Weakly Supervised Semantic Segmentation with Boundary Exploration

Liyi Chen, Weiwei Wu*, Chenchen Fu*, Xiao Han, and Yuntao Zhang

School of Computer Science and Engineering, Southeast University, China
{lychen, weiweiwu, 101012509, xiaohan}@seu.edu.cn, ytzhang01@foxmail.com

Abstract. Weakly supervised semantic segmentation with image-level labels has attracted a lot of attention recently because these labels are already available in most datasets. To obtain semantic segmentation under weak supervision, this paper presents a simple yet effective approach based on the idea of explicitly exploring object boundaries from training images to keep coincidence of segmentation and boundaries. Specifically, we synthesize boundary annotations by exploiting coarse localization maps obtained from CNN classifier, and use annotations to train the proposed network called BENet which further excavates more object boundaries to provide constraints for segmentation. Finally generated pseudo annotations of training images are used to supervise an off-the-shelf segmentation network. We evaluate the proposed method on PASCAL VOC 2012 benchmark and the final results achieve 65.7% and 66.6% mIoU scores on *val* and *test* sets respectively, which outperforms previous methods trained under image-level supervision.

Keywords: Weak supervision, semantic segmentation, deep learning.

1 Introduction

Deep learning and Convolutional Neural Networks (CNNs) have recently achieved great success in computer vision, such as semantic segmentation. Driven by the requirement of scene recognition, various models have been proposed [5, 34] to accurately segment foreground in images. However, manual annotation for training semantic segmentation network demands massive financial investments and is a time-consuming effort. To alleviate the heavy dependence on pixel-level annotations, weakly supervised learning for semantic segmentation is adopted, which uses weak annotations in semantic segmentation, including bounding boxes (*i.e.* information of instance location and dimension) [15], scribbles (*i.e.* a sparse set of pixels with a category label) [21] and image-level labels (*i.e.* information of which object classes are present/absent) [30, 13]. Among all the supervisions above, image-level label is widely used as it is available in most datasets (*e.g.* VOC and MS COCO).

* indicates co-corresponding authors.

Even though using image-level label is common and convenient in recent years, there exists a critical issue when classifying each pixel only with image-level class labels. The classification task requires translation *invariance* but the semantic segmentation task is position-sensitive and requires translation *variance*. To address this issue, *Class Activation Maps* (CAM) [35] is proposed to overcome the inherent gap between classification and segmentation by adding a global average pooling (GAP) in the top of fully convolutional network to get class localization maps. However, this architecture tends to activate most discriminative object regions and obtains incomplete segmentation results. In recent works [26, 13, 2, 30], CAM is usually taken as an initial localization technology followed by additional methods to refine it. However, most approaches focus on propagating foreground regions but do not consider the coincidence of segmentation and morphological boundary, which is as the main limitation for segmentation performance. In many cases, the segmentation might be stretched into irrelevant regions if its propagation is not properly constrained by object boundaries.

Although object boundary is an essential factor to influence the segmentation performance, it is difficult to excavate it without boundary annotation and prior knowledge. In this paper, we propose *boundary exploration based segmentation* (BES) approach, which explicitly explores object boundaries to refine semantic segmentation of training images. More concretely, we propose a simple scheme to obtain a small amount of boundary labels by filtering given localization maps, and then the network BENet trained by synthetic boundary labels is designed to explicitly explore more object boundaries. Finally, massive explored boundary information is used to provide constraints for localization maps propagation.

The main contributions of the paper are summarized as follows:

- The proposed BES approach can explicitly explore object boundaries with only image-level annotation.
- To get reliable initial localization maps, we introduce a module called attention-pooling to improve the performance of CAM.
- We demonstrate the effectiveness of BES by training DeepLab-v2 [5] with generated training image segmentations. BES achieves the state-of-the-art performance on PASCAL VOC 2012 benchmark [8] with 65.7% mIoU on *val* set and 66.6% mIoU on *test* set.

The remainder of this paper is organized as follows. We present related work in Section 2. The details of BES approach are described in Section 3. In Section 4, we investigate BES efficiency and compare it to the state-of-the-art approaches. Finally, Section 5 concludes this work.

2 Related Work

In this section, we firstly present the previous related researches in weakly semantic segmentation, and then describe the literatures of image-level supervision learning. Additionally, we summarize and analyse the common idea of image-level supervised semantic segmentation.

Weakly-supervised Semantic Segmentation. Weakly-supervised method for semantic segmentation attracts a large interest recently due to the simplicity and availability of its required labels compared to fully supervision segmentation learning. Various types of annotations are applied as supervision to address the data deficiency problem, including image-level label [1, 20, 31], bounding box [15, 23], scribble [21], and so on. In particular, image-level labels as a simplest supervision are popularly used since they demand minimum costs and can be obtained from most visual datasets.

Image-level Supervised Learning. Image-level class labels provide the multi-class classification information but no object localization cues. In early works, Graph-based models [33, 32] which consider superpixels similarity of images are adopted to do segmentation task. With the development of deep learning, the problem of semantic segmentation was transferred to the task of assigning class labels for each pixel in images so that neural network framework can be conveniently employed. The work in [25] utilizes multiple instance learning (MIL) to train the segmentation model. Papandreou et al. [23] employed the Expectation-Maximization algorithm to predict object pixels. Both of them were time-consuming and their performances were not satisfactory. Recently, some new approaches are proposed which make great progress in benchmark. In [31], an iterative learning method is adopted, the network is initially trained using simple images and corresponding saliency maps as labels, and the ability of segmentation is progressively enhanced by increasing complexity of train data. [17] proposed a method to train segmentation network with joint loss function (*i.e.* seeding, expansion and constrain-to-boundary).

Even if most state-of-the-art methods of weakly semantic segmentation have different implementation details, they share a similar strategy: *coarse-to-fine*. *coarse* here means to localize object in image and allow the existence of deviation from ground truth, these mismatch prediction will be refined in *fine* step.

Coarse-to-fine. Excavating localization cues with only classification annotation available is a challenging task. Early methods like saliency detection [27] had achieved huge progress. Most recently, researchers notice that CAM [35] is a good staring-point for segmentation from image-level annotation. The bridge of classification and localization is usually accomplished by adding a global average pooling (GAP) or re-designed calculation operation [17, 25] in the top of fully convolutional network. Object regions will be activated by network and these pixels contributed to classification score are regarded as foreground, Such localization technology are usually used in *coarse* step of weakly semantic segmentation.

However, CAM tends to focus on discriminative parts of object and other smooth-texture regions will not be activated or with a low response. To address this issue, additional processes to refine localization results are adopted which is called *fine* step. Incomplete object prediction and mismatch with object boundaries are two main limitations in the performance of *coarse*. In [13], coarse local-

ization maps are regarded as initial seed and then expanded, conditional random field (CRF) is adopted to keep prediction coincides with boundaries in expanding process. AffinityNet in [2, 1] is used to predict semantic affinities between pixels, where the pixel-wise affinity is further transformed into transition probability matrix to direct the propagation of CAM results. [30] uses initial CAM as erasing masks to force classification network discover more relative regions, and then fuse each part to get segmentation maps.

Although object boundary is absolutely essential for semantic segmentation, it is difficult to capture precise boundary information with only image class annotations. In [30, 31], authors did not take account of object boundary and just employ dCRF to make segmentation smooth. Some other approaches [1, 2] implicitly exploit boundary information by calculating affinities between pixels, which is not accurate enough and usually deviate from the ground truth. In this paper, we propose BES to explicitly predict object boundary for segmentation without extra annotations. And BES do not need additional information like more training samples [12] or salience prior [13, 27]. The most related work to ours is [1] which implicitly obtains boundaries by predicting affinity between pixels. However our BES can predict boundaries in a explicit way and achieves much more efficiency. Additionally, we provide experimental comparisons to above methods in Section 4.

3 The Proposed Approach

The BES framework for obtaining semantic segmentation for training images is illustrated in Figure 1. As the key component in the framework, *boundary exploration* targets to predict precise boundary maps with the original training images as input. Boundary labels are manually synthesized from localization maps and used to train BENet to predict boundaries. The boundary classification ability of BENet helps to explore massive boundaries which are then used for revising localization maps. The other component in framework called *attention-pooling CAM* is an improved CAM mechanism to obtain better initial object localization maps. The details of these two components are described in the following subsections.

3.1 Boundary Exploration for Localization Maps

In this subsection, we introduce the design of *boundary exploration* and how does it used to revise localization maps. We firstly exploit localization maps to synthesize a small amount of credible boundary labels, and then these are used to explore more boundary cues through training BENet.

Synthesizing Boundary Labels. As localization maps represented by pixel-wise class probability are inconvenience for boundary label synthesis, we firstly convert probability to a certain class label for each pixel. Suppose P_i^c is the

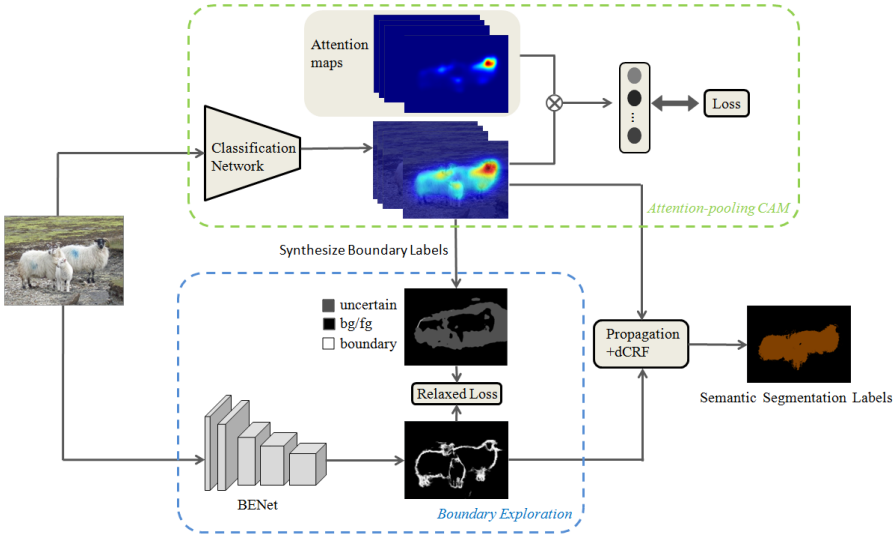


Fig. 1. The proposed BES framework. First, coarse localization maps are obtained through the attention-pooling CAM. Then, boundary labels are synthesized to train BENet which is able to excavate more object boundaries. Finally, semantic segmentation is generated by applying predicted boundary information to revise localization maps.

probability of pixel i belonging to class $c \in \mathcal{C}$. The classification result $\hat{y}_i$ for pixel i is obtained by a thresholding operation as following:

$$\hat{y}_i = \begin{cases} \arg\max_{c \in \mathcal{C}}(P_i^c) & \text{if } \max_{c \in \mathcal{C}}(P_i^c) > \theta_{fg} \\ 0 & \text{if } \max_{c \in \mathcal{C}}(P_i^c) < \theta_{bg} \\ 255 & \text{otherwise} \end{cases}, \tag{1}$$

where 0 and 255 mean background label and uncertain label, respectively, while θ_{fg} and θ_{bg} are thresholds for foreground and background respectively in order to filter uncertain pixels. An additional dCRF operation is used to refine reliable foreground and background regions.

We notice that though localization maps exist some misclassified pixels, it works well in regions near the boundaries. In other words, the classification results for pixels in the border of background and foreground are much reliable. According to the definition of boundary, we assume that there exist boundaries where adjacent regions contain approximate numbers of identical foreground and background pixels. Following this idea, we use a sliding window to count numbers of local identical pixels and employ statistics information to determine whether the pixel in the center of window is boundary or not. Given a sliding window centered at pixel i with size w, we use N_i^c to denote the number of pixels

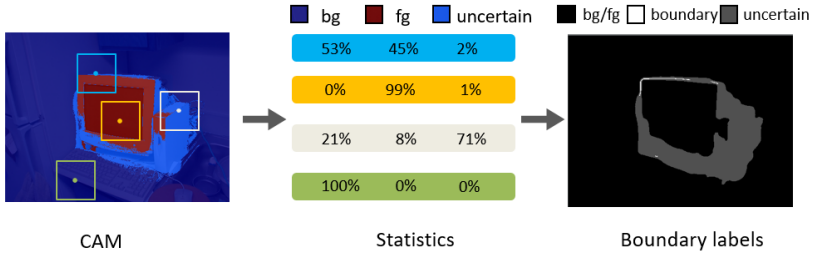


Fig. 2. Visualization of synthesizing boundary labels. A sliding window is used to capture local statistics information for each pixel, and then boundary labels are synthesized based on statistics.

assigned with label c in the window. The statistical proportion of the window for each class is denoted as S_i^c:

$$S_i^c = \frac{N_i^c}{w \times w} \quad \forall c \in \mathcal{C} \cup \{0\}. \tag{2}$$

The pixel will be regraded as boundary if it satisfies the following two conditions. First, There should exist sufficient identical pixels in the window to support inference. Second, the area size of foreground region and background region in the window need to be close enough. Formally, the boundary label $\hat{B}_i$ for pixel i is computed as follows:

$$\hat{B}_i = \begin{cases} 0 & \text{if } \min\{\max_{c\in\mathcal{C}} S_i^c, S_i^0\} > 2\ \theta_{scale} \\ & \quad \text{and } |\max_{c\in\mathcal{C}} S_i^c - S_i^0| \geq 2\ \theta_{diff} \\ 1 & \text{if } \min\{\max_{c\in\mathcal{C}} S_i^c, S_i^0\} > \ \theta_{scale} \\ & \quad \text{and } |\max_{c\in\mathcal{C}} S_i^c - S_i^0| < \ \theta_{diff} \\ 255 & \quad \text{otherwise} \end{cases}, \tag{3}$$

where θ_{scale} and θ_{diff} are thresholds for two conditions, respectively. 255 means that the pixel's boundary label is uncertain.

Pixels near boundaries are usually difficult to discriminate and may cause bad influence on BENet training. Therefore, the Equation (3) is designed as a discontinuous discriminant so that pixels near boundaries will be assigned with uncertain labels and not provide any supervision in the training phase. Figure 2 shows the process of synthesizing boundary labels. We calculate local statistics information for each pixel in the localization map and accordingly obtain boundary labels.

Relax Boundary Classification Loss for BENet Training. We employ BENet to directly predict a boundary map $\mathcal{B} \in [0,1]^{w\times h}$ for input image under the supervision of synthesized boundary labels. In training phase, the numbers of boundary, foreground and background pixels exist notable differences, and

the imbalance of training samples would make the model prediction tend to suppress boundary responses. To address this issue, we divide the training data into three parts: boundary pixels, foreground pixels and background pixels. The cross-entropy for each part is calculated respectively and aggregated into the final boundary loss function:

$$L_B = -\sum_{i\in\Phi_{bry}} \frac{W_i \log(P_i)}{|\Phi_{bry}|} - \frac{1}{2}\left(\sum_{i\in\Phi_c} \frac{\log(1-P_i)}{|\Phi_c|} + \sum_{i\in\Phi_{bg}} \frac{\log(1-P_i)}{|\Phi_{bg}|}\right), \quad (4)$$

where $\Phi_{bry} = \{i|\hat{B}_i = 1\}$, $\Phi_c = \{i|\hat{B}_i = 0, \max_{c\in\mathcal{C}} S_i^c > S_i^0\}$, $\Phi_{bg} = \{i|\hat{B}_i = 0, S_i^0 > \max_{c\in\mathcal{C}} S_i^c\}$ are pixel sets of boundary, foreground and background respectively, P_i is the boundary probability predicted by BENet for pixel i. To reduce the heavy dependence on synthetic boundary labels which may contain some misclassified samples, we relax the assumption that all boundary labels are reliable by adding a weight parameter W_i in the boundary loss term. In practice, W_i is set to $\sqrt{P_i}$ for online training. Consider a pixel labeled as boundary, i.e. $i \in \Phi_{bry}$, if its boundary probability estimated by BENet is a low value (which indicates that it is like background or foreground), its contribution weight for loss function will be decreased to reduce its influence for training.

Revising Localization Maps with Explored Boundaries. Trained BENet is able to excavate object massive boundaries. Then, we use predicted boundary maps as constraints to direct the propagation of coarse localization maps and get semantic segmentation. Compared to some flood-fill based propagation methods [13], random walk [3, 1] is an elegant method to revise localization maps. We adopt the methodology used in [1] to transfer boundary maps $\mathcal{B} \in [0,1]^{w\times h}$ to semantic affinity matrix. In the matrix, the affinity between pixel i and j is denoted as a_{ij}, which depends on the maximum boundary confidences in the path from i to j. To improve the computational efficiency, affinity will be treated positive only if distance of two pixels is less than a threshold γ:

$$a_{ij} = \begin{cases} (1 - \max_{k\in\Pi_{ij}} \mathcal{B}_k)^\beta & \text{if } P(i,j) < \gamma \\ 0 & \text{otherwise} \end{cases}, \quad (5)$$

where Π_{ij} is a set of pixels in the path from i to j, $P(i,j)$ is euclidean distance between i and j, and the hyper-parameter β is a value greater than 1 to control how conservative the random walk propagation is.

To simulate random walk process, element value a_{ij} in semantic affinity matrix is regraded as transition probability. Then propagation is accomplished by multiplying matrix to the localization maps, the simulation will be performed iteratively until predefined number of iterations is reached. Finally, dense Conditional Random Field (dCRF) [18] is applied to further slightly improves the quality of segmentation.

3.2 Attention-Pooling CAM to Obtain Localization Maps

Initial localization maps play an essential role in BES approach, which is not only used as seed region for revisiting, but also used to synthesize boundary labels. A common solution to obtain localization maps is to employ CAM technology which utilizes GAP after a fully convolutional network. However, as mentioned in [25, 7, 17], GAP used in CAM takes average value in the class localization map as classification score which implicitly assigns equal weights to each pixel vote. This will encourage response of irrelevant regions. Inspired by this motivation, we propose a new calculation called attention-pooling to replace GAP component in CAM.

Attention-Pooling for CAM. We preserve the fully convolutional network (FCN [22]) architecture used in CAM to capture object's spatial information, and change the calculation for obtaining classification scores by applying attention-pooling which can dynamically assign different weight for pixels. Let denote the localization map for class $c \in \mathcal{C}$ as M^c, which is embedded by convolutional layers, while M_i^c is the activation value at position i in M^c. The attention map A^c is generated for each class localization map M_i^c, the attention value in position i of A^c is denoted as A_i^c which represents the contribution weight of M_i^c to classification score s^c. s^c is the computed by summing up responses over all pixel positions, as follows:

$$s^c = \sum_i (M_i^c \times A_i^c) \quad \forall c \in \mathcal{C}. \tag{6}$$

The attention masks A^c is generated by utilizing the softmax function on relevant class localization map M^c:

$$A_i^c = \frac{\exp(kM_i^c)}{\sum\limits_{j \in \mathcal{J}} \exp(kM_j^c)} \quad \forall c \in \mathcal{C}, \tag{7}$$

where $\mathcal{J}$ is the set of pixels in M^c, and k is a hyper-parameter to adjust the intensity of attention. When k is set as zero, the attention-pooling will assign equal weight for every pixel. As the value k increases, the attention tends to focus on pixels with high response values. Generally, attention-pooling will be similar to global average pooling if k is close to zero, and be similar to global max pooling if k is high enough.

After training, the attention-pooling will be removed and localization maps are normalized to obtain class probability P_i^c:

$$P_i^c = \frac{M_i^c}{\max\limits_{j \in \mathcal{J}} M_j^c} \quad \forall c \in \mathcal{C}. \tag{8}$$

As described above, BES generates semantic segmentation for each training image. In Section 4, we demonstrate effectiveness of BES approach.

4 Experiments

In this section, we evaluate the efficacy of BES on the Pascal VOC 2012 semantic segmentation dataset [8] by training DeepLab-ASPP [5] with generated semantic segmentation of training images.

4.1 Implementation Details

Datasets. The benchmark of Pascal VOC 2012 has 21 different classes including background. The images in dataset are divided into three parts: 1464 images for training, 1449 images for validation and 1456 images for testing. Following common practice, we augment training part by adding training data from SBD [9]. In total, 10,582 images and relative image-level class labels are used for training. We use mean intersection-over-union (mIoU) as evaluation metric to measure semantic segmentation performance.

Attention-Pooling CAM Setting. In practice, we utilize ResNet50 as backbone network and replace the last fully connected layer with a 1x1 convolution layer. The stride in last stage is reduced from 2 to 1 so that more spatial information can be preserved. Therefore, resolution of output localization maps drop down to $\frac{1}{16}$ of input resolution after passing through the network.

Acceleration for Synthesizing Boundary Labels. Iteratively counting local numbers for every pixel would increase GPU memory usage and is very time-consuming. A simply way to compute statistics is to realize sliding window with the average pooling layer which has been implemented in common deep learning frameworks like PyTorch [24].

BENet Setting. BENet is based on ReNet50 [10] pretrained on ImageNet [6]. To fuse different level features of image, the features from shallow layers and deep layers are combined by concatenating resized feature maps from five stages with 1×1 convolution layers. In training phase, the parameters of network are optimized via batched stochastic gradient descent (SGD) for about 3,500 iterations, with a batch size 16. The learning rate is initially set to 0.1, and the momentum and weight decay are set to 0.9 and 0.0001 respectively. The training images are resized with a random ratio from (0.5, 1.5), and then augmented with horizontal flip after normalization, finally it is randomly cropped into size of 513×513.

Hyper Parameter Setting. Parameter k in Equation (7) is set to 0.001. θ_{fg} and θ_{bg} in Equation (1) are fixed to 0.30 and 0.07 respectively. To generate boundary labels, we set the size of search window w to 13, θ_{scale} to 0.35 and θ_{diff} to 0.10. For random walk parameters in revising stage, we use the setting given in [1] except setting the parameter β in Equation (5) to 5. For dCRF, we

Table 1. Ablation study on PASCAL VOC 2012 training set.

CAM	Attention-Pooling CAM	Boundary Exploration	dCRF	mIoU
✓				49.6
	✓			50.4
✓		✓		65.7
	✓	✓		66.4
	✓	✓	✓	**67.2**

follow the setting in [1] for CAM refinement, and θ_α is set to 20 for pseudo labels post-processing as shown in following equation:

$$k\left(f_i, f_j\right) = w_g \exp\left(-\frac{|p_i - p_j|}{2\theta_\alpha^2} - \frac{|I_i - I_j|}{2\theta_\beta^2}\right) + w_{rgb} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right). \quad (9)$$

4.2 Analysis of the BES

In this subsection, we firstly perform an ablation study to illustrate effective of each parts. Then, we evaluate the impact of various parameters values. Finally, we objectively analyse the quality of proposed BES approach.

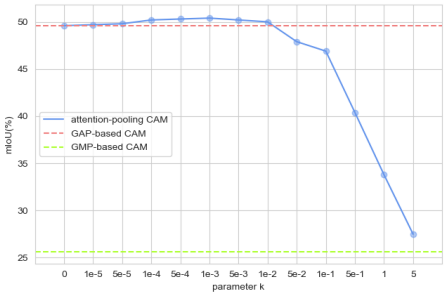
Ablation Experiment. To demonstrate the efficiency of the BES method, we report the influence of each step in Table 1 with mIoU as the evaluation metric. Comparing to CAM baseline which achieves 49.6%, the proposed attention-pooling CAM improves mIoU by 0.8% and this improvement is remained even after revising process. Applying boundary exploration, the localization maps performance is significantly improved from 50.4% to 66.4%. Additionally, post process dCRF brings extra 0.8% improvement. In experiments, we employ setting in last line which achieves 67.2% mIoU in VOC 2012 training set.

Hyper Parameter Effects. We evaluate the impact of hyper parameters for semantic segmentation in training set and report the results in Figure 3. For boundary label synthesis, we evaluate the influence of θ_{diff} and θ_{scale} in Equation (3) on the segmentation performance. As shown in Figure 3(a), BES approach is robust to various parameters values. For attention-pooling CAM, Figure 3(b) illustrates the effect of attention intensity parameter k in Equation (7). Compared to GAP-based CAM, the proposed attention-pooling CAM gets slightly improvement when attention intensity parameter k is set to a small value, but further increasing of k value makes performance drop down quickly.

Boundaries Evaluation. To better demonstrate the efficiency of boundary exploration, we evaluate the generated boundary maps on SBD benchmark [9], which contains semantic boundary annotations from 11355 images taken from

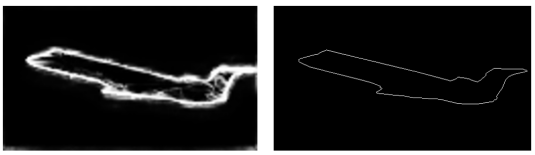
θ_{diff} \ θ_{scale}	0.20	0.25	0.30	0.35	0.40	0.45
0.10	65.4	66.0	66.0	**66.4**	66.1	65.6
0.15	65.4	65.7	66.2	66.1	66.2	65.9
0.20	65.0	65.6	66.1	66.0	**66.4**	65.7

(a)



(b)

Fig. 3. The performance when employing different parameter values in VOC 2012 training set.(a) The performance of semantic segmentation for different values of parameter θ_{diff} and θ_{scale} in boundary exploration. (b) The performance of attention-pooling CAM with various values of parameter k.



(a)

Precision(%)	Recall(%)	MF(%)
46.4	45.5	45.9

(b)

Fig. 4. Performance of boundary exploration, evaluated on the SBD *trainval* set. (a) The sample of predicted boundary map and the corresponding boundary label. (b) The boundary evaluation result.

the PASCAL VOC 2011 dataset. Since predicted boundary maps are class-agnostic and do not satisfy the semantic boundary requirement of SBD benchmark, we transform the ground-truth semantic boundary labels into class-agnostic labels. The precision, recall and maximal F-measure of the generated 11355 boundary maps are reported in Fig 4.

Quality Analysis. DeepLab-ASPP [5] trained by pseudo labels is employed to evaluate the proposed approach. Table 2 and 3 record concrete performance in VOC 2012 *val* and *test* set. We get substantial performance improvement compared to other methods, especially in large-scale object (e.g., bus and car) segmentation tasks.

We also analyse our failure cases and show it in Figure 5. The segmentation performance of BES is not good enough when dealing some objects (e.g., bike and chair) with complicated structures. We argue that it is due to two main factors. First, the BENet makes judgement for each pixel separately, and the discrete boundary prediction is hard to output consecutive boundary to provide effective constraint for propagation. Second, as the inherent limitation of neural

network, it is difficult to capture exact location of small object components like bike pedals or table legs.

Table 2. Semantic segmentation performance in PASCAL VOC 2012 *val* set.

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mean
EM-Adapt [23]	67.2	29.2	17.6	28.6	22.2	29.6	47.0	44.0	44.2	14.6	35.1	24.9	41.0	34.8	41.6	32.1	24.8	37.4	24.0	38.1	31.6	33.8
MIL+seg[25]	79.6	50.2	21.6	40.9	34.9	40.5	45.9	51.5	60.6	12.6	51.2	11.6	56.8	52.9	44.8	42.7	31.2	55.4	21.5	38.8	36.9	42.0
SEC [17]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
TPL[16]	82.8	62.2	23.1	65.8	21.1	43.1	71.1	66.2	76.1	21.3	59.6	35.1	70.2	58.8	62.3	66.1	35.8	69.9	33.4	45.9	45.6	53.1
PSA[2]	88.2	68.2	**30.6**	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	**73.7**	42.5	70.7	42.6	**68.1**	51.6	61.7
SSDD[26]	**89.0**	62.5	28.9	**83.7**	52.9	59.5	77.6	73.7	**87.0**	**34.0**	**83.7**	**47.6**	**84.1**	**77.0**	**73.9**	69.6	29.8	**84.0**	**43.2**	68.0	**53.4**	64.9
BES (Ours):																						
DeepLab-ASPP	88.9	**74.1**	29.8	81.3	**53.3**	**69.9**	**89.4**	**79.8**	84.2	27.9	76.9	46.6	78.8	75.9	72.2	70.4	**50.8**	79.4	39.9	65.3	44.8	**65.7**

Table 3. Semantic segmentation performance in PASCAL VOC 2012 *test* set.

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mean
EM-Adapt [23]	76.3	37.1	21.9	41.6	26.1	38.5	50.8	44.9	48.9	16.7	40.8	29.4	47.1	45.8	54.8	28.2	30.0	44.0	29.2	34.3	46.0	39.6
MIL+seg [25]	78.7	48.0	21.2	31.1	28.4	35.1	51.4	55.5	52.8	7.8	56.2	19.9	53.8	50.3	40.0	38.6	27.8	51.8	24.7	33.3	46.3	40.6
SEC [17]	83.5	56.4	28.5	64.1	23.6	46.5	70.6	58.5	71.3	23.2	54.0	28.0	68.1	62.1	70.0	55.0	38.4	58.0	39.9	38.4	48.3	51.7
TPL [16]	83.4	62.2	26.4	71.8	18.2	49.5	66.5	63.8	73.4	19.0	56.6	35.7	69.3	61.3	71.7	69.2	39.1	66.3	44.8	35.9	45.5	53.8
PSA [2]	89.1	70.6	**31.6**	77.2	42.2	**68.9**	79.1	66.5	74.9	29.6	68.7	56.1	82.1	64.8	**78.6**	**73.5**	**50.8**	70.7	47.7	63.9	**51.1**	63.7
SSDD [26]	**89.5**	71.8	31.4	79.3	47.3	64.2	79.9	74.6	84.9	**30.8**	73.5	**58.2**	**82.7**	73.4	76.4	69.9	37.4	80.5	54.5	**65.7**	50.3	65.5
BES (Ours):																						
DeepLab-ASPP	89.0	**72.7**	30.4	**84.6**	**47.5**	63.0	**86.8**	**80.7**	**85.2**	30.1	**76.5**	56.4	81.8	**79.9**	77.0	67.8	48.6	**82.3**	**57.2**	54.0	46.7	**66.6**

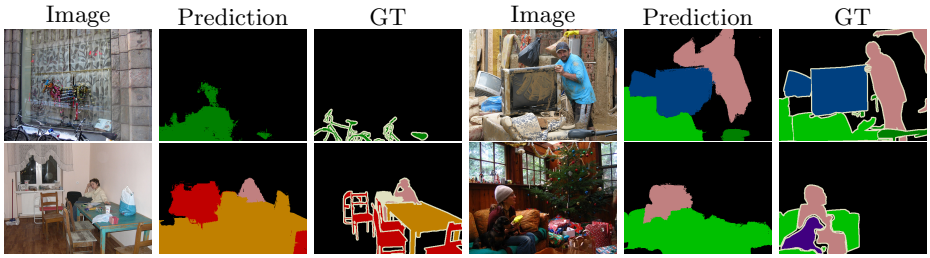


Fig. 5. Examples of failure cases in semantic segmentation.

4.3 Comparisons to the State-of-the-Art

In Table 4, we can observe that semantic labels generated by BES help the DeppLab-ASPP outperform all the listed image-level supervised methods both in *val* and *test* set. And even competitive with works [15] relying on bounding box. For fair comparison, we additionally provide DeepLab-CRF-LargeFOV [4] based result which uses VGG-16 [28] as backbone. We believe that other segmentation networks trained by our generated annotations can also achieve considerable performance.

Table 4. Comparison of semantic segmentation methods on PACAL VOC 2012 *val* and *test* set. The supervision types (Sup.) indicate: B–bounding box label, S–scribble label, F–pixel-level label, and I–image-level class label.

Methods	Sup.	Backbone	Training	val	test
WSSL [23]	B	VGG-16	10K	60.6	62.2
SDI [15]	B	VGG-16	10K	65.7	67.5
Scribblesup [21]	S	VGG-16	10K	63.1	-
FCN [22]	F	VGG-16	10K	-	62.2
DeepLab-v1 [4]	F	VGG-16	10K	67.6	70.3
PSPNet [34]	F	ResNet-101	10K	-	85.4
STC [31]	I	VGG-16	50k	49.8	51.2
TransferNet [11]	I	VGG-16	70K	52.1	51.2
AE_PSL [30]	I	VGG-16	10K	55.0	55.7
GAIN [20]	I	VGG-16	10K	55.3	56.8
CrawlSeg [12]	I	VGG-16	970K	58.1	58.7
MCOF [29]	I	ResNet-101	10K	60.3	61.2
DSRG [13]	I	ResNet-101	10K	61.4	63.2
IRNet [1]	I	ResNet-50	10K	63.5	64.8
FickleNet [19]	I	ResNet-101	10K	64.9	65.3
SSDD [26]	I	ResNet-38	10K	64.9	65.5
OOA [14]	I	ResNet-101	10K	65.2	66.4
BES (Ours):					
DeepLab-CRF-LargeFOV	I	VGG-16	10K	60.1	61.1
DeepLab-ASPP	I	ResNet-101	10K	**65.7**	**66.6**

Finally, we compare the BENet with IRNet, which predicts pixel affinity to implicitly capture boundary information. Figure 6 illustrates the training images and its corresponding semantic segmentation. Compared to IRNet results, boundary maps predicted by BENet precisely activates boundary pixels and suppresses foreground/background regions, which presents a better boundary discrimination ability. The semantic segmentation results demonstrate the benefit of explicitly exploring sufficient object boundaries.

5 Conclusion

To address the problem of weakly supervised semantic segmentation, we propose BES approach to explicitly explore object boundaries to refine coarse localization maps for training images. We design an attention-pooling CAM to get better object localization maps as seed region, then BENet is created to explore object boundaries and direct the propagation of semantic segmentation. On PASCAL VOC 2012 benchmark, BES performance outperforms the previous state-of-the-art methods. Extensive evaluation of ablation study and experiments validate the effectiveness and the robust of proposed BES. In the future, we plan to extend proposed approach to develop an end-to-end semantic segmentation framework.

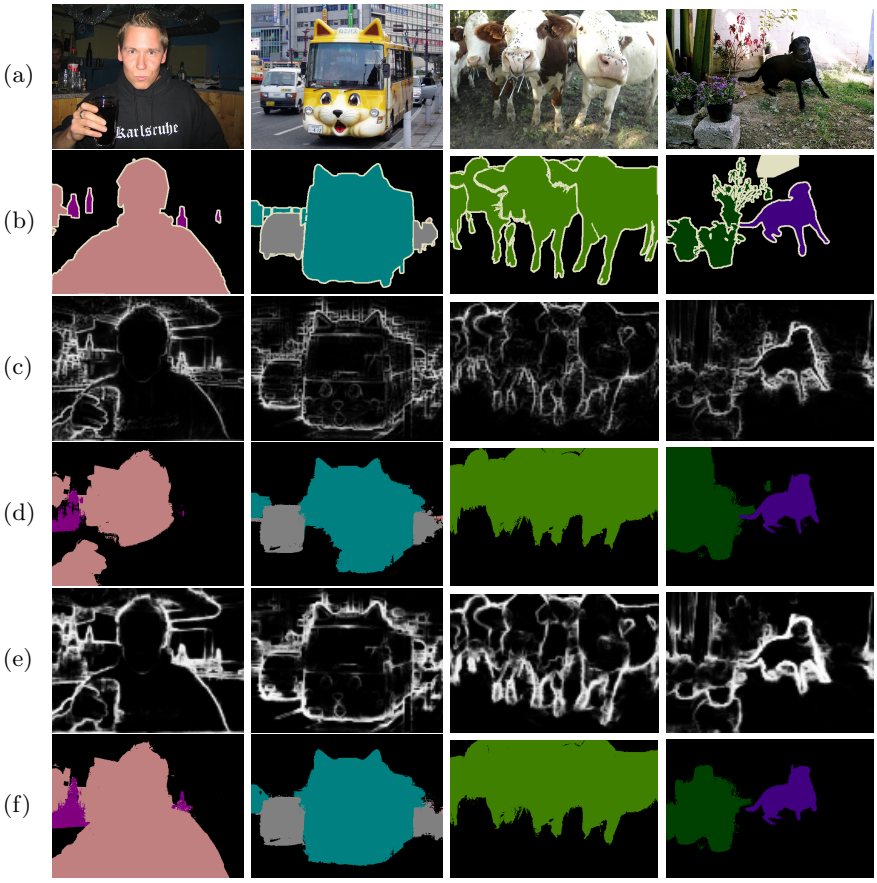


Fig. 6. Qualitative generated semantic segmentation on PASCAL VOC 2012 training set. (a) Original images. (b) Ground-truth. (c) Boundary maps of IRNet. (d) IRNet results. (e) Boundary maps of BES. (f) BES results.

Acknowledgments. This work was supported in part by the national key research and development program of China under grant No. 2019YFB2102200, Natural Science Foundation of China under Grant No. 61902062, 61672154, 61972086. and Jiangsu Provincial Natural Science Foundation of China under Grant No. BK20190332, Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education.

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2209–2218 (2019)

2. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4981–4990 (2018)
3. Bertasius, G., Torresani, L., Yu, S.X., Shi, J.: Convolutional random walk networks for semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 858–866 (2017)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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 (2017)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Durand, T., Mordan, T., Thome, N., Cord, M.: Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 642–651 (2017)
8. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
9. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 International Conference on Computer Vision. pp. 991–998. IEEE (2011)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Hong, S., Oh, J., Lee, H., Han, B.: Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3204–3212 (2016)
12. Hong, S., Yeo, D., Kwak, S., Lee, H., Han, B.: Weakly supervised semantic segmentation using web-crawled videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7322–7330 (2017)
13. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7014–7023 (2018)
14. Jiang, P.T., Hou, Q., Cao, Y., Cheng, M.M., Wei, Y., Xiong, H.K.: Integral object mining via online attention accumulation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2070–2079 (2019)
15. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 876–885 (2017)
16. Kim, D., Cho, D., Yoo, D., So Kweon, I.: Two-phase learning for weakly supervised object localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3534–3543 (2017)
17. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: European Conference on Computer Vision. pp. 695–711. Springer (2016)

18. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in neural information processing systems. pp. 109–117 (2011)
19. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5267–5276 (2019)
20. Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9215–9223 (2018)
21. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3159–3167 (2016)
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
23. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1742–1750 (2015)
24. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
25. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1713–1721 (2015)
26. Shimoda, W., Yanai, K.: Self-supervised difference detection for weakly-supervised semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5208–5217 (2019)
27. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
29. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1354–1362 (2018)
30. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1568–1576 (2017)
31. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2314–2320 (2016)
32. Zhang, L., Gao, Y., Xia, Y., Lu, K., Shen, J., Ji, R.: Representative discovery of structure cues for weakly-supervised image segmentation. IEEE Transactions on Multimedia **16**(2), 470–479 (2013)
33. Zhang, L., Yang, Y., Gao, Y., Yu, Y., Wang, C., Li, X.: A probabilistic associative model for segmenting weakly supervised images. IEEE Transactions on Image Processing **23**(9), 4150–4159 (2014)

34. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
35. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)