Semantic Image Segmentation with Pyramid Atrous Convolution and Boundary-aware Loss

Dongkyu Yu
20172320
Department of Computer Science and Engineering, POSTECH
dkyu92@postech.ac.kr

Abstract

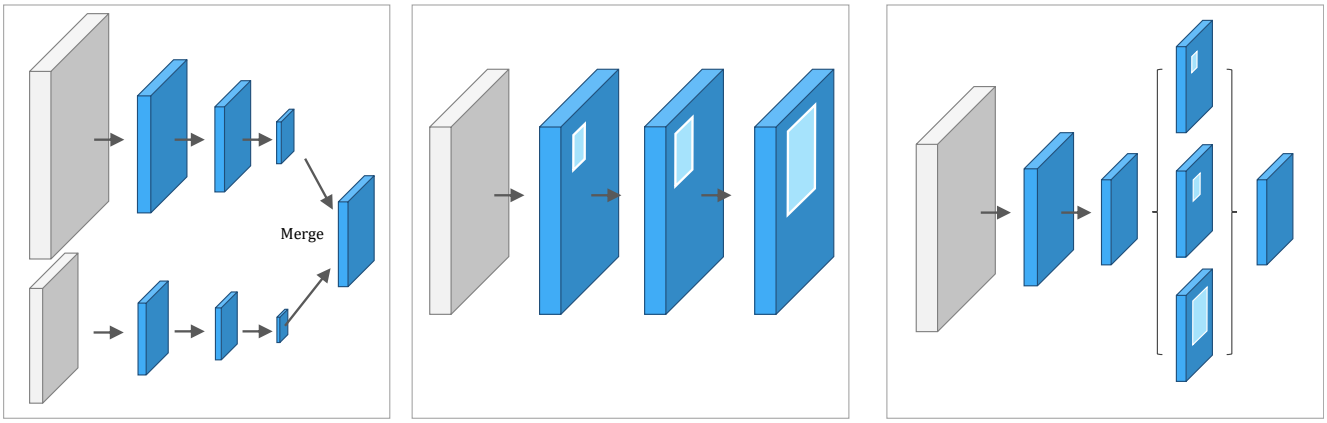
Semantic segmentation algorithm use features which have various receptive fields by multi-scale pooling or atrous convolution with various rates to encode multi-scale contextual information. These needs convolution filters for each scale or rate. Then each convolution filter only extracts features of specific scale, and it needs as many parameters as number of scales. Most of semantic segmentation algorithm adopt softmax loss which for classification do not reflect the characteristics of segmentation algorithm. In this work, we propose Pyramid Atrous Convolution (PAC) which consist of multiple atrous convolutions with different rates which share there parameters of filters, and Boundary-aware Loss (BAL), to network focus on boundary of objects in image which usually have large loss. We demonstrate the effectiveness of the proposed model on PASCAL VOC 2012, increasing the validation set performance of 1.24% then original DeepLabV3 [3] network from Google.

1. Introduction

In the task of semantic segmentation [6, 15, 4, 2], there are two mainstream network architectures. First, fully convolutional neural networks [14, 3, 19]. Fully Convolutional Networks (FCNs) [14] replace last fully connected layer of VGG network [18] to 1x1 convolution layer. FCNs upsample last feature maps to input image size by bilinear interpolation and training network by softmax cross entropy loss of upsampled feature maps and ground truth label. Second, Encoder-Decoder architecture [16, 17, 1], which consist of the encoder to encode an image into the features, and the decoder to decode encoded features into semantic segment image can learn how to reconstruct dense output by decoder layers. However Encoder-Decoder architectures are need so many parameters for deep encoder and decoder, and fully convolutional networks can not reconstruct dense prediction because of naive upsampling method and are not robust to various scales of objects compared to recent networks.

Figure 1: The images which are same class "Cat", but there are various size of objects in images.

Recent semantic image segmentation networks which consist of fully convolutional neural network [3, 19, 12] can effectively segment multi-scale objects from image like Fig. 1. Those networks exploits spatial pyramid pooling effect [9] by Pyramid Pooling Module (PPM) [19] which consist of pooling intermediate features in multi scales or use Atrous Spatial Pyramid Pooling (ASPP) [3] by atrous convolutions with various rates can capture large receptive field in multi-scale effectively. However those networks use multiple features for scale robustness. As a result, they needs a lot number of convolution filters for each scale, so number of parameters be increased and each filter only learn feature in specific scale. And the networks just use softmax cross entropy loss which for classification task did not reflect the characteristics of semantic segmentation algorithm.



(a) Image Pyramid (b) Deeper with Atrous Convolution (c) Spatial Pyramid Pooling

Figure 2: Several architectures to capture multi-scale contextual information.

In this work, we applying atrous convolution, which can efficiently capture large field-of-view with small number of parameters. We use atrous convolutions with multiple rates which share one convolution filter. So, number of parameters of the filter can drastically decreased and the filter can learn how to extract features from various scales. Furthermore, we propose Boundary-aware Loss (BAL) which simple yet effective loss function which makes network can effectively focus on hard region of semantic segmentation like complicated objects, objects boundaries, small objects, occluded or overlapped parts of objects.

2. Related Work

Model based on FCNs have demonstrated significant improvement on semantic segmentation benchmarks. DeepLab V3 [3] or PSPNet [19] which are model variants of FCNs have state-of-the-art performances with simple yet effective methods. Those use multi-scale contextual information aggressively for performance improvement by atrous convolutions with multiple rates or image pyramid. Atrous convolution enlarges the model's field-of-view to incorporate multi-scale context without increasing number of parameters. Image pyramid effectively capture multi-scale contexture information by simple resizing operations.

Image pyramid: Models which have multi-scale inputs can capture various scales of objects easily with simple methods. Features from small scale inputs capture long-range context with sparse abstracted information, and large scale inputs capture short-range context like small objects with dense detailed information. Ferabet at al. [7] transform input image through a Laplacian pyramid and feed it to convolution neural networks and merge all of features from all input scales.

Spatial pyramid pooling: Models such as DeepLab V3 [3] or PSPNet [19] perform spatial pyramid pooling at several scales by pooling features or apply atrous convolutions with several different rates. These models have shown promising results with simple modules like ASPP and PPM with various spatial contextual information.

Semantic segmentation loss: Almost every semantic segmentation methods [3, 19, 1, 14, 16, 15, 17] with fully convolutional networks use cross entropy loss for learning there networks. Which means, those treat semantic segmentation tasks as pixel-level classification. Zhu et al. [20] penalizing loss of top performers by additional generator and discriminator networks. Lin at al. [13] modify loss which focus on hard training examples by decrease factor of easy examples and increase factor of hard examples. But semantic segmentation tasks goal is predict class of each pixel in image which have spatial information. So, we need to modify semantic segmentation loss with spatial information.

In this work, we mainly explore atrous spatial pyramid module of [3] which can effectively capture short-range to long-range context without pooling features in small scales, and focal loss which have promising performance improvements in object detection task by focus on hard examples. Like ASPP module we propose Pyramid Atrous Convolution module which can effectively capture short-range to long-range than ASPP module with simple structures. Furthermore, like focal loss [13] in object detection, we propose Boundary-aware Loss which act like focal loss in semantic segmentation by focus on hard example of semantic segmentation tasks with a simple yet effective method.

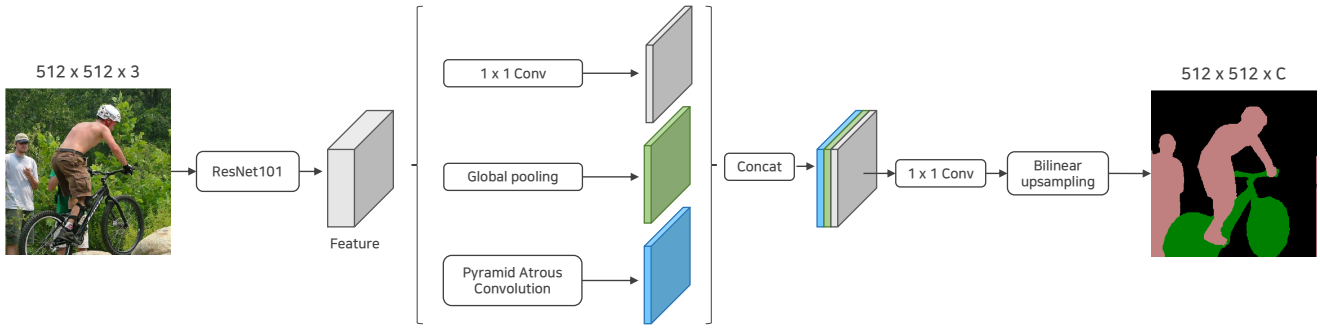


Figure 3: Overall architecture. In the end of network, concatenate 1x1 convolution feature, global average pooling feature and PAC feature and reduce dimension to number of classes by 1x1 convolution and bilinearly upsampling last features.

3. Methods

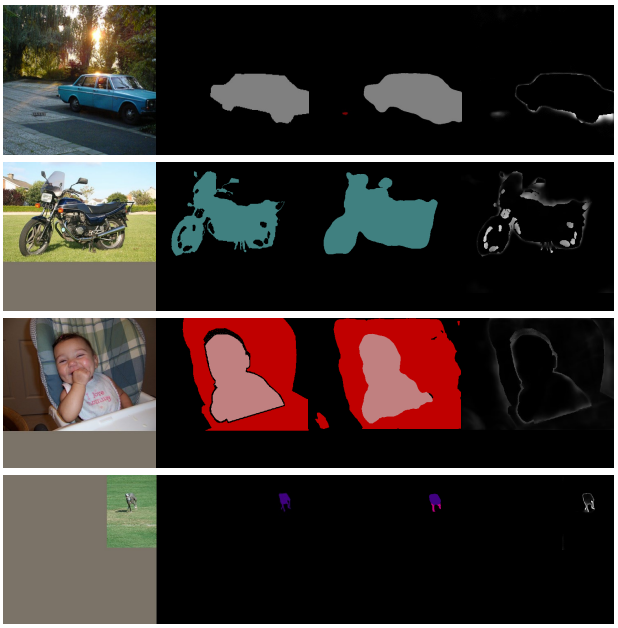
We start with our observations and analysis about loss image of semantic segmentation networks and existing modules for scale robustness like PPM and ASPP module. Loss image motivate us that how we can focus on spatial region where loss are significantly large. PPM and ASPP modules are simple and effective way to capture various scales of semantic information. However we think that we can further improve these modules more efficient.

3.1. Observations about loss

We observe and conduct analysis about loss image of semantic segmentation task. There shows significant loss at some specific cases (Fig. 4) like almost another semantic segmentation approaches [3, 19, 14, 16, 1, 17].

- Boundary of objects.
- Complicated objects.
- Occluded or overlapped objects.
- Small objects.

Because of lack of reconstruction power of fully convolutional networks, which just bilinearly upsample the feature to the desired spatial dimension, loss mostly appear at boundary of objects, as we can see at every row of Fig. 4. Second row of Fig. 4, most of detailed or thin parts of bike are overly smoothed or disappeared because deep convolutional neural networks (DCNNs) [11] decresing feature size to capture large semantic information and for computational efficiency. Third row, there are loss at the overlapped region of chair and baby. Last, most DCNNs hard to detect small objects as in last row. Output of DCNNs have very small feature maps which highly abstracted usually lack of dense information, so small objects detection are most challenging part of DCNNs.



(a) Image (b) Ground Truth (c) Prediction (d) Loss

Figure 4: Each row represents input RGB image, ground truth, prediction and loss respectively. The first row shows large loss at object boundary. The second row shows complicated object which structure have many thin components. The third row shows overlapped and occluded objects. The last shows small object. These cases have significant loss in prediction.

3.2. Boundary-aware Loss

Inspired by observations about loss, we propose simple yet effective loss function which called Boundary-aware Loss (BAL). In Fig. 4, most of loss are appear in boundary of objects. So, we extract edge of boundaries E_i by 2x2 filter f_E from semantic segmentation labels l_i for each class $\boldsymbol{i}$ (Eq. 1), then adopt gaussian blurring by gaussian filter f_G at boundary edge image and sum all of channels of results

E_G and add bias β (Eq. 3).

$$E_{i(x,y)} = \begin{cases} 0 & |(l_i \circledast f_E)_{(x,y)}| = 0 \\ 1 & |(l_i \circledast f_E)_{(x,y)}| > 0 \end{cases} \quad (1)$$

Where f_E:

$$f_E = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 0 \end{bmatrix} \quad (2)$$

$$\text{Gaussian edge } E_G = \sum\nolimits_i (E_i \circledast f_G) + \beta \quad (3)$$

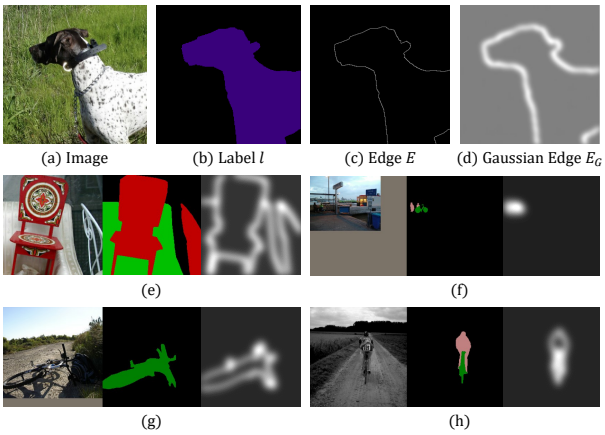


Figure 5: Input image, label l, boundary edge E and gaussian edge E_G. (e)-(h) are consist of image, label and gaussian edge respectevely. Gaussian edge E_G have high activation at corners and boundaries, complicated parts like bicycle handle in (g), overlapped or occluded region in (e) and (h), and small objects in (f).

In Fig. 5 gaussian edge E_G effectively focus on not only corners and boundaries of objects, but also complicated parts of objects, overlapped or occluded regions between objects, and small objects. So, we multiply gaussian edge E_G to original cross entropy loss $\boldsymbol{L}$, between prediction and ground truth, to amplify loss around boundaries and suppress loss of inner regions of each classes and we call it BAL. Where n is number of pixels in label l.

$$BAL = \frac{1}{n} \sum\nolimits_{(x,y)} E_{G(x,y)} \times L_{(x,y)} \quad (4)$$

This simple modification of cross entropy loss can deal with 4 problems we already mentioned, and networks focus effectively on hard regions of images in training process like Focal Loss [13] of object detection tasks.

3.3. Pyramid Atrous Convolution

With analysis about PPM and ASPP module, we introduce the Pyramid Atrous Convolution (PAC) module, which can effectively extracts features of various scales of objects with one convolution filter. The PAC consist of simple operations in Fig. 6, it consist of several atrous convolutions with multiple rates which use just one filter, and sum all of them to make output feature.

ASPP module consist of several atrous convolutions with different rates like PAC, but ASPP needs convolution filter for each convolution, so it needs lot more parameters than PAC module and each filter just learn features about specific scales, but PAC filter can learn multi-scales feature simultaneously. Furthermore, ASPP concatenate all of results of atrous convolutions in channel-wise can be consume more memory resources. On the other hand, output feature of PAC have same size as one atrous convolution features which drastically reduce output feature size.

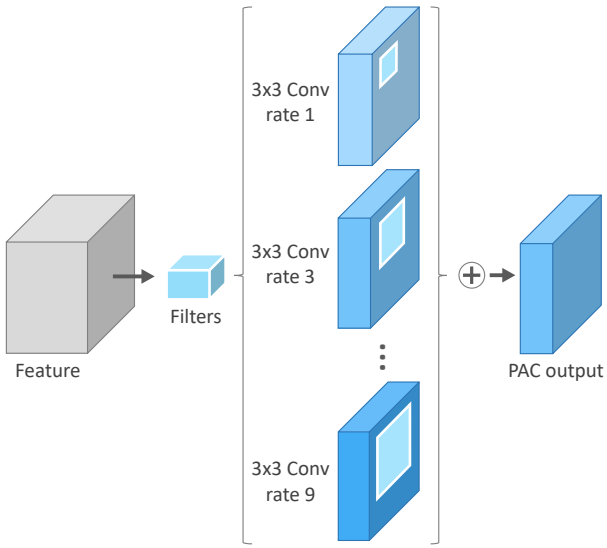


Figure 6: The Pyramid Atrous Convolution module, which consist of several atrous convolutions with different rates with same filter.

3.4. Overall Architecture

Our architecture Fig. 3 is consist of ResNet-101 [10] as backbone network, 1x1 convolution feature, global pooling feature, PAC feature and bilinear upsampling almost same as DeepLab V3 [3] architecture. 1x1 convolution feature efficiently encode the original 2048 channels feature to 256 channels feature. Global pooling feature effectively capture context about each channel of features by averaging each feature channel, and encoded by 1x1 convolution to 256 channels features. Furthermore, PAC module can extract feature, which robust to various objects scales in small size of features. These whole features are concatenated and encoded by 1x1 convolution which number of channel equal to classes to predict, and bilinealy upsample that feature for last prediction.

4. Experiments

We employ ImageNet-1k [5] pretrained ResNet-101 to extract dense feature maps by PAC module.

The proposed architecture is evaluated on the PASCAL VOC 2012 semantic segmentation benchmark [6] which contains 20 object classes and 1 background class. The original dataset contains 1,464 images for training, 1,449 images for validation and 1,456 images for test, which have pixel-level annotation. There are augmented dataset by the extra annotations provided by [8], which contains 10,582 training images. The performance is measured in pixel level intersection-over-union averaged across the 21 classes (mIoU). We follow the same training protocol as in [3]. We employ the same learning rate schedule, which use polynomial rate decay with power 0.9, and initial learning rate 0.00, crop image size 513 $\times$ 513, adjust image scales from 0.5 to 2.0 times with step size 0.25. Use batch normalization with batch normalization decay rate to 0.9997, weight decay rate of 0.0001, use momentum optimizer with moment of 0.9 and training network for 30k iterations.

4.1. Hyperparameters for BAL

We conduct several experiments about BAL to increase performance of networks in Table 1. In BAL, there are sevral hyperparameters have to modify. First, parameters for kernels. We use gaussian or linear kernels for BAL. Parameters of gaussian kernel f_G and linear kernel f_L have chosen by follow Eq. 5, 6. With kernel size n and center coordinate x_c, y_c.

$$f_{G(x,y)} = \exp\{\frac{-((x-x_c)^2 + (y-y_c)^2)}{n \times \sigma}\} \quad (5)$$

$$f_{L(x,y)} = n - (|x-x_c| + |y-y_c|) \quad (6)$$

Gaussian and linear kernels are normalized across 0 to 1. Second, we bilinearly downsampling original edge image E for increasing receptive fields of kernels and computational efficiency, and after applying convolution at edge image we bilinealy upsampling to original size. Third, β for BAL to control how much inner regions of classes will be contribute. We conduct experiments about kernels on DeepLabV3 [3] and we found gaussian kernels which size is 35, σ to 3.0, with β of E_G to 0.5 and down sample edge image by 4 times have best performance improvement.

4.2. PAC Design Choices

We experiment PAC module with various hyperparameters in Table 2. For number of atrous convolutions and there rates, number of filters and how to merge features.

We conduct several variations about number of atrous convolutions and there rates. At first, we just use same rates as in DeepLabV3 [3]. They use atrous convolutions with rates 6, 12, 18 which have best performance. So, we conduct experiment with rates 6, 12, 18 in PAC module but final performance is decreased. Then we conduct experiments with different combinations with atrous rates and number of atrous convolutions. We get best performance with 5 atrous convolutions with rates 1, 3, 5, 7, 9. We merge features of PAC with 1x1 convolution, sum all of features. For 1x1 convolution, first we concatenate all of features in channel-wise and squeeze by 1x1 convolution with desired number of channels. For summation, we just sum all of features from various atrous convolution features. Get better performance with sum all of features.

kernel	down	size	σ	β	ΔmIoU
linear	1	15		0.3	-0.35
linear	2	15		0.3	-0.12
linear	2	15		0.5	0.15
linear	4	15		0.5	0.32
linear	4	25		0.5	0.58
linear	4	35		0.5	0.68
gaussian	1	15	1.0	0.3	-0.82
gaussian	2	15	1.0	0.3	-0.43
gaussian	4	15	1.0	0.3	-0.14
gaussian	4	25	1.0	0.3	0.13
gaussian	4	35	1.0	0.3	0.24
gaussian	4	35	1.0	0.5	0.62
gaussian	4	35	2.0	0.5	0.81
gaussian	4	35	3.0	0.5	0.96

Table 1: Performance improvement comparisons about several parameters of BAL. **kernel** means which kernel to use, **down** means downsampling rate of edge image E, **size** means size of kernel, σ means σ of gaussian kernel and β means beta parameter of BAL.

# features	rates	concat	sum	ΔmIoU
3	6,12,18	✓		-0.12
4	6,12,18,24	✓		-0.18
3	6,12,18		✓	0.10
3	1,3,5	✓		0.14
5	1,3,5,7,9	✓		0.21
6	1,3,5,7,9,11	✓		0.11
5	1,3,5,7,9		✓	0.29

Table 2: Performance improvement comparions about several parameters of PAC. **# features** means number of convolution features to use, **rates** means atrous rates for each convolutions, **concat, sum** means merge convolution features by concatenation and 1x1 convolution or summation.

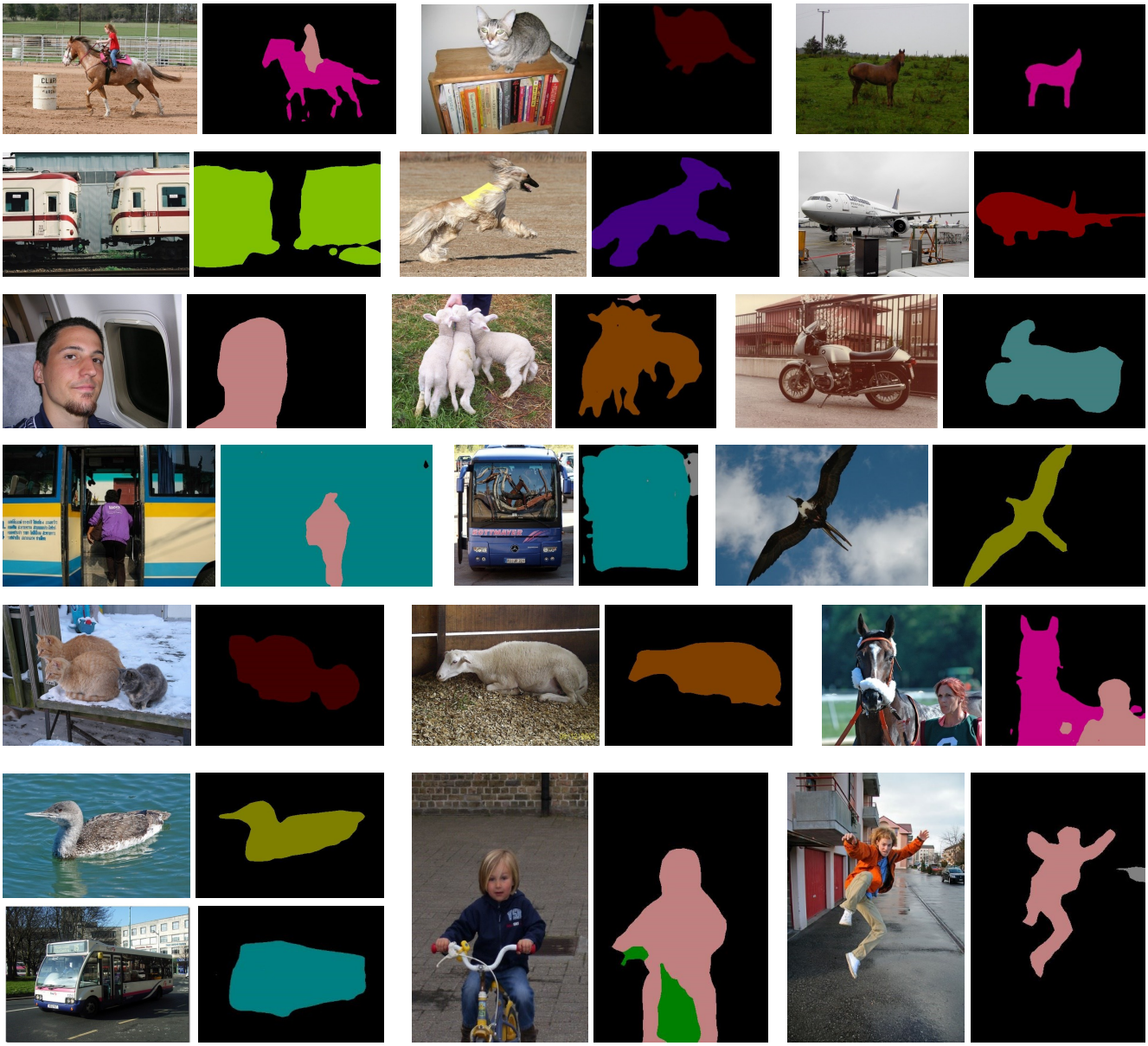


Figure 7: Visualization results on the val set when employing our best PAC + BAL model which trained on augmented dataset.

5. Conclusions

Our proposed PAC module and BAL significantly improves over previous DeepLab V3 network with simple yet effective methods. PAC module have performance improvements though it have simple structure with atrous convolutions with various rates and just sum all of them. BAL loss function is simple yet effective method which make networks effectively focus on hard region with simple edge extraction and gaussian edge image. Furthermore, BAL can increasing performance of general semantic segmentation networks without high cost.

Method	mIoU
DeepLab V3	76.68
ResNet + PAC	76.97
DeepLab V3 + BAL	77.64
ResNet + PAC + BAL	77.93

Table 3: Performance on PASCAL VOC 2012 val set. Results of DeepLab V3 is regenrated on 2 GPUs each have 8 images per batch per GPU.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[2] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. *CoRR, abs/1612.03716*, 5:8, 2016.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. 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] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[6] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[7] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013.

[8] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. 2011.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European conference on computer vision*, pages 346–361. Springer, 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[12] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang. Multi-scale context intertwining for semantic segmentation. In *Proc. Euro. Conf. on Computer Vision*, pages 603–619, 2018.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

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

[15] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

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

[17] O. Ronneberger, P. Fischer, and T. 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.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[19] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[20] X. Zhu, H. Zhou, C. Yang, J. Shi, and D. Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. *arXiv preprint arXiv:1809.00903*, 2018.