Neural Processing Letters (2021) 53:4177–4188
https://doi.org/10.1007/s11063-021-10587-7

Refine for Semantic Segmentation Based on Parallel Convolutional Network with Attention Model

Gang Peng[1] · **Shiqi Yang**[1] · **Hao Wang**[1]

Accepted: 9 July 2021 / Published online: 5 September 2021

Abstract
High-precision semantic segmentation methods require global information and more detailed local features. It is difficult for ordinary convolutional neural networks to efficiently use this information. In response to the above issues, this paper uses the attention to scale method and proposes a novel attention model for semantic segmentation, which aggregates multi-scale and context features to refine prediction. Specifically, the skeleton convolutional neural network framework takes in multiple different scales inputs, by which means the CNN can get representations in different scales. The proposed attention model will handle the features from different scale streams respectively and integrate them. Then location attention branch of the model learns to softly weight the multi-scale features at each pixel location. Moreover, we add an recalibrating branch, parallel to where location attention comes out, to recalibrate the score map per class. We achieve quite competitive results on PASCAL VOC 2012 and ADE20K datasets, which surpass baseline and related works.

Keywords Semantic segmentation · Parallel convolutional network · Attention model · Multi-scale · Multi-dilation

1 Introduction

With the advance of deep learning algorithms, significant progresses have been made in computer vision field. For instance, semantic segmentation, also known as image labeling or scene parsing which aims at giving label for each pixel, has made great breakthroughs in

Peng Ga ng , PhD, Assoc. Prof, IEEE member , Yang Shiqi (Co First Author),Master; Wang Hao (Corresponding Master graduate student).

✉ Hao Wang
wa_haoo@hust.edu.cn

Gang Peng
penggang@hust.edu.cn

[1] Key Laboratory of Image Processing and Intelligent Control Ministry of Education, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China

recent years. Efficient semantic segmentation can facilitate plenty of other missions such as image editing or image translation.

The state-of-the-art methods on semantic segmentation are all utilizing convolutional neural networks. And almost all these methods will deploy a pre-trained network originally designed for image classification, while image classification is a different task from semantic segmentation, which demands more location information. Attention models are most popular in the field of natural language processing (such as machine translation). In recent years, they have also emerged in the field of computer vision, such as image classification methods [1–3] using attention mechanisms, and generative adversarial networks[4] using attention models. Attention to Scale [5] introduced the attention mechanism to semantic segmentation for the first time, weighted fusion of multi-scale features, and improved accuracy. In order to capture and use context information on semantic segmentation tasks, the paper [6] introduced a context encoding module and semantic encoding loss function, which can selectively highlight class-related feature maps, making training more systematic and performance on small objects is often improved in practice. Similarly, the paper [7] proposed the Feature Pyramid Attention module and Global Attention Upsample module to obtain and utilize context information. Some papers[8, 9] adds two types of attention modules to the traditionally expanded FCN, which respectively simulate the semantic interdependence in the spatial and channel dimensions, and can adaptively integrate local features and their global dependencies.

In order to make better use of the spatial information of network features, this paper refers to the Attention to Scale method and proposes a new attention model for semantic segmentation. Specifically, the features at different locations of each branch-scale network are firstly upsampled to make their sizes consistent, and these features are directly fused across channels using dilation convolution with different dilation coefficients for different scale branch features. After such processing, the attention model can learn information from large scale branches to target regions with large spatial span, and also from small scale branches to target regions with small spatial span, thus obtaining feature representations containing more contextual information. Then obtaining the fused features, we designed two parallel small network branches to output adjustment information for each of the two kinds of main network predictions, including the coordinate attention branch and the fine-tuning calibration branch. The coordinate attention branch serves the same purpose as the ordinary attention model, while the fine-tuning calibration branch, which aims to find correlation information of adjacent targets or regions.

In this paper, we have mainly made two-fold contributions, as follows:

(1) We introduce a novel attention model into multi-scale streams semantic segmentation framework, the final mask prediction is produced by merging the predictions from multiple streams.
(2) The attention model utilizes fused features from different positions of CNN, which carry more contextual information, and has two branch outputs, where one is for location attention and another is for recalibrating.

2 Related Works

Recent approaches for semantic segmentation are all almost based on Fully Convolutional Network (FCN) [10], which outperforms the traditional methods by replacing the fully connected layers with convolutional layers in classification network. The follow-up works have extended the FCN from several points of view. Some works [11, 12] have introduced the coarse-to-fine structure with upsampling modules like deconvolution to give the final mask prediction. And due to the usage of pooling layer, spatial size has decreased largely, for which dilated (or atrous) convolution [13, 14] has been employed to increase the resolution of intermediate features and hold the same receptive field simultaneously.

Other works mainly focus on two directions. One is to post-process the prediction from the CNN through Conditional Random Field (CRF) to get smooth output. For example, in the DeepLab [13] method, a fully connected CRF treated each pixel in the network prediction graph as a graph node, and modeled the degree of association of all nodes on the graph, which can recover some details missed because of the slight spatial invariance of the convolutional network. Another direction is to ensemble multi-scale features. Because features from lower layers in CNN have more spatial information and ones from deeper layers have more semantic meaning and less location information, it is rational to integrate representations from various positions since location information is important for semantic segmentation. The first type method is to extract different visual features from the original image for fusion. For example, the paper [30] represented each image by extracting five different visual features. The second type method for multi-scale combines features from different stages with skip connection to get fused features for mask prediction. For example, the paper [31] devised a Hierarchical Deep Word Embedding (HDWE) model, which is a coarse-to-fine click feature predictor, which can utilize different levels of features. And another type is to resize input to several scales and pass each one with a shared network, it will produce final prediction using the fusion of multi-stream resulting features. There are also methods trying to exploit the capability of global context information, like ParseNet [15] which adds a global pooling branch to extract contextual features. And PSPNet [16] adopts a pyramid pooling module to embed global context information to achieve accurate scene perception.

Attention model has been all the rage in natural language processing area, such as [17], and it has also shown its effectiveness in computer vision and multimedia community recently[18–[21]. It allows model to focus on specific relevant features. Attention-to-scale [5] is the first approach to introduce attention model into semantic segmentation for multi-scale. It takes in different scale inputs. For each scale, the attention model produces a weight map to weight features at each location, and the weighted sum of score maps across all scales is then used for mask prediction. But it only utilizes the feature from specific layer to generate attention, which may omit many contextual details, and this can not ensure that the attention model can guide network to get precise results.

Referring to attention-to-scale, we propose a new attention model in this paper, which also takes in multi-scale inputs but integrates features from different layers, similar to hypercolumns [22]. The attention model has two branch outputs, i.e., one for location attention through which it drives network to focus on large objects or regions for small scale input and pay attention to small targets for large scale just like attention-to-scale, another branch is to recalibrate the score map per class since resulting features from several stages

carry contextual information. The outputs from attention model will be applied to multi-scale stream predictions, and final mask prediction is a weighted sum of all these streams.

In addition, the most relevant work to ours recently is the [23], which has applied multi-dilation strategy to Weakly- and Semi- Supervised Semantic Segmentation. Unlike that work, the multiple-dilation strategy in our proposed model aims to capture different spatial relation for different scale path, while the [23] deploys multiple-dilation for features from the same layers and fused the processed features.

3 Proposed Methods

In order to making the features learned by network as much global or contextual information as possible, and to improve the utilization of information with obtaining higher accuracy, we have done two main things, firstly multi-scale fusion with attention model have been combined to improve the network performance. Secondly, we design the attention model as two parallel small network branches to output adjusted information for each of the two predictions of the main network, which effectively improves the utilization of information by the network.

3.1 Attention Model with Multi-Scales

As discussed before, the higher-layer features contain more semantic information and lower ones carry more location information. Fusion of information from several spatial scales will improve the accuracy of prediction in semantic segmentation. In addition, multi-scale aggregation also capturemore contextual representations since some operations like pooling will dispose of the global context information, leading to local ambiguities which will be discussed later. It is the reason why multi-scale fusion gained a lot of popularity.

Since the backbone network in our work is extended from attention-to-scale [5], here we give a brief review on it. In attention-to-scale, the images are resized to several scales which will be fed to a weight-shared CNN, and the attention model takes as input the directly concatenating features from penultimate layer in each scale stream. The attention model consists of two convolutional layers and will produce n channels scores map, where n means the number of input scales. The attention model is expected to adaptively find the best weights on scales. But there exists some problems. The features from penultimate layer surely contain semantic representations, but they lack essential localization and global information fed to the attention model to achieve precise prediction. And we also posit that simply concatenating features from certain position is not conducive to lead the attention model to learn soft weight across scales. Seeing that the attention model is to put large weights on the large object or region in small-scale stream and gives large weights to the small targets in large-scale stream, we think it is rational to handle features from different scales respectively before integrating them.

Inspired by hypercolumns, we adopt the philosophy of it. Like depicted in Fig. 1, features from different stages in CNN get upsampled to same size and then we concatenate them all. To keep computation cost at bay, we choose the size of features after two pooling operation as the appointed resolution to do upsampling by bilinear interpolation. Through this way, the acquired features carry more localization and context information.

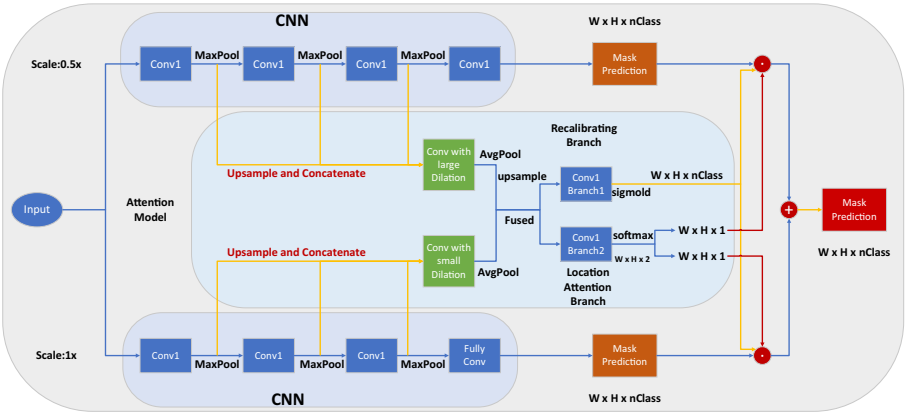

Fig. 1 Architecture of semantic segmentation framework with the proposed attention model. The attention model takes in features from different stages in CNN just like hypercolumns [22], and then it adopts convolutional layer with different dilation to process features for each scale respectively. Attention model produces two kinds of weight maps which are applied to multiple streams predictions. The final mask prediction is a sum of all streams

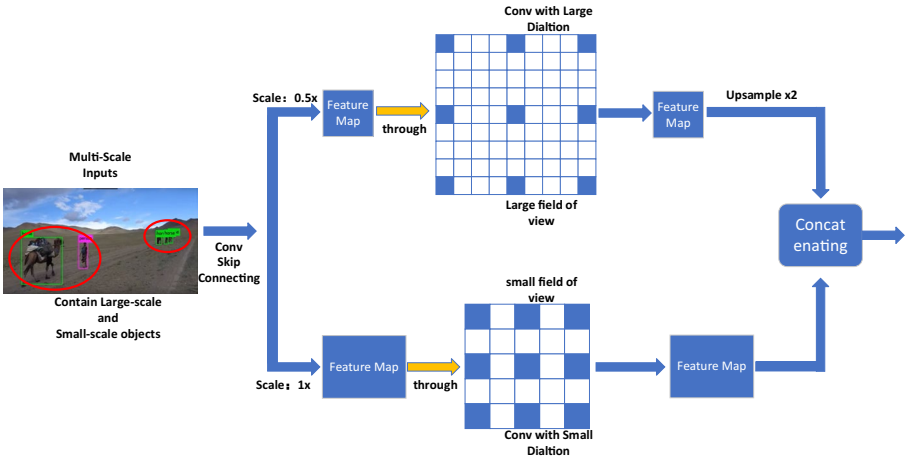

Fig. 2 Convolution with different dilation for different scale. Convolution with large dilation has large field of view while convolution with small dilation has small field of view

It is well-known that the structure of network has an impact on the range of pixels of the input image which correspond to a pixel of the feature map. In other words, filters will implicitly learn to detect features at specific scales due to the fixed receptive field. To accomplish our motivation of attention model which is to adaptively put weights on corresponding scale, we add a unique convolutional layer with unequal dilation for

each scale. This process is demonstrated in Fig. 2. Convolution with large dilation has large field of view (FOV) and is expected to catch the long-span interlink of pixels for large scale object or region in small scale stream, and small dilation convolution is deployed to encode target of small scale in large scale stream. After the dilated convolution, the features will be concatenated, resulting one contains much more abundant and context information, which has been proved by [23].

Moreover, the two-stream CNNs in Fig. 1 are actually the same one when implemented in practice, just like Siamese Network.

3.2 Two Branch Outputs of Attention Model

The concatenated features will go through two parallel convolutional branches: location attention branch and recalibrating branch.

In common with attention-to-scale, the attention model will produce soft weights for multiple scales (we refer to it as location attention). Assuming the number of input scale is n, and the size of mask prediction, which is denoted as P^s for scale s, is $W \times H$, $nClass$ means the class number of the objects. The location attention output by the model is shared across all channels. After the refinement of local attention, the mask predictions, denoted as M_i^s, are described as:

$$M_{i,c}^s = \sum\nolimits_{s=1}^{n} l_i^s \cdot P_{i,c}^s \tag{1}$$

The l_i^s is computed by:

$$l_i^s = \frac{\exp\left(wl_i^s\right)}{\sum_{j=1}^{n} \exp\left(wl_i^j\right)} \tag{2}$$

where wl_i^s is the score map produced by the location attention branch at position $i \in [0, W * H - 1]$ for scale s, before the softmax layer of course.

And since the fused features fed to the attention model contain context information, we want to make full use of them to eliminate some degrees of class ambiguity, i.e., to utilize contextual relationship to enhance the ability of classification. The lack of ability to collect contextual information may increase the chance of misclassification in certain circumstances. To take an example, neural network sometimes tends to take apart a large-scale object into several regions of different classes [24], or maybe classify a boat on the river as a car and so on in scene parsing [16] (these can be observed among visualization results in Sect. 4.1). To deal with these issues, we add a recalibrating branch parallel to location attention. It has the same architecture as location attention branch which means containing two convolutional layers, except that output channel changes to $nClass$ and sigmoid activation is deployed instead of softmax. This branch aims to find the interdependencies between adjacent objects or regions using the integrating features, and its output is used for recalibrating the score maps before the location attention refinement. Because the contextual relationship stay the same in different scale, the recalibrating outputs are shared across all scales. So the final mask prediction for each stream can be described as:

$$M_{i,c}^s = \sum\nolimits_{s=1}^{n} l_i^s \cdot [P_{i,c}^s \otimes wr_{i,c}] \tag{3}$$

where the $\otimes$ means element-wise multiplication and $wr_{i,c}$ means output in position i in channel $c \in [0, n-1]$ produced by recalibrating branch. Another choice for recalibrating branch is to predict bias per position in each channel instead of multiplication. But it will bring around 1% performance decrease according to our experiment.

And the ultimate mask prediction is as below, where M^s is the mask prediction of scale s:

$$M_{final} = \sum_{s=1}^{n} M^s \tag{4}$$

As for the loss function, we follow the setting of attention-to-scale, i.e., the total loss function is sum of $1 + S$ cross entropy loss functions for segmentation, where S symbolizes number of scales and one for final prediction.

4 Experimental Results

We experiment our method on two benchmark datasets: PASCAL VOC 2012 [25] and ImageNet scene parsing challenge 2016 dataset [26] (ADE20K).

For all training, we refer to the multi-scale scale setting method[5], only train the network with 2 scales, i.e., $1\times$upsample and $0.5\times$upsample. As for the different dilation strategy, we set it to 2 for small scale stream and 10 for large scale stream if not specified. And we also adopt the poly-learning rate policy [15], meaning current learning rate is computed by multiplying $\left(1 - \frac{iter}{max\_iter}\right)^{power}$ to base learning rate, where the *power* is set to 0.9. We refer to the layers in the last stage where gives mask prediction as decoder, layers previous to decoder are encoder. Learning rate of decoder is 10 times that of encoder. All experiments are implemented using PyTorch on a NVIDIA TITAN Xp GPU. The following experiments has a little improvement over our original conference paper.

4.1 PASCAL VOC 2012

The PASCAL VOC 2012 [25] segmentation dataset consists of 20 foreground object classes and a background class. The PASCAL VOC 2012 dataset we use is augmented with extra annotation by Hariharan et al. [27], resulting in 10,582 training images and 1449 validation set. In experiment we report performance results on original PASCAL VOC 2012 validation set.

DeepLab-LargeFOV [28] is chosen as base model. Since our work is extended from attention-to-scale, in order to compare fairly, we reproduce the DeepLabLargeFOV and attention-to-scale based on it by ourselves, following the set of attention-to-scale [5]. All

Table 1 Results on PASCAL VOC 2012 validation set. There exists 2 scale streams: $1\times$and 0.5x. The mIoU means mean intersection of union [10]

Method	mIoU
Baseline (DeepLab-LargeFOV)	61.40%
Merged with MaxPooling	63.88%
Merged with AvgPooling	64.07%
Attention-to-Scale	64.74%
Our method	**67.98%**

The bolded data only indicates the best results

Table 2 Ablation study for proposed method on PASCAL VOC 2012. The multi-stage means hypercolumns-like feature integration from different positions. Diverse dilation means utilizing different dilated convolution for multi-scale features. Extra branch means adding recalibrating branch. *-The base model is actually attention-to-scale. †-No diverse dilations means using standard convolution instead

Method	Multi-stage	Diverse dilations†	Location attention	Extra branch	mIoU
Base model*			√		64.74%
Base model +	√		√		65.80%
Base model + +	√	√	√		66.83%
Our method	√	√	√	√	**67.98%**

The bolded data only indicates the best results

these experiments use VGG16 [29] as skeleton CNN, which is pretrained on ImageNet. Our reproduction of them yields performance of 61.40% and 64.74% on the validation set respectively. The performance of attention-to-scale is lower than original paper, but the follow-up experiments still can verify effectiveness of our proposed method since ours is directly built on attention-to-scale. Noted that both of attention-to-scale and our work adopt extra supervision, meaning adding softmax loss function for each scale stream. The results of experiment are demonstrated in Table 1.

Merged with Pooling in Table 1 means adopting pooling operation as fusion approach for multi-scale stream instead of attention model. It can be seen that our method surpasses baseline and attention-to-scale by 6.58% and 3.24% respectively. Furthermore, we conduct additional experiments for ablation study of each module in our method. We cut off certain modules from our proposed method, re-train and report the performance of remainder, which is shown in Table 2. Please noted that base model without all these modules is

Fig. 3 Representative visual segmentation results on PASCAL VOC 2012 dataset. Images are from train and val set. GT means ground truth, and baseline means attention-to-scale approach. Our proposed method produces more accurate and detailed results

Table 3 Results on PASCAL VOC 2012 with different dilation rate, where 'S' means the dilated convolution block with the small dilation rate while 'L' denotes the large dilation block

Configuration of dilation rate	mIoU
S2 L12	67.98%
S4 L12	67.04%
S2 L10	**68.47%**
S4 L10	68.05%
S2 L8	67.71%
S4 L8	67.32%

The bolded data only indicates the best results

actually attention-to-scale approach. As you can see, the modules we design indeed take effect on segmentation task.

Since the attention-to-scale has verified the motivation which we share with by visualizing weight maps produced by the attention model, we don't replicate this experiment on our proposed model. Turning to qualitative results, some representative visual comparisons are provided between attention-to-scale and our method in Fig. 3. We observe that unlike attention-to-scale, our method can get finer contour in some cases and probability of breaking down a largescale object into several pieces decreases. Our results contain much more detailed structure and more accurate pixel-level categorization, which we posit it comes from the utilization of multi-scale and context information as well as the extra branch.

In addition, we also conduct experiments to investigate the effect of different dilation rate. The results are shown in Table 3. The 'S2 L12' is the setting in our original conference paper. The dilation rate directly influences the receptive field in the neural network, in other words the different dilated convolution can catch interlinks between different spatial regions. According to the results, the 'S2 L10' setting is better than the one of our original conference, which is 'S2 L12'.

4.2 ADE20K

ADE20K dataset first shows up in ImageNet scene parsing challenge 2016. It is much more challenging since it has 150 labeled classes for both objects and background scene parsing. It contains around 20 K and 2 K images in the training and validation sets respectively.

We deploy ResNet34-dilated8 [14] (not resnet50 because of limited GPU memory) as base CNN to investigate several different methods. Besides applying attention-to-scale and our proposed attention model, we also experiment on Pyramid Scene Parsing (PSP) [16] module as a comparison, which is a state-of-the-art approach on ADE20K dataset to

Table 4 Results on ADE20K validation set. *- Two multi-scale attention methods take as input two scale streams: 1× and 0.5×

Method	mIoU	Pixel accuracy
ResNet34-dilated8 (Baseline)	32.67%	76.41%
Baseline + attention-to-scale*	35.11%	76.82%
Baseline + PSP	36.43%	78.01%
Baseline + our attention model*	**37.31%**	**78.83%**
Baseline + our attention model* + PSP	**38.74%**	**79.39%**

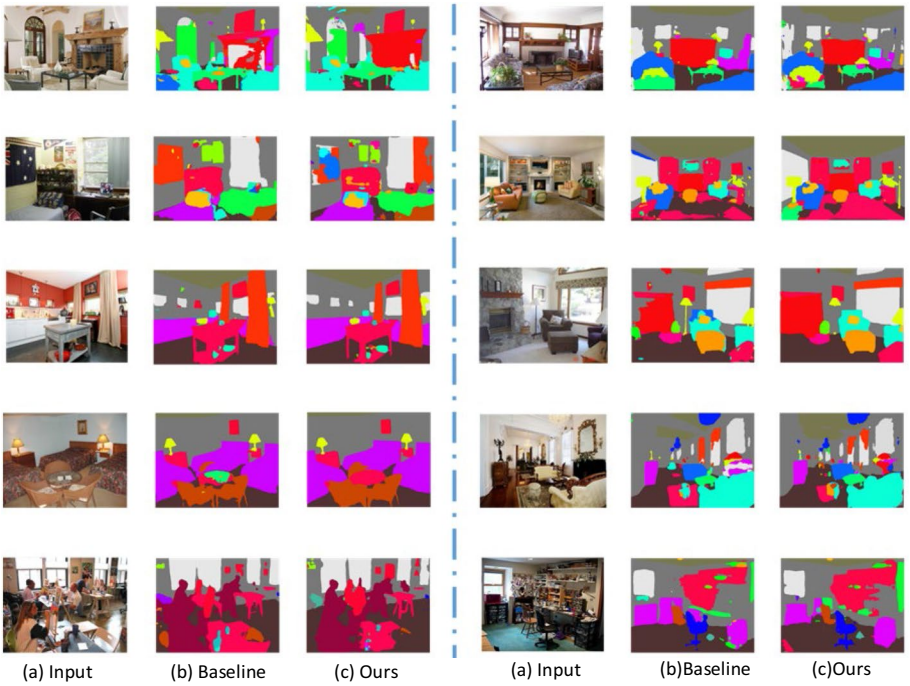

Fig. 4 Representative visual segmentation results on ADE20K dataset. Images are from train and val set. Baseline means attention-to-scale method

the best of our knowledge. The experiment results are presented in Table 4. The PSP here doesn't contain auxiliary loss in original paper. We can see that our proposed attention model outperforms other methods, and achieves 4.40% improvement on mIoU over baseline. Besides, we also embed both the PSP module and proposed attention module in baseline and it obtains further performance improvement.

Similarly, in order to further verify the effectiveness of the method, we also give a comparison of the actual effect graphs of the prediction of the baseline method and the attention-to-refine method, as shown in the Fig. 4. the baseline method is based on attention-to-scale and the method refers to the use of attention-to-refine, which does not include the PSP module. As shown in the figure, except for a few regions where attention-to-refine classifies incorrectly and attention-to-scale classifies correctly, the attention-to-refine method can effectively reduce category ambiguity and improve the accuracy of semantic segmentation in most cases.

5 Conclusion

High-precision semantic segmentation methods require the fusion of information from multiple different spatial scales. On the one hand, the convolutional neural network will lose some local information due to the pooling operation. On the other hand, the global or contextual information of an image is very important to eliminate local ambiguity. It is difficult for ordinary convolutional neural networks to efficiently use global information.

In response to the above issues, this paper uses the attention to scale method and propose a novel attention model for semantic segmentation. The whole CNN framework takes in multi-scale streams as input. Features from different stage of CNN are fused, then resulting one in each scale goes through convlutional layers with different dilation, which are expected to catch distinctive context relationship for different scales. After that, all these features get concatenated and resulting one is fed into two parallel convolution output branches of the attention model. One of the branches is location attention, aiming to pay soft attention to each location across channels. Another one is designed to fully utilize contextual information to deal with class ambiguity by recalibrating the prediction per location for each class. Experiments on PASCAL VOC 2012 and ADE20K show that proposed method make a significant improvement.

Acknowledgements The work of paper was supported by National Natural Science Foundation of China(No. 61672244), Hubei Province Natural Science Foundation of China(No.2019CFB526).

References

1. Wang F, Jiang M, Qian C, et al. (2017) Residual attention network for image classification[J]. arXiv preprint https://arxiv.org/abs/1704.06904
2. Zheng H, Fu J, Mei T, et al. (2017) Learning multi-attention convolutional neural network for fine-grained image recognition[C]. In Int. Conf. on Computer Vision. Venice, Italy, 6
3. Sun M, Yuan Y, Zhou F, et al. (2018) Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition[J]. arXiv preprint https://arxiv.org/abs/1806.05372
4. Chen X, Xu C, Yang X, et al. (2018) Attention-GAN for Object Transfiguration in Wild Images[J]. arXiv preprint https://arxiv.org/abs/1803.06798
5. Chen, Liang-Chieh, et al. (2016) "Attention to scale: Scale-aware semantic image segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition
6. Zhang, Hang, et al. (2018) "Context encoding for semantic segmentation." IEEE conference on Computer Vision and Pattern Recognition. June 18–23, 2018, Salt Lake City, USA
7. Li, Hanchao, et al. (2018) "Pyramid attention network for semantic segmentation." arXiv preprint https://arxiv.org/abs/1805.10180
8. Fu, Jun, et al. (2019) "Dual attention network for scene segmentation." IEEE Conference on Computer Vision and Pattern Recognition. June 15–20, 2019, Long Beach, USA
9. Liu, Yifu, et al. (2020) "Deep Dual-Stream Network with Scale Context Selection Attention Module for Semantic Segmentation." Neural Processing Letters: 1–19
10. Long, J., Shelhamer, E., Darrell, T. (2015): Fully convolutional networks for semantic segmentation. IEEE conference on computer vision and pattern recognition, Proceedings:3431–3440, June 7–12, 2015, Boston, USA.
11. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495
12. Noh, H., Hong, S., Han, B. (2015): Learning deconvolution network for semantic segmentation. IEEE International Conference on Computer Vision, Proceedings: 1520–1528, June 7–12 2015, Boston, USA
13. Chen, Liang-Chieh, et al. (2017) "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2017): 834–848
14. Yu, F., Koltun, V. (2015): Multi-scale context aggregation by dilated convolutions. arXiv preprint https://arxiv.org/abs/1511.07122
15. Liu, W., Rabinovich, A., Berg, A.C. (2015): Parsenet: Looking wider to see better. arXiv preprint https://arxiv.org/abs/1506.04579
16. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. IEEE Conference on Computer Vision and Pattern Recognition, Proceedings:2881–2890, July 21–26, 2017, Honolulu, USA
17. Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014) "Neural machine translation by jointly learning to align and translate." arXiv preprint https://arxiv.org/abs/1409.0473

18. Chen, Jingyuan, et al. (2017) "Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention." International ACM SIGIR conference on Research and Development in Information Retrieval. Aug 7–11, Shinjuku, Japan
19. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X. (2017): Residual attention network for image classification. IEEE Conference on Computer Vision and Pattern Recognition, Proceedings: 3156–3164, July 21–26, Honolulu, USA
20. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y. (2015): Show, attend and tell: Neural image caption generation with visual attention. International Conference on Machine Learning, Proceedings: 2048–2057, July 6–11, Lille, France,
21. Song, X., Feng, F., Han, X., Yang, X., Liu, W., Nie, L. (2018): Neural compatibility modeling with attentive knowledge distillation. arXiv preprint https://arxiv.org/abs/1805.00313
22. Hariharan, B., Arbel´aez, P., Girshick, R., Malik, J. (2015): Hypercolumns for object segmentation and fine-grained localization. IEEE conference on computer vision and pattern recognition, Proceedings:447–456, June 7–12, Boston, USA
23. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S. (2018): Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition, Proceedings: 7268–7277, June 18–23, Salt Lake City, USA
24. Li, X., Jie, Z., Wang, W., Liu, C., Yang, J., Shen, X., Lin, Z., Chen, Q., Yan, S., Feng, J.: Foveanet (2017): Perspective-aware urban scene parsing. IEEE International Conference on Computer Vision, Proceedings:784–792, Oct 22–29, 2017, Venice, Italy
25. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascalnetwork.org/challenges/VOC/voc2012/workshop/index.html
26. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A. (2016): Semantic understanding of scenes through the ade20k dataset. arXiv preprint https://arxiv.org/abs/1608.05442
27. Hariharan, B., Arbel´aez, P., Bourdev, L., Maji, S., Malik, J. (2011): Semantic contours from inverse detectors. Computer Vision (ICCV), 2011 IEEE International Conference on, Proceedings:991–998, Nov 6–13, Barcelona, Spain
28. Chen, Liang-Chieh, et al. (2014)"Semantic image segmentation with deep convolutional nets and fully connected crfs." arXiv preprint https://arxiv.org/abs/1412.7062
29. Simonyan, K., Zisserman, A. (2014): Very deep convolutional networks for large-scale image recognition. arXiv preprint https://arxiv.org/abs/1409.1556
30. Yu J, Rui Y, Tao D (2014) Click prediction for web image reranking using multimodal sparse coding. IEEE Trans Image Process 23(5):2019–2032
31. Yu, J., Tan, M., Zhang, H., Tao, D., & Rui, Y. (2019) Hierarchical deep click feature prediction for fine-grained image recognition. IEEE transactions on pattern analysis and machine intelligence