International Core Journal of Engineering Volume 7 Issue 6, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0041

Optimizing Semantic Segmentation based on Correlation Between Pixels

Nengyuan Liu[1,a], Xiaohong Shi[1,b], Wenjun Lu[1,c] and Zhuangzhuang Li[1,d]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China.

[a]201930310119@stu.shmtu.edu.cn, [b]xhshi@shmtu.edu.cn, [c]luwenjun0207@stu.shmtu.edu.cn, [d]201930310269@stu.shmtu.edu.cn

Abstract

In the process of image segmentation based on pixel-level semantic segmentation, it is very important to make full use of context information. At present, the mainstream network structure used in semantic segmentation mainly considers multi-scale fusion. Although it can effectively increase spatial details and semantic information, it lacks effective context information. Therefore, some pixels are incorrectly classified in the process of image segmentation, which leads to inaccurate edge segmentation between classes and over segmentation, and its accuracy is reduced. In order to solve the above problems, a standard strategy based on inter pixel correlation evaluation is proposed to optimize the accuracy of inter class edge segmentation. We define outliers based on credibility and pixel correlation. The outliers of credibility are determined by the credibility of semantic segmentation network, and outliers based on pixel correlation are defined by qualitative mapping algorithm. If a pixel not only meets the credibility outliers but also the pixel correlation outliers, then the pixel is divided into the category of background or excluded from the segmentation result. We use images of Pascal VOC 2012 dataset to verify and evaluate the proposed method. The results show that this method is helpful to the optimization of semantic segmentation.

Keywords

Semantic Segmentation; Pixel Correlation; Outliers; Qualitative Mapping.

1. Introduction

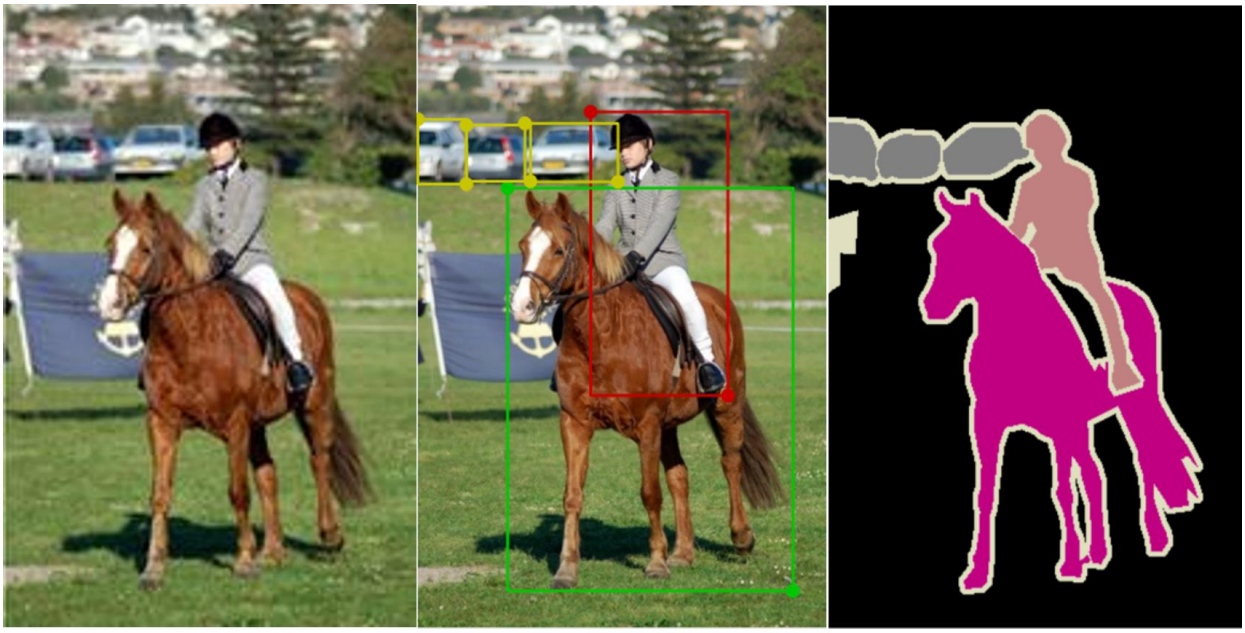
In recent years, the topic of artificial intelligence is very hot, especially in computer vision. The development of deep convolution neural network also pushes the performance of computer vision system to a new height, and image semantic segmentation is a very important research direction in the field of computer vision.

The simple understanding of semantic segmentation is the classification at the pixel-level. All pixels belonging to the same category should be classified into one category, so the semantic segmentation is to understand the image from the pixel-level. To understand the image, we need to obtain information from the classification of pixels and the spatial position of pixel categories to describe the image. Its task is to mark categories at the pixel-level of an image and apply directly to the field of computer vision [1] Image semantic segmentation can not only predict different categories in images, but also locate different semantic categories. This is different from target detection. Target detection only takes all the objects of interest in an image, including two subtasks of object location and object classification. At the same time, it determines the category and position of objects. However, each object of interest in the image is detected at the level of its minimum boundary

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0041

rectangle, and semantic segmentation needs further subdivision, which is based on pixel-level detection. The segmentation results are more accurate and intuitive.

The object detection and semantic segmentation are compared through an example, as shown in 0. The object detection covers the boundary box of the overlapping area of people, horses and vehicles, but the semantic segmentation is based on the pixel-level to identify people, horses and vehicles, and determine the obvious boundaries between each other.



(a) (b) (c)

Figure 1. Comparison of semantic segmentation and object detection. (a) represents input image; (b) represents object detection; (c) represents semantic segmentation.

Because the semantic segmentation process of image is pixel by pixel classification, and the edge pixels between heterogeneous regions are prone to misjudge the category or over segmentation. As in the two heterogeneous regions of man and horse in Fig. \ref{fig:1}, the edge is rough or the object that does not appear may be predicted and segmented due to the misclassification of pixel classification or over segmentation.

In this paper, we propose a method to optimize the accuracy of semantic segmentation. In order to effectively solve the problem of false classification or over segmentation in boundary points in the process of image semantic segmentation, we define outliers based on credibility and pixel correlation. The credibility data is derived from the classification value of each pixel in the image predicted by the full convolution neural network model, and the pixel correlation's data is based on the HSI color space and through qualitative mapping to determine the strength of the correlation between pixels. By judging and evaluating these two outliers, if a pixel is not only outliers based on credibility, but also outliers based on pixel correlation, the pixel will be removed or classified into the category of background. However, for one pixel that are not outliers based on pixel correlation, it is necessary to classify the pixel again by comparing the correlation between pixels, thereby optimizing edge segmentation and improving semantic segmentation accuracy.

2. Related Work

In the past research, texture primitive forest and random forest classifier are generally used to solve the semantic segmentation problem [2,3]. With the development of deep learning model, convolutional neural network (CNN) is used as the basic network in image semantic segmentation.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0041

In 2014, the Fully Convolutional Network (FCN) proposed by Long et al. of the University of California, Berkeley, promoted the original CNN structure and can make dense predictions without a fully connected layer [4,5]. FCN successfully transforms semantic segmentation into pixel-level tagging task. After that, different improved model structures such as ResNet [6], SegNet [7], U-Net [8] Based on CNN model structure appeared. Generally, the deeper the network is, the higher the level of features it has and the stronger expression ability. However, in fact, with the deepening of the network, there will be gradient disappearance and performance degradation problems. These improved model structures are to solve the gradient problem and the lack of feature information to varying degrees, but they did not effectively consider the reclassification of edge pixels between classes.

The basic idea of the algorithm based on integrated context information is to integrate features of different scales and to seek the optimal balance between local information and global information. Semantic segmentation is a pixel by pixel classification task, which needs to integrate different scale features to get local information and global information. From two aspects, local information can effectively improve the accuracy of pixel-level classification, and global information is also needed to deal with local ambiguity. The method based on the integration of context information has derived many algorithms, such as conditional random field [9], convolution with holes [10] and multi-scale prediction. There are many methods to improve the accuracy of edge segmentation. The general detection methods mainly use operators to mark the edge position as accurately as possible. The common edge detection operators are Sobel operator [11], Robert operator [12], Canny operator [13], etc. There are also seed optimization and simple linear iterative clustering methods to determine the category of pixels [14].

In this paper, based on the correlation between pixels to optimize the accuracy of semantic segmentation, a network structure model of multi-scale pooling and qualitative mapping module is designed. According to the credibility of the model prediction, whether the credibility value is based on the outliers value of credibility is analyzed, and the boundary points of the reclassification are obtained. On this basis, through the qualitative mapping method, we get the correlation data between pixels, redefine the boundary points, and finally get the semantic segmentation results.

3. Model

3.1 Overview

As shown in Figure 2, from the original image to the output semantic segmentation result image comparison, it can be found that if the network model predicts that the classification of the pixel is wrong, it is very easy to cause rough edge segmentation, and more seriously, it will predict objects that the image does not have. But because the pixel classification is wrong, it is judged that there is a certain type of object, and the object of this type is segmented.

(a) (b)

Figure 2. The accuracy of semantic segmentation is rough. (a) represents original image; (b) represents output image.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0041

The reason for this kind of error in model analysis and prediction is that although the model subsampled improves the semantic information, it also loses the spatial information. At the same time, because the correlation of pixels is not combined, the edges between categories appear rough, and objects that are not in the image are segmented. In order to effectively solve this problem, a method to optimize the accuracy of semantic segmentation is proposed.

3.2 Network structure

In order to accurately judge whether our proposed this method is effective, we adopted the mainstream network model framework of semantic segmentation: Fully Convolutional Neural Network (FCN) model, and made improvements on this basis. By understanding some of the problems of FCN, and in order to improve the accuracy of model prediction, we designed a multi-scale pooling (MSP) and qualitative mapping module. The basic framework of the network model is shown in Figure 3.

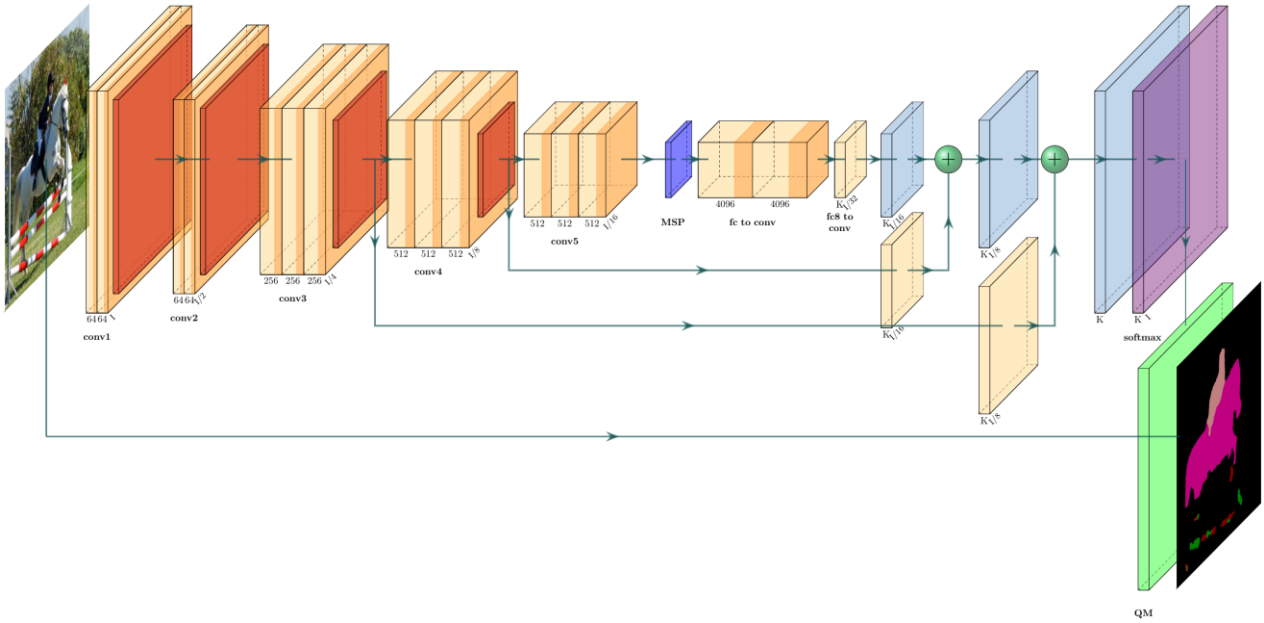


Figure 3. Basic frame diagram of the model. The framework of this model is fcn8s in the total convolution neural network. In the figure, MSP represents multi-scale pooling module; QM represents qualitative mapping module.

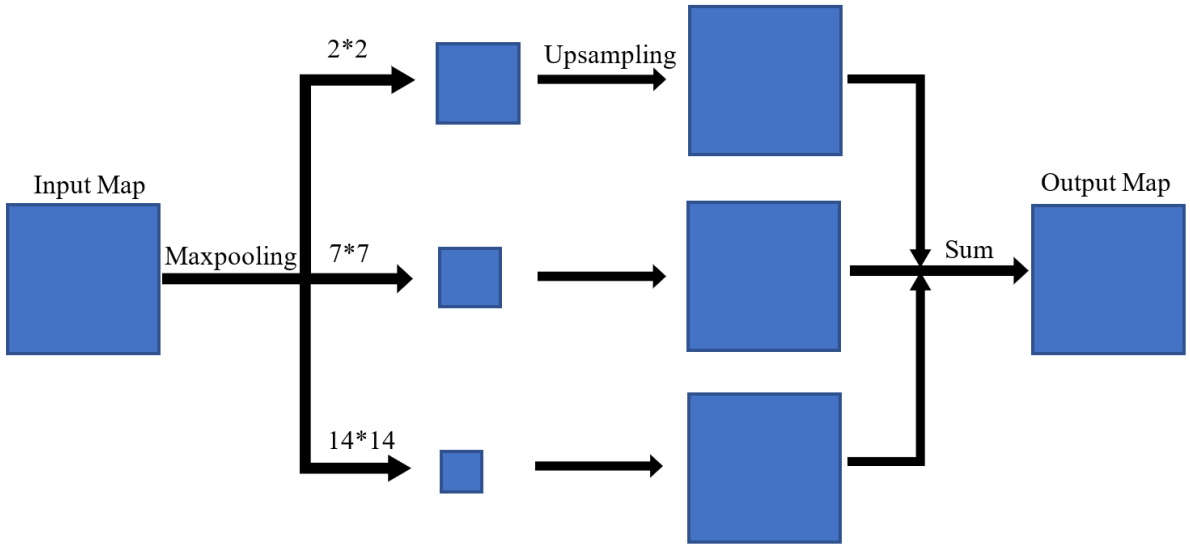


Figure 4. Multi-scale pooling structure diagram

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0041

The multi-scale pooling here uses several different pooling templates for extraction and fusion to obtain multiple scale features, thereby enhancing the power of image representation. We get the feature map output by the convolutional neural network, perform three pooling of different sizes to obtain three feature maps of different specifications. Then we upsampling the three feature maps, and use the weighted sum method to fuse the output feature maps. The detailed design is shown in Fig. 4.

3.3 Outliers of credibility

The confidence value is based on the output of the neural network model to evaluate the classification of each pixel, so each pixel corresponds to an *N*-dimensional vector:

$$P_{scores}=\{P_{score}^{1}, P_{score}^{2}, \ldots, P_{score}^{N}\} \quad (1)$$

P_{score}^{i} represents the probability that the pixel *P* belongs to category *i*, and *N* represents the number of categories in the dataset, that is, the total number of categories. The network model predicts the category of a pixel based on the category pointed to by the maximum credibility of the pixel. In order to obtain outliers of credibility, it can be concluded that when the maximum value of the pixel *P* credibility is very close to the maximum value of the remaining *N*-1 dimensional vector, it can be judged that the point is a outlier based on credibility. For example, the maximum credibility value of pixel *P* is P_{score}^{f}, followed by P_{score}^{s}, and the two values are very close, which shows that pixel *P* is likely to belong to s category, but the network model will predict and judge that the pixel belongs to class *f*.

Therefore, outliers based on credibility can be defined:

$$C\_outliers = \frac{P_{score}^{f} - P_{score}^{s}}{P_{score}^{f}} \quad (2)$$

Then the threshold *T* of $C\_outliers$ is set to judge whether each pixel is based on the outliers of credibility.

3.4 Outliers of pixel correlation

According to the qualitative mapping theory [15], we map the attributes of pixels to HSI color space model qualitatively, Therefore, each pixel *x (i, j)* attribute can be represented by the eigenvectors of H (Hue), S (Saturation) and I (Intensity). These three eigenvectors constitute a comprehensive attribute *X (i, j)*. The term *X (i, j)* can be given as

$$X(i,j) = H(i,j) \wedge S(i,j) \wedge I(i,j) \quad (3)$$

x(h,s,i) is the quantity value of attribute *X (i, j)*, and *h*, *s* and *i* are the quantitative eigenvalues of *H (i, j)*, *S (i, j)* and *I (i, j)* respectively. *h (i, j)* is a property of attribute *H (i, j)*, and so on. Here, the qualitative basis of *h (i, j)*, *s (i, j)* and *i (i, j)* property is as follows

$$\tau = [\alpha, \beta] \mid [\alpha, \beta] \quad (4)$$

Where [α, β] is a super long cube, which is the qualitative basis of the integrated property *X (i, j)*, as shown in Figure 5. The term [α, β] in (4) can be given as

$$[\alpha, \beta]=[h_1, h_2]\times[s_1, s_2]\times[i_1, i_2] \quad (5)$$

The mapping $f: X \times \tau \rightarrow \{0,1\} \times p$ is a qualitative mapping of *x = (h,s,i)* based on three-dimensional cube [α, β]. Here *p* represents the set of properties. If for any $x \in X$, there exists $[\alpha, \beta] \in \tau$ and the property $X(i,j) \in X_0(i,j)$ with [α, β] as qualitative basis, make (6) set up.

$$f(x, [\alpha, \beta]) = f(x \in ? [\alpha, \beta]) = f_h(h) \wedge f_s(s) \wedge f_i(i) \quad (6)$$

Where $\in ?$ definition indicates whether it belongs to, the term $f_h(h)$ in (6) can be given as

$$f_h(h) = \begin{cases} 1, h(i,j) \in [h_1, h_2] \\ 0, h(i,j) \notin [h_1, h_2] \end{cases} \quad (7)$$

$f_s(s)$ and $f_i(i)$ are similar to (7), They are the truth values of property propositions *H (i, j)*, *S (i, j)* and *I (i, j)*, respectively.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0041

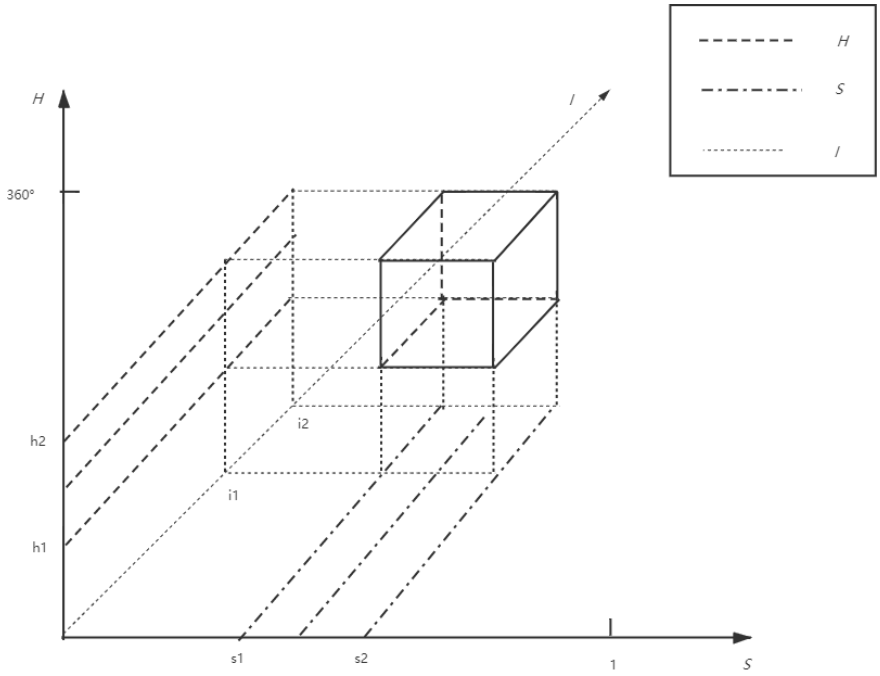


Figure 5. Super long cube graph with qualitative basis mapping in HSI color space

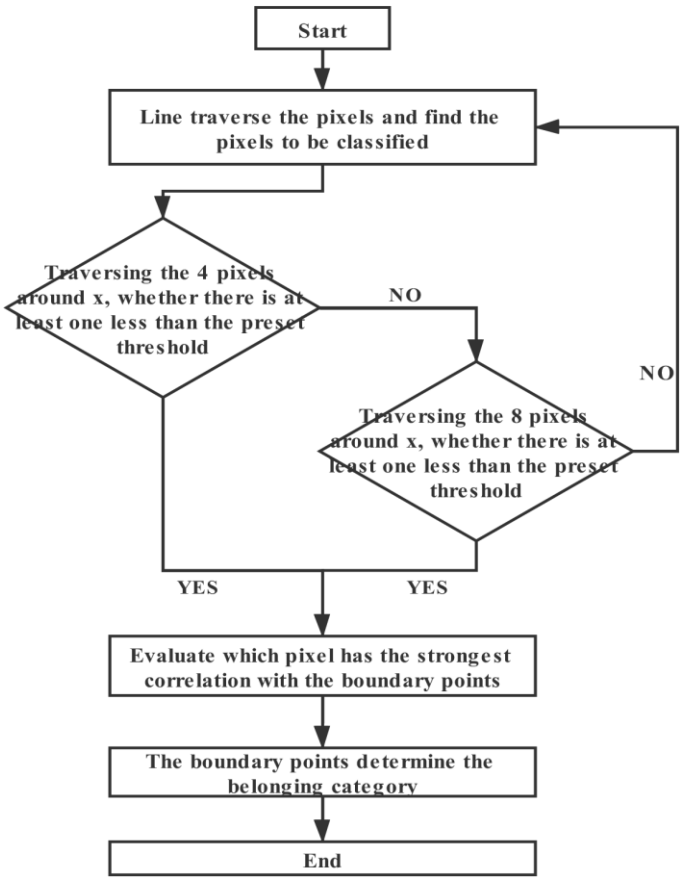


Figure 6. Flow chart of qualitative mapping algorithm

We can use the network model to determine which pixels are based on the outliers of credibility, and mark these pixels. Then we use the qualitative mapping method to further subdivide these marked pixels. The pixels here are actually the edge pixels between classes.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0041

The following describes how qualitative mapping determines the correlation between pixels:

According to the difference between the three representation vectors of pixel *x(i, j)* in the outliers of credibility and the eigenvectors of four or eight known pixels around it, we can compare whether it is within the given threshold, and judge whether it is the strongest correlation with a certain point, that is, the minimum difference. Therefore, it is determined that the pixel belongs to the same category as the pixel with the strongest correlation.

The specific algorithm flow is as follows, and its flow chart is shown in Figure 6:

Step1. The pixels in the image are line traversed to find the boundary point *x(i, j)* to be classified.

Step2. Firstly, four pixels around the boundary point *x(i, j)* are traversed and compared with their eigenvectors to determine whether at least one of them is less than the preset threshold. If so, skip to step 4, otherwise continue to step 3.

Step3. The eight pixels around *x(i, j)* are traversed and compared with their eigenvectors to determine whether at least one of them is less than the preset threshold. If not, then it is satisfied that both the outliers of credibility and the outliers of pixel correlation are satisfied. Therefore, the pixel is divided into the category of background. And return to step 1 to find the next boundary point.

Step4. The difference between the feature vectors of the boundary point and which pixel is the least.

Step5. The points to be classified are evaluated and classified to determine the category.

When we judge the correlation between pixels, we first judge whether it can be qualitative, that is, whether the occurrence of qualitative change can be controlled under the condition of quantitative change. Here, we use the starting threshold of the three attribute components of the comprehensive attribute *X* as a qualitative benchmark. Because these three components can be processed separately and are independent of each other in the HSI color space. We use the absolute value of the difference between the three feature vectors between pixels to create a color triangle. The three differences are $|\Delta s|$, $|\Delta i|$ and $|\Delta h|$. $|\Delta s|+1$ and $|\Delta i|+1$ are the two sides of the triangle, and $|\Delta h|+1$ is the angle between the two sides. Here the value is increased by one, the purpose is to prevent the occurrence of 0 value. By comparing the size of the triangle area to determine the correlation intensity. The area of the triangle is calculated as follows

$$A = 1/2(|\Delta s| + 1)(|\Delta i| + 1)\sin(|\Delta h| + 1) \quad (8)$$

4. Experiments and Results

4.1 Datasets

Pascal VOC [16] 2012 dataset is widely used in FCN model, which mainly provides label data for image visual tasks. Because we are improving based on the FCN model, In order to determine whether the method proposed in this paper is feasible or not, we use this dataset.

4.2 Experimental details

We use mean intersection over union (MIoU) as the measurement standard.

In the experiment, we compared the thickness of the boundary line by comparing the different thresholds of marking the boundary points to be reclassified, and finally determined that the threshold *T* was 0.2. Therefore, the boundary points based on the outliers of credibility are obtained. In the process of reclassification, we set the threshold values of three color components as $dh = 30$, $ds = 0.2$, $di = 20$. Through qualitative mapping evaluation, the correlation between pixels is judged, and then the marked boundary points are reclassified to get the semantic segmentation results.

4.3 Results analysis

Since the purpose of this paper is to improve the edge segmentation accuracy of semantic segmentation, we test 21 types of images in Pascal VOC 2012 data set in the experiment, and compare the data before and after using this method, as shown in Table 1.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0041

From the experimental data, it can be concluded that the MIoU before the boundary point re classification is about 63.1%, and after the screening and re classification based on two outliers is about 65.4%, which is a certain improvement compared with 62.6% of fcn8s.

We do semantic segmentation through the dataset or some other images, and the result shows the contrast graph, as shown in Figure 7. From the semantic segmentation result graph of the model output, It can be seen that this method helps to classify edge pixels between classes and make the segmentation edge result more smooth and accurate.

Table 1. IoU DATA PER CATEGORY

Class	Before(IoU(%))	After(IoU(%))
Person	74.7	78.4
Bird	72.4	73.6
Cat	75.3	79.5
Cow	69.4	72.1
Dog	67.7	70.2
Horse	69.6	70.4
Sheep	72.4	74.6
Aeroplane	74.9	77.8
Bicycle	29.4	30.8
Boat	50.6	54.3
Bus	79.1	79.8
Car	72.7	76.8
Motorbike	70.9	70.9
Train	71.7	72.7
Bottle	63.3	64.9
Chair	20.6	22.4
Table	43.8	46.8
plant	50.1	52.4
Sofa	48.6	49.3
Tv/monitor	59.5	62.1
Background	89.1	92.9

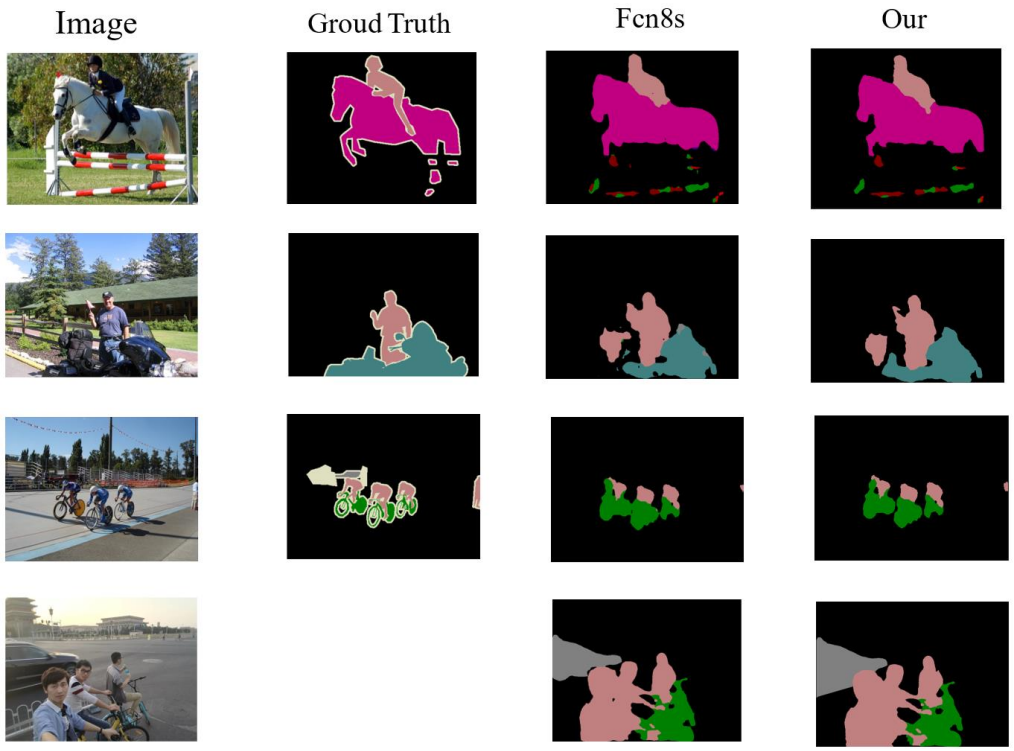


Figure 7. The results of fcn8s and this model are obtained by using the pictures in three datasets. The last row is not a dataset image, the main content of which can be clearly seen as bicycles, cars and humans.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0041

5. Conclusion

From the experimental results, we can see that multi-scale pooling and qualitative mapping proposed by us can improve the segmentation accuracy. This qualitative mapping can map a research object into different coordinate systems through some of its attributes. Here, we map pixels to HSI color space, that is to map the one-dimensional pixels of the picture to the three-dimensional coordinate system, and get the qualitative basis of the comprehensive attribute through these three attributes, so as to determine the strength of the correlation between the pixel to be classified and the classified pixel.

References

[1] E. Romera and J. M. Álvarez and L. M. Bergasa and R. Arroyo.: ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. In:IEEE Transactions on Intelligent Transportation Systems, 19(1):263-272(2018)

[2] Hay,G.J. and Niemann, K.O. and Mclean,G.F.: An object-specific image-texture analysis of H-resolution forest imagery. In: Remote Sensing of Environment,55(2):108-122(1996)

[3] J. Shotton et al.: Real-time human pose recognition in parts from single depth images. In: CVPR, pp.1297-1304, Providence, RI(2011)

[4] J. Long and E. Shelhamer and T. Darrell.: Fully Convolutional Networks for Semantic Segmentation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence,39(4):640-651(2015)

[5] Evan Shelhamer, Jonathan Long, Trevor Darrell.: Fully Convolutional Networks for Semantic Segmentation. In: IEEE Computer Society(2017)

[6] K. He and X. Zhang and S. Ren and J. Sun.: Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770-778, Las Vegas, NV(2016)

[7] V. Badrinarayanan and A. Kendall and R. Cipolla.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12):2481-2495(2015)

[8] Ronneberger O, Fischer P, Brox T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention. Springer International Publishing, pp.234--241(2015)

[9] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc, pp.282–289(2001)

[10] Yu, Fisher, and Vladlen Koltun.: Multi-scale context aggregation by dilated convolutions. In: arXiv preprint arXiv:1511.07122(2015)

[11] Lin Hong, Yifei Wan and A. Jain.: Fingerprint image enhancement: algorithm and performance evaluation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence,20(8):777-789(1998)

[12] Rosenfeld,Azriel. : The Max Roberts Operator is a Hueckel-Type Edge Detector. In: IEEE Transactions on Pattern Analysis and Machine Intelligence,3(1):101-103(1981)

[13] Canny, John.: The complexity of robot motion planning. MIT Press (1987)

[14] Arunkumar, Manonmani and Pushparaj, Vijayakumari.: Semantic segmentation Using Seed Picking Crossover Optimization algorithm. In: IET Image Processing, 14(11):2503-2511(2020)

[15] J. Feng.: Attribute Grid Computer based on Qualitative Mapping and its application in pattern Recognition. In: IEEE International Conference on Granular Computing, pp.154-161(2009)

[16] Everingham, M., Van Gool, L., Williams, C.K.I. et al.: The PASCAL Visual Object Classes (VOC) Challenge. In: International Journal of Computer Vision, 88(2):303-338(2010)