Semantic segmentation based on semantic edge optimization

Hao Hu
School of Electronic and Information Engineering
Changchun University of Science and Technology
Changchun,China
1653583559@qq.com

Hua Cai*
School of Electronic and Information Engineering
Changchun University of Science and Technology
Changchun,China
caihua@cust.edu.cn
* Corresponding author

Zhiyong Ma
Department of Urology, the First Hospital of Jilin University
Changchun China
mazhiyong@jlu.edu.cn

Weigang Wang
Department of Urology, the First Hospital of Jilin University
Changchun China
wangwgzt@163.com

2021 International Conference on Electronic Information Engineering and Computer Science (EIECS) | 978-1-6654-1674-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/EIECS53707.2021.9587939

Abstract—Semantic segmentation based on deep learning is to extract semantic features by convolution, and then classify each pixel. In the process of feature extraction,the subsampled operation used to expand the receptive field will cause a large amount of detail information loss, resulting in the loss of small-scale objects in the image and blurred segmentation edge. In order to solve this problem, this paper proposes a segmentation model of semantic edge optimization, which uses the end-to-end semantic edge detection network to learn the semantic edge features of the image, and then fuses the semantic edge features with the semantic segmentation features, so that more edge information can be retained in the final segmentation image. On the camvid and cityscape datasets, our optimization model improve over original semantic segmentation model by 1.4% and 1.5% in terms of mean IoU.

Keywords - Semantic segmentation ; Semantic edge ; Edge detection; Feature fusion

I. INTRODUCTION

Semantic segmentation is to classify every pixel in the image according to the semantic information contained in the image, and annotate the objects in the image semantically[1]. It plays a key role in the research of computer vision, and is widely used in autonomous driving, medical diagnosis, intelligent security and other fields.

Traditional image segmentation is limited to the ability of computer, it can only extracting the color, spatial structure, texture information and other shallow features of the image to segment the image into several disjoint regions, so that these features show consistency or similarity in the same domain, and can not achieve semantic segmentation[2]. However, only through these low-level semantic information for image segmentation, the segmentation accuracy is far from meeting the requirements when the environment is more complex . With the development of computer, many researchers introduce deep learning into image segmentation. Convolutional neural network has strong learning ability and can learn the deep features of image [3], which makes segmentation more robust. Convolution is further used to learn the features of the object and train the classifier on the target region segmented by the traditional method, so as to realize the semantic annotation of the image [4].

In particular, the full convolution networks (FCN)[5] creatively replaces the full connection layer with deconvolution operation, so that a prediction is generated for each pixel, and the image level classification is extended to pixel level classification, which greatly improves the segmentation speed and accuracy, and greatly promotes the subsequent development of semantic segmentation. However, FCN and its subsequent segmentation algorithm still have some defects, because it uses pooling operation to expand the receptive field when extracting features. With the deepening of the network level, the semantic level is gradually improved, but the resolution is gradually reduced, resulting in the loss of a large number of target location and edge contour information, which makes the effect of the segmented edge blur.

In order to solve the problem of edge detail information loss, this paper proposes a semantic segmentation algorithm based on semantic edge optimization. Convolution network is used to learn the semantic edge information end-to-end, and the obtained semantic edge information is integrated into the semantic segmentation network to retain more edge information, so as to optimize the final segmentation result and segment clear edges.

II. RELATED WORK

A. Edge Detection

Before deep learning, most edge detection methods distinguish edges according to the gradient changes of image edges, such as Sobel operator [6] and Canny operator [7]. These methods can get good results in simple scenes. Because these traditional methods only extract the shallow object information, the edge detection effect is not ideal when there are other interferences such as background and light in the scene, The robustness is not high. With the emergence of deep learning, there are many edge detection methods based on convolutional neural network, such as DeepContour [8], HED [9].

DeepContour network uses deep convolutional neural network (DCNN) to learn the discriminant features of contour

978-1-6654-1674-0/21/$31.00 ©2021 IEEE

detection, and uses structured forest as the contour and non contour classifier of depth features. HED uses full convolution neural network and depth supervision network to fully extract the shallow and deep information of the image, which makes the edge extraction more accurate and the detection effect more stable.

B. Edge optimization for semantic segmentation

Effective use of edge information can improve the performance of segmentation. Deeplab[10] uses CRF to optimize the result of semantic segmentation, so that the edge of segmentation is more accurate. Aiming at the complex problem of CRF calculation,Chen et al.[11] uses domain transform to preserve the edge information of image, which improves the segmentation speed without losing the segmentation accuracy. However, because CRF only uses shallow features such as color and texture to optimize the segmentation result, the robustness is not high.Huang et al.[12] added an edge detection module to the original semantic segmentation structure. The input image first enters the semantic segmentation module, and then obtains the target shape details from the segmentation module.

III. APPROACH

The innovation is to extract the semantic edge information of the object and integrate it into the semantic segmentation network, so that the segmentation results retain more edge information, Moreover, edge with semantics can effectively improve the problem of no difference between classes which caused by the similarity of external features of adjacent objects,enhance the extraction of semantic features.

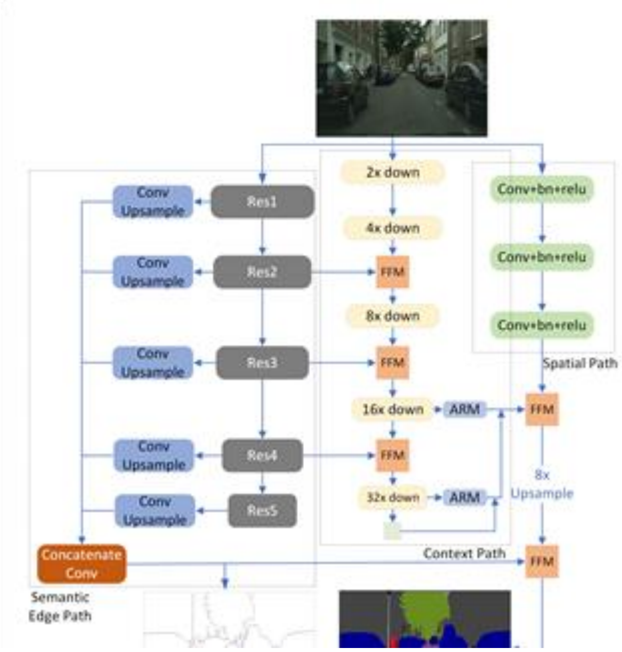

Figure 1. Overview of Our Method

The semantic segmentation model is Bisenet[13], which has good segmentation performance. A path to extract semantic edge is added on the basis of the original network, and the Bisenet network is fine tuned. The extracted semantic edge features are integrated into the context path to improve the accuracy of semantic segmentation. The network model proposed in this paper is shown in Figure 1, which is composed of three paths.

A. Semantic Edge Path

In this paper, HED network is used for semantic edge extraction, and ResNet18[14] is selected as the baseline model. The pooling layer in ResNet18 is removed and the structure is slightly adjusted to avoid excessive loss of edge detail information, retain more edge location information, and further improve the detection accuracy. On the side of each layer of ResNet18 network, there is an output layer for monitoring.The output of all layers is connected to realize the fusion of shallow contour information and deep contour information, and the network is supervised in depth to obtain more accurate edge detection results.

B. Semantic Segmentation Path

Compared with the accuracy of semantic segmentation model, the segmentation speed of the model is also particularly important. Bisenet designs a dual path network, one is spatial path, and the other is context path. Spatial path has more channels and shallow network, which is used to retain rich spatial information. The context path uses fewer channels and deeper network to get enough receptive field and extract deep semantic information. Then the feature maps extracted from the two paths are fused by a feature fusion module, and the final result is output. Realize the balance between speed and accuracy of the model.

Our model has made some modifications in the context path, using the feature fusion module to extract the edge information from the semantic edge path and the features extracted from the context path, so as to enhance the retention of the edge information, and the semantic information of the edge can also optimize the feature extraction of the context path, making the features extracted from the model more accurate.

Because the features extracted by these two paths are not the same, the feature of spatial path is shallow feature, while the context path is deeper feature. The feature graph extracted by them can not be fused by simple weighting. In order to realize the effective fusion of the two feature graphs, a feature fusion module (FFM) and attention refinement module (ARM) are designed.

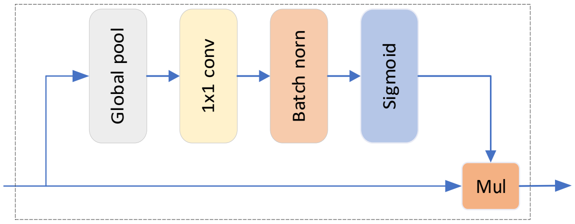

Figure 2. Attention Refinement Module

Attention Refinement Module(ARM), is used to refine the features in context. Global average pooling is used to obtain global semantic information, and an attention vector is calculated to refine the learned features, so that global semantic information can be obtained without up sampling.

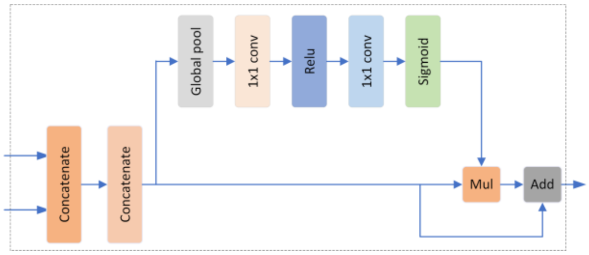

Figure 3. Feature Fusion Module

Feature fusion module(FFM),first concatenate the output features of the two paths and then use batch normalization to balance scales of features. Next, the feature pool is transformed into feature vector and calculate a weight vector. Using the weight vector to re-weight the feature which equivalent to feature selection and fusion.

C. Loss function

In this paper, a main loss function and three auxiliary loss functions are used to supervise the network training. The principal loss function is used to supervise the output of the whole network, and two auxiliary functions are used to supervise the output of the context path,these three loss functions all use softmax loss,as Equation 1 shows.

$$\text{loss} = \frac{1}{N}\sum_{\mathrm{i}} L_{\mathrm{i}} = -\frac{1}{N}\sum_{\mathrm{i}} \log\left(\frac{\mathrm{e}^{\mathrm{p_i}}}{\sum_{\mathrm{j}} \mathrm{e}^{\mathrm{p_j}}}\right) \qquad (1)$$

where p is the output prediction of the network.

The detection of semantic edge can be regarded as a multi-classification problem. The multi-class loss function is used in the semantic edge path,because the proportion of edge pixels in all pixels is very small, in order to avoid the problem of type imbalance, the weight is increased in the loss, as Equation 2 shows.

$$loss_{\mathrm{edge}} = \sum_{\mathrm{k}}\sum_{\mathrm{p}} \left\{ \begin{array}{l} -\theta \overline{Y}_{\mathrm{k}}(p) \log Y_k(p|I;W) \\ -(1-\theta)(1-\overline{Y}_k(\mathrm{p})) \log(1\text{-}\mathrm{Y}_k(p|\mathrm{I};\mathrm{W})) \end{array} \right\} \qquad (2)$$

where k is the number of semantic edge classes, $\overline{Y}$ is the label of image semantic edge, θ is the proportion of non semantic edge pixels to total pixels.

Furthermore,we use the parameter α and β to adjust the proportion of auxiliary functions in the total loss,as Equation 3 presents.

$$L(\mathrm{X};\mathrm{W}) = l_p(X;W) + \alpha\sum_{i=2}^{K} l_i(X_i;W) + \beta l_{edge}(E;W) \qquad (3)$$

where l_p is the principal loss of the whole network, l_i is the loss of context path and l_{edge} is the loss from semantic edge path. In depth supervision of network through joint loss function.

IV. Experiments

A. Evaluation Metrics

In this paper, mIoU(mean Intersection over Union) is used as the evaluation standard of segmentation effect. IoU is used to calculate the ratio of intersection and union between the prediction results of each class and the labels of the network output, and then mIoU is obtained from them,as Equation 3 shows.

$$\mathrm{m}I\mathrm{oU} = \frac{1}{K+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \qquad (4)$$

where k+1 is the number of classes ,i is the real value, j is the predicted value, $\mathrm{p_{ij}}$ means that the sample originally i is predicted as j.

B. Experimental Results

As shown in Table I，is our method statistic accuracy result on CamVid dataset.We use training data set and validation data set to train our model, using 960x720 resolution. From the data, we can see that the improved model is 1.4% higher than the original model in mIoU, and has different degrees of improvement in each category.

TABLE I. Results on the CamVid dataset

Method	Building	Tree	Sky	Car	Sign	Road	Pedestrain	Fence	Pole	Sidewalk	Bicyclist	Mean Iou(%)
SegNet	88.8	87.3	92.4	82.1	20.5	97.2	57.1	49.3	27.5	84.4	30.7	55.6
Enet	74.7	77.8	95.1	82.4	51.0	95.1	67.2	51.7	35.4	86.7	34.1	51.3
BiseNet（Xception）	82.2	74.4	91.9	80.8	42.8	93.3	53.8	49.7	25.4	77.3	50.0	65.6
EDANet	82.5	75.0	90.8	81.0	43.7	93.3	54.6	44.4	28.5	78.3	57.9	66.4
BiseNet（Res18）	83.0	75.8	92.0	83.7	46.5	94.6	58.8	53.6	31.9	81.4	54.0	68.7
Ours1 (Xception)	82.5	75.2	92.1	81.2	43.7	93.7	56.3	50.9	27.3	79.6	51.7	**66.9**
Ours1 Res18	83.2	76.4	92.3	84.4	48.2	94.8	59.6	55.2	33.6	83.3	57.1	**70.1**

As shown in Table II, the performance comparison between different strategies on Cityscape dataset.We achieve the best performance compared with these methods.In particular outperform Bisenet by 1.5%,which show that our method boost the performance.

TABLE II. Results on the cityscapes dataset

Method	BaseModel	Mean IOU(%)	
		val	test
DeepLab	VGG16	-	63.1
FCN-8s	VGG16	-	65.3
RefineNet	Res101	-	73.6
BiseNet	Xception39	72.0	71.4
BiseNet	Res18	78.6	77.7
BiseNet	Res101	80.3	78.9

As show in figure 4, the example results of our method and Bisenet based on ResNet18 model on Cityscapes dataset. By contrast, our method can make the edge segmentation more accurate. For example, in the first picture, the motorcycle and bicycle overlap,Bisenet fails to completely segment the front part of the motorcycle. In the second picture, ours method can well segment the distant lamp posts. In the third picture, the edge of the lamp post is more continuous and retains a more complete road surface.

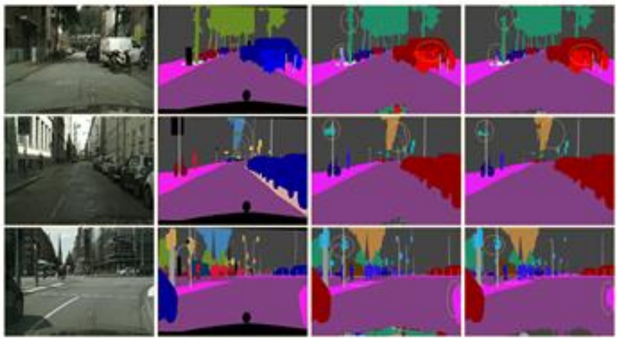

Figure 4. Example results of the output

V. Conclusion

The semantic edge optimization model proposed in this paper can effectively solve the problem of fuzzy edge segmentation. The edge detection network is used to extract the semantic edge information of the image, which not only solves the edge fuzzy problem of segmentation results, but also improves the inter class error caused by the similar shape of adjacent objects. Through the comparative experiment, we can see that our method has been improved in the segmentation accuracy, and we can also see that the edge of the object has been well segmented through the image.

References

[1] Jain A K,Dubes R C.“Algorithms for clustering data”. Technometrics, 1988, 32(2) : 227 － 229.

[2] LATEEF F,RUICHEKY.“Survey on semantic segmentation using deep learning techniques”.Neurocomputing, 2019,338: 321-348.

[3] WANGYR , CHEN Q L , WU JJ. “Research on image semantic segmen-tation for complex environments”. Computer Science, 2019, 46(9) : 36-46.

[4] Gu,J.,Wang,Z.,Kuen,J.,Ma,L.,Shahroudy,A.,Shuai,B.,Liu,T.,Wang,X.,Wang,L.,Wang,G.and Cai,J.,2015.“Recent advances in convolutional neural networks”.Ar Xiv preprint arXiv:1512.07108.

[5] Long J, Shelhamer E, Darrell T. “Fully convolutional networks for semantic segmentation”[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.K. Elissa, “Title of paper if known,” unpublished.

[6] KITTLER J. “On the accuracy of the Sobel edge detector”. Image and Vision Computing, 1983,1(1) : 37-42.

[7] CANNY J. “A computational approach to edge detection”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, PAMI-8（6） : 679-698.

[8] SHEN W , WANG X , WANG Y , et al,DeepContour:“A deep convolutional feature learned by positive-sharing loss for contour detection” ［C］/ / Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Ｒecognition. 2015: 3982-3991.

[9] XIE S,TU Z.“Holistically-nested edge detection” ［C］/ / Proceedings of the IEEE International Conference on Computer Vision. 2015: 1395-1403

[10] Chen LC , Papandreou G , Kokkinos I ,et al. Deep Lab : “Semantic image segmentation with deep convolutional nets , atrous convolution , and fully con-nected CRFs”. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 40(4) : 834-848.

[11] Chen LC , Barron J T,Papandreou G,et al.“Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform”[C]//Proceedings of IEEE Conference on ComputerVision and Pattern Recognition. Las Vegas , USA , 2016 : 4545-4554.

[12] Huang Q, Xia C, Zheng W, et al. “Object boundary guided semantic segmentation”[C]//Asian Conferenceon Computer Vision.Springer, Cham, 2016: 197-212

[13] Yu Yu C, Wang J, Peng C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation ［C］/ /Proceedings of the European Conference on Computer Vision (ECCV) , 2018: 325 － 341.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016. 1, 2, 3