# Bounding Box based Annotation Generation for Semantic Segmentation by Boundary Detection

Xiaolong Xu, Fanman Meng, Hongliang Li, Qingbo Wu, Yuwei Yang and Shuai Chen
*School of ICE, University of Electronic Science and Technology of China*, Chengdu, China

***Abstract*—This paper proposes a new method to generate pseudo-annotations from manual bounding boxes for semantic segmentation. Different from traditional local data driven based methods such as Conditional Random Field (CRF) and GrabCut, we aim at using class-agnostic bounding box based segmentation models. To this end, we propose a new segmentation network, which formulates segmentation task as a sparse boundary point detection task rather than dense pixel label prediction task, and therefore can provide new type of pseudo-annotations. Furthermore, we detect object boundary based on direction, and use multiple directions to handle various shapes of objects. Moreover, we further enhance the pseudo generation by combining different types of segmentation masks. Classical Fully Convolutional Networks (FCN) network based on dense prediction is also modified to generate diverse foreground masks. A simple fusion method based on intersection operation is proposed to combine the two types of pseudo-annotations. We verify the effectiveness of our method on PASCAL VOC 2012 validation dataset. The mIoU value is 67.9%, which outperforms the state-of-the-art method by 1.1%.**

***Index Terms*—Weakly Supervised Semantic Segmentation, Pseudo-annotation, Bounding Box, Boundary Prediction**

## I. Introduction

Semantic segmentation requires a large number of pixel-level annotations [1]. However, manually generating the annotations is an extremely time-consuming task. To this end, researchers replace pixel-level manual annotations with simple annotations such as windows, scribbles, and image-level labels [2], [3]. An efficient method is to generate pseudo-annotations from manual bounding boxes, and then uses the pseudo-annotations to train the segmentation model [4].

Good pseudo-annotation generation for bounding box based method relies on accurate transformation from window region to object region. By observing the fact that single segmentation method is hard to ensure the robust transformation, researchers propose fusion method that combines different types of pseudo-annotations generated by diverse methods to form more accurate pseudo-annotations. Experiments show that fusion method improves the quality of pseudo-annotation well [4]. However, the performance of fusion method depends on the diversity of initial pseudo-annotations.

Based on the fusion strategy, we try to build class-agnostic segmentation models to generate different types of pseudo-annotations. Two pseudo-annotation generation networks are proposed. One is boundary point detection network which formulates segmentation task as sparse boundary point detection task. Such method can not only provide new type of pseudo-annotation, but also reduce the prediction cost by three times. Meanwhile, the boundary point prediction is implemented in multiple directions, and their combination can ensure the robust segmentation of objects with diverse shapes. The other is classical FCN model which is dense prediction of pixel labels, and achieves segmentation by learning the mapping between windows and annotations. Since the two kinds of pseudo-annotation models formulate segmentation problem as sparse boundary point detection problem and dense pixel label prediction problem respectively, with different segmentation processes and segmentation results, their results are totally different, and their fusion can ensure high-quality generation of pseudo-annotations. Experiments demonstrate the effectiveness of the proposed method.

## II. The Proposed Method

### *A. Overview*

The proposed method consists of two steps: foreground mask generation and foreground mask combination. Two class-agnostic segmentation networks are firstly introduced. Then, the fusion methods for both multiple bounding boxes and multiple semantic segmentation annotations are proposed to form the final pseudo-annotations.

### *B. Foreground Mask Generation*

*1) Boundary Point Detection Network:* The pipeline and details of the proposed segmentation network are shown in Fig. 1, where the idea is to formulate the segmentation problem as the boundary point detection problem. Our method predicts the boundary points based on a certain direction, such as the horizontal direction. Moreover, in order to achieve sparse prediction for low computation burden, we uniformly sample $n$ lines in the direction and predict two edge points (such as left and right point for horizontal direction) on each line. Therefore, the segmentation problem is transformed into a prediction problem of $2 \times n$ points. By setting $n$ to be a small value, it is possible to achieve sparse and fast prediction. By connecting the boundary points, the object region is obtained.

Meanwhile, single direction cannot predict inner boundary points of concave shape. To overcome such shortcoming, we employ multiple directions, such as horizontal and vertical directions. The advantage of using multiple predictions is that the edge of failure in one direction will be compensated by other directions. Better boundary can be generated by combining multiple directions. Specifically, we use the prediction method to generate boundary points on each direction. Then, we combine the results of all directions to produce the final

978-1-7281-3038-5/19/$31.00 ©2019 IEEE

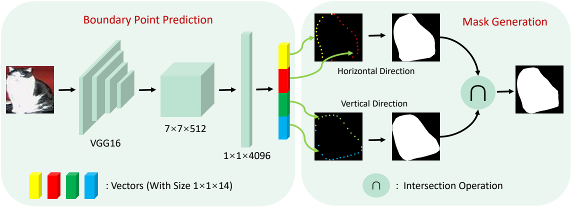
Boundary Point Prediction
VGG16
7×7×512
1×1×4096
: Vectors (With Size 1×1×14)
Mask Generation
Horizontal Direction
Vertical Direction
∩
∩ : Intersection Operation

Fig. 1. The pipeline of our proposed boundary point detection network.

TABLE I
THE MIOU VALUES ON PASCAL VOC 2012 VALIDATION DATASET. OUR BACKBONE NETWORK IS SET TO VGG-16 FOR FAIR COMPARISON.

| Methods | mIoU |
|---|---|
| BoxSup$_{Box}$(ICCV'2015) [6] | 52.3 |
| WSSL$_{CRF}$(ICCV'2015) [7] | 60.6 |
| GAIN (CVPR'2018) [8] | 55.3 |
| MCOF (CVPR'2018) [9] | 56.2 |
| AffinityNet(CVPR'2018) [10] | 58.4 |
| DSRG(CVPR'2018) [2] | 59.0 |
| MDC(CVPR'2018) [11] | 60.4 |
| FickleNet(CVPR'2019) [3] | 61.2 |
| SDI$_{M+G}$(CVPR'2017) [4] | 65.7 |
| BCM-FR$_{CRF}$(CVPR'2019) [12] | 66.8 |
| Ours | **67.9** |

object boundary. We simply use the intersection of regions to combine the regions of multiple directions. We use mean square error loss function to train the model.

*2) FCN Network:* FCN Network adopts the classical FCN in [5], with simple modification that the input is the cropped regions, and the output is binary mask. The binary cross entropy loss function is used to train the model.

### C. Forming semantic segmentation Annotation

For each network, we next form the semantic segmentation annotations based on the results of the bounding boxes. Our idea is to put the segmentation result into the image mask according to the position of the window. Note that some pixels may be segmented into different foregrounds (For example, when the "Cat" is on the "Sofa", the Cat's area will be assigned to the labels of "Sofa" and "Cat" simultaneously). For such case, we merge the results according to the size of the region and assign pixel with the label of small object. Experiments show that such method can ensure high-quality training.

### D. Combining the Annotations

Given two pseudo-annotations, we merge them by keeping their same labels, and filtering out the different labels (ignored when calculating losses). We use the pseudo-annotations generated above to train and establish the semantic segmentation model.

## III. EXPERIMENTAL RESULTS

We verify our method on PASCAL VOC 2012 validation dataset. For fair verification, we use all the images in MS COCO 2017 except the 20 classes in PASCAL VOC 2012 to train the two class-agnostic segmentation networks. Each window region is resized to 224 × 224. We set $n = 14$. Horizontal and vertical directions are used empirically. The performance is measured by the mean intersection-over-union (mIoU) value.

We compare our proposed method with several semantic segmentation methods which are based on bounding boxes or other weak annotations. The mIoU values are shown in Table I. It is seen that the mIoU value of the proposed method is 67.9%, which is 1.1% larger than the mIoU value 66.8% of the state-of-the-art method such as BCM-FR$_{CRF}$.

## IV. CONCLUSION

This paper proposes a method to generate pseudo-annotations from bounding boxes for weakly supervised semantic segmentation. Boundary Point Detection network and FCN network are firstly proposed to generate regions. Then simple fusion strategies are proposed to form pseudo-annotations. The experimental results on PASCAL VOC 2012 dataset demonstrate the effectiveness of the proposed method.

## ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 61871087, Grant 61502084, Grant 61831005, and Grant 61601102, and supported in part by Sichuan Science and Technology Program under Grant 2018JY0141.

## REFERENCES

[1] 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," *PAMI*, vol. 40, no. 4, pp. 834–848, 2017.

[2] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *CVPR*, 2018, pp. 7014–7023.

[3] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in *CVPR*, 2019, pp. 5267–5276.

[4] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *CVPR*, 2017, pp. 876–885.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *PAMI*, vol. 39, no. 4, pp. 640–651, 2014.

[6] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *ICCV*, 2015, pp. 1635–1643.

[7] G. Papandreou, L.-C. Chen, K. Murphy, and A. Yuille, "Weakly-and semi-supervised learning of a dcnn for semantic image segmentation (2015)," *arXiv preprint arXiv:1502.02734*.

[8] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *CVPR*, 2018, pp. 9215–9223.

[9] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *CVPR*, 2018, pp. 1354–1362.

[10] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *CVPR*, 2018, pp. 4981–4990.

[11] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *CVPR*, 2018, pp. 7268–7277.

[12] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation," in *CVPR*, 2019, pp. 3136–3145.

978-1-7281-3038-5/19/$31.00 ©2019 IEEE