DISCRIMINATIVE FEATURES RECONSTRUCTION NETWORK FOR SEMANTIC SEGMENTATION

Qiuhao Zhou *Yanbo Ma* *Haihua Lu* *Xuesong Chen* *Yong Zhao*

School of Electronic and Computer Engineering, Shenzhen Graduate School of Peking University
Shenzhen, China

ABSTRACT

Thanks to the development of convolutional neural networks (CNNs), researchers have proposed lots of effective semantic segmentation models. However, there are still two problems disturbing researchers, one of which is objects misidentification on the image level and another one is poor performance on details, especially the boundary of objects. To tackle these two problems, we propose Discriminative Features Reconstruction Network (DFR) containing two modules: Second-order Pyramid Features Reconstruction Module (SPFR) and Second-order Boundary Attention Module (SBA). Specifically, SPFR fuses different scales features to gain pyramid receptive field. Besides, SPFR extracts second-order statistics data to retrieve more discriminative features. Furthermore, we put forward SBA that is helpful to refine the segmentation results. On SBA, low-level features recover localization details under the high-level feature guidance. Our DFR achieves state-of-the-art performance on PASCAL VOC 2012 dataset with mIoU accuracy 81.1% without pre-training COCO dataset and post-processing.[1]

Index Terms— Semantic segmentation, second-order statics, attention mechanism, image processing, encoder-decoder network

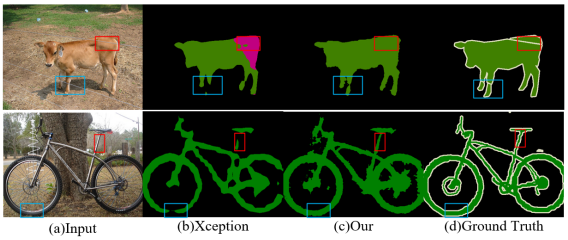

Fig. 1. Visualization results on Pascal VOC dataset 2012. Red rectangles show objects misidentification problem while blue ones point out poor boundary performance. On the first row of (b), because the size of receptive field is inappropriate, a pixel of cow with a red rectangle receptive field can be predicted to be hoarse, regarded as objects misidentification problem. We cannot distinguish it from hoarse or cow under receptive field of red rectangles. On the second row of (b), we can see that pixels on the boundary with different labels has heavily overlapping receptive field, which also make objects misidentification problem. Both first and second row of (b) show poor performance on boundary. Our Discriminative Features Reconstruction Network are designed to acquire multi-scale discriminative features and recover details automatically and effectively. Our results are as shown in (c).

1. INTRODUCTION

Semantic segmentation is one of the most important task of computer vision , which requires dense, pixel-accurate prediction. With the rapid development of deep convolutional neural networks (DCNNs)[1], researchers make a remarkable progress on most of computer vision problems including the semantic segmentation task[2][3][4]. However, these methods still meet two difficult problems. First, the patches, which share the same label, may be categorized to different labels. As shown in figure1(b), the network categorizes some patches of cow to hoarse. We regard this problem as objects misidentification problem. Second, the network cannot precisely outline the boundary of two adjacent patches that have different semantic labels. Deep convolution neural network cannot accurately outline the boundary of objects and lose detailed information, which is called as poor boundary performance.

To tackle these two problem, we rethink the semantic segmentation task. Semantic segmentation is a pixel wise classification task. First, different from single classification problem, semantic segmentation task faces a problem that the size of object varies greatly. It requires network predicts the category of each pixel in an appropriate receptive field automatically. Second, the label of a boundary pixel is different from the label of pixel on the other side but their corresponding receptive fields overlay heavily. Moreover, the excellent performance of CNN relies on image pooling and other down-sampling operations which inevitably cause the detail loss.

Our Discriminative Features Reconstruction Network (DFR) involves two components: Second-order Pyramid Features Reconstruction Module (SPFR) and Second-order Boundary Attention Module (SBA). SPFR aims to retrieve

[1]Our results can be seen at http://host.robots.ox.ac.uk:8080/anonymous/IKHF7D.html

discriminative multi-scale features while SBA is to refine the coarse result by providing boundary detail information with guidance of high-level features. In summary, there are three contributions in our paper:

First, we analyze the problems of semantic segmentation, finding that there are the objects misidentification problem and poor boundary performance influencing the semantic segmentation performance. Thus, we put forward DFR and validate it on Pascal VOC 2012 dataset[5].

Second, we design SPFR to get multi-scale information. Most importantly, it can acquire more discriminative and more representative multi-scale features.

Third, we propose SBA, which introduces attention mechanism and second order statics to refine the result. SBA get more precise and smooth prediction by utilizing the low-level features under high-level features guidance.

2. RELATED WORK

FCN[2] is the first effective semantic segmentation model based on CNN and many CNN approaches achieve excellent performance[3][6][7][4].

Multi-scale context: multi-scale context is necessary for multi-scale objects segmentation. ParseNet[8] just simply applies global pooling operation to attain global information. The PSPNet[6] and Deeplab[4] extent it to the multi-scale pooling or multi-rate atrous convolution and reach a new peak in the semantic segmentation task. However, these methods treat all branches equal, do not maintain details information and mining more representative statics data, which lower models ability to tackle objects misidentification problem and result in poor boundary performance.

Encoder-decoder Network: state-of-the-art segmentation frameworks shows encoder-decoder structure effectiveness. It is useful to integrate different levels features to recovers the reduced spatial information caused by down-sampling operation. For example, SegNet[3] recover information by recording the pool indices while u-net[9] directly concatenates the high-level features and low-level features. However, these methods just sum up the features of adjacent stages without considering that different levels features is not equal and responsible for different parts. Commonly, the improvement on result of these encoder-decoder structure is little.

Attention mechanism: Attention mechanism makes the network focus on what we want. Many recent works show the great ability of attention mechanism on many image tasks, including classification[10] and image segmentation[11].

3. METHOD

In this section, we first introduce our proposed Second-order Pyramid Features Reconstruction Module (SPFR) and Second-order Boundary Attention Module (SBA). We explain how these two modules deal with two important problems described above. Finally, we elaborate our complete Discriminative Features Reconstruction Network (DFR) architecture.

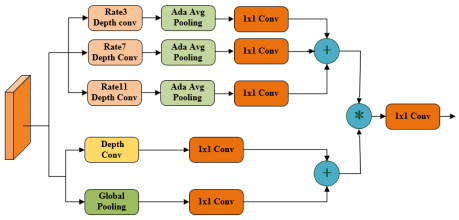

Fig. 2. Second-order Pyramid Features Reconstruction Module. Ada Avg Pooling means adaptive average pooling. + means element-wise adding operation. * means second-order features extractor.

3.1. Second-order Pyramid Features Reconstruction Module

Recently, many models, such as[12][13][6], apply the spatial pyramid pooling(SPP), ASPP module or Feature Pyramid Attention (FPA) to get multi-scale features. However, all these modules meet local information missing and grids problem, which will lower their ability to tackle objects misidentification problem and blur prediction. More importantly, all branches are equal. Pooling branches or multi-scale branches will hamper the models performance and influence the feature representation on detail. It causes poor boundary performance.

Our Second-order Pyramid Features Reconstruction Module consists multi-scale branch, local details branch and second-order features extractor as shown in figure2. First of all, we need multi-scale branch to encode the different scales information. It can acquire appropriate receptive field as shown on the first row of figure1(c). We introduce multi-scale features extraction. We perform atrous convolution on feature map. However, only some points are computed on atrous convolution, which influences the utilization rate of features[14] and model robustness. To deal with this problem, we apply adaptive pooling operation after atrous convolution. Our multi-scale branch can make full use of features and improve the robustness of network. Besides, our local details branch aims to maintain the local details features. As described above, the atrous convolution has smooth effect and will make the network lose details information. However, the local features without global information are also inaccurate. Under the premise of keeping local features details, we introduce global convolution to local details branch. We design second-order features extractor as follow:

$$x_{i,j} = F(i, j, f_A, f_B) = f_A(i, j)^T f_B(i, j)$$

$x_{i,j}$ representation second order feature on the location i, j while f_A and f_B are the feature maps from multi-scale branch

and local details branch. It is inevitably lose details information when we directly mix up features come from local details branch and multi-scale branch. Furthermore, the heavy overlaying of adjacent pixels' receptive field requires us acquire more discriminative features. It is known that higher rank statics contains more discriminative information[15]. So, our second-order features extractor can extract second-order features and make features more discriminative.

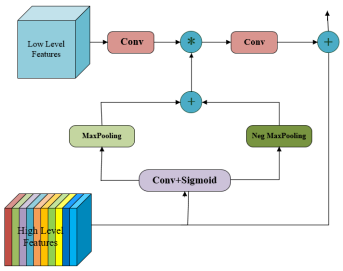

Fig. 3. Second-order Boundary Attention Module.

3.2. Second-order Boundary Attention Module

Decoder architecture is useful to recover the detailed information. FCN adopts bilinear upsampling gradually while PSPNet directly upsamples the result. Exfuse[16] and Deeplabv3+ prefer to concatenate different levels features and then make a prediction. Both naive decoder and multi-level feature fusion decoder neglect the fact that low-level features have more details information and treat them equal. Actually, the low-level and high-level features have diverse representation. Here, we detail our Second-order Boundary Attention Module(SBA). On the one hand, it is helpful to recover detail information by introducing low-level but high resolution features to high-level features. As we describe, the pooling operation and convolution operation with the stride of 2 cause detail loss. Moreover, the overlaying of receptive field will be alleviated when we process high resolution features map. On the other hand, high level features should be the guidance of low-level features to select where the localization details should be recover. As we know, high-level features are abundant with semantic information while low-level features have more details features but lack semantic information. Inspired by[17] and attention mechanism we design SBA, shown in figure3. First, we perform a convolution on coarse results and then we apply the boundary extractor to get boundary semantic features. The boundary extractor formula is shown below:

$$X = Maxpool(X) + Maxpool(-X)$$

Maxpool means max-pooling operation while X represents the feature maps. Through boundary extractor, the values of pixels not around the boundary are close to zero while others are not. After that, we apply second-order features extractor on lower features and boundary semantic features. From this decoder, we think the lower features just fine-tune the boundary under the guidance of high-level features. In a way, Second-order Boundary Attention Module takes the advantage of attention mechanism. Because of boundary extractor, second-order features of SBA is only none-zero on the boundary of pixels which makes the SBA pay attention on the boundary and modify the boundary result. However, the naive decoder and other decoder will add lower features on entire higher features, lowing the weight of high semantic features and adding too many detailed information in the interior of an object.

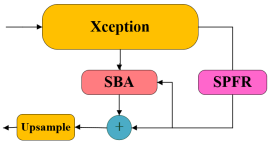

Fig. 4. Overview of the Discriminative Features Reconstruction Network. We use Xception to extract dense features and then we perform SPFR to acquire coarse pixel prediction. At last, we use SBA to recover localization details and extract precise result.

3.3. Complete Network Architecture

The Xception model[18] has promising image classification performance and Modified Aligned Xception used in deeplabv3+ show its strong potential for the task of semantic image segmentation. Therefore, we use modified aligned Xception as a backbone. With our designed SPFR and SBA, we propose our Discriminative Features Reconstruction Network as shown in figure4. Xception is used as a feature extractor. SPFR encodes the features and predicts a coarse result. We believe that the coarse result has good category prediction but lackluster performance on boundary. SBA, as a decoder module, refines the uncertain patches of the coarse result, which is common in boundary.

4. EXPERIMENTAL RESULTS

We evaluate our network on PASCAL VOC 2012 semantic segmentation dataset which contains 20 foreground object classes and one background class. The original dataset contains 1464 training images, 1449 validation images and 1456 testing results. We augment the dataset by the Semantic Boundaries Dataset[19], containing 10,582 images for training. We conduct a complete ablation study at first and finally report our excellent performance on PASCAL VOC 2012. The performance is mainly measured by 21 classes average pixel intersection-over-union (mIOU).

Our training protocol is the same as [13]. We employ learning rate schedule with poly policy and initial rate 7e-3. We train the network using mini-batch stochastic gradient

Method	IoU(%)	Pixel Accuracy(%)
Xception baseline	76.42	94.91
Xception with SPFR	79.45	95.50
Xception with ASPP	79.17	\
SPFR without AAP	79.00	95.40

Table 1. Detailed result of ablation study of second-order pyramid features reconstruction module. 'ASPP' means Atrous Spatial Pyramid Pooling. 'SPFR' means Second-order Pyramid Features Reconstruction Module. 'AAP' means adaptive average pooling.

Method	IoU(%)	Pixel Accuracy(%)
SPFR	79.45	95.50
SPFR with FCN decoder	79.59	95.45
SPFR with UNet decoder	80.25	95.60
SPFR with SBA	80.93	95.79

Table 2. Detailed result of ablation study of second-order boundary attention module.

descent (SGD) with batch size 20. Our image crop size is 512×512 and our loss function is the cross-entropy function averaged over each pixel. We adopt random scale data augmentation, random flipping and rotation during training.

4.1. Ablation study for SPFR Module

We first use the Xception as our base feature network and just directly upsample the output without using a decoder. Then, we test the essential of our proposed SPFR. Our SPFR improves the performance from 76.42% to 79.45% as table1 shows.

Ablation for second feature extractor: As shown in table 1, our base line has the performance of 76.42% mIoU on the validation set. First, we compare ASPP module on Xception result with our SPFR, which is the most important module of DeeplabV3+. The performance of ASPP is 79.17%. Then, we perform our SPFR and we get higher perfomance of 79.45% IoU. From these experiment, we show that our second feature extractor is useful.

Ablation for adaptive pooling: For the prymaid structure, we adopt atrous convolution with different atrous rates, which has low utilization rate of features. To prove our adaptive average pooling performance, we conduct an experiment without adaptive average pooling. As the results show in table, the performance without adaptive average pooling fall by 0.4%.

4.2. Ablation study for SBA Module

Since SPFR gets precise pixel-wise prediction on the stride of 16, SBA pays attention to recover the detail and fine-tune the result at a higher resolution. Specifically, we perform projection on coarse result computed by SPFR and then apply boundary extractor on it. We add lower features with the stride of 4 and coarse result. To prove effectiveness of our SBA, we evaluate our SBA network in the base of Xception with PSF. In detail, we first test the result of Xception with PSF as shown on the first row of table 1. Then, we add FCN decoder, which just add results of different resolutions. The FCN decoder just gets around 0.1% improvement. After that, we perform UNet decoder, which is similar to SBA without boundary extractor. At last, we test the SBA decoder. SBA decoder do not treat different levels features equally. It is shown that our boundary extractor is effective. The result tells us that low-level features can more accurately refine the coarse result with the guidance of high-level features. The total results are shown in table2. We can find that our SBA can improve the performance from 79.45% to 80.93%.

4.3. PASCAL VOC 2012

Combined our SPFR module and SBA module, we perform our Discriminative Features Reconstruction Network on PASCAL VOC 2012 test set. In evaluation, we further fine-tune our network on PASCAL VOC 2012 trainval set for evaluation on the test set. We compare our results with FCN[2], DeepLabv2, DeconvNet[20], DPN[21] and Piecewise[22]. Details results can be seen in the table3. As the results shown, our network outperforms lots of state-of-the-art models.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mobike	person	plant	sheep	sofa	train	tv	mean
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
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
Our	**94.5**	**66.1**	**83.9**	67.5	**81.5**	**92.7**	**89.7**	**95.0**	36.2	**88.2**	**66.4**	**90.3**	**90.2**	**90.7**	**88.4**	**67.9**	**89.9**	**62.0**	**89.2**	**77.7**	**81.1**

Table 3. IoU(%) results on PASCAL VOC 2012 test set.

5. CONCLUSION

We put forward a novelty discriminative features reconstruction network for semantic segmentation with remarkable performance. It contains two modules, second-order pyramid features reconstruction module and second-order boundary attention module. Second-order pyramid features module aims to acquire discriminative multi-scale features while second-order boundary attention module helps to recover pixel detailed information. Our experimental results show that our method has state-of-the-art performance on the semantic segmentation benchmark of PASCAL VOC 2012.

This work was supported by Science and Technology Planning Project of Shenzhen(No. NJYJ20170306091531561), Science and Technology Planning Project of Shenzhen (No. JCYJ2016050617265 1253), and National Science and Technology Support Plan, China(No. 2015BAKO1B04)

6. REFERENCES

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

[2] 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*, 2015, pp. 3431–3440.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*, 2018.

[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*, vol. 88, no. 2, pp. 303–338, 2010.

[6] 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)*, 2017, pp. 2881–2890.

[7] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation.," in *Cvpr*, 2017, vol. 1, p. 5.

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

[9] 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*. Springer, 2015, pp. 234–241.

[10] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, vol. 7, 2017.

[11] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[12] 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*. Springer, 2014, pp. 346–361.

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

[14] Chen-Wei Xie, Hong-Yu Zhou, and Jianxin Wu, "Vortex pooling: Improving context representation in semantic segmentation," *arXiv preprint arXiv:1804.06242*, 2018.

[15] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[16] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Dazhi Cheng, and Jian Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," *arXiv preprint arXiv:1804.03821*, 2018.

[17] Golnaz Ghiasi and Charless C Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 519–534.

[18] François Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint*, pp. 1610–02357, 2017.

[19] 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*, 2015, pp. 447–456.

[20] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation," in *IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[21] Ziwei Liu, Xiaoxiao Li, Ping Luo, and Chen Change Loy, "Semantic image segmentation via deep parsing network," in *IEEE International Conference on Computer Vision*, 2015, pp. 1377–1385.

[22] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid, "Efficient piecewise training of deep structured models for semantic segmentation," pp. 3194–3203, 2015.