Computer Science and Information Systems 19(3):1409–1426 https://doi.org/10.2298/CSIS220321040T

BiSeNet-oriented Context Attention Model for Image Semantic Segmentation

Lin Teng and Yulong Qiao*

College of Information and Communication Engineering,
Harbin Engineering University
Harbin 150001, China
tenglinheu@163.com
qiaoyulong@hrbeu.edu.cn

Abstract. When the traditional semantic segmentation model is adopted, the different feature importance of feature maps is ignored in the feature extraction stage, which results in the detail loss, and affects the segmentation effect. In this paper, we propose a BiSeNet-oriented context attention model for image semantic segmentation. In the BiSeNet, the spatial path is utilized to extract more low-level features to solve the problem of information loss in deep network layers. Context attention mechanism is used to mine high-level implied semantic features of images. Meanwhile, the focus loss is used as the loss function to improve the final segmentation effect by reducing the internal weighting. Finally, we conduct experiments on open data sets, and the results show that pixel accuracy, average pixel accuracy, and average Intersection-over-Union are greatly improved compared with other state-of-the-art semantic segmentation models. It effectively improves the accuracy of feature extraction, reduces the loss of feature details, and improves the final segmentation effect.

Keywords: image semantic segmentation, BiSeNet, context attention, focus loss.

1. Introduction

The different elements in the image are formed by combining various pixels together. Therefore, the method of classifying these pixels by elements is called image semantic segmentation. As a core technology in computer vision research, semantic segmentation has many advantages in pixel-level prediction and image classification by using advanced semantic features of images [1,2]. At present, it has been widely used in medical and health care, storage management, traffic safety and many other fields, it has important research value and significance. Image semantic segmentation based on deep learning is a hot topic in recent years. As a large number of deep learning methods which have been successful in image classification, object detection, natural language processing and other fields have been improved and migrated to the field of semantic segmentation. The semantic segmentation technology has made great breakthrough and gradually changed the development trend of various industries [3,4].

In the process of image semantic segmentation, difficulties and challenges are usually faced in the aspects of target, category and background [5]. For the target, even if it is the

* Corresponding author

same target, if the illumination, Angle of view and distance are different or in the static and moving state, the images will be different. And even there will be mutual occlusion between adjacent targets. In terms of categories, there are still differences between the same category, and there are also similarities between the different categories. As for the background, the background in the real scene is relatively complex, which brings great difficulties to the semantic segmentation [6]. For traditional semantic segmentation such as gray segmentation and conditional random fields, the underlying features of the image are usually used to divide the region of the image, and its segmentation accuracy needs to be further improved. At present, with the development of convolutional neural network algorithm (CNN) and its application in semantic segmentation, a large number of semantic segmentation models based on deep learning have been proposed, which can solve the problem of difficult feature selection in traditional semantic segmentation [7].

The application of convolutional neural network (CNN) has made rapid progress in image semantic segmentation. Various semantic segmentation networks based on convolutional neural networks have been proposed. At present, there are roughly three kinds of researches in the field of image segmentation. 1) Improving the segmentation performance by improving the structure of the convolutional network [8,9] and combining with deeper neural networks. Russo et al. [10] proposed an image segmentation algorithm with low parameter number based on convolutional neural network. By improving the deep-level neural network and applying multi-scale dilated convolution, the algorithm increased the scale standardization layer and optimized the network to improve the segmentation effect while reducing the number of parameters. 2) Based on the encoder-decoder architecture [11,12], a variety of methods are adopted to extract feature information to improve the resolution of feature maps and thus improve the segmentation effect. Tian et al. [13] proposed an improved DUpsampling algorithm based on DeepLabV3+ architecture. In the decoding module, a DUpsampling method was adopted to replace bilinear interpolation, which improved the segmentation accuracy and reduced the computational complexity while restoring the size of the feature graph. 3) The attention mechanism [14,15] is used to model the target feature information, so as to highlight the detail information of the feature map and improve the segmentation effect. Li et al. [16] proposed a pyramid attention network (PAN) for image segmentation, which combined feature pyramid attention network (FPA) and global up-sampling attention network (GAU) to replace the dilated space convolution pooling pyramid (ASPP) structure for feature extraction. In the face of complex scenes, the above methods are prone to large segmentation errors. Therefore, in order to solve the above problems and improve the utilization rate of high-level and low-level feature information, a new method based on BiSeNet and context attention model is proposed in this paper.

The rest of this paper is organized as follows. In the second section, we reviewed more related work. In the third section, the description method of the proposed image sematic segmentation is introduced. Then, in the fourth section, we conduct an experimental analysis. Finally, we summarize the paper.

2. Related works

Traditional image segmentation algorithms are based on the color, texture information and spatial structure of the image, and the semantic information of the same region is consis-

tent, but the attributes of different regions are also different. There are many segmentation methods, mainly including simple broad-value segmentation, region growth, edge feature detection and graph division [17]. Reference [18] proposed to use the structured forest method to generate edge probability, and utilize watershed algorithm to transform edge probability into initial cut blocks. In order to avoid over-segmentation, the hypermetric contour graph algorithm was used to select appropriate broad values to generate segmentation blocks to obtain more accurate contour information, and random forest was used to train segmentation blocks to obtain semantic segmentation results. Reference [19] proposed a hierarchical graph partition method, namely, the Oriented Image Foresting Transform (OIFT), which could be customized for the target object group according to its boundary polarity. This method had a small number of image partitions and could accurately isolate the desired target region with known polarity. The local contrast of the image region was used to make it robust to illumination variation and non-uniformity effect. Because no data training was required, the calculations were relatively simple. However, if the segmentation task was difficult, the performance of segmentation should be further improved. Reference [20] proposed an image segmentation method combining global image features with complete convolutional networks. The method used the parameter learning process of the unified deep learning model embedded in the full convolutional network to encode the whole image content and make the segmentation more reasonable and accurate. This kind of method basically obtained the underlying features through the use of artificial design features, and its segmentation efficiency could not well meet the actual requirements.

The semantic segmentation method based on deep learning automatically learns data features instead of using artificial data features, which is different from traditional image segmentation methods. End-to-end semantic segmentation prediction can be completed by using deep neural networks [21]. The three most important processes in deep learning include feature extraction, semantic segmentation and post-processing. After that, many models such as FCN, VGG16, ResNet or deep network semantic segmentation are developed. Reference [22] proposed a method based on ResNet network to fuse shallow feature image information with deep feature image by defining parallel branches. The features were extracted and fused by parallel dilated convolution with different sampling rates, so as to effectively extract the features and context information of different layers. In order to improve the stability of parameter tuning, batch normalized calculation was introduced into the new module. The defect of the convolutional network was that it was insensitive to image details due to its low spatial resolution, and the edge of segmentation was relatively rough. Reference [23] proposed a weakly supervised learning algorithm with size constraints based on improved deep convolutional neural network for image segmentation. Compared with the existing complete supervision methods, the image segmentation process only used the image-level label and the boundary box label to guide, which was easier to implement. Its disadvantage was that the target information was not enough, and the context information would be lost, so that the boundary could not be accurately located. Reference [24] proposed a novel DenseGram network, which could reduce gaps and segmented degraded images more effectively than traditional strategies. Experimental results showed that the proposed dense-Gram network produced the latest semantic segmentation performance on degraded images using PASCAL VOC 2012, SUNRGBD, CamVid and CityScapes data sets. At present, there is no mechanism or structure that can

make the current network deliberately learn the differences between different categories, which also leads to the high-level semantic features sometimes share the information of the target and its own background. The segmentation of the target is not accurate. In reference [25], a multi-scale semantic segmentation model based on deep residual network was proposed. It was mainly used to enhance the segmentation accuracy of remote sensing image of different scale objects in small sample remote sensing image dataset. Although the end-to-end semantic segmentation model structure was implemented, due to the emphasis on feature understanding and target category prediction, the problem of inaccurate positioning between target and background or the boundary of different targets was caused.

Jiang et al. [26] proposed a deep architecture that could run in real time, using residual connection and decomposition convolution to maintain high efficiency and good accuracy. Yi et al. [27] proposed an efficient spatial pyramid module (ESP) based on extended convolution, enabling it to perform efficiently in terms of computation, memory and accuracy. Grant-Jacob et al. [28] proposed a new context-guided network (CGNet), which could effectively learn the joint features of local features and surrounding context, and further improve the joint features through the surrounding context features, so as to improve the real-time performance and accuracy of the network. Li et al. proposed a single lightweight backbone network to aggregate and identify features respectively through sub-network and sub-cascade, so as to reduce the number of parameters and still obtain enough receptive fields, thus enhancing the learning ability of the model and achieving a balance between speed and segmentation performance. BiSeNet(Bilateral Segmentation Network) [29] divided the Segmentation task into two parallel modules (spatial path module and context path module), which took advanced features and receptive fields into account and significantly improved the detection speed of the network. In conclusion, due to the limitations of the image segmentation method based on manual design features, this paper proposes a BiSeNet semantic segmentation network based on context content. Firstly, the overall structure of the improved segmentation network and differences from the original bilateral segmentation network are described, and the role of the proposed feature fusion module in the context path is emphasized. Secondly, the subnetwork feature fusion module which is used to aggregate features of different depth is described in detail. At the same time, focal loss is used as the loss function to solve the problems of unbalanced sample number of different categories and different object differentiation difficulties, so as to improve the accuracy of target recognition and improve the segmentation efficiency. Finally, experiments are carried out on public image data and comparison with other state-of-the-art networks. The results prove the effectiveness of the proposed method.

3. Proposed Image Sematic Segmentation Model

Semantic segmentation technology is one of the main tasks of computer vision. It is based on the pixel level of the image to some regions of the image corresponding semantic labels. In recent years, in order to meet the requirement of semantic segmentation accuracy, semantic segmentation model technology has made some progress. However, the current mainstream real-time semantic segmentation model acceleration methods are compromise accuracy for speed. For example, in image processing, it cuts the original image or

directly changes the size of the original image to limit the network input size to reduce the computational complexity. Although these methods are simple and effective in improving network speed, the loss of spatial detail still affects the detection effect, especially the boundary part, resulting in a decrease in measurement and visualization accuracy.

BiSeNet consists of two components: Spatial Path (SP) and Context Path (CP). The former solves the problem of spatial information loss in deep network by acquiring more low-level features. The latter mainly solves the problem of receptive field constriction. The BiSeNet is shown in figure 1.

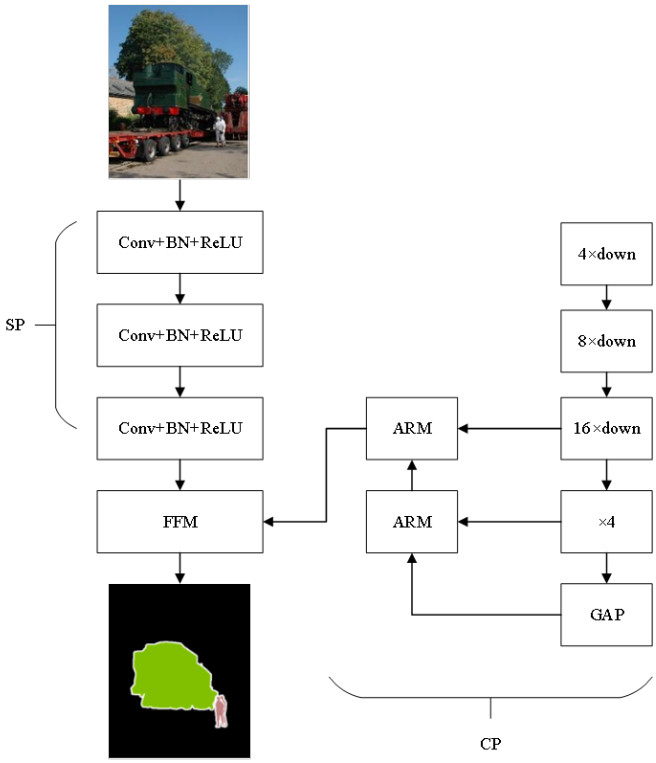

Fig. 1. BiSeNet structure

The overall structure of the improved BiSeNet semantic segmentation network model is also divided into two branches: Spatial Path (SP) and Context Path (CP).SP module is used to obtain high-resolution feature maps and obtain more accurate Spatial information. Its structure consists of three convolutional layers, each of which contains a convolutional layer with step size of 2. After batch standardization processing and ReLU nonlinear activation, the size of the output image through this path is 1/8 of the original image.

CP module enables the network to obtain a larger receptive field. In order to ensure the accuracy and improve the computational speed, Xception model is adopted in the backbone network as a lightweight feature extraction network [30]. Xception can perform fast

down-sampling operation to obtain a large receptive field. The improved BiSeNet semantic segmentation network model is shown in figure 2. The features of different depths are aggregated to obtain the sub-network feature fusion module, and the advanced features are further processed to refine the advanced features. At the same time, feature maps of the same size at each stage of the backbone network are fused to make the context path module possess more low-level features and spatial information, retain spatial details of image structure, and improve its judgment ability of large-scale targets and fine structure edges.

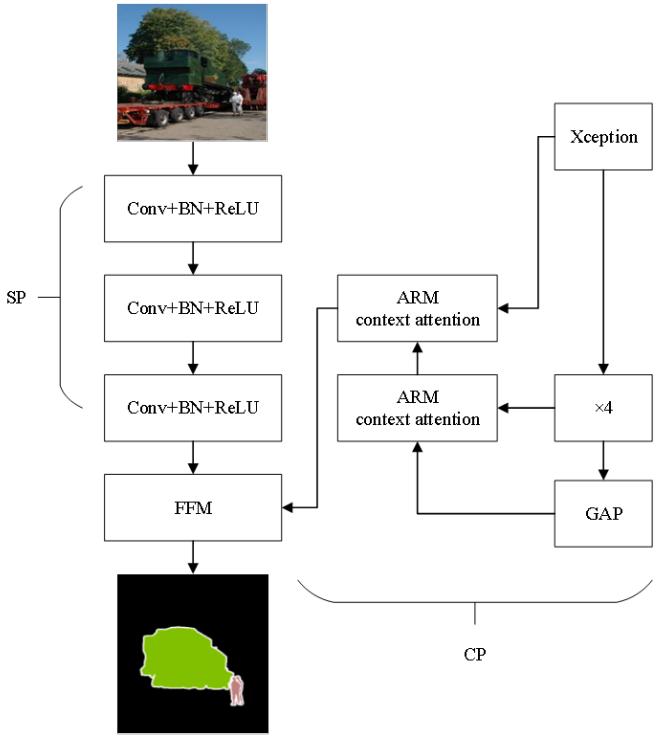

Fig. 2. Overall structure of proposed network model

The sub-network feature fusion module uses the output of the main Xception network as input to refine the features and further improve the network performance. Then, a Global Average Pooling (GAP) layer is added to the tail of the subnetwork feature fusion module to obtain larger receptive fields. Then, using the Content Attention Module (CAM), On the basis of obtaining the global context of the original image through global average pooling, CAM further calculates the attention vector to guide feature learning, and then up-sampling through bilinear interpolation is used to make the size of feature map and SP spline feature the same size. The Feature Fusion Module (FFM) is used to connect the output features of spatial path and context path. Then, the connected feature pool is converted into a feature vector and a weight vector is calculated by batch nor-

malization and balancing the scale of features. This weight vector can be re-weighted to achieve the feature output combining SP and CP.

3.1. Xception model

Xception model is a network based on deeply separable convolution, which is implemented by replacing Inception module with deeply separable convolution on the basis of Inception-v3. Xception model shows good image classification results in ImageNet [31], and the calculation speed is very fast. Chen et al. [32] proposed a encoder-decoder with detachability convolution for semantic image segmentation network. The encoder-decoder introduces Xception model to complete the task of semantic image segmentation, and improves the Xception model by combining TensorFlow deeply detachability convolution. A more dense feature graph is extracted by using depth-separable convolution instead of dilated convolution, and the structure of the Xception model is shown in figure 3. Where A, B, C denote entry flow, middle floe and exit flow respectively.

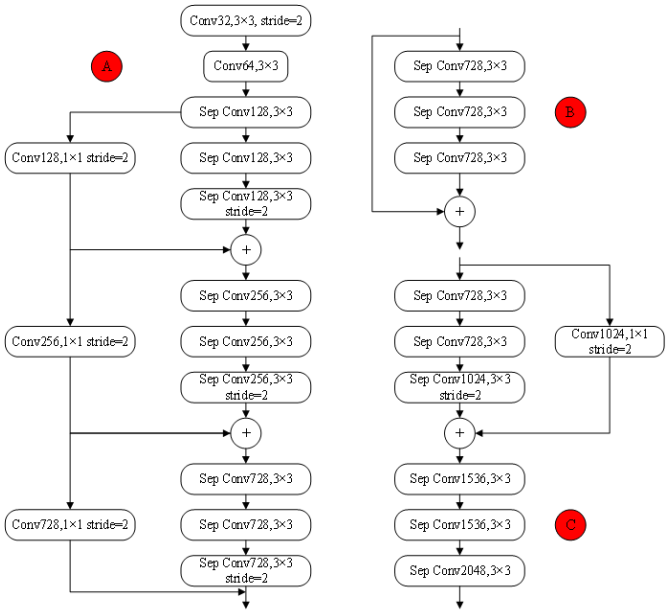

Fig. 3. Structure diagram of Xception module

On the basis of the original model, more layers are added and the network structure of inlet flow is not modified, which achieves fast computation and high storage efficiency. All maximum pooling operations are replaced by step-depth detachable convolution in order to extract feature images by using voidable convolution at arbitrary resolution. Additional batch normalization and ReLU activation are added after each 3×3 deep convolution.

3.2. Context attention model

In the practical application of image semantic segmentation, there is a large amount of data, so the calculation cost can be reduced by establishing a k-nearest neighbor graph $G = (V, E)$ to represent local areas [33]. $V = 1, 2, \cdots, N$ is the point set, and $E \subseteq V \times \phi_i$ is the adjacent edge of the adjacent point pair. ϕ_i is the set of points in the neighborhood of point x_i. In order to prevent the point set from being affected by rotation transformation, the coordinate x_{ij} of the point in the local area is transformed into the relative coordinate of the central point x_i. The obtained edge characteristics are expressed as follows:

$$Fy_{ij} = (x_i, x_{ij} - x_i), x_i \in \Re^F. \tag{1}$$

$$\forall x_{ij} \in Neighbors(x_i), x_i \in V, x_{ij} \in \phi_i. \tag{2}$$

In order to fully mine fine-grained details and multi-scale context information of images, context attention convolution layer is established based on Point Net. The encoding methods mainly include attention coding and context recurrent neural network coding. Attention encoding mainly learns fine-grained features in local regions. Context recurrent neural network coding learns multi-scale context geometry features between local regions. Where, multi-layer perceptron operation is represented by $MLP(*)$, and the number of convolution kernels is represented by $*$.

The attention encoding mechanism generally selects MLP first [34], and the output channel of the selected MLP is F1. Then, the selected MLP is used to map the original point features and edge features to the feature space with higher dimensions. The following is the specific representation.

$$u_i' = \sigma\Theta(\kappa(f_{F\times 1}(x_i))). \tag{3}$$

$$h_i' = \sigma\Theta(\kappa(f_{F\times 1}(y_i))). \tag{4}$$

Where, the nonlinear activation function is denoted as σ after parameterization. The set of parameters that can be learned in the convolution kernel is expressed as Θ. κ stands for batch normalization. f is convolution operation. The subscript $F \times 1$ is the size of the convolution kernel. In this experiment, the value of F1 is 16, that is, the number of feature channels is 16. MLP is used to process u_i' and h_i', and the self-attention coefficient and neighborhood attention coefficient of x_i are respectively generated. By combining the two coefficients, the attention coefficient c_{ij} from the center point x_i to k neighboring points in the neighborhood can be obtained as:

$$c_{ij} = Selu(\sigma\Theta(\kappa(f_{1\times 1}(u_i')))) + \sigma\Theta(\kappa(f_{1\times 1}(h_i')))). \tag{5}$$

Where, the nonlinear activation function is $Selu()$. Softmax function is used to normalize the attention coefficient, so that the convergence efficiency of the model is improved.

$$a_{ij} = exp(c_{ij}) / \sum_{j=1}^{k} exp(c_{ij}). \tag{6}$$

In order to mine fine-grained local features, attention coefficient a_{ij} is multiplied by local image feature h'_{ij}. At this moment of attention as feature selector, in describing the point x_i, on which the concentration coefficient can identify ability of neighborhood characteristics of adaptive capacity, to strengthen the neighborhood features such as noise, meaningless effectively suppressed, thus mining the fine-grained detail information fully and effectively.

By inputting the feature sequence $S_k = s_k^1, \cdots, s_k^t, \cdots, s_k^T$ of the sampling points into BiSeNet, the correlation between the sampling points in different scale neighborhoods is obtained. In order to fully mine context information, the hidden layer d is used to encode neighborhood feature vectors of different scales in sampling points successively. In addition, BiSeNet is used to encode neighborhood feature vectors of different scales of sampling point x_i, and the state of the hidden layer will be updated successively. Specific updates are as follows:

$$d_t = \zeta(d_{t-1}, s_{k-1}^t), t \in [1, T]. \tag{7}$$

In the equation, ζ is a nonlinear activation function; s_{k-1} is the $t-1$ neighborhood feature vector, and d_{t-1} is the hidden layer state of s_{k-1}. Then the $t-th$ neighborhood feature vector in the sampling point is s_k. When BiSeNet is used to encode s_k, the corresponding output o_t is:

$$o_t = \omega_a d_t. \tag{8}$$

Where, the weight matrix that can be learned is ω_a. All feature sequences will get the hidden layer state after learning, and the hidden layer state is denoted as d_T. The multi-scale context geometric feature o_T of the sampling point can be obtained by multiplying ω_a and d_T.

The introduction of attention encoding is certainly helpful to improve the network's ability to capture fine-grained details in local areas to a certain extent. But it does not pay attention to the contextual geometry information between local areas. This is extremely important for image semantic segmentation [33]. The advantage of context BiSeNet encoding is that it can fully mine the high-level features of multi-scale context, which makes it possible to compensate each other for the fine-grained local features of a relatively low level and the multi-scale context geometric features of a relatively high level. By selecting Selu function, all fine-grained local features at different levels in the sampling point are fused into context geometric features. After the fusion of the two features, the sample point size of context fine-grained geometric features can be obtained as $N \times F2$. Before feature fusion, the $N \times 128$ image is sampled by interpolation operation on $R \times 128$ image. $F_{\sum_i}$ after fusion is calculated as follows:

$$F_{\sum_i} = Selu(o_T + l_i). \tag{9}$$

where l_i is the local fine granularity feature.

3.3. Focal Loss

The data set is considered unbalanced, if the samples of a certain class of targets are greatly superior in number to those of other classes. This imbalance will lead to two problems. 1) Low training efficiency. Since most samples are simple targets, these samples

provide the model with less useful information during training. 2) The advantages of simple sample size will affect the training of the model and degrade the model performance. Guo et al. [34] proposed focal Loss function to solve the problem of category imbalance by reducing the internal weighting.

There are many kinds of target objects, the size and shape of objects of the same type are also different. It contains very few individual objects that stand out. CE loss function can not balance the learning of a small number of samples well, so focal Loss is introduced as a loss function to solve the sample imbalance problem in the segmentation task. Focal Loss is an improvement on the cross-drop function. By modifying the cross drop function and adding the sample difficulty weight adjustment factor $(1-p_t)^\gamma$, the imbalance of sample categories and sample classification difficulty is alleviated and the model accuracy is improved. The mathematical expression is:

$$L_{FL}(p_t) = -(1-p_t)^\gamma log p_t. \tag{10}$$

We add a category weight α, and equation (7) is rewritten as:

$$L_{FL}(p_t) = -\alpha_t(1-p_t)^\gamma log p_t. \tag{11}$$

Where α is the weight parameter between categories. $(1-p_t)^\gamma$ is the simple/difficult sample regulator. γ is the focusing parameter. When the prediction of a class is accurate, that is, when p_t is close to 1, the value of $(1-p_t)^\gamma$ is close to 0. When the prediction of a category is inaccurate, that is, when p_t approaches 0, the value of $(1-p_t)^\gamma$ approaches 1. Set $\gamma = 2$, $\alpha = 0.25$.

4. Experiments and Analysis

Experiments are carried out on PASCAL VOC 2012 benchmark data set [35]. The dataset is published by the International Computer Vision Challenge for image classification, detection or semantic segmentation. It contains 20 foreground object classes and one background class, including people, animals, traffic vehicles and indoor household items. There are 1464 images for training set, 1449 images for validation set and 1456 images for testing set. The experiment is implemented on TensorFlow, a deep learning framework. The operating system used in the experiment is Windows 11, and the graphics card is NVIDIA RTX3060. A dense feature graphs are extracted using pre-trained Xception by ImageNet. Adam optimizer and Poly learning strategy are adopted. In the experiment, the image is cut to 256×256 for training. In the initial training process, a small learning rate is used to achieve smooth start. Set the initial learning rate as 1×10^{-4}, momentum as 0.9, and select iteration training as 50000 times.

4.1. Evaluation index

Mean intersection over Union (MIOU), pixel accuracy (PA) and mean pixel accuracy (MPA) are used to evaluate the segmentation effect of the proposed method on the data set. The higher values of MIOU PA, and MAP denote the better image semantic segmentation effect.

Given that there are $k+1$ segmentation classes in the image (including k target classes and 1 background class). p_{ij} (False Positives) represents the number of pixels that belong to class i but are predicted to be class j. p_{ji} (False Negatives) represents that the number of pixels that belong to class j but are predicted to be class i. p_{ii} (True Positives) indicates the true number of pixels.

Pixel accuracy (PA) is defined as follows:

$$PA = \frac{\sum_i^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k} p_{ij}. \tag{12}$$

Mean pixel accuracy (MPA) is defined as:

$$MPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^k p_{ij}}. \tag{13}$$

Mean intersection over Union (MIOU) is defined as:

$$MIOU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^k p_{ij} + \sum_{j=0}^k (p_{ji} - p_{ii})}. \tag{14}$$

Where formula (12) represents the proportion of correctly classified pixel points and all pixel points in the image. Formula (13) represents the proportion of correctly classified pixel points of each category and all pixel points of that category, and then calculates the average value. Formula (14) represents the intersection of the predicted region and the actual region in the image divided by the union of the predicted region and the actual region.

4.2. Results and Analysis

During the experiment process, a context attention mechanism is introduced to improve the accuracy of feature images, filter background information and reduce the loss of details. In the context attention module, two convolution levels $1 \times k$ and $k \times 1$ are applied to the high-level features to obtain spatial concerns. The segmentation results of different k values are shown in Table 1. As can be seen from Table 1, different convolution operations on feature graphs result in significantly different segmentation results. When $k = 8$, MIOU and MPA are the highest. Therefore, the training model with $k = 8$ is finally selected for verification.

Then, we conduct experiments on the PASCAL VOC 2012 dataset. Compared with different image semantic segmentation algorithms including PSPNet [36], DANet [37], DeepLabV3+ [38], SANet [39], DUpsamling and reference [40-42]). The experimental results are shown in Table 2. As can be seen from Table 2, the MIOU value of the proposed algorithm is 5.88% higher than PSPNet algorithm, 3.48% higher than DANet algorithm, 3.30% higher than DeepLabV3+ algorithm, and 1.24% higher than SANet algorithm. Compared with the , it is improved by 2.41%, 2.18%, 1.50% and 1.04% compared with DUpsamling algorithm, reference [40], reference [41], and reference [42] respectively. Also, in terms of the AP and MAP, the proposed method obtains the better results.

Table 1. MIOU and MPA values with different k

k	MIOU%	MPA%
1	79.18	82.54
2	80.22	82.67
3	81.74	83.54
4	82.06	83.96
5	82.78	84.63
6	83.47	85.21
7	84.73	85.88
8	88.91	89.25
9	86.46	87.25
10	85.41	86.93

Table 2. Comparison results of different image semantic segmentation algorithms

Method	AP%	MAP%	MIOU%
PSPNet	89.21	90.33	79.67
DANet	90.42	91.25	82.07
DeepLabV3+	90.88	91.35	82.25
DUpsamling	91.26	91.87	83.14
Reference [40]	91.78	92.06	83.37
Reference [41]	92.34	92.58	84.05
SANet	93.54	93.88	84.31
Reference [42]	93.89	94.12	84.51
Proposed	94.56	95.71	85.55

Table 3. Comparison between BiSeNet and proposed method

Method	MPA%	MIOU%
BiSeNet	90.23	82.25
Proposed	93.61	85.55

In order to better verify the performance of the proposed algorithm, we make comparison between BiSeNet and proposed method, as shown in Table 3. As can be seen from Table 3, the MIOU value of the proposed algorithm is improved, also the MPA is improves by 3.38% compared with BiSeNet.

Based on BiSeNet algorithm, ablation experiments are performed to verify the better results of the proposed method. The image semantic segmentation results of different combination methods are shown in Table 4.

Table 4. Ablation experiments

Number	BiSeNet	Xception	Context attention	Focal loss	MIOU%	MAP%
a	Yes	No	No	No	82.25	89.31
b	Yes	Yes	No	No	82.67	90.24
c	Yes	No	Yes	No	82.29	89.57
d	Yes	Yes	Yes	No	85.03	92.45
e	Yes	Yes	Yes	Yes	85.55	95.71

From the comparison of results a and b in Table 4, it can be seen that the MIOU value increases by 0.42% by fine-tuning the Xception model and adding a low-level feature extraction path. The comparison between the results of a and c shows that the MIOU value increases by 2.16% when the attention mechanism is introduced on the basis of the original network, indicating that the accuracy of feature extraction is effectively improved by the attention mechanism. The comparison of c and d results shows that multi-path extraction of low-level features can increase MIOU value by another 0.62%. According to the comparison of results of d and e, the use of Focal Loss improves MIOU value by 0.52%.

The training loss curves of BiSeNet algorithm and the proposed algorithm are shown in figure 4. The x-axis is training time and the y-axis is loss value. As can be seen from figure 4, the loss value is relatively high at the beginning of the model training. With the increase of training times, the loss curve gradually stabilizes. Compared with the original algorithm, the loss value of the proposed algorithm decreases greatly, indicating that the proposed algorithm can effectively reduce the loss of feature information and improve the final segmentation effect.

Different algorithms are used to segment the test set in PASCAL VOC 2012 dataset. The experimental comparison results are shown in figure 5. As can be seen from the figure, compared with reference[40] algorithm and reference [41] algorithm, the proposed algorithm has clearer target segmentation boundary. Reference [42] algorithm obviously has the problem of unbalanced bicycle segmentation. Compared with the Reference [42] algorithm, the segmentation results of the proposed algorithm are significantly more balanced. Experimental results show that the proposed algorithm has a significant improvement in the boundary segmentation effect of the target background, refines the target boundary, improves the segmentation effect of the target object, and has a better object resolution ability.

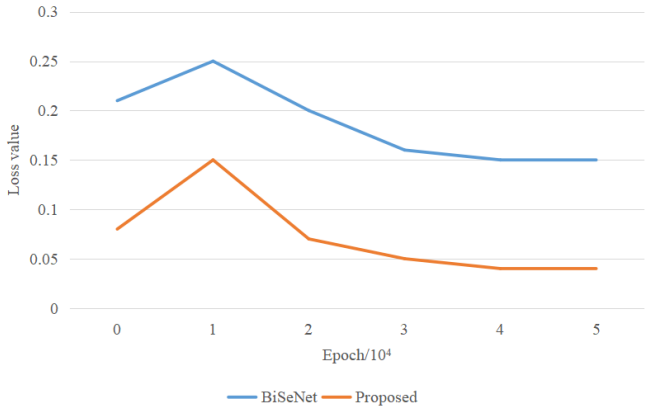

Fig. 4. Loss values with different methods

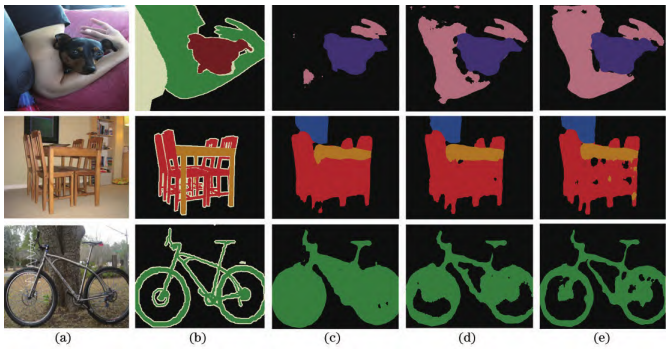

Fig. 5. Comparison of segmentation results. (a) Original images; (b) Ground Truth; (c) reference [41]; (d) reference [42]; (e) Proposed algorithm

5. Conclusion

In this paper, an image semantic segmentation algorithm based on BiSeNet and context attention mechanism is proposed. A low-level feature extraction path is added in BiSeNet to increase feature information and expand receptive fields. Without affecting the network speed, it improves the segmentation accuracy. Context content attention mechanism is introduced to extract high-level features and low-level features, and the two features are fused to obtain rich context information effectively, filter background information, and obtain more detailed feature maps. In order to solve the problem that the sample number of different categories is not balanced and the difficulty of distinguishing different objects is different, the focus loss function is used instead of the cross drop loss function to reduce the loss of feature details. Experiments are carried out on PASCALVOC 2012 data set, and the experimental results show that the proposed algorithm has a great improvement in image segmentation accuracy compared with other algorithms. In the future, we will apply image semantic segmentation to a wider range of fields, such as aerospace, remote sensing, medicine. At the same time, we will also develop more advanced methods to further improve accuracy.

References

1. Zhang G, Zhao K, Hong Y, et al. "SHA-MTL: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification," *International Journal of Computer Assisted Radiology and Surgery,* vol. 16, pp. 1719-1725, (2021).
2. H. Zhang et al. "Multiscale Visual-Attribute Co-Attention for Zero-Shot Image Recognition," *IEEE Transactions on Neural Networks and Learning Systems,* (2021). doi: 10.1109/TNNLS.2021.3132366.
3. X. Lei and H. Ouyang. "Kernel-Based Intuitionistic Fuzzy Clustering Image Segmentation Based on Grey Wolf Optimizer With Differential Mutation," *IEEE Access,* vol. 9, pp. 85455-85463, (2021).
4. Fan Wang, Chen Chen, Haitao Zhang and Youhua Ma. "Short-term Load Forecasting Based On Variational Mode Decomposition And Chaotic Grey Wolf Optimization Improved Random Forest Algorithm," *Journal of Applied Science and Engineering,* Vol. 26, No. 1, pp. 69-78, (2020).
5. Fung D, Liu Q, Zammit J, et al. "Self-supervised deep learning model for COVID-19 lung CT image segmentation highlighting putative causal relationship among age, underlying disease and COVID-19," *Journal of Translational Medicine,* vol. 19, no. 1, (2021).
6. Xian S, Cheng Y, Chen K. "A novel weighted spatial T-spherical fuzzy C-means algorithms with bias correction for image segmentation," *International Journal of Intelligent Systems,* vol. 37, no. 2, (2022)
7. Zhang L, X Hu, Zhou Y, et al. "Memristive DeepLab: A hardware friendly deep CNN for semantic segmentation," *Neurocomputing,* vol. 451, pp. 181-191 (2021).
8. H. -Y. Han, Y. -C. Chen, P. -Y. Hsiao and L. -C. Fu. "Using Channel-Wise Attention for Deep CNN Based Real-Time Semantic Segmentation With Class-Aware Edge Information," *IEEE Transactions on Intelligent Transportation Systems,* vol. 22, no. 2, pp. 1041-1051, (2021).
9. Jisi A and Shoulin Yin. "A New Feature Fusion Network for Student Behavior Recognition in Education," *Journal of Applied Science and Engineering,* vol. 24, no. 2, pp. 133-140. (2021)
10. Russo G. "On Unsupervised Methods for Medical Image Segmentation: Investigating Classic Approaches in Breast Cancer DCE-MRI," *Applied Sciences,* vol. 12, no. 1. (2022)

11. Gurita A, Mocanu I G. "Image Segmentation Using Encoder-Decoder with Deformable Convolutions," *Sensors,*, vol. 21, no. 5, pp. 1570. (2021)
12. C. Lyu, G. Hu and D. Wang. "HRED-Net: High-Resolution Encoder-Decoder Network for Fine-Grained Image Segmentation," *IEEE Access,* vol. 8, pp. 38210-38220, (2020)
13. Z. Tian, T. He, C. Shen and Y. Yan. "Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3121-3130, (2019). doi: 10.1109/CVPR.2019.00324.
14. Cai W, Zhai B, Liu Y, et al. "Quadratic Polynomial Guided Fuzzy C-means and Dual Attention Mechanism for Medical Image Segmentation," *Displays,* vol. 70, no. 6, pp. 102106. (2021)
15. A. Bera, Z. Wharton, Y. Liu, N. Bessis and A. Behera. "Attend and Guide (AG-Net): A Keypoints-Driven Attention-Based Deep Network for Image Recognition," *IEEE Transactions on Image Processing,* vol. 30, pp. 3691-3704, (2021).
16. Yang T, Yoshimura Y, Morita A, et al. "Pyramid Predictive Attention Network for Medical Image Segmentation," *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences,* vol. E102, no. A(9), pp. 1225-1234. (2019)
17. Al-Huda Z, Zhai D, Yang Y, et al. "Optimal Scale of Hierarchical Image Segmentation with Scribbles Guidance for Weakly Supervised Semantic Segmentation," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 35, no. 10, (2021)
18. Lsel P D, Kamp T, Jayme A, et al. "Introducing Biomedisa as an open-source online platform for biomedical image segmentation," *Nature Communications,* 11(5577). (2020)
19. Guo H, Yang D. "PRDNet: Medical image segmentation based on parallel residual and dilated network," *Measurement,*vol. 173, no. 1, pp. 108661. (2020)
20. Huang M, Huang S, Zhang Y, et al. "Medical Image Segmentation Using Deep learning with Feature Enhancement," *IET Image Processing,* vol. 14, no. 5. (2020)
21. Olimov B, Sanjar K, Din S, et al. "FU-Net: fast biomedical image segmentation model based on bottleneck convolution layers," *Multimedia Systems,* vol. 27, no. 4, pp. 637-650, 2021.
22. Zheng T, Duan Z, Wang J, et al. "Research on Distance Transform and Neural Network Lidar Information Sampling Classification-Based Semantic Segmentation of 2D Indoor Room Maps," *Sensors,* vol. 21, no. 4, pp. 1365. (2021)
23. Shoulin Yin, Hang Li, Desheng Liu and Shahid Karim. "Active Contour Modal Based on Density-oriented BIRCH Clustering Method for Medical Image Segmentation," *Multimedia Tools and Applications,* Vol. 79, pp. 31049-31068, (2020).
24. Wech T, Ankenbrand M J, Bley T A, et al. "A data-driven semantic segmentation model for direct cardiac functional analysis based on undersampled radial MR cine series," *Magnetic Resonance in Medicine,* vol. 87. (2022)
25. Jiang, D., Li, H., Yin, S. "Speech Emotion Recognition Method Based on Improved Long Short-term Memory Networks," *International Journal of Electronics and Information Engineering,* Vol. 12, No. 4, pp. 147-154. (2020)
26. Jiang M, Zhai F, Kong J. "Sparse Attention Module for optimizing semantic segmentation performance combined with a multi-task feature extraction network," *The Visual Computer,* vol. 12. (2021)
27. R. Yi, Y. Huang, Q. Guan, M. Pu and R. Zhang. "Learning From Pixel-Level Label Noise: A New Perspective for Semi-Supervised Semantic Segmentation," *IEEE Transactions on Image Processing,* vol. 31, pp. 623-635, (2022).
28. Grant-Jacob J A, Praeger M, Eason R W, et al. "Semantic segmentation of pollen grain images generated from scattering patterns via deep learning," *Journal of Physics Communications,* vol. 5, no. 5, 055017 (11pp). (2021)
29. Yu C, Wang J, Peng C, et al. "BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation," *ECCV 2018. Lecture Notes in Computer Science,* vol. 11217, pp. 334-349. Springer, Cham. (2018).

30. Polat Z. "Detection of Covid-19 from Chest CT Images using Xception Architecture: A Deep Transfer Learning based Approach," *Sakarya University Journal of Science,* vol. 25, no. 3, pp. 813-823, (2021)
31. Xiaowei Wang, Shoulin Yin, Ke Sun, et al. "GKFC-CNN: Modified Gaussian Kernel Fuzzy C-means and Convolutional Neural Network for Apple Segmentation and Recognition," *Journal of Applied Science and Engineering,* vol. 23, no. 3, pp. 555-561, (2020).
32. George B, Assaiya A, Roy R J, et al. "CASSPER is a semantic segmentation-based particle picking algorithm for single-particle cryo-electron microscopy," *Communications Biology,* vol. 4, no. 1. (2021)
33. Dai, Y., Xu, B., Yan, S., Xu, J.: Study of cardiac arrhythmia classification based on convolutional neural network. Computer Science and Information Systems, Vol. 17, No. 2, 445-458. (2020), https://doi.org/10.2298/CSIS191229011D
34. Ge, Y., Zhu, F., Huang, W., Zhao, P., Liu, Q.: Multi-Agent Cooperation Q-Learning Algorithm Based on Constrained Markov Game. Computer Science and Information Systems, Vol. 17, No. 2, pp. 647-664. (2020), https://doi.org/10.2298/CSIS191220009G
35. Wong C C, Yeh L Y, Liu C C, et al. "Manipulation Planning for Object Re-Orientation Based on Semantic Segmentation Keypoint Detection," *Sensors,* vol. 21, no. 7, 2280. (2021)
36. Guo X, Xiao R, Lu Y, et al. "Cerebrovascular Segmentation from TOF-MRA based on Multiple-U-net with Focal Loss Function," *Computer Methods and Programs in Biomedicine,* vol. 202, no. 3, pp. 105998. (2021)
37. Liu R, He D. "Semantic Segmentation Based on Deeplabv3+ and Attention Mechanism," *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). IEEE,* (2021).
38. H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia. "Pyramid Scene Parsing Network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2017, pp. 6230-6239, doi: 10.1109/CVPR.2017.660.
39. J. Fu et al. "Dual Attention Network for Scene Segmentation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2019, pp. 3141-3149, doi: 10.1109/CVPR.2019.00326.
40. Chen LC., Zhu Y., Papandreou G., Schroff F., Adam H. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," *ECCV 2018. Lecture Notes in Computer Science, vol 11211. Springer, Cham.* (2018)
41. Z. Zhong et al. "Squeeze-and-Attention Networks for Semantic Segmentation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 13062-13071, (2020). doi:10.1109/CVPR42600.2020.01308.
42. Li X, Chen J, Ye Y, et al. "Fast Semantic Segmentation Model PULNet and Lawn Boundary Detection Method," *Journal of Physics: Conference Series,* vol. 1828, no. 1, pp. 012036 (16pp). (2021)
43. Trajanovski S, Shan C, Weijtmans P, et al." Tongue Tumor Detection in Hyperspectral Images Using Deep Learning Semantic Segmentation," *IEEE Transactions on Biomedical Engineering,* vol. 68, no. 4, pp. 1330-1340. (2021)
44. Wang K, Xiang K, Yang K. "Polarization-driven Semantic Segmentation via Efficient Attention-bridged Fusion," *Optics Express,* vol. 29, no. 4. (2021)

Lin Teng is a doctoral student at the School of Information and Communication Engineering, Harbin Engineering University. Her research interests include image processing, image segmentation.

Yulong Qiao is a professor at the School of Information and Communication Engineering, Harbin Engineering University. His main research areas: statistical image processing, image/video processing and applications.

Received: March 21, 2022; Accepted: September 10, 2022.