Multi-branch Semantic Segmentation Network

LiHua Wei, YingDong Ma*
College of Computer Science
Inner Mongolia University
Hohhot, China
csmyd@imu.edu.cn

***Abstract*—Effective semantic segmentation requires both spatial details and object-level semantic information. Meanwhile, context information is also important for complex scene understanding. However, it is hard to meet these demands simultaneously in the top-down CNN structure. In this paper, we tackle this problem with a Multi-branch Semantic Segmentation Network (MSS Net). The proposed MSS Net consists of three parts, including a spatial network, a semantic network and a context network. The spatial network utilizes convolutional layers with small stride and a spatial pyramid pooling module to extract multi-scale spatial features. In the semantic network, multiple level features are combined to enhance semantic information. The context network integrates different scales contextual information to facilitate objects localization in complex scene. The proposed semantic segmentation framework has been evaluated on the CamVid and the Cityscapes datasets. Experimental results demonstrate that the MSS Net achieves state-of-the-art performance.**

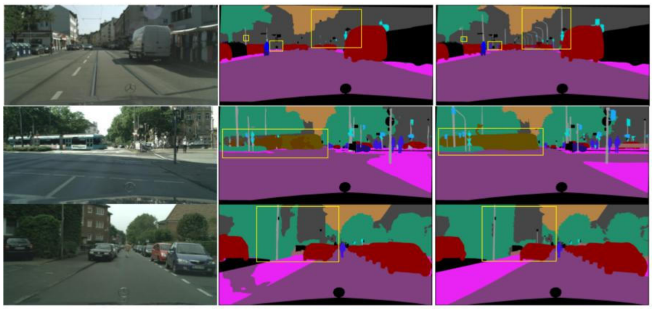
***Contribution*—We propose a novel multi-branch network for effective semantic segmentation.**

Keywords—Semantic segmentation, convolutional neural network, contextual information.

I. Introduction

Semantic segmentation is one of the fundamental tasks in computer vision, which assigns labels to each pixel of input images. Most computer vision applications, including automatic driving, indoor navigation, human-computer interaction, and virtual reality, rely on precise semantic segmentation results. With the fast development of convolutional neural networks (CNNs), impressive results have been reported in semantic segmentation due to the network ability of automatic learning of hierarchical features [1, 2, 3, 4, 5].

As a typical CNN structure, the fully convolution network (FCN) [1] has been widely used in segmentation applications. The FCN downsamples input images progressively to yield low resolution feature maps. Although final feature maps provide rich semantic information, the top-down CNN structure suffers spatial information loss. As shown in Fig. 1, some objects are not predicted or predicted incorrectly by the FCN network [1]. For example, some small parts (e.g. street lamps in the first row) are missing. It indicates that FCN-based methods lack the ability to make prediction on small objects.



(a) Image (b) FCN (c)Ground Truth

Figure 1. Visualization examples on Cityscapes dataset. The second column shows that the FCN network has difficulty in predicting small parts and object details due to spatial information loss.

Recent works on semantic segmentation show that there are mainly three approaches to combine spatial features with high-level semantic information. Firstly, the FCN series replaces the fully connected layers with the convolutional layers to form a full convolutional network. Object and scene segmentation are conducted from deep layer feature maps which can be enhanced by fusing previous layer features. Nevertheless, prediction based on high-level feature maps may lead to lower accuracy. Secondly, image pyramid-based methods [4] transform input images into different scales and apply CNN model on each level image to generate multiple level features. Predictions from different levels are combined to obtain the final output. Although these methods are simple, they might lose spatial details in the downsampling process and increase computational cost. Thirdly, some segmentation works adopt the Encoder-Decoder structure [2,6]. These methods use low-level features to restore spatial resolution in the decoder stage. However, some detail structure information might be lost in the downsampling stage and only part of them can be recovered from the upsampling stage.

Context relationship has been proved to be important especially for complex scene understanding. In the second and third rows of Fig. 1, some parts of bus and building are predicted incorrectly. This problem can be remedied by using the contextual information. In [5], global context is obtained from global average pooling. Although global pooling is commonly used in various visual applications, global descriptors are not suitable for semantic segmentation due to loss of spatial relationship. The works in [3] extracts contextual information by using dilated convolutions. Dilated convolutions play an important role in maintaining spatial resolution of deep feature maps. Meanwhile, some algorithms, such as DeepLab [4], adopt Atrous Spatial Pyramid Pooling

This work was supported by the National Natural Science Foundation of China under Grant 61461039.

(ASPP) to encode global context. However, segmentation based on these methods lack of an effective scheme to combine global and local context information.

Inspired by these works, we propose a multi-branch semantic segmentation network to alleviate these problems. The proposed MSS Net consists of a spatial network, a semantic network and a context network. The semantic network is designed to enhance multiple level semantic information, while the spatial network focuses on learning spatial details from high resolution feature maps. The context network combines multiple scale features to capture local and global context information.

Specifically, in the spatial network, input images are transformed by using convolutional layers with small stride. Spatial information of different regions is aggregated by a Multi-scale Pyramid Pooling (MPP) module. To integrate low-level features with high-level semantic information, we introduce a Feature Enhancement Module (FEM) in the semantic network to capture multi-level semantic information. Different level feature maps are combined to form shallow features and deep features in the context network. A Context Pooling Aggregation Module (CPAM) is developed to collect local and global context information from these combined features. The final predictions are obtained by concatenating outputs of spatial network and context network. The main contributions of this paper are summarized as follows:

- We introduce a multi-branch network for semantic segmentation. The proposed approach combines spatial details, multi-level semantic clues, and multi-scale context information with a three-branch structure. The spatial network learns spatial details by aggregating spatial information from different regions and the semantic network combines multiple level feature maps to obtain semantic features. The context network collects local and global context information to further improve segmentation performance.
- We propose the feature enhancement module and a multi-scale pyramid pooling module to enhance network representative capability of low-level features and high-level semantic information. A context pooling aggregation module is introduced to obtain multiple scale context information.

II. RELATED WORK

A. *CNN-based Semantic Segmentation*

The FCN [1] is one of the most commonly used architectures in large-scale computer vision tasks. By replacing the fully-connected layers with convolution layers, it achieves great progress in semantic segmentation. Nevertheless, segmentation based on high-level feature maps only may lead to lower accuracy. Some semantic segmentation methods adopt the encoder-decoder structure to restore spatial resolution. The work [7] presents a new model to upgrade the FCN architecture. The model has a contracting path and an expansive path, in which missing spatial information can be extrapolated by upsampling operators in the expansive path. Vijay et al. proposed the SegNet to improve segmentation accuracy [2]. The SegNet consists of an encoder network and a decoder network, which maps low resolution features to high resolution feature maps for pixel-wise classification.

Recently, some modified encoder-decoder structures have been reported. In [8], Bilinski and Prisacariu proposed dense decoder shortcut connections that allows decoder blocks to use previous level semantic feature maps.

B. *Contextual and Spatial Information*

Semantic segmentation not only relies on spatial details to obtain pixel-level predictions but also requires context information to make reliable prediction in complex scene and classification of objects with similar appearance. The work [9] improves segmentation performance with patch-patch context and patch-background context information. The pyramid scene parsing network [5] aggregates scene context features from different sub-regions which strength network capability of global scene understanding. The DeepLab [4] obtains image-level features through global average pooling. The atrous spatial pyramid pooling is adopted to concatenate dilated convolution features for collecting context information. In [10], Yang et al. proposed the DenseASPP which connects a set of atrous convolutional layers to obtain dense feature maps for autonomous driving.

C. *Attention Mechanism*

The attention mechanism plays an important role in the process of human perception as we have the ability to pay more attention to salient regions in a cluttered visual scene. Recently, attention mechanism has been successfully applied in many computer vision tasks, such as image classification, and image segmentation [11,12]. In the Pyramid attention network [11], Li et al. combines attention mechanism with pyramid structure. The method extracts dense features from lower level feature maps based on high-level semantic guidance. The DANet [12], which is developed based on the self-attention mechanism, adaptively integrate local features and global dependencies to capture context information.

D. *Multi-branch Networks*

One of the main limitations of a typical top-down CNN architecture is that high resolution features obtained from shallow layers are lack of high-level semantic meaning which can only be extracted from deep layers. To alleviate the problem, some semantic segmentation approaches adopt the multi-branch architecture. The main purpose of multi-branch models is that different level features can be obtained separately from various network paths.The Refinenet [6]uses a multi-path refinement network to implement semantic segmentation. Feature maps of multiple level Resnet blocks are combined in different paths so that semantic features captured from deep layers can be refined using fine-grained features from earlier convolutions. The Bisenet [13] employs a spatial path to preserve spatial information for generating high-resolution features.Meanwhile, a context path with fast

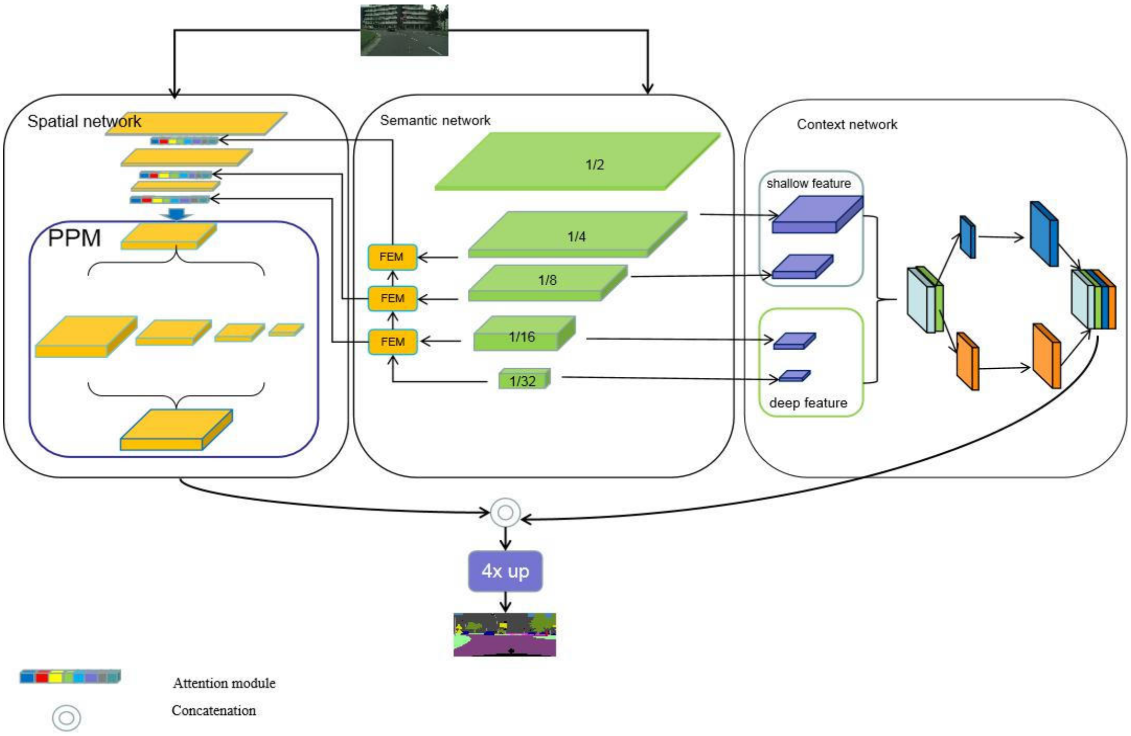


Figure 2. The overall structure of the Multi-branch Semantic Segmentation Network. FEM: Feature Enhancement Module. CPAM: Context Pooling Aggregation Module. PPM: Pyramid Pooling Module.

downsampling is used to yield sufficient receptive fields. The ContextNet [14] adopts a similar structure with the Bisenet, in which a deep branch captures global context information with a shallow branch that focuses on spatial details. To implement semantic segmentation on high-resolution images, Zhao et al. proposed the image cascade network, in which input images with different resolutions are processed in multi-resolution branches [15]. Different level feature maps are combined in the cascade feature fusion unit to integrate semantic information with image features. The two-branch Fast-SCNN [16] proposes a learning to downsample module which extracts low-level features through multi-branch efficiently and, meanwhile, it uses a deep branch to compute global features.

III. Multi-branch Semantic Segmentation Network

A. *Overview*

In the task of semantic segmentation, both spatial details and object-level semantic information are crucial to achieve high accuracy. In addition, local and global contextual information is also necessary for complex scene understanding. However, for a typical top-down CNN structure, it is difficult to meet these demands simultaneously. In this work, we introduce a multi-branch semantic segmentation network to solve the problem. The proposed MSS Net consists of three main components: the spatial network, the semantic network and the context network. Multi-level features are integrated in the semantic network to capture semantic information and the context network collects multiple scale context information. The overall framework is shown in Fig. 2.

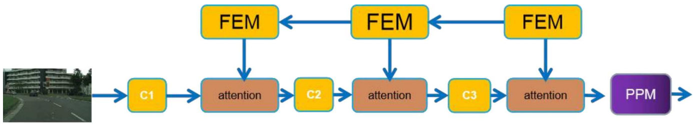


Figure 3. Structure of the spatial network. FEM: Feature Enhancement Module. PPM: Pyramid Pooling Module.

B. *Spatial Network*

For semantic segmentation applications, both fine-level spatial details and object-level semantic information are necessary. Recently, lots of approaches have been developed to solve the problem of missing spatial information. Some methods attempt to encode spatial information with dilated convolution [3]. The U-shape structure [2,7] is also widely utilized, in which deep features are combined with features of shallow layers to increase spatial details.

Motivated by these works, we propose the spatial network to enhance fine-level features. The spatial network consists of three convolution layers to preserve spatial details. The batch normalization and the rectified linear unit (ReLU) are utilized for normalization and activation. An attention module is appended to each convolution layer, which is designed to adjust output response for regions of interesting. These layers are followed by a Pyramid Pooling Module (PPM) to integrate multi-scale spatial information. Fig. 3 illustrates the overall structure of the spatial network.

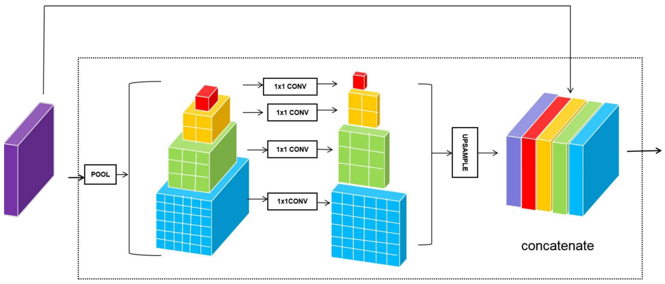


Figure 4. The Pyramid Pooling Module.

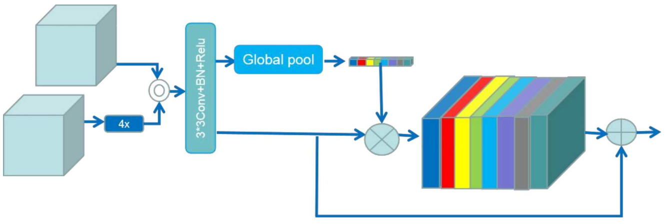


Figure 5. The Attention Module.

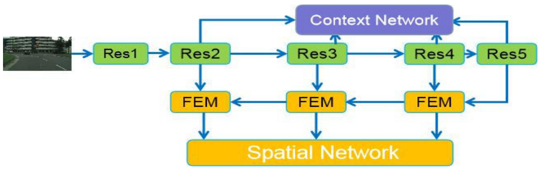


Figure 6. Structure of the semantic network.

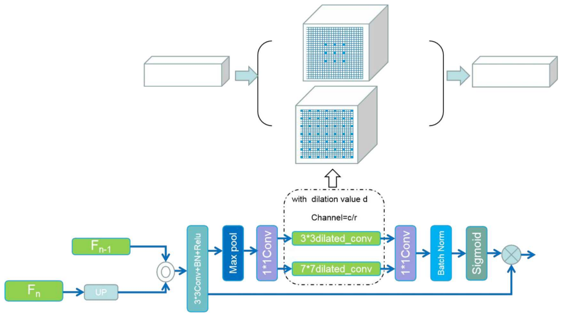


Figure 7. The Feature Enhancement Module.

Multi-Scale Pyramid Pooling Module: In the spatial network, we introduce a pyramid pooling module to capture multiple scale spatial information. As shown in Fig. 4, the PPM fuses features from four pyramid scales. Let input feature maps have spatial size of W×H×D, where D is the number of feature channels. The pyramid pooling separate feature maps into four levels. Features in each pyramid level has spatial size of $\frac{W}{k}\times\frac{H}{k}, k=[1,2,3,6]$. We use a 1×1 convolution to reduce feature channels to D/4. In the next step, feature maps with smaller sizes are upsampled to $W\times H$ by using bilinear interpolation. Features of different pyramid scales are then concatenated with input feature maps as the output pyramid pooling features which have spatial size of $W\times H\times 2D$.

Attention Module: In the spatial network, an attention module is applied to learn weight for multiple level features, as illustrated in Fig. 5. Input features of the attention module contain the output feature maps of the feature enhancement module and the spatial network convolution layer. The smaller size feature maps are up-sampled to match the sizes of different input features. These features are concatenated to form the combination features (CF) and followed by a 3×3 convolution with batch normalization and the rectified linear unit. The global pooling is utilized to generate global context which is multiplied by CF. Finally, weighted features are added with the combination features as the output features of the attention module. The attention module not only combines features of the spatial network and semantic network but also makes feature refinement by learning feature weight to improve segmentation accuracy.

C. Semantic Network

In the semantic segmentation task, object level knowledge and the scene clues are important for complex scene understanding and classification of objects with similar appearance. To utilize different level semantic clues, some approaches [5] use global average pooling to get high-level semantic information. As the representation capabilities of each block in CNNs are different, the U-shape structure [1,2,7,17] is widely adopted to combine multiple level information with global semantic knowledge.

In the proposed method, the ResNet is adopted as backbone in the semantic network. The semantic network has two main purposes. Firstly, a feature enhancement module (FEM) is introduced to combine multiple level features and provide enhanced features to the spatial network. Secondly, feature maps of different blocks are concatenated to yield deep feature block and shallow feature block for the context network. Fig. 6 shows the overall structure of the semantic network.

Feature Enhancement Module: To utilize multiple level features, we introduce the multi-level FEM module to combine different level features. In the FEM module, concatenation is adopted to combine multiple level feature maps. As shown in Fig. 7, features of two levels, F$_n$ and F$_{n+1}$ are concatenated to obtain combined feature F. Feature F is transformed by a 3×3 convolution with batch normalization and the rectified linear unit to obtain feature F'. After reducing feature channels by using a 1×1 convolution, a 3×3 dilated convolution and a 7×7 dilated convolution are utilized in parallel to expand receptive field. The module has two hyperparameters: dilation rate (d) and reduction ratio (r). The dilation rate determines size of the receptive field and helps to aggregate context information. The reduction ratio controls computational overheads. Through experimental verification, we set $\{d=4, r=16\}$. The final feature M is obtained as：

$$\begin{aligned}M &= F'\otimes\sigma(BN(Con_{1\times1}(f_{3\times3}(Con_{1\times1}(Maxpool(F')))+\\ &\quad f_{7\times7}(Con_{1\times1}(Maxpool(F')))))) \end{aligned} \tag{1}$$

where σ is the sigmoid function, BN is batch normalization, $f_{3\times3}$ and $f_{7\times7}$ are 3×3 dilated convolution and 7×7 dilated convolution, respectively.

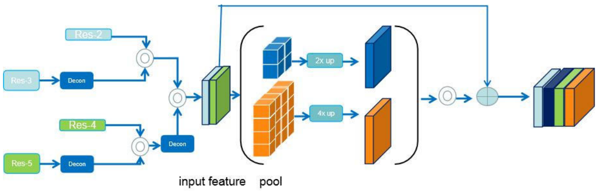


Figure 8. The Context Pooling Aggregation Module. The kernel size of the two pooling layers are 2×2 and 4×4, respectively.

D. Context Network

Some visual recognition works explored the capability of context information in object detection and semantic segmentation [4,5,9,10]. These works have shown that utilizing scene context clues make the final prediction more reliable. To further improve segmentation performance, we present a context pooling aggregation module which combines different scales context information to facilitate exact location of objects in complex scene.

As shown in Fig. 8, the deep feature block and the shallow feature block are combined to form input features of the CPAM module. We use deconvolution to unify the size of features of different blocks. Let the $W \times H$ input feature maps have D channels. The pyramid pooling fuses feature maps under two pyramid levels with kernel size of 2×2 and 4×4, respectively. Then we upsample the smaller size features to get the same size features as the original feature maps. Finally, two pyramid level features are concatenated with the input features to get the output features with spatial size of $W \times H \times 2D$. The CPAM module strengthens each spatial location of input feature maps with local context at different scale space. Combination of different pooling layer features not only expands the receptive field, but also lead to better utilization of multi-scale contextual information.

IV. EXPERIMENTS

The proposed MSS Net is evaluated on two widely used segmentation datasets, the CamVid [18] and the Cityscapes [19]. We first introduce the implementation details. Then we conduct ablation experiments on the CamVid validation set and report performance comparison between the proposed method and state-of-the-arts on CamVid and Cityscapes.

A. Implementation details

In this work, the ResNet-50 network is adopted as the baseline model. The baseline parameters are learned on the ILSVRC 2012 dataset using a desktop computer with a single Nvidia 1080Ti GPU. Our implementation is based on the public platform TensorFlow using the Root Mean Square prop (RMSprop) optimization algorithm. The initial learning rate is 0.0001 and the decay is 0.995. For experiments on CamVid, the number of training iterations is 110k. The training iterations is 71k when training on the Cityscapes. The input RGB images are randomly cropped to patches with the size of 512×512 pixels when training on the Camvid. The input images are 768×768 pixels when training on the Cityscapes. We use the mean intersection-over-union (Mean IoU) and class average accuracy (ClassAvg) to evaluate performance on different datasets. The results reported in section Ⅳ-B are evaluated using the validation datasets of Camvid. All other experiments use the testing datasets of two benchmarks.

B. Ablation Study

1) Learning of two hyper-parameters: Experimental results of different dilation rate (d) and reduction ratio (r) are summarized in Table Ⅰ and Table Ⅱ. The dilation value determines the size of receptive fields. Table Ⅰ shows the results of using three different dilation rates. The best result is obtained when the dilation rate is 4. The reduction rate is directly related to the number of channels. In Table Ⅱ, the best accuracy of 90.21% is achieved when setting the reduction ratio to 16 (with fixed dilation rate of 4). Based on the above results, in the following experiments, we set dilation rate to 4 and the reduction ratio to 16.

2) Feature Enhancement Module: Table Ⅲ shows the results of using different number of FEM modules. In Table Ⅲ, 1FEM means we use only one FEM module to combine Res4 and Res5 features. Likewise, 2FEM indicates that two FEM modules are employed to combine Res3 and Res4 features, Res4 and Res5 features, respectively. In the third experiment, we add another FEM module between Res2 and Res3 features (3FEM). In these experiments, the spatial network and the context network are removed and FEM features are upsampled as the network output. Experimental results demonstrate that utilization of FEM to integrate multiple level features improves performance significantly. Compared to the baseline, using of three FEM modules increases Mean IoU from 60.47% to 66.53%.

TABLE I. ABLATION STUDY OF VARIOUS DILATION RATE (D), THE REDUCTION RATIO (R) IS SET TO 16.

Value	Parameters(M)	MeanIoU(%)	ClassAvg(%)
2	164.50	65.93	89.99
4	**164.50**	**66.53**	**90.21**
6	164.50	65.27	89.84

TABLE II. ABLATION STUDY OF DIFFERENT REDUCTION RATIO (R), THE DILATION RATE (D) IS SET TO 4.

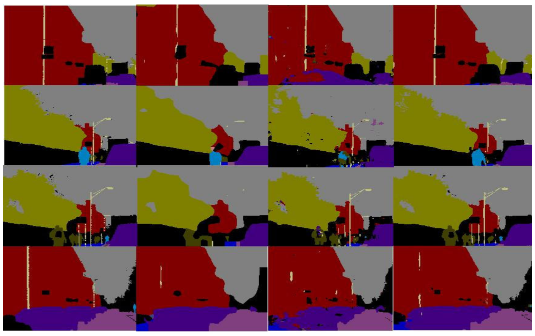
Value	Parameters(M)	MeanIoU(%)	ClassAvg(%)
8	183.49	64.96	88.72
16	**164.50**	**66.53**	**90.21**
32	155.01	66.31	90.04

TABLE III. ABLATION STUDY OF THE FEATURE ENHANCEMENT MODULE. NFEM: THE NUMBER OF FEATURE ENHANCEMENT MODULE USED IN MSS NET.

Method	Parameters(M)	MeanIoU(%)	ClassAvg(%)
1FEM	148.79	63.93	89.42
2FEM	163.53	65.76	90.05
3FEM	**164.50**	**66.53**	**90.21**
baseline	89.81	60.47	88.37

TABLE IV. ABLATION STUDY OF DIFFERENT STRUCTURE.

FEM	Att	PPM	CPAM	MeanIoU(%)
√				66.53
√	√			67.64
√	√	√		69.26
√	√	√	√	72.47



(a)Ground Truth (b)Baseline (c)PSPNet (d)MSS Net

Figure 9. Visualization results on CamVid test set.

3) Attention Module : In the spatial network, the attention module is appended to each convolution layer to compute feature weight. Results of experiments with or without attention modules are listed in the first two rows of Table Ⅳ. In the first experiment, attention modules are removed from the spatial network. Experimental results show that we obtain 66.53% Mean IoU without attention modules. The second experiment achieves 67.64% Mean IoU, in which output features of the FEM modules are enhanced by attention modules. These experiments demonstrate that the attention module selects important spatial information for better scene segmentation.

4) Multi-Scale Pyramid Pooling Module: The multi-scale pyramid pooling module is applied in the spatial network to integrate global semantic information with multi-scale spatial details. In the experiment without the PPM module, feature maps of the last convolution layer are combined with the FEM features as the output feature maps of spatial network. As shown in Table Ⅳ, the MSS Net has 67.64% Mean IoU without the PPM module whereas the segmentation performance increased to 69.26% by using the proposed multi-scale PPM module. The experiment proves that collecting multiple scales spatial context information is necessary for semantic segmentation applications. Please note that the context network is removed in all these experiments for better performance comparison.

5) Context Pooling Aggregation Module: The third and fourth rows of Table Ⅳ shows the ablation experiment with and without the context pooling aggregation module. The context pooling aggregation module is implemented for encoding different scales context information. Experimental results show that utilizing CPAM module increases performance from 69.26% to 72.47%. We obtain the highest performance of 73.25% when using ResNet-101 as the backbone model. As more spatial detail and multiple scale context clues are collected from different modules, we observe more detailed structures and accurate location of different size objects.

C. Experimental Results on CamVid

We compare performance of the proposed approach with state-of-the-arts on the CamVid dataset. The results are shown in Table Ⅴ. Fig. 9 presents some visual examples of the proposed method. We obtain 71.6% Mean IoU when using the ResNet-50 as backbone network. Our method outperforms most state-of-the-arts, e.g. we observe accuracy improvement about 2.5% and 6.9% over two recent segmentation methods, the PSPNet50 [5] and the DFANet A [20], respectively. When the ResNet-101 is used as the backbone model, The MSS Net achieves 72.4% Mean IoU. We note that the algorithm has better results on some classes with ResNet-50 backbone, such as tree and pedestrian, whereas ResNet-101 has better results on classes including building and sky. The main reason is that networks with fewer layers have more spatial information, thus they achieve better segmentation results on complex objects. Fig. 9 shows that most performance improvements come from accurate location of small objects as some small parts, such as traffic sign, pedestrian, fence, pole and bicyclist, are missed in experiments using the baseline model only. Experimental results demonstrate that the proposed multi-branch model captures more spatial detail and encodes multiple scale context clues for better semantic segmentation.

TABLE V. PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON CAMVID TEST DATASET.

Method	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Bicyclist	Mean IoU (%)
Segnet[2]	68.7	52.0	87.0	58.5	13.4	86.2	25.3	17.9	16.0	60.5	24.8	46.4
FCN8[1]	77.8	71.0	88.7	76.1	32.7	91.2	41.7	24.4	19.9	72.7	31.0	57.0
ENet[17]	74.7	77.8	95.1	82.4	51.0	95.1	67.2	51.7	35.4	86.7	34.1	51.3
Dilation8[3]	82.6	76.2	89.0	84.0	46.9	92.2	56.3	35.8	23.4	75.3	55.5	65.3
Dilation8 + FSO[21]	84.0	77.2	91.3	85.6	49.9	92.5	59.1	37.6	16.9	76.0	57.2	66.1
Tiramisu[22]	83.0	77.3	93.0	77.3	43.9	94.5	59.6	37.1	37.8	82.2	50.5	66.9
Gadde et al.[23]	83.0	77.3	93.0	77.3	43.9	94.5	59.6	37.1	37.8	82.2	50.5	66.9
Chandra et al.[24]	81.2	75.1	90.3	85.2	48.3	93.9	57.7	39.9	15.9	80.5	54.8	65.7
DFANet A[20]	n/a											64.7
MSSNet (ResNet50)	86.8	**83.3**	96.3	83.7	**73.6**	95.9	**71.2**	**86.1**	**59.1**	87.5	85.4	71.6
MSSNet(ResNet101)	**90.5**	82.9	**96.4**	**86.2**	72.7	**96.4**	68.5	85.8	53.3	**87.9**	**85.6**	72.4

TABLE VI. PERFORMANCE COMPARISON ON CITYSCAPES TEST SET.

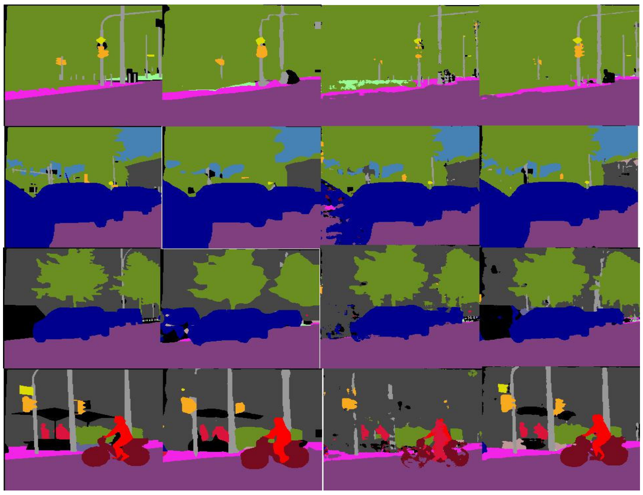
Methods	Backbone	Mean IoU	FPS	GPU
Segnet[2]	VGG16	60.1	14.6	TitanX
ENet[17]	-	58.3	76.9	TitanX
Deeplab-v2[4]	ResNet101	63.1	0.25	TitanX
RefineNet[6]	ResNet101	73.6	-	-
Contextnet[14]	-	66.1	65.5	TitanX
Fast-SCNN[16]	-	68.0	-	-
BiSeNet[13]	ResNet101	78.9	45.7	TitanXp
ICNet[15]	-	69.5	30.3	TitanX
PSPNet[5]	ResNet101	78.4	0.78	TitanX
PAN[11]	ResNet101	78.6	-	-
CANet[25]	ResNet101	78.6		
DABNet[26]	-	70.1	27.7	1080Ti
DenseASPP[10]	DenseNet161	80.6	-	-
SVCNet[27]	-	81.0	-	-
MSS Net	ResNet101	81.3	18.6	1080Ti

D. Experimental Results on Cityscapes

In this section, we conducted experiments to evaluate the effectiveness of our approach on the Cityscapes dataset. Performance comparison with some recent segmentation works on the Cityscapes dataset are shown in Table Ⅵ. Fig. 10 illustrates visualization results of our approach on the Cityscapes dataset. The proposed MSS Net achieves 81.3% Mean IoU on the Cityscapes test set. To improve segmentation accuracy, some segmentation works use larger input images [13] or extra training data [5,7]. Different from these methods, our network uses finely annotated images as training data. The performance improvement is due to the multi-branch network structure, which utilizes different modules, including multi-scale pyramid pooling module, context pooling aggregation module and feature enhancement module, to learn representative multi-level features for semantic segmentation.

V. CONCLUSION

In this paper, we present a multi-branch network for semantic segmentation. The proposed segmentation approach consists of three branches. Specifically, multiple level features are integrated in the semantic network branch to extract high-level semantic information. The spatial details are enhanced in the spatial network branch by aggregating spatial features from different regions. The context information is collected in the context network to further improve segmentation accuracy. The feature enhancement module and the context pooling aggregation module are also introduced to enhance network representative capability and embed multiple level context information. Experimental results on the CamVid and Cityscapes datasets demonstrate superior semantic segmentation performance of the new framework as compared with state-of-the-arts.



(a)Ground Truth (b)Baseline (c)PSPNet (d)MSS Net

Figure 10. Visualization results on Cityscapes test set.

REFERENCES

[1] Long, J., Shelhamer, E., and Darrell, T, "Fully convolutional networks for semantic segmentation," *In Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 3431-3440, 2015.

[2] Badrinarayanan, V., Kendall, A., and Cipolla, R, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE transactions on pattern analysis and machine intelligence, 39.12(2017), pp. 2481-2495.

[3] Yu, F., and Koltun, V, (2015), "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122.

[4] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L, "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), pp. 834-848.

[5] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J, "Pyramid scene parsing network," *In Proceedings of the IEEE conference on computer vision and pattern recognition* , pp. 2881-2890, 2017.

[6] Lin, G., Milan, A., Shen, C., and Reid, I, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," *In Proceedings of the IEEE conference on computer vision and pattern recognition* , pp. 1925-1934, 2017.

[7] Noh, H., Hong, S., and Han, B, "Learning deconvolution network for semantic segmentation," *In Proceedings of the IEEE international conference on computer vision* , pp. 1520-1528, 2015.

[8] Bilinski, P., and Prisacariu, V, "Dense decoder shortcut connections for single-pass semantic segmentation," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio*n, pp. 6596-6605, 2018.

[9] Lin, G., Shen, C., Van Den Hengel, A., and Reid, I, "Efficient piecewise training of deep structured models for semantic segmentation," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3194-3203, 2016.

[10] Yang, M., Yu, K., Zhang, C., Li, Z., and Yang, K, "Denseaspp for semantic segmentation in street scenes," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3684-3692, 2018.

[11] Li, H., Xiong, P., An, J., and Wang, L. (2018). "Pyramid attention network for semantic segmentation." arXiv preprint arXiv:1805.10180.

[12] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., and Lu, H, "Dual attention network for scene segmentation," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3146-3154, 2019.

[13] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., and Sang, N, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," *In*

Proceedings of the European Conference on Computer Vision, pp. 325-341, 2018.

[14] Poudel, R. P., Bonde, U., Liwicki, S., and Zach, C, (2018), “Contextnet: Exploring context and detail for semantic segmentation in real-time,” arXiv preprint arXiv:1805.04554.

[15] Zhao, H., Qi, X., Shen, X., Shi, J.,and Jia, J, “Icnet for real-time semantic segmentation on high-resolution images,” *In Proceedings of the European Conference on Computer Vision*, pp. 405-420, 2018.

[16] Poudel, R. P., Liwicki, S., and Cipolla, R, (2019), “Fast-SCNN: fast semantic segmentation network,” arXiv preprint arXiv:1902.04502.

[17] Paszke, A., Chaurasia, A., Kim, S., and Culurciello, E, (2016), “Enet: A deep neural network architecture for real-time semantic segmentation,” arXiv preprint arXiv:1606.02147.

[18] Brostow, G. J., Shotton, J., Fauqueur, J., and Cipolla, R, (2008, October), “Segmentation and recognition using structure from motion point clouds,” *In European conference on computer vision*, pp. 44-57, Springer, Berlin, Heidelberg.

[19] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... and Schiele, B, “The Cityscapes dataset for semantic urban scene understanding,” *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213-3223,2016.

[20] Li, H., Xiong, P., Fan, H., and Sun, J, “Dfanet: Deep feature aggregation for real-time semantic segmentation,” *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pp. 9522-9531, 2019.

[21] Kundu, A., Vineet, V., and Koltun, V, “Feature space optimization for semantic video segmentation,” *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3168-3175, 2016.

[22] Jégou, S., Drozdzal, M., Vazquez, D., Romero, A, and Bengio, Y, “The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation,” *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 11-19, 2017.

[23] Gadde, R., Jampani, V., and Gehler, P. V, “Semantic video cnns through representation warping,” *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 4453-4462, 2017.

[24] Chandra, S., Couprie, C., and Kokkinos, I, “Deep spatio-temporal random fields for efficient video segmentation,” *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8915-8924, 2018.

[25] Liu, M., and Yin, H, “Cross Attention Network for Semantic Segmentation,” *In 2019 IEEE International Conference on Image Processing* , pp. 2434-2438,(2019, September) ,IEEE.

[26] Li, G., Yun, I., Kim, J., and Kim, J, (2019), “DABNet: Depth-wise Asymmetric Bottleneck for Real-time Semantic Segmentation,” arXiv preprint arXiv:1907.11357.

[27] Ding, H., Jiang, X., Shuai, B., Liu, A. Q., and Wang, G , “Semantic correlation promoted shape-variant context for segmentation,” *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pp. 8885-8894, 2019.