Signal, Image and Video Processing (2021) 15:1031–1039
https://doi.org/10.1007/s11760-020-01828-8

ORIGINAL PAPER

Fully convolutional network with attention modules for semantic segmentation

Yunjia Huang[1] · Haixia Xu[1]

Received: 20 July 2020 / Revised: 23 October 2020 / Accepted: 23 November 2020 / Published online: 2 January 2021

Abstract
Fully convolutional network is a powerful end-to-end model for semantic segmentation. However, it performs prediction pixel by pixel to pose weak consistency on intra-category. This paper proposes fully convolutional network with attention modules for semantic segmentation. Based on the framework of fully convolutional network, the post-processing attention module and skip-layer attention module are introduced to enhance the relevancy among pixels. Post-processing attention module is to calculate the similarity among pixels to obtain global information. Skip-layer attention module is designed to combine semantic information from a deep, coarse layer with contour information from a shallow, fine layer to produce the feature with high resolution and strong semantic information. Loss function, obtained by cross-entropy between estimated probability and label, is to optimize the network. Extensive experiments demonstrate that the proposed approach is superior to DeepLab and other models in performance of mean IoU with moderate computational complexity

Keywords Semantic segmentation · Fully convolutional network · Attention module

1 Introduction

Semantic segmentation is to assign consistent labels to pixels with similarly semantic attribute. It has essential applications in the field of unmanned driving, medical image recognition and intelligent safeguard systems, etc.

In the early stage, conventional methods [1,2], such as threshold optimization and watershed algorithm, were used to segment the image into regions, and the regions were classified and annotated with geometric shapes and textures. Then, probability models [3] and machine learning methods [4] were developed in semantic segmentation in vehicle license plate recognition and medical image segmentation.

Recently, deep Learning [5] has been leading the visual task. Fully convolutional network (FCN) [6] adopted classification networks into the fully convolutional network and transferred the learned representative to pixel-wise classification, followed by many robust networks such as U-Net [7], DeepLabv3 [8] and so on. Thanks to the effects of transpose convolution and skip-layer, the prediction result of FCN not only has the same size as the input but also ensures the robustness and accuracy of FCN.

FCN serves as the foundation and baseline of modern semantic segmentation methods. However, it performs prediction pixel by pixel to pose weak consistency on intra-category. It is helpful to enhance the connection among pixels to improve the performance of FCN. Therefore, it is necessary to increase the discriminative ability of feature representations for pixel-wise recognition. The practical and direct approach is to use the information of adjacent pixels or spatial correlation in the convolution process.

In OCNet [9] and PSPNet [10], the pyramid pooling module partitions the feature maps into multiple regions, and the pixels in each region are regarded as the context of the pixel belonging to the region. The Atrous Spatial Pyramid Pooling module in DeepLabv3 considers spatially sampled pixels at different atrous rates as the context of the centre pixel. In DANet [11], self-attention mechanism [12] is used to capture feature dependence in spatial dimension and channel dimension, respectively.

Inspired by PSPNet, this paper proposes FCN with post-processing attention module (PPAM) and skip-layer attention module (SAM) for semantic segmentation to deal with the

✉ Haixia Xu
xhxia2002@126.com

Yunjia Huang
yuichi0507@163.com

[1] School of Automation and Electronic Information, Xiangtan University, Xiangtan, China

problem of poor consistency on intra-category and similarities on intra-category among pixels. PPAM is to capture the spatial dependence of any two pixels in the feature map. SAM is to generate the feature with high resolution and strong semantic information through fusing high-level feature with the low-level feature. The two modules work together to enhance the connection among pixels.

FCN with PPAM and SAM shows an excellent performance of 72.01% mIoU with moderate computational complexity and hardware on PASCAL VOC 2012 .

The main contributions are summarized below.

(1) We propose the FCN with attention modules, which is used to improve the consistency of semantic prediction.
(2) We design the PPAM and SAM to improve the performance of segmentation. PPAM is used to compute similarity among pixels, and extract precise pixel-wise contextual information from high-level feature map in FCN. SAM is to fuse semantic information from a high level with contour information from a low level.
(3) We formulate the loss function by the sum of Branch-1 loss function and Branch-2 loss function. Compared with the conventional method with only one branch cross-entropy loss function, Branch-2 loss we designed provides fault tolerance for network, add extra gradient flow during back-propagation, thereby helping to reduce gradient vanishing problem.

2 Related work

There are a variety of networks proposed to carry out semantic segmentation tasks. Firstly, the network represented by DeepLab system [13] introduces atrous convolution and spatial pyramid pooling with different void rates in different branches to obtain multi-scale image representation. Secondly, changing the encoder–decoder structure to produce better quality results is also a hot topic. Besides, more and more networks are beginning to add attention mechanisms to capture context information. These methods make use of the context information as much as possible by changing the structure of the network.

Spatial Pyramid Pooling: The pyramid pooling can extract and aggregate feature maps from different sizes to improve the robustness of neural networks. Spatial pyramid pooling has been widely employed to provide a good descriptor for overall scene interpretation, especially for various objects in multiple scales. Besides, spatial pyramid pooling can well solve the problem of large computing time of R-CNN [14].

Encoder–decoder: The purpose of this method is to solve the problem of image size reduction and contour information loss caused by continuous convolution and pooling. Most of the best semantic segmentation frameworks are based on the encoder–decoder network [15], which have also been successfully applied to many computer vision tasks, including object detection [16,17], panoptic segmentation [18]. However, most methods ignore context information when combining high-level features with low-level features.

Self-attention module: The attention mechanism from natural language processing can effectively capture the useful areas in the image, and the overall network performance can be improved. The work [19] is the first to propose the self-attention mechanism and apply it to action recognition task in the video. The self-attention operation can effectively capture the long-range dependency between different positions. Therefore, each position can attain the global field of vision without causing degradation of the feature map. The work [11] extends the self-attention mechanism in the task of scene segmentation and carefully designs two types of attention modules to capture rich context information for better feature representations with intra-class compactness.

The remainder of this paper is organized as follows. Section 3 discusses the proposed PPAM, SAM and loss function for segmentation. In Sect. 4, experiments and discussion are given. The final section presents concluding remarks as well as future work.

3 The proposed method

In this section, we propose FCN with PPAM and SAM and first introduce our model and detail the proposed modules and loss function.

3.1 Overall architecture

The proposed model is composed of base dilated FCN and attention modules of PPAM, SAM. The overall model architecture is shown in Fig. 1, and the algorithm process is that:

Based on baseline dilated FCN, first perform feature extraction of four levels on input image. Then capture the context from the last level Res-4 with PPAM (detailed in Fig. 3), fuse features with the guidance of SAM (shown in Fig. 4). Finally output the prediction of segments.

Base dilated FCN [20]: We take ResNet-50 [21] as backbone network for feature extraction in the FCN architecture, which gets 4 level features from Res-1 to Res-4 with the atrous rates {1,1,2,2} and the strides {1,2,1,1}. In addition, the Res-4 output size of the feature map from ResNet-50 is 1/8 of the input image.

PPAM: PPAM is integrated into FCN and placed after the backbone of feature extraction. It is designed to collect the context information of the last layer of the feature map, to secure an accurate prediction of each pixel.

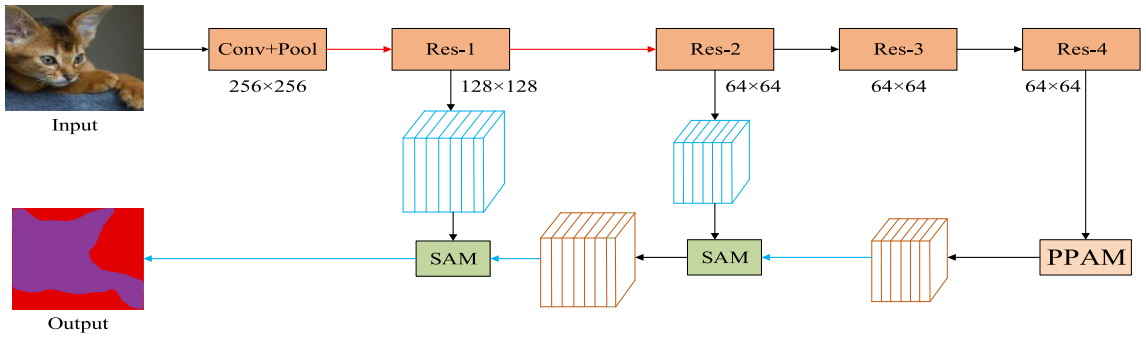

Fig. 1 Overall architecture of the proposed model. 'Conv+Pool' represents convolution layer and pooling layer, 'Res-1,2,3,4' represents residual modules in dilated ResNet-50, 'PPAM' represents post-processing attention module, which is shown in Fig. 3, and 'SAM' represents skip-layer attention module, which is shown in Fig. 4. The red and blue lines represent the down-sample and up-sample operators, respectively

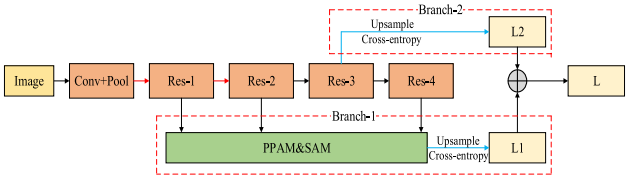

Fig. 2 The calculating of the proposed loss function. The red and blue lines represent the down-sample and up-sample operators, respectively

SAM: The introduction of SAM into the skip-layer ensures that the high-level feature map containing semantic information is well integrated with low-level feature map showing contour information.

3.2 PPAM

It is the key for scene understanding to collect the discriminative feature representation. Inspired by OCNet [9], we design the PPAM, as shown in Fig. 3, which uses up-sample position attention module (UPAM) in multiple parallel branches to obtain pixel-dependent feature map. The UPAM encodes a wide range of semantic information into the receptive field to enhance the semantic consistency among pixels, which use position attention module [11].

As illustrated in Fig. 3, given a feature map X which is $64 \times 64 \times 2048$, we first feed it into a convolution layer and take four down-sample operations by bilinear interpolation to generate four new feature maps a, b, c and d, where $\{a, b, c \text{ and } d\}$ have the same channels. Then, we use UPAM to extract context information from $\{a, b, c, d\}$, and get $\{A, B, C, D\}$. After that, we concatenate X, A, B, C and D on channels to obtain the final output Y. As is widely regarded, each channel of high-level feature can be viewed as a category-specific response and correlated with different semantic responses.

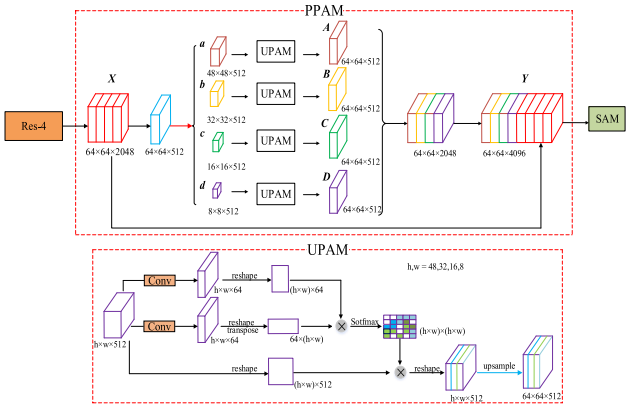

Fig. 3 The details of post-processing attention module. '64×64, 48×48, 32×32' represents the resolution of feature maps. The red and blue lines represent the down-sample and up-sample operations with bilinear interpolation, respectively

The concatenation operation on four branches is used to get more context information through collecting information of different receptive fields, which aims to enhance the dependency between pixels and the differentiation between classes by integrating the refine features (local information) and global features (context information).

3.3 SAM

The architecture of SAM is shown in Fig. 4. It is to fuse semantic information from high-level features with contour information from low-level features.

Low-level feature information and high-level feature information are concatenated on channels to restore contour information gradually. Inspired by U-Net, good results can be obtained by fusing decoder with encoder. The encoding process, namely the down-sample process, is to extract abstract features from the input image through the pooling layer or

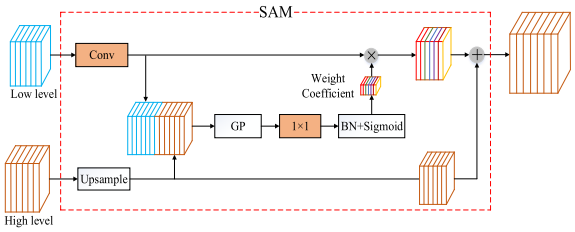

Fig. 4 The details of skip-layer attention module. 'Conv' represents convolution layers. 'GP' represents global pooling. '1 × 1' represents 1 × 1 convolution. 'BN+Sigmod' represents batch normalization and nonlinear activation function sigmoid. '×' represents the Hadamard product of low-level feature and weight coefficient. '+' represents the sum of the corresponding pixels of low-level feature and high-level feature

the atrous convolution layer. The decoding is the up-sample process to recover the position information gradually.

As illustrated in Fig. 4, the 'Conv' is to reduce the number of channels of low-level feature map and to increase receptive fields through three sets of convolution layers $\{1\times1, 3\times3, 3\times3\}$. The 'Upsample' is to resize the high-level feature map to the size of the low-level feature map. Global pooling (GP) provides maximum receptive fields with global context information. After that, 1 × 1 convolution is used to conduct feature map channel compression, so that the concatenation feature map can be used as a weight coefficient. The operation '×' means weight coefficients are applied to low-level feature maps along the channel. Finally, the high-level feature and low-level feature are fused by the sum of corresponding pixels. The fusion result serves as the high-level information of the next skip-layer, which use high-level information to guide for low-level information.

3.4 Loss function

We define the loss function as shown in Fig. 2, which as a sum of cross-entropy of Branch-1 and that of Branch-2, and formulate it by

$$L = L_1 + \lambda \times L_2 \tag{1}$$

$$L_1 = -\sum_{j=0}^{C-1}\sum_{i=0}^{n-1} p[i,1]\log(Y[i,j]) \tag{2}$$

$$L_2 = -\sum_{j=0}^{C-1}\sum_{i=0}^{n-1} p[i,1]\log(y[i,j]) \tag{3}$$

where n is equal to $B \times \mathrm{h} \times \mathrm{w}$, B is batch size, C is the number of categories, h and w are the height and width of the image, p is ground truth label with the manner of one-hot. Y in Eq. (2) is the estimated probability of overall segmentation (Branch-1), and y in Eq. (3) is the estimated probability of Res-3 (Branch-2), respectively.

Table 1 The comparison of computational complexity

Method	BaseNet	FLOPs	Params
Dilated FCN	ResNet-50	4.110G	25.557M
DeepLabv3	ResNet-50	10.004G	39.115M
DeepLabv3+	ResNet-50	34.947G	40.295M
PSPNet	ResNet-50	75.534G	65.585M
OCNet	ResNet-50	39.075G	43.594M
Our	ResNet-50	38.772G	42.454M

In order to calculate the loss, the prediction image and ground truth label are arranged in the form of matrix. We traverse all pixels in two-loop structure to calculate the loss of each pixel and finally accumulate the loss of the whole. The Branch-2 Loss function (B2L) is paralleled with Branch-1 Loss function (B1L) to reduce the vanishing gradient problem for earlier layers stabilize the training. The weight factor λ balances the terms L_1 and L_2 .

3.5 Computational complexity

In this section, we compare computational complexity with dilated FCN, DeepLabv3/v3+, PSPNet and OCNet. As shown in Table 1, all results are achieved with backbone ResNet-50 and output stride 8. The floating-point operations (FLOPs) and parameters (Params) are computed with the input size 224 × 224. From Table 1, our network is lighter than PSPNet and OCNet in computation.

4 Experiments

In this section, the proposed model is evaluated on dataset Cityscapes [22] and PASCAL VOC 2012 [23] and compared with the state-of-the-art semantic segmentation networks. The state-of-the-art networks are DeepLabv3 [8], which designs serial and parallel convolution modules with atrous and uses various atrous rates to get hold of multi-scale information; DeepLabv3+, which proposes the Atrous Spatial Pyramid Pooling module to exploit the convolution features of different scales and encodes the features of global information to improve the segmentation effect; OCNet [9], which proposes a pixel-by-pixel object context module that contains the information of objects of the same category as the pixel; PSPNet [10], which offers global-scene-level feature maps to obtain sufficient context information and global information of different sensory fields.

Table 2 Employing B2L and PPAM for dilated FCN on PASCAL VOC 2012 *val* set. 'B2L' represents the loss function of Branch-2

BaseNet	B2L	PPAM	mIoU (%)
ResNet50			63.56
ResNet50	✓	✓	69.11
ResNet101	✓	✓	70.49

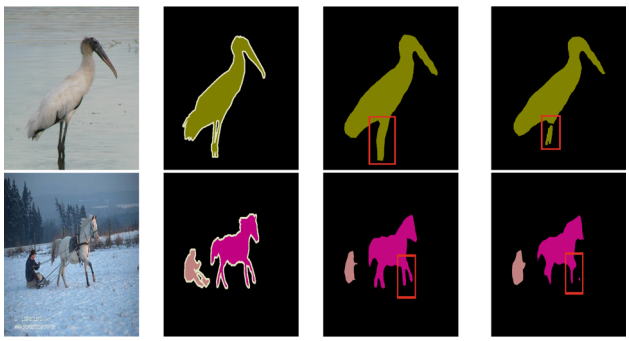

Fig. 5 Visualization results of PPAM on PASCAL VOC 2012 *val* set

Table 3 The segmentation performance with PPAM and SAM on PASCAL VOC 2012 *val* set

BaseNet	B2L	PPAM	Res-1	Res-2	mIoU (%)
ResNet50	✓	✓			69.11
ResNet50	✓	✓	✓		70.40
ResNet50	✓	✓		✓	69.13
ResNet50	✓	✓	✓	✓	70.89

'Res-1,2' represents the first and the second residual modules in backbone network

Table 4 Comparison between different strategies on PASCAL VOC 2012 *val* set

MG	DA	MS	mIoU (%)
			70.89
✓			71.06
✓	✓		71.85
✓	✓	✓	72.01

4.1 Parameters setup

All the experiments are implemented with Pytorch 1.1.0 and performed on the PC with 2 NVIDIA GPUs GTX-1080Ti and 22GB memory, running Ubuntu 18.04 system.

According to Mask R-CNN [24], a learning strategy named Poly in training adopts the stochastic gradient descent (SGD) [25] with batch size 8, momentum 0.9 and weight decay 0.0001. The learning rate is reduced by the initial learning rate of LR be multiplied by $\left(1-\frac{\text{cur\_iter}}{\text{max\_iter}}\right)^{0.9}$.

Table 5 Per-class results on Cityscapes testing set

Method	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic light	Traffic sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	mIoU (%)
Dilated FCN	91.6	73.6	79.5	26.2	36.5	41.7	52.4	57.3	81.7	61.6	81.9	69.4	43.7	84.9	28.6	40.9	38.8	44.9	59.1	57.56
DeepLabv3+	96.3	75.2	84.2	42.7	41.3	43.5	48.7	61.2	82.7	63.3	88.1	73.7	53.7	87.6	56.5	58.3	51.4	51.6	62.7	64.26
PSPNet	97.2	76.2	82.9	46.8	45.8	40.9	50.6	64.7	77.4	59.3	86.1	76.8	57.3	85.9	50.7	61.5	60.6	54.7	63.4	65.15
OCNet	97.5	78.3	88.3	45.4	45.4	46.7	56.0	65.4	89.0	68.5	92.3	77.9	57.9	90.2	55.6	65.6	55.6	55.8	66.9	68.33
Our	96.1	76.4	88.4	40.8	49.8	49.7	65.3	69.4	88.7	66.7	91.1	77.8	57.1	88.5	40.5	50.7	44.4	53.3	67.6	66.53

The above experimental results are all performed on the same experimental equipment with the same parameters

Table 6 Per-class results on the PASCAL VOC 2012 *val* set

Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Person	Plant	Sheep	Sofa	TV	Train	mIoU (%)
Dilated FCN	75.9	30.2	77.1	57.3	65.3	80.8	73.6	77.9	20.6	70.6	52.4	70.9	62.9	68.8	74.8	45.3	63.6	39.1	73.7	61.5	63.56
DeepLabv3	76.7	21.0	80.3	61.8	62.9	84.0	78.4	80.0	19.6	77.8	53.5	75.5	69.4	71.3	74.4	44.8	68.5	41.2	79.3	61.5	65.29
DeepLabv3+	77.5	37.3	83.9	64.3	66.5	85.9	78.9	80.4	27.1	80.8	59.3	79.5	74.1	74.2	77.4	44.6	73.0	37.2	77.4	59.5	68.02
PSPNet	83.9	36.8	77.8	61.1	71.8	85.9	77.7	82.7	36.3	77.0	51.2	76.5	75.7	78.9	79.2	48.2	80.6	51.2	76.5	59.9	69.35
OCNet	83.4	49.7	81.7	62.7	71.7	77.9	81.3	82.0	30.6	73.3	57.7	77.9	76.0	82.3	79.9	58.2	78.4	50.2	77.4	61.8	70.60
Our	87.2	37.5	82.3	53.7	73.5	91.4	86.0	86.8	40.7	72.5	56.0	79.6	78.2	82.3	82.3	53.7	80.0	52.0	81.7	66.1	72.01

The above experimental results are all performed on the same experimental equipment with the same parameters

The 'cur_iter' is the current number of iterations, and the 'max_iter' is the maximum number of iterations in the training process. The weight factor λ of the loss function in Eq. (1) is 0.4. It should be noted that due to the limitation of the experiment equipment, batch size 16 and ResNet-101 are not available for our experiments. Therefore, the replicated experimental results cannot reach the results given in the paper.

For dataset PASCAL VOC 2012, the cropped size is set as 513×513 pixels, and the initial learning rate LR is 0.007, and for dataset Cityscapes, the cropped size of image is 768×768 pixels, and the initial learning rate LR is 0.01. It is worth mentioning that the two initial learning rate LR are consistent with those set in DeepLab, PSPNet and OCNet.

At the same time, we conduct data augmentation with a horizontal flip and random scale in training, and randomly scale the image with a scale rate of 0.5–1.5 for inference.

4.2 Ablation experiments

We evaluate how each of these factors, along with loss function of B2L, PPAM and SAM, affects *val* set performance. Ablation segmentation is performed on dataset PASCAL VOC 2012 *val* set.

B2L and PPAM are employed based on base dilated FCN (short for B_FCN); the segmentation performance is illustrated in Table 2. We can see from Table 2 that the proposed network achieves 69.11% mean IoU on ResNet-50, and 70.49% mean IoU on ResNet-101. Compared with B_FCN, B_FCN with B2L improves by 1.57% mIoU and with B2L and PPAM improves by 5.55% mean IoU.

It can be seen from Fig. 5 that under the same network structure and parameters, adding PPAM can effectively improve the performance of semantic segmentation and enhance the relevancy among pixels.

SAM is utilized based on B_FCN with B2L and PPAM (short for B_FCN_BP); the segmentation performance is shown in Table 3, and it achieves 70.89% mIoU, increased by 1.78% compared with B_ FCN_BP. We find that SAM proposed in our work is feasible and effective. Besides, due to the limitation GPU memory, there is no further study of the experiment of backbone ResNet-101 which needs more hardware resources.

We adopt strategies to improve performance further. Multi-grid (MG) [8]: We employ a hierarchy of grids of different sizes {4, 8, 16} in the Res-4. Data augmentation (DA): we use a horizontal flip and random scale. Multi-scale (MS): We average the segmentation probability maps from 5 image scales {0.5, 0.75, 1, 1.25, 1.5} for inference.

MG, DA and MS are exerted based on B_FCN_BP with SAM (short for B_FCN_BPS); the performance is illustrated in Table 4. We adopt MG to obtain better feature representations of pre-trained network, which further achieves 0.17%

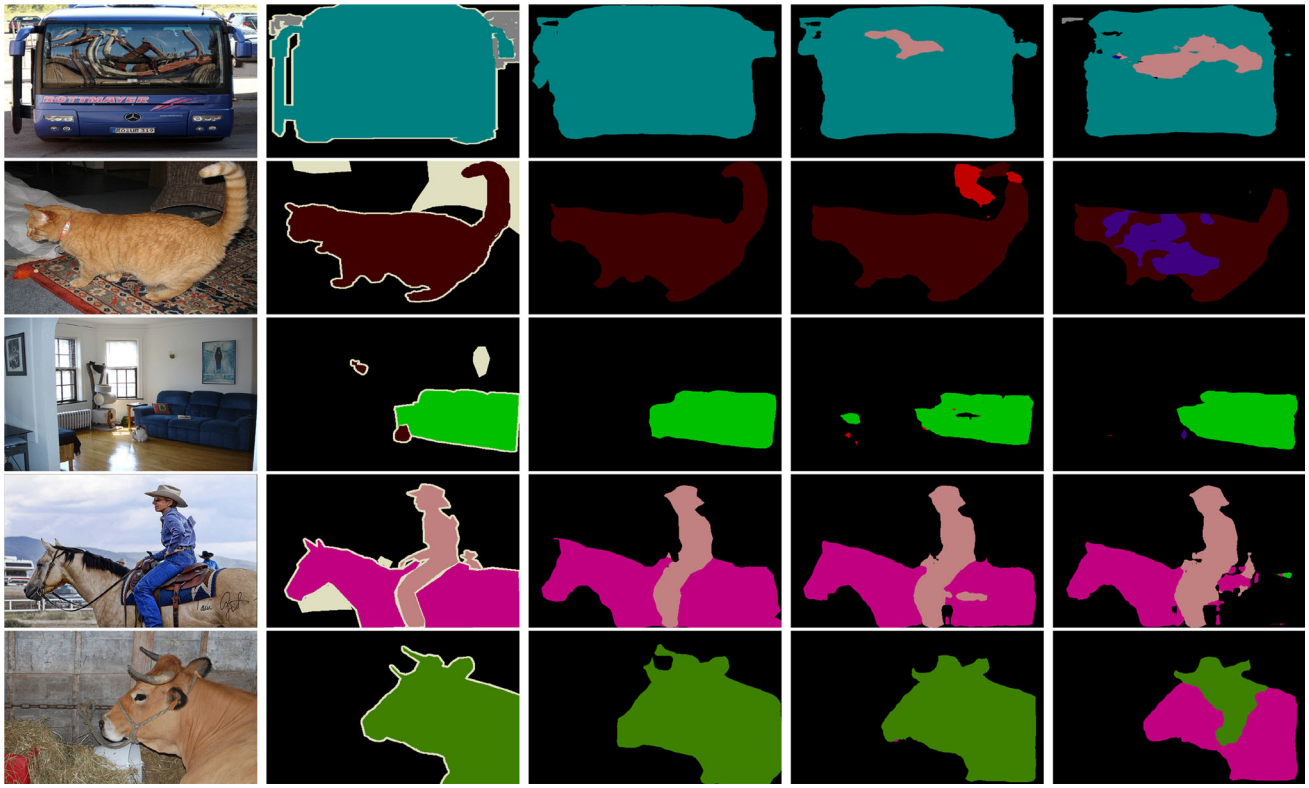

Fig. 6 Visualization results on the PASCAL VOC 2012 *val* set. The first column is the original image, the second column is the ground truth label, and the third to fifth columns are the results of 'Our', OCNet and PSPNet

improvements compared with B_FCN_BPS. DA with random scaling improves the performance by 0.79%, which shows that network benefits from enriching scale diversity of training data. Finally, segmentation map fusion further improves the performance to 72.01%.

4.3 Cityscapes

Cityscapes [22] is a large-scale dataset that focuses on the understanding of urban street scene from 50 different European cities, which contains 30 classes and only 19 classes of them are used for scene parsing evaluation. The dataset contains about 5000 fine annotated images and 20,000 coarsely annotated images. The provided set has 2975 training images, 500 validation images and 1525 test images.

We further compare our proposed network with those methods on the dataset Cityscapes, and the results are shown in Table 5. Here, we use the fine annotation dataset. It can be seen from Table 5 that 'Our' performs better than the state-of-the-art models dilated FCN, DeepLabv3, DeepLabv3+ and PSPNet with 66.53% mean IoU on fine annotation dataset.

4.4 PASCAL VOC 2012

PASCAL VOC 2012 [23] is a famous dataset on visual object classes challenge which includes 20 object categories and one background, concerns three main tasks: classification, detection and segmentation. We focus on the task of segmentation. The dataset is split into a training set 1464 images, a validation set 1449 images with annotation and a test set without annotation.

We evaluate the proposed model on PASCAL VOC 2012 and compare it with DeepLabv3 [8], OCNet [9], PSPNet [10]. The model of FCN with PPAM and SAM is trained with augmented data that are described in Sect. 4.1.

We conduct dilated FCN, DeepLabv3, DeepLabv3+, PSPNet, OCNet with the same parameters setup, and comparative experiment results are shown in Table 6. It can be seen from Table 6 that 'Our' is nearly 3% mIoU more than PSPNet and 2% mIoU more than OCNet.

It can be seen from the comprehensive analysis from Tables 1, 5 and 6 that our proposed network has lower computational complexity than that of PSPNet and OCNet and has better prediction than those of two networks on the same experimental platform and datasets. The proposed network achieves moderate computational complexity and an excellent prediction accuracy (72.01%).

4.5 Visualization results

In this section, we visualize the prediction image of each model to compare and analyze the performance of each model, and the visualization results are shown in Fig. 6, in which the first column is the original image, the second column is the ground truth label, and the third to fifth columns are the prediction image of 'Our', OCNet and PSPNet.

Comparing the data from Table 6, the segmentation performance of 'Our' is better than PSPNet and OCNet as a whole. In detail, 'Our' is better at predicting the edges of objects and it can be able to predict the pixels of within class.

5 Conclusions

In this paper, we propose a modified FCN semantic segmentation method based on the attention module to enhance the semantic consistency. Evaluation and empirical results demonstrate that **the FCN with PPAM and SAM** achieves the superior performance over dilated FCN, DeepLabv3, DeepLabv3+, PSPNet and OCNet. The attention modules are sufficient to enhance the relevancy among pixels and to produce the semantic consistency prediction.

In the future, we need to consider working out the problem of procuring different sizes through adaptive pooling and replace ResNet with MobileNet.

Acknowledgements This work was supported in part by the Joint fund for regional innovation and development of NSFC (U19A2083), by the Science and Technology Plan Project of Hunan Provinc (2016TP1020), open fund project of Hunan Provincial Key Laboratory of Intelligent Information Processing and Application for Hengyang normal university(IIPA20K04).

References

1. Rother, C., Kolmogorov, V., Blake, A.: GrabCut-interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. **23**, 309–314 (2004)
2. Jian-Feng, Xia: Segmentation and recognition of cancer cells based on mathematical morphology. Electron. Sci. Technol. **29**, 36–38 (2016)
3. He, X., Zemel, R.S., Ray, D.: Learning and incorporating top down cues in image segmentation. In: Proceedings of the 9th European Conference on Computer Vision. Graz, Austria, 7–13 May, pp. 338–351 (2006)
4. RavD, Bober M., Farinella, G.M., Guarnera, M., Battiato, S.: Semantic segmentation of images exploiting DCT based features and random forest. Pattern Recogn. **52**, 260–273 (2016)
5. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**, 504–507 (2006)
6. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Patten Anal. Mach. Intell. **39**, 640–651 (2017)
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 Oct, pp. 234–241 (2015)
8. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR, arXiv:1706.05587 (2017)
9. Yuhui, Y., Jingdong, W.: OCNet: object context network for scene parsing. CoRR, arXiv:1809.00916 (2019)
10. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico, USA, 25–30 June, pp. 2881–2890 (2017)
11. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, Long Beach CA, USA, 16–20 June, pp. 3146–3154 (2019)
12. Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K.: Recurrent models of visual attention. In: Advances in Neural Information Processing, June 24, pp. 2204–2212 (2014)
13. Chen, Liang-Chieh, Papandreou, George, Kokkinos, Iasonas, Murphy, Kevin, Yuille, Alan L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 834–848 (2018)
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, Columbus, America, 24–27 June, pp. 580–587 (2014)
15. Badrinarayanan, Vijay, Kendall, Alex, Cipolla, Roberto: Segent: deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 2481–2495 (2017)
16. Lin, T.-Y., Dollr, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, America, 25–30 June, p. 4 (2017)
17. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 18–22 June, pp. 8759–8768 (2018)
18. Liu, H., Peng, C., Yu, C., Wang, J., Liu, X., Yu, G., Jiang, W.: An end-to-end network for panoptic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, Long Beach CA, USA, 16–20 June, pp. 6172–6181 (2019)
19. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 18–22 June, pp. 7794–7803 (2018)
20. Miao, S., Piat, S., Fischer, P., et al.: Dilated FCN for multi-agent 2D/3D medical image registration. In: AAAI Conference on Artificial Intelligence, New Orleans, USA, 2–7 Feb, pp. 4694–4701 (2018)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 27–30 June, pp. 770–778 (2016)
22. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 27–30 June, pp. 3213–3223 (2016)
23. Everingham, Mark, Van Gool, Luc, Williams, Christopher K.I., Winn, John, Zisserman, Andrew: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**, 303–338 (2010)
24. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 834–848 (2018)

25. Krizhevsky, A., Sutskever, I., Hinton, G.E., et al.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, Lake Tahoe, USA, 3–6 Dec, pp. 1097–1105 (2012)