.9
√	√	√	×	66.7
√	√	√	√	68.8

The basic U-Net model was improved in this paper, and attention mechanism was introduced according to the feature extraction characteristics, such as spatial attention module, dual attention module and channel attention module. To realize a semantic segmentation network model of urban street images with both speed and accuracy, the traditional convolution is replaced by the depthwise separable convolution. The experimental results show that the feature representation capability of the feature extraction network was improved due to the combination of the attention module, and our improved U-Net model can get better performance. Through test on the Cityscapes dataset, the *PA* is 90.3% and the *MIoU* is increased by 9.2%. Meanwhile, the use of depthwise separable convolution reduced the number of parameters of the original U-Net model from 36M to 13M, which greatly reduces the computational complexity and increases the speed of the network. To be specific, the improved U-Net network model has a computing time of 38 ms/frame and can process images up to 27 frames per second, which meets the real-time requirements for semantic segmentation of autonomous driving traffic scenes.

Statements and Declarations

The authors declare that there are no conflicts of interest related to this article.

References

[1] ZHOU J M, LI B J, CHEN S Z. A real-time segmentation method of road scene based on multi-layer feature fusion[J]. Surveying and mapping bulletin, 2020, (1)：10-15.

[2] MO Y, WU Y, YANG X, et al. Review the state-of-the-art technologies of semantic segmentation based on deep learning[J]. Neurocomputing, 2022, 493：626-646.

[3] BAI J, HAO P H, CHEN S H. Traffic scene understanding using lightweight convolutional neural network image semantic segmentation[J]. Journal of automotive safety and energy, 2018, 9(04)：433-440.

[4] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(4)：640 - 651.

[5] LIU W M, XIN Y L, JIANG X Y. Semantic segmentation of residual network image combined with jump connection[J]. Information technology, 2020, 44(06)：5-9.

[6] YANG C J. Image semantic segmentation based on convolutional neural network[D]. Lanzhou：Northwest Normal University, 2020：25.

[7] BADRINARAYANAN V, KENDALL A, CIPOLLA R, et al. SegNet：a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12)：2481-2495.

[8] YU F, KOLTUN V, FUNKHOUSER T. Dilated residual network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Hawaii, USA. New York：IEEE, 2017：472-480.

[9] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[EB/OL]. (2014-12-22) [2022-06-20]. https：//arxiv.org/abs/1412.7062.

[10] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab：semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 40(4)：834-848.

[11] ZHANG Y H, LIU H, TIAN W, et al. A method of rain cloud cluster segmentation in Tibet based on DeepLabV3[J]. Journal of computer applications, 2020, 40(09)：2781-2788.

[12] KUMAR P, SHANKAR H A. Convolutional neural network with batch normalisation for fault detection in squirrel cage induction motor[J]. IET electric power applications, 2021, 15(1)：39-50.

[13] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA, USA. New York：IEEE, 2019：3146-3154.

[14] RONNEBERGER O, FISCHER P, BROX T. U-Net：convolutional network for biomedical image segmention[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention, October 5-9, 2015, Munich, Germany. Berlin, Heidelberg：Springer-Verlag, 2015：234-241.

[15] CHEN Z, LI D, FAN W, et al. Self-attention in reconstruction bias U-Net for semantic segmentation of building rooftops in optical remote sensing images[J]. Remote sensing, 2021, 13(13)：2524.

[16] XIAO J Q. Semantic segmentation of road scene based on deep learning[D]. Changchun：Jilin University, 2019：23-27.

[17] WU T. Research on road scene semantic segmentation algorithm based on fully convolutional neural network[D]. Chongqing：Southwest University, 2020：14-16.

[18] YU F. Research and implementation of multi-scene image semantic segmentation based on fully convolutional neural network[C]//3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019), March 29-30, 2019, Dalian, China. Paris：Atlantis Press, 2019：156-161.

[19] CHEN Z, LI D, FAN W, et al. Self-attention in reconstruction bias U-Net for semantic segmentation of building rooftops in optical remote sensing images[J]. Remote sensing, 2021, 13(13)：2524.

[20] LUO P F. Research on semantic segmentation of autonomous driving city scene[D]. Wuhan：Wuhan University, 2019：16-22.

[21] YUAN X, SHI J, GU L. A review of deep learning methods for semantic segmentation of remote sensing imagery[J]. Expert systems with applications, 2021, 169：114417.

[22] ZHANG L, HU X, ZHOU Y, et al. Memristive DeepLab：a hardware friendly deep CNN for semantic segmentation[J]. Neurocomputing, 2021, 451：181-191.