Pyramid Attention Network for Semantic Segmentation

Hanchao Li[1]
lihanchao@bit.edu.cn

Pengfei Xiong[2]
xiongpengfei@megvii.com

Jie An[3]
jie.an@pku.edu.cn

Lingxue Wang[1]
neobull@bit.edu.cn

[1] School of Optics and Photonics
Beijing Institute Of Technology

[2] Megvii Inc. (Face++)

[3] Key Laboratory of Machine Perception
Peking University

Abstract

A *Pyramid Attention Network*(PAN) is proposed to exploit the impact of global contextual information in semantic segmentation. Different from most existing works, we combine attention mechanism and spatial pyramid to extract precise dense features for pixel labeling instead of complicated dilated convolution and artificially designed decoder networks. Specifically, we introduce a Feature Pyramid Attention module to perform spatial pyramid attention structure on high-level output and combining global pooling to learn a better feature representation, and a Global Attention Upsample module on each decoder layer to provide global context as a guidance of low-level features to select category localization details. The proposed approach achieves state-of-the-art performance on PASCAL VOC 2012 and Cityscapes benchmarks with a new record of mIoU accuracy 84.0% on PASCAL VOC 2012, while training without COCO dataset.

1 Introduction

With the recent development of convolutional neural networks (CNNs)[11], remarkable progress are proposed in pixel-wise semantic segmentation tasks due to rich hierarchical features and end-to-end trainable framework[19][18][1][33]. However, during encoding high dimension representations, original pixel-wise scene context suffers spatial resolution loss. As shown in Figure 1, our FCN baseline lacks ability to make prediction on small parts. The sheep beside the cow is made another wrong category on the second row. And on the first row the bicycle handle is missing. We take two main challenges as consideration.

The first issue is that the existence of objects at multiple scales cause difficulty in classification of categories. To solve this problem, PSPNet[33] or DeepLab system[2] performs spatial pyramid pooling at different gird scales or dilate rates(called Atrous Spatial Pyramid Pooling, or ASPP). In ASPP module dilated convolution is a kind of sparse calculation which may cause grid artifacts [28]. And the pyramid pooling module proposed in PSPNet may lose pixel-level localization information. Inspired by SENet[8] and Parsenet[18], we

(a) Image   (b) Ground Truth   (c) FCN(baseline)   (d) Proposed FPA   (e) Proposed PAN
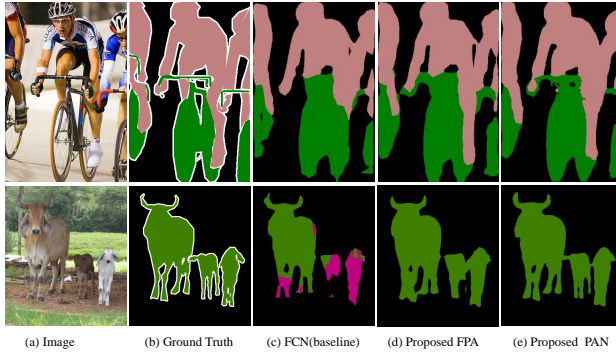
Figure 1: Visualization results on VOC dataset[5]. As we can see, FCN base model is difficult to make prediction on small parts of objects and details. On the first row the bicycle handle is missing and the animal is predicted to be another wrong category on the second row. Our Feature Pyramid Attention(FPA) module and Global Attention Upsample(GAU) module are designed to increase respective field and recover pixel localization details effectively.

attempt to extract precise pixel-level attention for high-level features extracted from CNNs. And Figure 1 shown that our proposed Feature Pyramid Attention (FPA) module is capable to increase respective field and classify small objects effecitvely.

Another issue is that high-level features are skilled in making category classification, while weak in restructuring original resolution binary prediction. Some kind of U-shape networks, such as SegNet[1], Refinenet[13], Tiramisu proposed in [10] and Large Kernel Matters[22] perform complicate decoder module which use low-level information to help high-level features recover images detail. However, they are time consuming. To solve this issue, we proposed a effective decoder module named Global Attention Upsample(GAU), which can extract global context of high-level features as guidance to weight low-level feature information without causing too much computation burden.

In summary, there are three main contributions in our paper. Firstly, We propose a Feature Pyramid Attention module to embed different scale context features in an FCN based pixel prediction framework. Then, We develop Global Attention Upsample, an effective decoder module for semantic segmentation. Lastly, Combined Feature Pyramid Attention module and Global Attention Upsample, our Pyramid Attention Network architecture archieves state-of-the-art accuracy on VOC2012 and cityscapes benchmark.

2 Related Work

In the following, we review recent developments in semantic segmentation tasks. Since models based on Fully Convolutional Networks(FCNs)[19] have achieved significant improvement on several segmentation tasks[19][1]. There is a lot of research focused on exploring the network structure to make better use of the contextual information[23][28][33][31][2].
Encoder-decoder: State-of-the-art segmentation frameworks are mostly based on encoder-decoder networks[23][1][29], which have also been successfully applied to many computer vision tasks, including human pose estimation[20], object detection[15][16], image stylization[26],

portrait matting[25][24], image Super-Resolution[32][27], and so-on. However, most methods attempt to combine the features of adjacent stages to enhance low-level features, without consideration of their diverse representation and the global context information.

Global Context Attention: Inspired from ParseNet[17], global branch is adopted in several methods[33][29] to utilize global scene context. Global context easily enlarge the receptive field and enhance the consistency of pixel-wise classification. DFN[29] embeds global average pooling branch in the top to extend the U-shape architecture to a V-shape architecture. EncNet[31] introduces an encoding layer with a SENet[8] like module to capture the encoded semantics and predict scaling factors that are conditional on these encoded semantics. All of them results in great performance in different benchmarks. In this paper, we apply global pooling operator as an accessory module adding to the decoder branches to selective the discriminative multi-resolution feature representations, which is proved to be effective.

Spatial Pyramid: This kind of models apply paralleled spatial pyramid pooling to exploit the multi-scale context information. Spatial pyramid pooling [7][30] has been widely employed to provide a good descriptor for overall scene interpretation, especially for various objects in multiple scales. Based on this, PSPNet[33] and Deeplab series[3][2] extend the global pooling module to the Spatial Pyramid Pooling and Atrous Spatial Pyramid Pooling, respectively. However, these models have shown high quality segmentation results on several benchmarks while usually need huge computing resources.

3 Method

In this section, we first introduce our proposed Feature Pyramid Attention(FPA) module and Global Attention Upsample(GAU) module. Then we describe our complete encoder-decoder network architecture, Pyramid Attention Network designed for semantic segmentation task.
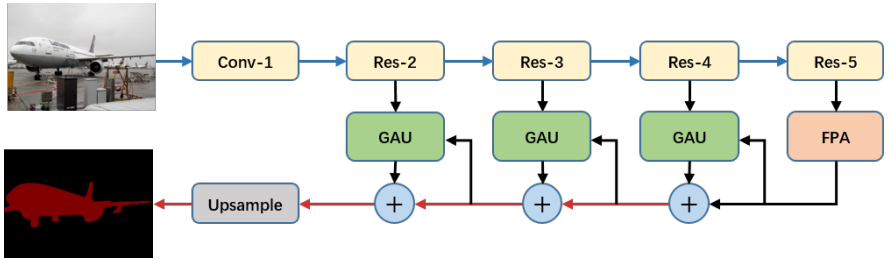

Figure 2: Overview of the Pyramid Attention Network. We use ResNet-101 to extract dense features. Then we perform FPA and GAU to extract precise pixel prediction and localization details. The blue and red lines represent the downsample and upsample operators respectively.

3.1 Feature Pyramid Attention

Recent models, such as PSPNet[33] or DeepLab[2], performed spatial pyramid pooling at several grid scales or apply ASPP module. Dilated convolution may result in local information missing and 'grids' which could be harmful for the local consistency of feature maps.

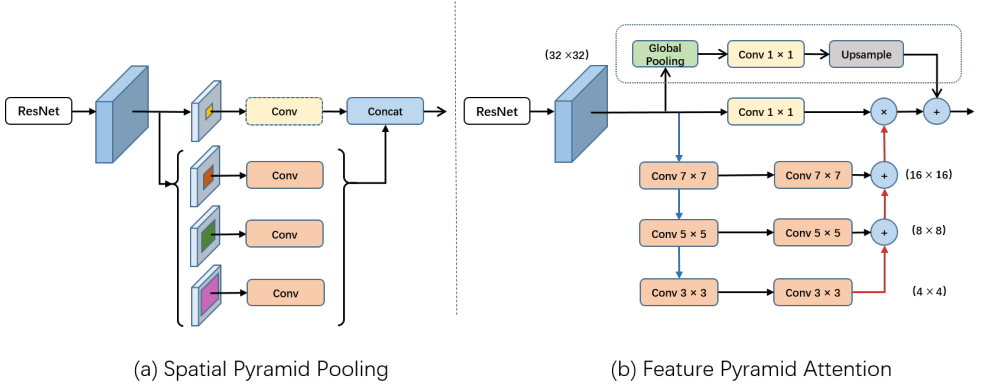Pyramid pooling module proposed in PSPNet loses pixel localization during different scale pooling operations.

Figure 3: Feature Pyramid Attention module structure. (a) Spatial Pyramid Pooling structure. (b) Feature Pyramid Attention module. '4×4, 8×8, 16×16, 32×32' means the resolution of feature map. The dotted box means the global pooling branch. The blue and red lines represent the downsample and upsample operators respectively.

Inspired by Attention Mechanism, we consider how to provide precise pixel-level attention for high-level features extracted from CNNs. In the current semantic segmentation architecture, the pyramid structure can extract different scale of feature information and increase respective field effectively in pixel-level, while this kind of structure lacks global context prior attention to select the features channel-wisely as in SENet[8] and EncNet[31]. On the other hand, using channel-wise attention vector is not enough to extract multi-scale features effectively and lack pixel-wise information.

With above observation, we propose Feature Pyramid Attention (FPA) module. The pyramid attention module fuses features from under three different pyramid scales by implementing a U-shape structure like Feature Pyramid Network. To better extract context from different pyramid scales, we use 3×3, 5×5, 7×7 convolution in pyramid structure respectively. Since the resolution of high-level feature maps is small, using large kernel size doesn't bring too much computation burden. Then the pyramid structure integrates information of different scales step-by-step, which can incorporate neighbor scales of context features more precisely. Then the origin features from CNNs is multiplied pixel-wisely by the pyramid attention features after passing through a 1×1 convolution. We also introduce global pooling branch concatenated with the output features, which improve our FPA module performance further. The final module structure is shown in Figure 3.

Benefit from spatial pyramid structure, Feature Pyramid Attention module can fuse different scale context information and produce better pixel-level attention for high-level feature maps in the meantime. Unlike PSPNet or ASPP concatenates different pyramid scale feature maps before channel reduce convolution layer, our context information is multiplied with original feature map pixel-wisely, which doesn't introduce too much computation.

3.2 Global Attention Upsample

There are several decoder architecture designs in the current semantic segmentation networks. PSPNet[33] or Deeplab[2] uses bilinearly upsample directly which can be seen as a naive decoder. DUC[28] uses large channel convolution combined with reshaping as one-step decoder module. Both the naive decoder and one-step decoder lack different scales of low-level feature map information and could be harmful to recover spatial localization to origin resolution. And the common encoder-decoder networks mainly consider using different scales of feature information and gradually recover sharp object boundaries in the decoder path. In addition, most methods of this type usually use complicate decoder blocks, which cost plenty of computation resource.

Recent research has shown that combining CNNs with well-designed pyramid module can obtain considerable performance and capability to obtain category information. We consider that the main character of decoder module is to repair category pixel localization. Furthermore, high-level features with abundant category information can be used to weight low-level information to select precise resolution details.
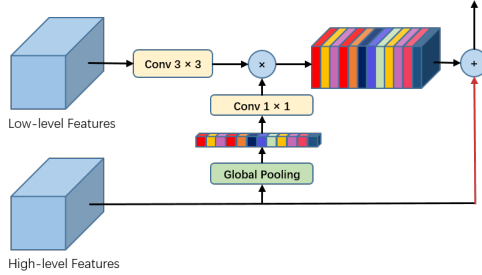

Figure 4: Global Attention Upsample module structure

Our Global Attention Upsample module performs global pooling to provide global context as a guidance of low-level features to select category localization details. In detail, we perform 3×3 convolution on the low-level features to reduce channels of feature maps from CNNs. The global context generated from high-level features is through a 1×1 convolution with batch normalization and nonlinearity, then multiplied by the low-level features. Finally high-level features are added with the weighted low-level features and upsampled gradually. This module deploys different scale feature maps more effectively and uses high-level features provide guidance information to low-level feature maps in a simple way.

3.3 Network Architecture

With proposed Feature Pyramid Attention(FPA) and Global Attention Upsample(GAU), we propose our Pyramid Attention Network(PAN) as Figure 2. We use ResNet-101 pretrained on ImageNet with the dilated convolution strategy to extract the feature map. In detail, the dilated convolution with rate of 2 is applied to *res5b* blocks, so the output size of feature maps from ResNet is 1/16 of the input image like DeepLabv3+. We also replace the 7x7 convolutional layer in the original ResNet-101 by three 3×3 convolutional layers like PSP-Net and DUC. We use the FPA module to gather dense pix-level attention information from

Method	mean IoU(%)	Pixel Acc.(%)
ResNet101	72.60	93.90
ResNet101+SE	75.74	94.48
ResNet101+C333+MAX	77.13	94.79
ResNet101+C333+AVE	77.54	94.89
ResNet101+C333+MAX+GP	77.29	94.81
ResNet101+C333+AVE+GP	77.61	94.88
ResNet101+C357+MAX	77.40	94.77
ResNet101+C357+AVE	78.19	95.00
ResNet101+C357+MAX+GP	77.03	94.71
ResNet101+C357+AVE+GP	78.37	95.03

Table 1: Detailed performance of Feature Pyramid Attention with different settings. 'SE' means using SENet attention module to replace the pyramid structure. For the pyramid structure used in Feature Pyramid Attention module, 'C333' represent all the kernel size of convolution is 3×3. 'C357' means the kernel size of convolution is 3×3, 5×5, 7×7 respectively, as Figure 3 shown. 'MAX' and 'AVE' represent max pooling and average pooling operations. 'GP' means the global pooling branch.

the output of ResNet. Combined with global context, the final logits are follow by GAU module to generate the final prediction maps.

We treat Feature Pyramid Attention module as center block between Encoder and Decoder structure. Without Global Attention Upsample module, Feature Pyramid Attention module can also provide enough precious pix-level prediction and class identification, as we show in Section 4 below. After implementing Feature Pyramid Attention module, we perform Global Attention Module as a fast and effective decoder structure, which use high-level features to guide low-level information and combine both precisely.

4   Experimental Results

We evaluate our approach on two main segmentation datasets: PASCAL VOC 2012 semantic segmentation[5] and urban scene dataset Cityscapes[4]. We first conduct a complete ablation study on VOC 2012 dataset, and finally report the state of art performances.

For a practical deep learning system, devil is always in the details. We use the "poly" learning rate policy where the initial rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$ with power 0.9 and initial rate $4e - 3$, and train the network using mini-batch stochastic gradient descent (SGD) with batch size 16, momentum 0.9 and weight decay 0.0001. The cross-entropy error at each pixel over the categories is applied as our loss function. We adopt randomly left-right flipping and random scaling between 0.5 and 2 for all datasets during training.

4.1   Ablation Experiments

The PASCAL VOC 2012 contains 20 foreground object classes and one background classes. The original dataset involves 1,464 images for training, 1,449 images for validation and 1,456 images for testing. The dataset is augmented by the Semantic Boundaries Dataset[6], resulting in 10,582 images for training. In this subsection, we use PASCAL VOC 2012

validation set for the evaluation and our crop size is 512×512. The performance is measured in terms of pixel intersection-over-union (IOU) averaged across the 21 classes with the single-scale input.

4.1.1 Feature Pyramid Attention

First, we experiment the performance of the base ResNet-101 with dilated convolution mentioned above and directly upsample the feature network's output. To evaluate our Feature Pyramid Attention(FPA) module, we perform FPA module after ResNet with direct upsampling as the same as the baseline. In detail, we conduct experiments with several setting, including pooling types of max and average, attention with pyramid structure or just one branch using global context similar to SENet, different kernel size in pyramid structure, with or without global pooling branch.

Ablation for pooling type: We notice that average pooling works better than max pooling in all settings. For using all the convolution with 3×3 kernel size, 'AVE' setting improves the performance from 77.13% to 77.54% compared to 'MAX' setting. So we adopt average pooling in our final module.

Ablation for pyramid structure: As shown in Table 1, our baseline model achieves mIoU of 72.60% on the validation set. Based on the observation in Section 3, we firstly implement the pyramid structure with 'C333' setting and average pooling , which improves the performance from 72.6% to 77.54%. And we also perform the SENet attention module to replace the pyramid structure to evaluate the performance compared with our pyramid attention module. As shown in Table 1, compared to SENet attention module, 'C333' and 'AVE' setting improves the performance by almost 1.8%.

Ablation for kernel size: For the pyramid structure using average pooling, we use large kernel convolution 'C357' to replace 3×3 kernel size, shown in Figure 3. As mentioned in Section 3.1, the resolutions of feature maps in the pyramid structure are 16×16, 8×8, 4×4 respectively, so using large kernel setting doesn't bring too much calculation burden. This improves performance from 77.54% to 78.19%.

Ablation for global pooling: We further add global pooling branch in the pyramid structure to improve performance. Finally, the best setting yields results 78.37/95.03 in terms of Mean IoU and Pixel Acc. (%). The results show that our pyramid attention can extract pix-level attention information effectively.

When using directly upsampling as naive decoder, our FPA module also shows notable progress compared to PSPNet and DeepLabv3 under the same output stride (*stride*=16). Since PSPNet didn't report performance on VOC validation set, we implement the pyramid pooling module proposed in PSPNet combined with our base model. And we use the same training protocol and keep the same output stride for fair comparison. The results are shown in Table 2. Our FPA module is more capable than the other two modules. Compared to the pyramid pooling module and DeepLabv3, our FPA module outperforms DeepLabv3 by almost 1.2% under the same output stride.

4.1.2 Global Attention Upsample

Since Feature Pyramid Attention module provides precise pix-level prediction, Global Attention Upsampling (GAU) focus on using low-level features to recover pixel localization. To be specific, we perform global pooling and 1×1 convolution used to generate global context as guidance. The 3×3 convolution used to reduce the channels of the low-level feature

Method	mean IoU(%)	Pixel Acc.(%)
ResNet101+PSPNet* [33]	76.82	94.62
DeepLab V3[2] (*stride*=16)	77.21	–
ResNet101+FPA	78.37	95.03

Table 2: Comparison with other state of art methods. 'ResNet101+PSPNet*' means that we implement the pyramid pooling module on our base model(*stride*=16).

map from encoder module.

We first evaluate GAU module combined with ResNet101 baseline, then we experiment FPA and GAU together on VOC 2012 *val* set. As for the design of decoder module, we evaluate three different design respectively, (1) only using low-level features from skip-connection without global context attention branch, (2) using 1×1 convolution to reduce channels of low-level features in GAU module, (3) replace 1×1 convolution with 3×3 convolution to preform channel reduction. As shown in Table 3, without global context attention branch, our decoder module merely improves performance from 72.60% to 73.56%. Then we add global pooling operation to extract global context attention information, which improves performance from 73.56% to 77.84% significantly.

Method	GP	1×1 Conv	3×3 Conv	mean IoU(%)	Pixel Acc.(%)
ResNet101	—	—	—	72.60	93.90
ResNet101+GAU			√	73.56	94.15
ResNet101+GAU	√	√		77.48	94.89
ResNet101+GAU	√		√	77.84	94.96

Table 3: Detailed performance with different setting of decoder module.

As we take GAU as a slight and effective decoder module, we also compare with Global Convolution Network[22] and Discriminate Feature Network(DFN)[29]. The results are shown in Table 4. The structure of 'DFN(ResNet101+RRB)' is a ResNet101 combined with proposed Refinement Residual Block(RRB) as decoder module. Our decoder module outperforms RRB by 1.2%. It is worth noted that Global Convolution Network used extra COCO dataset combined with VOC dataset for training and obtained 77.50%, while our decoder module can achieve 77.84% without COCO dataset for training.

Method	mean IoU(%)	Pixel Acc.(%)
DFN(Res-101+RRB) [29]	76.65	–
Global Convolution Network [22]	77.50	–
ResNet101+GAU	77.84	94.96

Table 4: Comparison with other state of art decoder module.

4.2 PASCAL VOC 2012

Combined our best setting on Feature Pyramid Attention and Global Attention Upsampling module, we experiment the complete network architecture—Pyramid Attention Network

(PAN) on PASCAL VOC 2012 *test* set. In evaluation, we apply the multi-scale inputs (with scales= $\{0.5, 0.75, 1.0, 1.25, 1.5, 1.75\}$) and also left-right flipped the images in evaluation, as listed in Table 5. Since the PASCAL VOC 2012 dataset provides higher quality of annotation than the augmented datasets [9], we further fine-tune our model on PASCAL VOC 2012 *trainval* set for evaluation on the test set. More performance details are listed in Table 6. Finally our proposed approach achieve performance of 84.0% without MS-COCO [14] and Dense-CRF post-processing[2].

Method	MS	Flip	mean IoU(%)	Pixel Acc.(%)
PAN			79.38	95.25
PAN	√		80.77	95.65
PAN	√	√	81.19	95.75

Table 5: Performance on VOC 2012 *val* set.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean IoU(%)
FCN[ ]	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
DeepLabv2[ ]	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
CRF-RNN[ ]	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	78.2	60.4	80.5	77.8	83.1	80.6	59.5	82.8	47.8	78.3	67.1	72.0
DeconvNet[ ]	89.9	39.3	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	80.7	65.0	72.5
DPN[ ]	87.7	59.4	78.4	64.9	70.3	89.3	83.5	86.1	31.7	79.9	62.6	81.9	80.0	83.5	82.3	60.5	83.2	53.4	77.9	65.0	74.1
Piecewise[ ]	90.6	37.6	80.0	67.8	74.4	92.0	85.2	86.2	39.1	81.2	58.9	83.8	83.9	84.3	84.8	62.1	83.2	58.2	80.8	72.3	75.3
PSPNet[ ]	91.8	71.9	94.7	71.2	75.8	95.2	89.9	95.9	39.3	90.7	71.7	**90.5**	**94.5**	88.8	**89.6**	72.8	89.6	64.0	85.1	76.3	82.6
EncNet[ ]	94.1	69.2	**96.3**	**76.7**	**86.2**	96.3	90.7	94.2	38.8	90.7	**73.3**	90.0	92.5	88.8	87.9	68.7	92.6	59.0	86.4	73.4	82.9
PAN(ours)	**95.7**	**75.2**	94.0	73.8	79.6	**96.5**	**93.7**	94.1	**40.5**	**93.3**	72.4	89.1	94.1	**91.6**	89.5	**73.6**	**93.2**	62.8	**87.3**	**78.6**	84.0

Table 6: Per-class results on PASCAL VOC 2012 *test* set. PAN outperforms the state of art approaches and achieves 84.0% without pre-training on COCO dataset.

4.3 Cityscapes

The Cityscapes contains 30 classes, and 19 of them are considered for training and evaluation. The dataset contains 5,000 finely annotated images and 19,998 images with coarse annotation. In detail, the fine annotated images are split into training, validation and testing sets with 2,979, 500 and 1,525 images respectively. During training, we didn't use coarse annotation dataset. And our crop size of image is 768×768. We also used ResNet101 as base model like in Section 4.1. The performance on the test set are reported in Table 7.

Method	mean IoU(%)
DPN [18]	66.8
DeepLabv2 [3]	70.4
RefineNet [13]	73.6
DUC [28]	77.6
PSPNet [33]	78.4
PAN(ours)	78.6

Table 7: Performance on Cityscapes testing set without coarse annotation dataset.

5   Conclusion

We have proposed a notable Pyramid Attention Network for semantic segmentation. We designed Feature Pyramid Attention module and an effective decoder module Global Attention Upsample. Feature Pyramid Attention module provides pixel-level attention information and increases respective field by performing pyramid structure. Global Attention Upsample module exploits high-level feature map to guide low-level features recovering pixel localization. Our experimental results show that the proposed approach can achieve comparable performance with other state-of-art models on the PASCAL VOC 2012 semantic image segmentation benchmark.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[6] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *european conference on computer vision*, pages 346–361. Springer, 2014.

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017.

[9] Varun Jampani, Martin Kiefel, and Peter V Gehler. Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4452–4461, 2016.

[10] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1175–1183. IEEE, 2017.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[12] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[13] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.

[16] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. *arXiv preprint arXiv:1803.01534*, 2018.

[17] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[18] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 1377–1385. IEEE, 2015.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[20] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[21] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[22] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. *arXiv preprint arXiv:1703.02719*, 2017.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[24] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. In *Computer Graphics Forum*, volume 35, pages 93–102. Wiley Online Library, 2016.

[25] Xiaoyong Shen, Ruixing Wang, Hengshuang Zhao, and Jiaya Jia. Automatic real-time background cut for portrait videos. *arXiv preprint arXiv:1704.08812*, 2017.

[26] Abhinav Shrivastava, Rahul Sukthankar, Jitendra Malik, and Abhinav Gupta. Beyond skip connections: Top-down modulation for object detection. *arXiv preprint arXiv:1612.06851*, 2016.

[27] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4809–4817. IEEE, 2017.

[28] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. *arXiv preprint arXiv:1702.08502*, 2017.

[29] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. *arXiv preprint arXiv:1804.09337*, 2018.

[30] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[31] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. *arXiv preprint arXiv:1803.08904*, 2018.

[32] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. *arXiv preprint arXiv:1802.08797*, 2018.

[33] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[34] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.