Enlarging Effective Receptive Field of Convolutional Neural Networks for Better Semantic Segmentation

Yifan Gu, Zuofeng Zhong, Shuai Wu and Yong Xu
Shenzhen Graduate School,Harbin Institute of Technology
Bio-Computing Research Center
Shenzhen, China
guyifan@stmail.hitsz.edu.cn,zfzhong2010@gmail.com,shuaiwu9@gmail.com,yongxu@ymail.com

***Abstract*—Recently, convolutional neural networks have shown powerful capability in different fields of computer vision, and have become the most effective means for dense prediction problems such as semantic segmentation. However, methods based on fully convolution network(FCN) are inherently limited to the size of the receptive field for each pixel, which leads to the bad performance of predicting object boundary. In this paper, we propose a novel deep neural network module, namely group dilated convolution(GDC), to effectively enlarge the receptive field, and a top-to-down pathway network is exploited simultaneously. The idea is that dilation convolution with different ratios can cover features of different scales, which shows a significant Mean IOU improvement in comparison with the baseline network.**

***Keywords*-semantic segmentation; receptive field; group dilated convolution; top-down pathway network;**

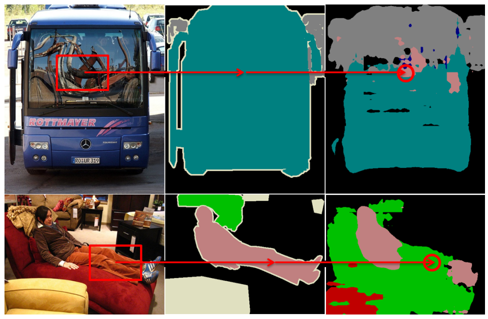

Figure 1. Illustration of the misclassification examples.Images in the left column are the original images, the center column shows the ground truth,and the right column shows the prediction result of the baseline model.

I. Introduction

Semantic segmentation is a fundamental research of computer vision. The main challenge of this task is to precisely recognize the irregular shape region for every appointed object. This is a pixel-level recognition task which is more difficult and meaningful than the object detection task. Semantic segmentation is widely applied on augmented reality, home-automation devices, and self-driving vehicles. Therefore developing preeminent segmentation algorithm is of great significance and strong demand.

In recent years, convolutional neural network (CNN) has achieved significant success on many computer vision tasks, such as image classification [1],face recognition [2], object detection [3], [4]. The neural network is able to learn effective features from repeated convolution and pooling operations based on the large-scale dataset. These features convert the original image into a discriminative subspace and show important property of simple transformation invariance including scaling, translation, and rotation. Owing to the powerful semantic information of feature maps in CNN, it is also widely used in semantic segmentation [5].

One factor that affects the performance of the semantic segmentation task is the narrow effective receptive field. Although the theoretical receptive filed in a common neural network is larger than the whole input image, Luo W et. *al* found that the effective receptive field only occupies a small fraction of the theoretical receptive filed [7], and a lot of researchers tend to learn more global context information to enlarge receptive field [8], [9]. The features do not contain enough context information, the final prediction results will be strongly interfered by the noise especially when objects of different categories are too close to each other. Fig. 1 shows the misclassification examples. It is clear that pixels in the center of the front window of the bus are misclassified. This is mainly because that features corresponding to the center pixel (the area illustrated by the red rectangle) do not show any information about the bus. If the effective receptive field of such features contains the whole bus, the prediction results may be correct. The images in the second row show a similar situation and people can recognize the thigh of the person in the image, but from the view of computer vision, the part is confused with the background. Therefore, to enlarge the effective receptive field is an effective way to improve the semantic segmentation performance.

Besides to add extra convolution layers to enlarge the receptive field [10], some other methods are also proposed to solve this problem. Liu et.*al* proposed to use global average feature of a layer to augment the features at each location [11], and Zhao et. *al* improved this idea and proposed pyramid pooling module to cover global context information by different-region-based context aggregation [12]. Dilated convolution [13] is also widely used in semantic segmentation to solve the problem. Chen et. *al* used dilated convolution to convert original classification network into a network with less stride operation [6], [14]. Yu et. *al* used it to aggregate multi-scale context information [8]. Wang et. *al* changed dilation rates of convolution layer [15]. In this paper, we also explore the

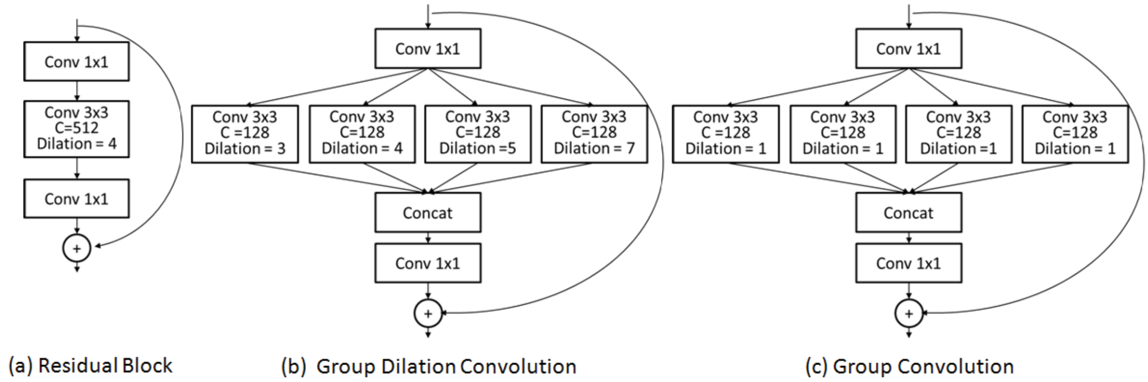

Figure 2. Illustration of the process to convert residual block to group dilation convolution with four branches. (a) shows the original residual block in the residual network [6] (b) is the group dilation convolution module which is convert from (a), and (c) shows the group dilation with similar parameters of (b).

usage of dilation convolution to enlarge the receptive field. The changes to dilation rate can effectively enlarge the kernel size without any extra computation, it thus offers an efficient mechanism to control the receptive field.

The semantic segmentation task requires to assign each pixel of the image with a special category label, and the final feature map requires to have the same size as the origin input image, which conflicts with the common design of neural networks. Low-level features of CNN contain more detail information whereas high-level features contain more semantic information, both of them are significant for the semantic segmentation task [10]. FCN [5]up-sampled the last layer directly to the origin image size and exploited the expanded feature map to perform classification.However, the coarseness of the final expanded feature map strongly limits the improvement of the final result. Thus, several kinds of neural network architectures have been proposed to solve this problem. For example,FCN [5] combined coarse-to-fine prediction probabilities from multiple layers by averaging segmentation probabilities. Hypercolumns [16] and HyperNet [17] combined features from multiple layers before making the prediction. Refinenet [10] used a cascade strategy to combine multi-scale features. SegNet [18], Decovnet [19], used the deconvolution operation to up-sample feature maps step by step to recall more local information at the cost of much extra computation and many parameters. But the local information is lost in the forward pass, the performance of recalling detailed information is not as good as expected. DeepLab [6], [14] methods proposed to use dilated convolution to reduce the times of down sampling and showed the promising efficiency. But the feature maps of such networks will be larger several times than the origin which causes extremely memory consuming, so that we can not directly put the networks in GPU memory without any sub-sampling operation. The authors made a trade-and-off between feature map size and memories. They down-sampled the networks twice in previous layers to fit the requirement of memory. On this condition, the size of the final feature map is small and but too coarse to predict object boundary and requires extra pose processing, such as conditional random field(CRF). Therefore,in this paper, we propose a top-to-down pathway network to up-sample the DeepLab network for refining the feature map. The final feature will contain both detailed information to classify the object boundary and multi-scale context information to predict label for every pixel.

II. Our Approach

A. Group Dilation Convolution (GDC)

In this section, we describe the formulation of our proposed framework in detail. We start by discussing how to enlarge the receptive field. As pointed out previously, dilation convolution is our main tactics. We explore how to use dilation convolution with different ratios to enlarge the receptive field, and propose the small size structure in Fig. 2, which is called group dilation convolution.

Our proposed module is shown in Fig. 2(b) and is inspired by the atrous Spatial Pyramid Pooling (ASPP) module of Deeplab-v2 [6]. It also applies the parallel dilated convolution with different ratios which are expected to cover different regions of multiple scales and enlarge the receptive field. We may apply the idea to the origin convolutions(Fig. 2(a) shows a residual block in [6]) in the residual neural network. However, when ASPP module is used, the number of parameters will increase linearly with the number of dilation rates. To control the numbers of parameters in the neural network and reuse the pre-trained model, we propose a novel structure which is shown in Fig. 2(b). Three convolution layers 1×1, 3×3, 1×1 are stacked. The 1×1 convolution layers are responsible for adjusting the dimensions, while 3×3 convolution is used to learn nonlinear features. We split the 3×3 convolution layer into k branches equally along the channel axis. In each branch, there is a special dilation rate to extract special features. The feature map size in each branch is equal and we can directly fuse them by simple concatenation. The total channel number of all branches is the same as the original residual block, and no extra parameters are needed. We can directly use the original pre-trained model to initialize the networks. The implement of such structure

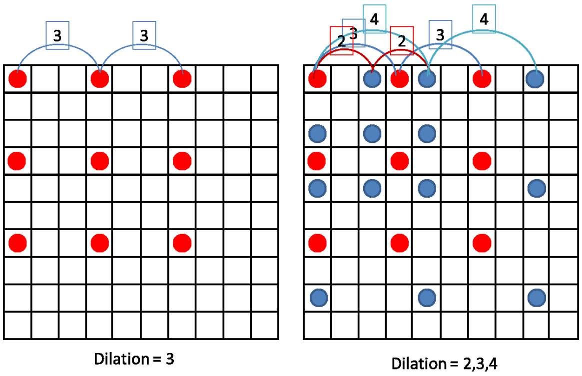

Figure 3. Illustration of activations of the top left kernels on the feature map. The left image shows activations of kernels with one dilation rate, and the right image shows the union of activations of kernels with different dilation rates.

produces a group convolution(Fig. 2(c)) with four groups if dilation rate becomes 1, which is called Group Dilation Convolution.

There is another advantage of this structure. If one dilation convolution has kernel size of $k \times k$, and dilation rate is r, the receptive field of which is equal to a normal convolution with kernel size $d = k + (k-1) \times (r-1)$. The connections to previous feature map of the dilated convolution are much sparer than the normal convolution, although they have the same receptive field in theory. When we use dilated convolution to replace the stride operation, it loses some detailed information in the process. In our structure, different dilation rates are corresponding to the different receptive fields, they connect activations from different positions in the feature map, and several activations which don't be connected in convolution with a special dilation rate are connected in our module. For example, supposing that we operate dilated convolution on a 10×10 feature map. If the original dilation rate of all channels is 2, activations with the first kernel in each channel is shown as the left image of Fig. 3. The 1×1 convolution also obtains information from 9 activations. When dilatation rates of different channels are changed, such as 2,3,4, the union activations of the corresponding first kernel in each channel are shown in the right image of Fig. 3. The 1×1 convolution can obtain information from more activations(22 activations), which means that the proposed method is ability to achieve more useful information.

B. Top-to-down pathway network

Enlarging the receptive field still has a limitation of predicting small objects to some extent, therefore, we explore multi-scale features for high-resolution prediction. Section 2.1 mainly focus on collecting more global context information to improve the prediction performance. In this section, we hope also to obtain more local information when the receptive field is expanded. Our design illustrated in Fig. 4 is a independent structure to the baseline network. In other words, it could be embedded in any base bone network.

The main structure (which is shown in the left column of Fig. 4) can be divided into 5 stages according to a original residual network. Only the first three stages need a down-sample operation and reduce the resolution of feature maps (conv1, res2a, res3a). The network increases the dilatation rate by 2 in stage-4 and stage-5 to fit the original receptive field in residual network. So the dilation rates in stage-4 and stage-5 are 2 and 4 respectively. We choose the last convolution layer in each stage to represent features of different scales. The structure of the top-to-down pathway network is described as follows. Step 1, from one stage, the receptive field is expanded and parameters are adjusted by the stack convolution model. Step 2, an adaptive convolution is applied to fit the feature map size of the previous stage. If the feature maps in two stages are different, bilinear interpolation is used to upsampling the feature map to fit the previous feature map size. Step 3, features from the two stages are fused by element adding operation in each location. Step 4, the previous three steps are repeated to obtain a refined feature map gradually. Step 5, a chain convolution module is used to matain learning features and to learn nonlinear features. The whole process can be trained end-to-end and step by step. In each stage, we add an extra convolution layer to generate prediction map and apply an extra supervisory loss to help training. The loss weight of different stages is set individual.

We also employ the multi-stages method to predict each stage result after a stacked convolution module with equal loss weights and then combine the results of different stages directly. Unfortunately, the result is not satisfied, the accuracy of each stage from top to down becomes worse and worse, The best result in the multi-stages method is worse than the method that only predicting result of the top stage. We confirm that information from different stages contains diverse contents. Some of them are conflicting and some of them are complementary.

Thus we choose to use the cascade method to add more useful information. Features of the top stage contain the most powerful information about the object category, based on which we add more localization information which is contained in previous stages to recognize the object boundary better. Stacked convolution and chain convolution modules give us another chances to achieve context information of larger region and aggregate multi-scale features.

Stacked Dilated Convolution (SDC). SDC is a single extension of the residual block in [20], but we removed the batch normalization layer. We choose to use two stacked convolutions with a 3×3 kernel because dilation convolution need to be applied when the kernel size is larger than 1. We use different dilation rates in each convolution to cover multi-scale features and the channel number of each stage is the same as the channel number of convolutions to the corresponding stage. This idea is the same as the ASPP module [6], but we do not use ASPP module directly. When using the SDC module, we can obtain information of each dilated convolution, and the

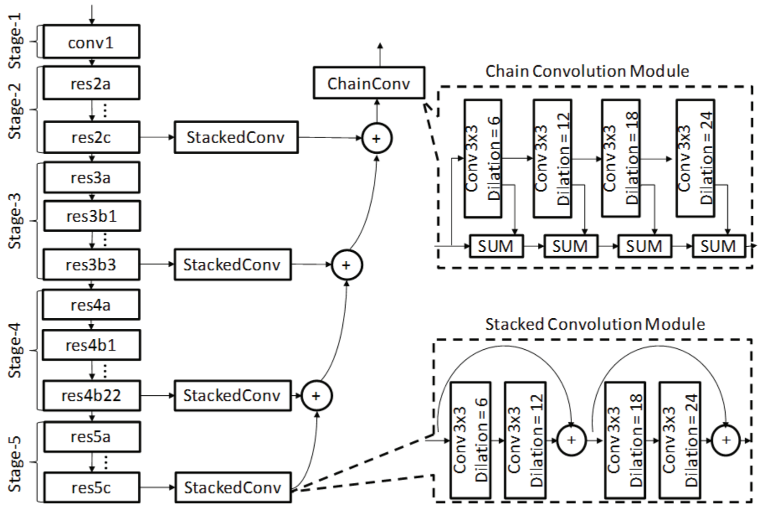

Figure 4. Illustration of our proposed network, which fuse features from top-to-down to obtain a high resolution prediction.

information when combining these dilated convolutions is also obtained. Compared with ASPP module which uses parallel dilated convolutions, we can obtain more information with similar structure. Experiments in Section 3 also prove it.

Chain Dilated Convolution (CDC). In the last part of our framework, we hope to matain the useful information we have learned and continue to learn new information when considering multi-scale information. This module is inspired by the success of skip connection. Skip connection is the core concept of the residual network which combines lower layer to higher layer directly and passes the gradient directly to earlier layer in the backward process. So the module passes original information through all processes when keep adding newly learned information by element adding. The different scale sizes contain not only the original size of convolution, but also the combination areas of two or three or more convolutions. This is another advantage of this module, which can be extended by convolution of any numbers. If we use k convolutions to construct the module, we call it the k-chains module. In Fig. 4, it shows a module with 4-chains. Each convolution in this part has the same channel. It will learn more nonlinear features from fusion feature map and generate the final prediction.

III. Experiment

We verify our method on the PASCAL VOC 2012 dataset [21] to show its effectiveness. We implement our framework using the public available Caffe Library [22] and build it based on the public available implementations of DeepLab [14] and DeepLab-v2 [6]. We also attempt to optimize the runtime memory according to the implement of GBDNet [23] and make some other simple tricks, which is beneficial to build deeper networks and train networks with larger batch size.

A. PASCAL VOC Dataset

The PASCAL VOC2012 segmentation benchmark contains 1464 training images, 1449 validation images, and 1456 test images. Using the extra annotations provided by [24], the training set is augmented to have 10582 images. The dataset has 20 object categories and 1 background category with pixel-level annotation.

B. Baseline Model

We use the DeepLab method as our baseline model by replacing VGG network [25] to the ResNet-101 network [20]. Specifically, the network has a down-sampling rate of 8, and dilation convolution with the rates of 2 and 4 are applied to res4b and res5b blocks. At the top of the network, there is a convolution layer with kernel size is 3, dilation rate 12 and the total channels of this layer equal to the number of dataset categories. The top convolution is used to generate final prediction map with a stride of 8 compared to the original image. Training labels are down-sampled by a factor of 8 to supervise the loss of each image. We apply the cross-entropy error at each pixel over the categories and then all pixel locations of the output map will be accumulated. We optimize this objective function using standard Stochastic Gradient Descent (SGD) in the training phase.

We train the network with image patch size 321×321 (randomly cropped from the original image), and randomly scale image of factors between 0.5 and 1.5 to augment the training data. The initial learning rate is set to 1e-3, and a poly learning rate with power=0.9 is applied (as in [6]). Weight decay and momentum are set to 0.0005 and 0.9 respectively. Finally, it achieves mean IOU of 71.72% on the validation set, which is similar to the paper [6].

C. GDC

In this section, we focus on verifying the effectiveness of group dilation convolution (GDC) structure. The only thing we change is the dilation rates of 3x3-kernel convolution in res5 layers. To be more specific, we split each convolution of 3x3 into 4 groups along the channel with different dilation rates. As pointed out before, this strategy could increase the effective receptive field and is more

Table I
EXPERIMENTS ABOUT DILATION RATES IN GDC MODULE

Model	DilationRates	Mean IoU	Pixel Acc.
Baseline	[4,4,4,4]	71.72	93.55
GDC-small	[3,4,4,5]	72.55	93.80
GDC-middle	[3,4,5,7]	73.14	93.93
GDC-large	[3,5,9,17]	**73.74**	**94.02**

Figure 5. Visualization results on PASCAL VOC 2012 validation dataset.

Table II
EXPERIMENTS ABOUT TOP-TO-DOWN PATHWAY NETWORK

Model	Mean IoU	Pixel Acc.
Baseline	73.74	94.02
Res4	75.86	94.53
Res2	**76.46**	**94.69**

effective than the original design, and original convolution in res5b could also be considered as four groups with the same dilation rates of 4 respectively.

As illustrated in Tab I, we evaluate the dilation rates with different values in res5 layers. It is clear that all of them could improve the segmentation result. We experiment with three kinds of dilation rates. the small one is the nearest to the original rates, and when we continue to enlarge the distance of the original dilation rates, the accuracy is increased.

D. Top-to-down pathway network

In this section, we use the GDC-large model as a start point to explore the effective of top-to-down pathway network. We train the network step by step, and in each step, we train a sub-network based on the previous step model. You also can train the whole network end-to-end, but there are a lot of parameters which will be initialized randomly. We find that we will obtain better network if we have a nice pre-trained model, and the process tells us features in each layer benefit to the semantic segmentation task. In each sub-network, we both have SDC and CDC module, the only difference is the layers we fusion. Results are shown in Tab II, Res-N means that we merge features from the top stage to the stage-N to predict the result.

Table III
ABLATION STUDIES ABOUT SDC AND CDC MODULES

Model	Dilation Rates	Mean IoU	Pixel Acc.
Baseline	[12]	71.72	93.55
ASPP	[6,12,18,24]	72.93	93.89
SDC	[6,12,18,24]	**73.34**	**93.97**
CDC	[6,12,18,24]	73.04	93.90

To evaluate the efficiency of stacked dilated convolution (SDC) and chain dilated convolution module (CDC), we embed our proposed module in the ASPP module in DeepLab-v2 and compared the results with the original ASPP module. We apply two stacked dilation models with dilation rates of 6,12,18,24, which will have similar parameters with ASPP module. The same as chained dilation convolution module, we use the 4-chained module with the same dilation rates. The comparison results are shown in Tab III. SDC module shows better result than other two modules, and SDC and CDC module both have a better result than ASPP module.

IV. CONCLUSION

We propose simple yet effective convolutional operations for improving semantic segmentation systems. A group dilated convolution module and a top-to-down pathway network are designed to effectively enlarge the receptive field and generate a refined prediction. This idea can also be extended to improve other models.

ACKNOWLEDGEMENT

This paper is partially supported by Guangdong Province high-level personnel of special support program (No. 2016TX03X164).

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] K. Guo, S. Wu, and Y. Xu, “Face recognition using both visible light image and near-infrared image and a deep network,” *CAAI Transactions on Intelligence Technology*, vol. 2, no. 1, pp. 39–47, 2017.

[3] Y. Li, K. He, J. Sun *et al.*, “R-fcn: Object detection via region-based fully convolutional networks,” in *Advances in Neural Information Processing Systems*, 2016, pp. 379–387.

[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.

[5] 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*, 2015, pp. 3431–3440.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *arXiv preprint arXiv:1606.00915*, 2016.

[7] W. Luo, Y. Li, R. Urtasun, and R. Zemel, “Understanding the effective receptive field in deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, 2016, pp. 4898–4906.

[8] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” *arXiv preprint arXiv:1511.07122*, 2015.

[9] F. Yu, V. Koltun, and T. Funkhouser, “Dilated residual networks,” in *Computer Vision and Pattern Recognition*, 2017.

[10] G. Lin, A. Milan, C. Shen, and I. Reid, “Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation,” *arXiv preprint arXiv:1611.06612*, 2016.

[11] W. Liu, A. Rabinovich, and A. C. Berg, “Parsenet: Looking wider to see better,” *arXiv preprint arXiv:1506.04579*, 2015.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” *arXiv preprint arXiv:1612.01105*, 2016.

[13] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, “A real-time algorithm for signal analysis with the help of the wavelet transform,” in *Wavelets*. Springer, 1990, pp. 286–297.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” *arXiv preprint arXiv:1412.7062*, 2014.

[15] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, “Understanding convolution for semantic segmentation,” *arXiv preprint arXiv:1702.08502*, 2017.

[16] B. Hariharan, P. Arbeláez, R. Girshick, and J. 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.

[17] T. Kong, A. Yao, Y. Chen, and F. Sun, “Hypernet: towards accurate region proposal generation and joint object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 845–853.

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

[19] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[20] 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*, 2016, pp. 770–778.

[21] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” *arXiv preprint arXiv:1408.5093*, 2014.

[23] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang *et al.*, “Crafting gbd-net for object detection,” *arXiv preprint arXiv:1610.02579*, 2016.

[24] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, “Semantic contours from inverse detectors,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 991–998.

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