Adaptive Spatial-BCE Loss for Weakly Supervised Semantic Segmentation

Tong Wu[1][0000−0003−4463−4139], Guangyu Gao*[1][0000−0002−0083−3016], Junshi Huang*[2][0000−0002−8395−1463], Xiaolin Wei[2][0000−0002−3983−047X], Xiaoming Wei[2][0000−0002−7471−8344], and Chi Harold Liu[1][0000−0002−0252−329X]

[1] Beijing Institute of Technology, China
guangyugao@bit.edu.cn
[2] Meituan, China
huangjunshi@meituan.com

Abstract. For Weakly-Supervised Semantic Segmentation (WSSS) with image-level annotation, mostly relies on the classification network to generate initial segmentation pseudo-labels. However, the optimization target of classification networks usually neglects the discrimination between different pixels, like insignificant foreground and background regions. In this paper, we propose an adaptive Spatial Binary Cross-Entropy (Spatial-BCE) Loss for WSSS, which aims to enhance the discrimination between pixels. In Spatial-BCE Loss, we calculate the loss independently for each pixel, and heuristically assign the optimization directions for foreground and background pixels separately. An auxiliary self-supervised task is also proposed to guarantee the Spatial-BCE Loss working as envisaged. Meanwhile, to enhance the network's generalization for different data distributions, we design an alternate training strategy to adaptively generate thresholds to divide the foreground and background. Benefiting from high-quality initial pseudo-labels by Spatial-BCE Loss, our method also reduce the reliance on post-processing, thereby simplifying the pipeline of WSSS. Our method is validated on the PASCAL VOC 2012 and COCO 2014 datasets, and achieves the new state-of-the-arts. Code is available at https://github.com/allenwu97/Spatial-BCE.

Keywords: WSSS, Spatial-BCE, pseudo-labels, adaptive threshold

1 Introduction

Semantic segmentation aims to allocate labels to each pixel of an image, which has made significant progress driven by the Deep Neural Networks (DNNs). However, fully-supervised semantic segmentation requires pixel-level annotations, which are costly and time-consuming. Thus, researchers are motivated to use cheaper alternatives, including bounding boxes [23,30], scribbles [26], points [3] and image-level labels [31]. Among them, the image-level labels can be directly obtained from existing datasets, or when constructing large-scale datasets, web

[*] Corresponding Authors.

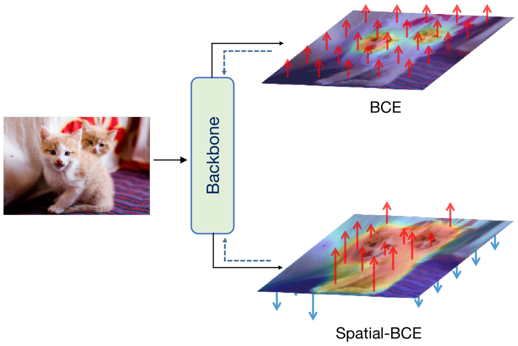

Fig. 1. The optimization directions of pixels when the network is trained by different loss function. The blue and red arrows correspond to the directions towards foreground and background, respectively. Best viewed in color.

search engines can automatically provide images and corresponding category tags. Thus the image-level labels based methods have lowest cost and become the mainstream of WSSS.

Most image-level labels based WSSS methods rely on Class Activation Maps (CAMs) [44] to generate the initial pseudo-labels for the training of semantic segmentation networks. Since the characteristics of the classification network, the CAMs tend to highlight the discriminative object regions, rather than the complete object regions, which deviates from the requirement of semantic segmentation. In this case, most previous WSSS works are motivated to extend the active regions of CAMs. One common solution is to obtain and fuse multiple highlight regions by multiple CAMs, including using multiple dilated convolutional blocks [36] and networks in different epochs [19]. These methods are intuitive, but easily to be over-activated or under-activated when faced with unconventional object sizes and shapes, and also their effects highly depend on hyper-parameters which harms the generalization.

In practice, the size of the foreground regions in the initial pseudo-label can be adjusted by the threshold, and a small threshold often refers to large foreground regions. However, classification networks usually ignore the feature discrimination between pixels, especially pixels in non-discriminative foreground regions and background regions. It is difficult and even unreasonable to divide precise target object boundaries by a global threshold overall images. Therefore, the crucial point to improve the quality of the initial pseudo-labels is enhancing the discrimination between foreground and background pixels, so that we can easily identify the object boundaries during inference.

For each positive category, the traditional BCE loss is calculated based on the average of the whole probability map, thus all pixels are optimized in the same direction, which reduces the discrimination between the foreground and background. Thus, we propose a novel Spatial Binary Cross-Entropy (Spatial-BCE) loss, which optimizes the foreground and background pixels in different directions, as shown in Fig. 1. Spatial-BCE Loss is designed to calculate the loss independently for each pixel. The optimization direction of each pixel depends on its probability in the prediction map and the threshold which divides the foreground and background. If its value is larger than the threshold, the parameters are optimized toward the foreground, otherwise, the parameters are optimized toward the background. With only image-level labels, the network lacks guidance on the division of foreground and background at the pixel level. Therefore, we introduce an auxiliary task to constrain the distribution of pixel probability, making Spatial-BCE Loss correctly assign the optimization direction for each pixel. Noticed that this auxiliary task is self-supervised, there is no requirement of the additional data. With this auxiliary task, the network which is trained by Spatial-BCE Loss, can correctly divide the foreground and background pixels and increase the discrimination between them. In that case, we can further obtain high-quality initial pseudo-labels with not only complete foreground regions but also precise boundaries.

In Spatial-BCE, we need a threshold to divide the foreground and background pixels. Although it can be a fixed threshold, a better way is to generate it through the network attentively for diverse inputs. Therefore, we design an alternate training strategy to alternately optimize this threshold. Previous methods determine such an optimal threshold through trial-and-error. However, it inevitably causes performance degradation when there is a lack of pixel-level labels for comparison required for trial-and-error in a real scenario of WSSS. Due to Spatial-BCE Loss allowing the network to generate the threshold adaptively, we can directly use it during inference. This training strategy makes us maintain stable performance even if without pixel-level labels as a comparison.

Our main contributions can be summarized as follows:

- We propose a novel Spatial-BCE Loss for WSSS, which can optimize the foreground and background pixels separately to increase the discrimination between them. With the help of an auxiliary self-supervised task, Spatial-BCE allows the network to generate high-quality initial labels.
- We design an alternate training strategy to allow the network to adaptively generate foreground and background dividing threshold. In that case, the network can generate the dividing threshold by itself during inference, reducing the impact of hyper-parameters on performance.
- Our method greatly improves the quality of the initial pseudo-labels, which in turn benefits the trained semantic segmentation network. Even if we remove post-processing that relies on additional networks or data, we still achieve competitive results. Under the same experimental configuration, we achieve the new state-of-the-arts of PASCAL VOC 2012 and MS-COCO 2014.

2 Related Work

2.1 Pseudo-label Generation

For WSSS of image-level labels, the common pipeline is to generate pseudo-labels through a classification network, and then use the pseudo-labels to train a semantic segmentation network. Most methods use CAM as the initial segmentation pseudo-labels. Since CAM only highlights the discriminative regions, the previous works are motivated by expanding the focus regions of CAM. Hou *et al.* [17] proposed a self-erasing network, which forces the network to pay attention to the remaining part of the object by erasing discriminative regions. Erasing is intuitive and effective for expanding CAM, but it is difficult to give reliable termination conditions the iterative erasure. Wei *et al.* [36] expand the receptive field of discriminative features by adding dilated convolutions with different dilated rates. Different receptive fields can generate CAMs with different distributions of highlight regions. After the fusion of multiple CAMs, the purpose of expanding the highlight regions can be achieved. A similar method also includes the generation and fusion of multiple CAMs through different epoch parameters [19] and different layers [25]. This idea of merging multiple CAMs is mostly based on the designer's prior knowledge rather than guiding the network to learn to focus on the non-discriminative regions. Therefore, the extended range is relatively fixed and the robustness is low.

Most of the latest methods tend to guide the network to focus on non-discriminative regions autonomously by adding auxiliary tasks. SEAM [35] introduced consistency regularization to provide self-supervision for network learning; MCIS [33] explore the semantic relationships across images; Chang *et al.* [5] add sub-categories for each image through clustering, allowing the network to focus on more fine-grained features. EDAM [37] effectively combines semantic segmentation tasks and classification tasks, letting the classification network directly predict the mask of each category. EPS [24] use additional saliency maps as supervision during training, so that the predictions can have accurate object boundaries and discard co-occurring pixels. The above auxiliary tasks are based on the classification task, and the loss function of the classification task is BCE Loss. However, we found that BCE Loss has the disadvantage when it is used for WSSS. BCE Loss does not distinguish between foreground pixels and background pixels, which leads to the confusion of foreground and background in CAM. Therefore, we upgraded the BCE Loss to Spatial-BCE Loss, make the network can optimize the foreground and background separately, solve the problem from the roots.

2.2 Pseudo-label Refinement

Besides improving the quality of the initial pseudo-labels, some works design additional networks to refine the initial pseudo-labels. SEC [20] proposed the three principles of seed, expansion, and constraint, which are followed by many subsequent works. DSRG [18] combines the seeded region growing with DNN, uses CAM as the initial cue, and continuously expands the seed regions. AffinityNet [2] generates a transition matrix by learning the similarity between pixels, and implements the semantic propagation by random walks. Ahn *et al.* [1] used the high-confidence foreground and high-confidence background to generate the supervision of the object boundaries, which further improves the accuracy when refining CAMs. Some works [13,19,24,33,36,37] use additional saliency maps as background cues to enhance the details of the object boundaries. Due to class-agnostic, these post-processings not only increase the complexity of the pipeline,

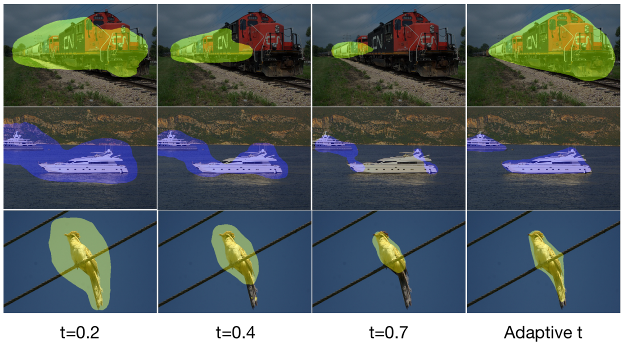

Fig. 2. Visualization of pseudo-labels from CAMs with different background threshold and our adaptive threshold. Best viewed in color.

but also easily introduce new noise. Our method can achieve competitive results without using additional saliency maps or training additional networks.

3 Approach

3.1 Overview

For an image with a specific image-level category, there is only part of pixels that strictly belong to such category, which we call target pixels, and the remained parts are the non-target pixels. In a multi-label classification task with traditional BCE Loss, the features of all pixels are averaged for the prediction of image probabilities belonging to certain categories. In this case, all pixels are assigned with the same optimization direction. Therefore, the discrimination of features between different pixels will be ignored, especially for the non-target pixels (e.g., pixels of background) and non-discriminative target pixels. The neglect of feature discrimination makes it difficult to get the precise object regions in the segmentation task. In Fig.2, we present the pseudo-labels with various thresholds to distinguish the target pixels and the non-target pixels from the classification network. When setting a large threshold (0.7 or 0.4), the pseudo-labels only activate partial objects. To recall the rest target parts by reducing the threshold to 0.2, irrelative non-target regions surrounding the boundary are also activated.

Considering the feature discrimination of different pixels, we propose the Spatial-BCE Loss to optimize the prediction of pixels individually. Let's denote $\boldsymbol{F}$ the feature map of image $\boldsymbol{I}$, and $P^c = Sigmoid(Linear^c(GAP(\boldsymbol{F})))$ represents the probability of $\boldsymbol{I}$ belonging to the c^{th} category, where GAP is the Global Average Pooling. In the Spacial-BCE Loss, we generate the probability map $\boldsymbol{P}^c = Sigmoid(Linear^c(\boldsymbol{F}))$ to calculate the loss of c^{th} category, where $\boldsymbol{P}^c = \{p_i^c\} \in \mathcal{R}^{w \times h}$, and p_i^c is probability of i^{th} pixel belonging to c^{th} category. Generally, the function of Spatial-BCE Loss is written as:

$$\mathcal{L}_{Spatial-BCE} = \sum_{c=1}^{C} \sum_{i=1}^{h \times w} (y^c R(p_i^c) - (1 - y^c) log(1 - p_i^c)) \tag{1}$$

where $R(\cdot)$ is the pixel-wise re-factoring strategy for positive categories, and y^c is the image-level label of c^{th} category. The details of Spatial-BCE loss are proposed in Sec. 3.2. To prevent the collapsing of Spatial BCE loss, an auxiliary self-supervision task is introduced in Sec. 3.3. Different from most existing works [13,19,24,33] using the same threshold for all categories through trial-and-error, our adaptive threshold strategy is specified in Sec. 3.4.

3.2 Spatial-BCE Loss

Assuming the category-specific threshold t^c is pre-defined, we treat the pixels whose probabilities exceed the threshold t^c as target candidates ($p_i^c > t^c$), otherwise as non-target candidates ($p_i^c \leq t^c$). The main idea of $\mathcal{L}_{Spatial-BCE}$ is to divide the target candidates from non-target ones, and penalize the *uncertain* candidates by $R(\cdot)$. To this end, we propose three principles to design $R(\cdot)$:

1. The *uncertainty* of target candidates monotonically decreases with the increment of predicted probabilities, and reaches 0 if the probability is 1.
2. Likewise, the *uncertainty* of non-target candidates decreases monotonically as the predicted probabilities decrease and is 0 with the probability of 0.
3. The *uncertainty* of regions reach maximum if the predicted probability is surrounding the aforementioned threshold.

Based on principle 1&3, we define the uncertainty of target candidates as:

$$R_{tg}(p_i^c) = -(p_i^c - t^c)^2 + (1 - t^c)^2 \tag{2}$$

where $p_i^c \in (t^c, 1)$. As a monotonically decreasing function over p_i^c, $R_{tg}(1) = 0$ (satisfying principle 1), and reaches the maximum value when p_i^c approaches t^c (satisfying principle 3). In practice, we use an α-balanced variable of $R_{tg}(p_i^c)$:

$$R_{tg}(p_i^c) = -\alpha(p_i^c - t^c)^2 + 1 \tag{3}$$

where $\alpha = (1 - t^c)^{-2}$ is a variable to normalize the maximum value to be 1.

Similarly, for pixels in non-target candidates ($p_i^c \in (0, t^c]$), following principle 2&3, we define the uncertainty of non-target candidates as:

$$R_{ntg}(p_i^c) = -\beta(p_i^c - t^c)^2 + 1 \tag{4}$$

where $\beta = t^{c-2}$ is a normalized variable for non-target pixels.

Finally, the $R(p_i^c)$ can be written as a *differentiable* piecewise function:

$$R(p_i^c) = \begin{cases} R_{tg}(p_i^c) & p_i^c > t^c \\ R_{ntg}(p_i^c) & p_i^c \leq t^c \end{cases} \tag{5}$$

In Fig. 3, we visualize the pixel uncertainty of positive samples by BCE Loss and Spatial-BCE Loss respectively. The solid lines represent $R(p_i^c)$ with different t^c, and the dashed line refers to the $\log p_i^c$ which is the uncertainty metric in BCE Loss. Different from BCE Loss, which regards samples with low probability as

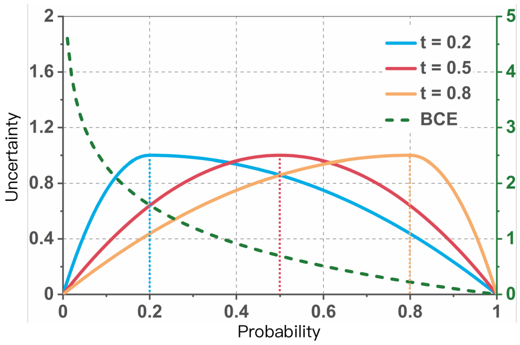

Fig. 3. The curve of uncertainty for the positive samples with different prediction probabilities. Best viewed in color.

the main optimization object, $R(p_i^c)$ turns to consider it as a confident non-target pixel. Furthermore, $R(p_i^c)$ increases the uncertainty of p_i^c when the value surrounds t^c, allowing the network to focus more on the hard-examples. Since the value of t^c is adaptively learned (sec Sec. 3.4), $R(p_i^c)$ has strong generalization when assigning optimization directions to pixels on different images.

3.3 Auxiliary Supervision

As we may observe in the experiment, the Spatial-BCE Loss can easily collapse to a trivial solution that the probabilities of all pixels are zero. Therefore, we import the distribution of target and non-target pixels as an extra constraint for the training of Spatial-BCE Loss. Since this pixel distribution is lacking in image-level labels, we design a self-supervised method without additional data to obtain it.

Given the image-level labels, we first train a classification network by BCE Loss to generate rough pseudo-labels. After using dense Conditional Random Field (dCRF) [21] for enhancement, we calculate Q^c for each image as:

$$Q^c = \frac{\sum_{i=1}^{h\times w} |l_i = c|}{\sum_{i=1}^{h\times w} |l_i \neq c|} \tag{6}$$

where l_i is the predicted category of i^{th} pixel after dCRF, and Q^c represents Target and Non-Target (T/NT) proportion of the c^{th} category. Note again that Q^c is calculated independently for each image.

With the initialized Q^c, we integrate our self-supervised task into the network as an extra constraint. We first predict the T/NT proportion of images by:

$$\widehat{P^c} = \frac{\sum_{i=1}^{h\times w} p_i^c}{\sum_{i=1}^{h\times w} (1 - p_i^c)} \tag{7}$$

Then, we urge the predicted $\widehat{P^c}$ by current network to approach Q^c. To this end, the Kullback-Leibler (KL) Divergence is used as the loss function:

$$\mathcal{D}_{KL} = \sum_{c=1}^{Pos} y_c \widehat{P^c} log(\frac{\widehat{P^c}}{Q^c}) \tag{8}$$

where Pos is the positive categories of corresponding image. Although the initial Q^c is not accurate, it still provides sufficient constrain for the distribution of pixels, thereby preventing the model from collapsing. In our experiment, we update Q^c of training images by the on-training network at every 8k iterations.

Finally, we use $\mathcal{L} = \mathcal{L}_{Spatial-BCE} + \mathcal{D}_{KL}$ as the whole losses. $\mathcal{D}_{KL}$ provides the constraint for the distribution of target and non-target pixels. $\mathcal{L}_{Spatial-BCE}$ converts the blurry soft predictions into polarized results, thus reducing the uncertainty of target and non-target pixels. The conjunction of $\mathcal{L}_{Spatial-BCE}$ and $\mathcal{D}_{KL}$ for joint training induces the network gradually generates more accurate pseudo-labels, as well as T/NT proportion.

3.4 Adaptive Threshold

We assume the category-specific threshold t^c is pre-defined in Sec. 3.2. In practice, rather than treating t^c as a fixed hyper-parameter, we generate t^c adaptively during the training process, and the algorithm is presented in Alg. 1.

Specifically, we separate every γ iterations as a *training phase*. At the first θ iterations of each phase, we initialize the threshold t based on T/NT proportion Q for input images. In detail, we estimate the coarse proportion of target pixels in the image for each positive category c by $\frac{Q^c}{1+Q^c}$, and use the $\frac{Q^c}{1+Q^c}$-th percentile of predicted probabilities $\boldsymbol{P^c}$ as the image-specific category threshold t^c. In this case, the network is only updated by the gradient of loss w.r.t $\boldsymbol{P}$. This strategy provides a relatively reliable value for t_c and thus stabilizes the network training.

To chase a better threshold, we generate t_c by the network through convolution layers and update it via the Spatial-BCE loss in the rest $(\gamma - \theta)$ iterations of each *training phase*. Note that the gradient of loss w.r.t $\boldsymbol{P}$ is detached in this sub-phase, since simultaneous learning of t and $\boldsymbol{P}$ usually result in the instability of network training. As aforementioned, we update the T/NT proportion Q based on the on-training network parameters at the last iteration of each phase.

Pseudo-label Generation. During inference, we directly use t generated by network to as the image-specific threshold of foreground and background. Generally, the pixel category l_i for specific image is determined by:

$$l_i = \begin{cases} \arg\max_c(S_i^c) & \text{if} \max(S_i^c) > 0 \\ \text{background} & \text{otherwise} \end{cases} \tag{9}$$

where $S_i^c = \{p_i^c - t^c, \forall c \in Pos\}$ is the de-biased probability set of pixels for c^{th} positive category. This strategy of adaptive threshold allows our method to select reliable dividing threshold (see Fig. 2), and thus alleviates the degradation of performance caused by artificial hyper-parameters.

Algorithm 1 Training with Adaptive Threshold

```
Input: Initial $T/NT$ proportion $Q$; Hyper-parameter $\theta$, $\gamma$
Output: Probability map $\boldsymbol{P}$; Thresholds $t$
  for iter < Iteration do
    if $iter\%\gamma < \theta$ then
      $t = \frac{Q}{1+Q}$-th percentile of $\boldsymbol{P}$
      $\boldsymbol{P} = \boldsymbol{P}.update()$
    else
      $t = t.update()$
      $\boldsymbol{P} = \boldsymbol{P}.detach()$
      if $iter\%\gamma == \gamma - 1$ then
        update $Q$
      end if
    end if
  end for
```

4 Experiments

4.1 Datasets and Evaluation Metric

We train and validate our method on PASCAL VOC 2012 [12] and MS-COCO 2014 [27]. The PASCAL VOC 2012 contains 21 categories (including a background category). Following the previous works, we use the augmented training set [15] containing $10,582$ images for training. The validation set and test set of PASCAL VOC 2012 contain $1,449$ and $1,456$ images respectively. The MS-COCO 2014 data set contains 81 categories including a background category, and the training set and validation set contain $82,081$ and $40,137$ images, respectively. We have excluded images that do not contain any foreground categories in MS-COCO 2014 as [9,24]. As some pixels in the ground-truth of MS-COCO overlap with multiple categories, we use the annotation of COCO-Stuff [4] as an alternative, which shares the same image set with MS-COCO.

We validate our method on the validation and test set of PASCAL VOC and the validation set of MS-COCO. The performance of the test set of PASCAL VOC is obtained by submitting the results to the official evaluation website. For all experiments, the mean IoU (mIoU) is used as the evaluation metric.

4.2 Implementation Details

We use ResNet38 [38] as the backbone of the classification network in our method. We randomly crop each image to the size of 368×368 and use RandomAug [10] as the data augmentation during training. During training, we also set learnable category-specific logit scales, which are initialized to $ln(10)$. A large logit scale can decrease the uncertainty that the pixel belongs to the foreground or background and speed up the convergence. During inference, we drop out logit scales to make prediction smoother, which is convenient for finding the optimal threshold. We perform multi-scales fusion when generating the initial pseudo-labels. After generating the pseudo-labels, we respectively apply IRN [1] and saliency maps generated by PoolNet [28] as post-processing. In different configurations, we use Deeplab-LargeFOV [6] and Deeplab-ASPP [7] as the semantic segmentation network, and their backbones are ResNet101 [16] and VGG16 [32] respectively. All backbones are pretrained on ImageNet [11].

We train our model on 4 Tesla V100 GPUs with 16 GB memory. We use the SGD optimizer and set the initial learning rate to 10^{-2}. In the initialization phase, the learning rate (LR) will drop from 10^{-2} to 10^{-3}. After each iteration of Q^c, the LR will drop from 10^{-3} to 0. Other important hyper-parameters are set as follows: batch-size is 24, weight-decay is $5e^{-4}$, and momentum is 0.9.

4.3 Determining Thresholds

To determine the optimal threshold t^c, most of the previous methods need to traverse the range of $[0, 1]$ to generate pseudo-labels with different thresholds. The final threshold is obtained by calculating mIoU with the pixel-level ground-truth of the training set. Our Spatial-BCE loss allows the network to adaptively generate the t^c, as mentioned in Sec. 3.4. In that case, we can do inference without comparing with the pixel-level ground-truth, more strictly following the requirements of WSSS. In the subsequent experiments, we use * to indicate the results of using the adaptive threshold. For a fair comparison with previous methods, we also report the performance of manually choosing the optimal threshold.

4.4 Comparisons to State-of-the-arts

Table 1. The mIoU (%) of the initial pseudo-labels (Init) and after refining by dCRF (+dCRF), on PASCAL VOC 2012 training set.

Method	Init	+dCRF
Baseline	53.0	60.2
Chang *et al.* [CVPR2020] [5]	50.9	55.3
SEAM[CVPR2020] [35]	55.4	56.8
AdvCAM[CVPR2021] [22]	55.6	62.1
Ours*	**65.3**	**66.3**
Ours	**68.1**	**70.4**

In Tab. 1, we compare the quality of pseudo-labels of different methods. *Baseline* means method with pseudo-labels generated by ResNet38 based classification network using multi-scale inference through DA Layer [37], and the mIoU on training set of PASCAL VOC is 53.0%. The different ways to generate thresholds are mentioned in Sec. 4.3. After iteratively updating Q^c by 3 times, our initial pseudo-labels achieve mIoU of 65.3% with the adaptive thresholds, which is 9.7% higher than the previous SoTA of AdvCAM [22]. Since we have already used dCRF before updating Q^c, we also list results after dCRF for fair comparisons. It shows that our method is 4.2% higher than the previous SoTA, even if we do not need pixel-level labels during inference. When we manually choose the best threshold as previous works, our results can be 8.3% higher than the previous SoTA. The significant improvements on the initial pseudo-labels quality laid the foundation for us to simplify the post-processing process.

Table 2. Comparison to previous state-of-the-art approaches for WSSS on PASCAL VOC 2012 validation and test sets. $\mathcal{I}$**: Image-level labels**, $\mathcal{S}$**: Saliency maps**.

Method	Sup	Val	Test
IRNet[CVPR2019] [1]	$\mathcal{I}$	63.5	64.8
BES[ECCV2020] [8]	$\mathcal{I}$	65.7	66.6
CONTA[NeurIPS2020] [43]	$\mathcal{I}$	66.1	66.7
RRM[AAAI2020] [42]	$\mathcal{I}$	66.3	66.5
MBMNet[MM2020] [29]	$\mathcal{I}$	66.2	67.1
ECS-Net[ICCV2021] [34]	$\mathcal{I}$	66.6	67.6
AdvCAM[CVPR2021] [22]	$\mathcal{I}$	68.1	68.0
Ours*	$\mathcal{I}$	**68.5**	**69.7**
Ours	$\mathcal{I}$	**70.0**	**71.3**
OAA$^+$[ICCV2019] [19]	$\mathcal{I}, \mathcal{S}$	65.2	66.4
CIAN[AAAI2020] [14]	$\mathcal{I}, \mathcal{S}$	64.3	65.3
MCIS[ECCV2020] [33]	$\mathcal{I}, \mathcal{S}$	66.2	66.9
ICD[CVPR2020] [13]	$\mathcal{I}, \mathcal{S}$	67.8	68.0
AuxSegNet[ICCV2021] [39]	$\mathcal{I}, \mathcal{S}$	69.0	68.6
NSROM[CVPR2021] [41]	$\mathcal{I}, \mathcal{S}$	70.4	70.2
EDAM[CVPR2021] [37]	$\mathcal{I}, \mathcal{S}$	70.9	70.6
EPS[CVPR2021] [24]	$\mathcal{I}, \mathcal{S}$	71.0	71.8
Ours*	$\mathcal{I}, \mathcal{S}$	**70.5**	**71.6**
Ours	$\mathcal{I}, \mathcal{S}$	**71.8**	**73.4**

Tab. 2 shows the results on the PASCAL VOC 2012 validation and test set. The mainstream post-processing in previous methods is divided into two groups: methods with only image-level labels (such as AffinityNet [2], IRN [1]) or using additional saliency map. When there are only image-level labels, the post-processing relies on dCRF to construct the supervision for training extra network. Due to the training of extra network also requiring pixel-level ground-truth to finetune the hyper-parameters, we only use dCRF when we adaptively generate thresholds. It can be seen that under the premise of removing additional network and optimal threshold, our method can still achieve the new SoTA. When our method uses IRN as post-processing, the results are higher than the SoTA of the same post-processing by 1.9% and 3.3% on the validation set and test set, respectively.

Another mainstream post-processing method is using a pre-trained saliency detection network to generate saliency maps, and use saliency maps as the background cues, and the predictions as the foreground cues. Saliency map has accurate boundaries, but it can only highlight salient objects in images that are

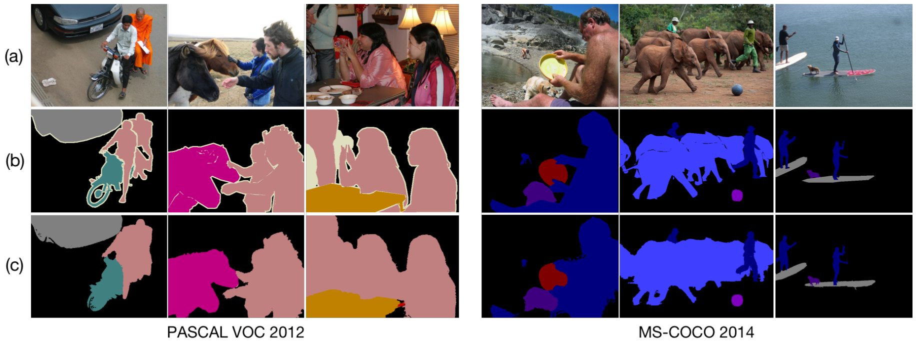

Fig. 4. Qualitative examples of segmentation results on the validation set of PASCAL VOC 2012 and MS-COCO 2014. (a) Original images, (b) Ground truth, (c) Our predictions.Best viewed in color.

class-agnostic. To reduce the noise, we introduce the approach mentioned in EDAM [37], modifying the misjudgment regions in the saliency maps based on the predicted probability maps. In the end, our performance based on the saliency maps is 71.8% and 73.4% on the validation set and test set, also achieving the new SoTA of this experimental configuration.

Table 3. Comparison to previous SoTA approaches of weakly-supervised semantic segmentation on MS-COCO 2014 validation set.

Method	Backbone	Sup	Val
SEC[ECCV2016] [20]	VGG16	$\mathcal{I}+\mathcal{S}$	22.4
DSRG[CVPR2018] [18]	VGG16	$\mathcal{I}+\mathcal{S}$	26.0
ADL[TPAMI2020] [9]	VGG16	$\mathcal{I}+\mathcal{S}$	30.8
CONTA[NeurIPS2020] [43]	ResNet38	$\mathcal{I}$	32.8
SGAN[ACCESS2020] [40]	VGG16	$\mathcal{I}+\mathcal{S}$	33.6
AuxSegNet[ICCV2021] [39]	ResNet38	$\mathcal{I}+\mathcal{S}$	33.6
EPS[CVPR2021] [24]	VGG16	$\mathcal{I}+\mathcal{S}$	35.7
Ours*	VGG16	$\mathcal{I}$	**35.2**
Ours	VGG16	$\mathcal{I}+\mathcal{S}$	**38.6**

Tab. 3 shows the results of our method on MS-COCO 2014. On the MS-COCO dataset, we use VGG16 as the backebone, and use original CAMs as initial pseudo-labels. When we only use image-level labels and adaptively generate thresholds, we can achieve comparable results with previous SoTA which use extra saliency maps. If We use the same post-processing strategy which is mentioned above for saliency maps. On the validation set, our segmentation network achieves mIoU of 38.6%, which is 2.9% higher than EPS [23].

4.5 Ablation Studies

Iteratively update of Q^c. Since the model generates more accurate T/NT proportion during the training process, we update Q^c at every 8k iterations to

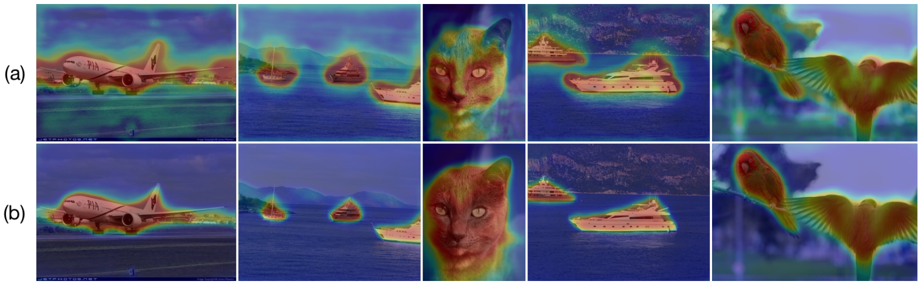

Fig. 5. Visualization of heatmaps on PASCAL VOC 2012. (a) Heatmap generated by our baseline. (b) Our final heatmap. Best viewed in color.

make Q^c gradually approaches ground-truth. In Tab. 4, we show the mIoU of the pseudo-labels after each phase. It can be seen that the improvement in the first iteration is most significant, and then the improvements gradually weaken. Considering the time cost, we terminate the training process after the third phase. Based on two different strategies for determining the threshold of pseudo-labels, we improve the performance of 6.1% and 10.2% respectively compared with the baseline. It is worth noting that our method does not depend on the specific network structure, in theory, any current model can be further finetuned by our method to achieve performance enhancement.

Table 4. The mIoU on PASCAL VOC 2012 training set when each time Q^c is updated.

iteration	Init	+dCRF
Baseline	53.0	60.2
Phase1*/Phase1	61.7/64.2	64.4/68.6
Phase2*/Phase2	64.5/67.5	66.1/70.7
Phase3*/Phase3	65.3/68.1	66.3/70.4

Contribution of Components. In Tab. 5, we measure the effect of different loss functions on the performance. After generating the initial Q^c based on the same results, we show the results of using only $\mathcal{D}_{KL}$ and $\mathcal{L}_{Spatial-BCE} + \mathcal{D}_{KL}$. When Q^c is updated after the first phase, the gap between the two sets is not obvious, only 1.2%. However, when Q^c is updated after the second phase, the gap increases to 3.8%. It can be concluded that without Spatial-BCE to optimize the distribution of predicted probability, $\mathcal{D}_{KL}$ will soon fall into the bottleneck, and the performance cannot be continuously improved.

In Tab. 6, we show the performance of pseudo-labels when the T/NT proportion is obtained in different ways. When t^c is determined by Q, the T/NT proportion of prediction is fixed, so the performance is difficult to improve. When the t^c is adaptively generated based on P, the network has the opportunity to optimize T/NT proportion of prediction, which leads to better results.

Discrimination between foreground and background. In Fig.5, we compare heatmaps generated by our method with heatmaps generated by our baseline. It can be seen that after the fineturn by Spatial-BCE, the highlight regions of the heatmaps are more complete, and the activations of the background re-

Table 5. The mIoU on PASCAL VOC 2012 training set when the network is trained by different loss function.

Loss	Init	Phase1	Phase2
$\mathcal{D}_{KL}$	60.2	63.0	63.7
$\mathcal{L}_{Spatial-BCE} + \mathcal{D}_{KL}$	60.2	64.2	67.5

Table 6. The mIoU on PASCAL VOC 2012 training set when the T/NT proportion is determined in different ways.

Method	Init	Phase1	Phase2
Fixed	60.2	59.7	60.5
Adaptive	60.2	61.7	64.5

gions are significantly reduced. The initial heatmaps are also an active part of the background regions while highlighting the regions of the object. Our heatmap can be found that the activation degrees of the pixels on the boundary are nearly the same, which makes us completely divide the foreground regions. At the same time, near the boundary of the object, the activation degree of the pixels drops rapidly (the color changes from red to blue), which means that it hardly divides the background into the foreground when generating pseudo-labels, avoiding over-activation. This is the contribution of Spatial-BCE which increases the discrimination between the pixels of the foreground and background.

5 Conclusion

In this paper, we propose the novel loss function, Spatial-BCE Loss, to improve the quality of initial pseudo-labels for weakly-supervised semantic segmentation. Spatial-BCE Loss is re-factoring from the traditional BCE Loss, and is urged to assign the different optimization directions for the target and non-target pixels of each positive category. Through our alternate training, Spatial-BCE Loss can not only improve the feature discrimination between foreground and background pixels, but also allow the network to adaptively generate dividing threshold. The trained classification network can generate initial pseudo-labels without additional networks or data with accurate boundaries. Benefiting from high-quality initial pseudo-labels, we achieve new state-of-the-art of PASCAL VOC 2012 and MS-COCO 2014 datasets under various experimental configurations.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No. 61972036, and in part by the Science and Technology on Optical Radiation Laboratory (China) under Grant No. 61424080213.

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. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: Proceedings of the European Conference on Computer Vision. pp. 549–565 (2016)
4. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1209–1218 (2018)
5. Chang, Y.T., Wang, Q., Hung, W.C., Piramuthu, R., Tsai, Y.H., Yang, M.H.: Weakly-supervised semantic segmentation via sub-category exploration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8991–9000 (2020)
6. 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)
7. 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 pp. 834–848 (2017)
8. Chen, L., Wu, W., Fu, C., Han, X., Zhang, Y.: Weakly supervised semantic segmentation with boundary exploration. In: Proceedings of the European Conference on Computer Vision. pp. 347–362 (2020)
9. Choe, J., Lee, S., Shim, H.: Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. IEEE transactions on pattern analysis and machine intelligence **43**, 4256–4271 (2020)
10. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703 (2020)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
12. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
13. Fan, J., Zhang, Z., Song, C., Tan, T.: Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4283–4292 (2020)
14. Fan, J., Zhang, Z., Tan, T., Song, C., Xiao, J.: Cian: Cross-image affinity net for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 10762–10769 (2020)
15. 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 (2011)

16. 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)
17. Hou, Q., Jiang, P., Wei, Y., Cheng, M.M.: Self-erasing network for integral object attention. Advances in Neural Information Processing Systems **31** (2018)
18. 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)
19. 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)
20. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 695–711 (2016)
21. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. Advances in neural information processing systems **24** (2011)
22. Lee, J., Kim, E., Yoon, S.: Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4071–4080 (2021)
23. Lee, J., Yi, J., Shin, C., Yoon, S.: Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2643–2652 (2021)
24. Lee, S., Lee, M., Lee, J., Shim, H.: Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5495–5505 (2021)
25. Lee, S., Lee, J., Lee, J., Park, C.K., Yoon, S.: Robust tumor localization with pyramid grad-cam. arXiv preprint arXiv:1805.11393 (2018)
26. 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)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755 (2014)
28. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3917–3926 (2019)
29. Liu, W., Zhang, C., Lin, G., Hung, T.Y., Miao, C.: Weakly supervised segmentation with maximum bipartite graph matching. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2085–2094 (2020)
30. 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)
31. 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)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

33. Sun, G., Wang, W., Dai, J., Van Gool, L.: Mining cross-image semantics for weakly supervised semantic segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 347–365 (2020)
34. Sun, K., Shi, H., Zhang, Z., Huang, Y.: Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7283–7292 (2021)
35. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12275–12284 (2020)
36. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7268–7277 (2018)
37. Wu, T., Huang, J., Gao, G., Wei, X., Wei, X., Luo, X., Liu, C.H.: Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 16765–16774 (2021)
38. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition pp. 119–133 (2019)
39. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Sohel, F., Xu, D.: Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6984–6993 (2021)
40. Yao, Q., Gong, X.: Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. IEEE Access pp. 14413–14423 (2020)
41. Yao, Y., Chen, T., Xie, G.S., Zhang, C., Shen, F., Wu, Q., Tang, Z., Zhang, J.: Non-salient region object mining for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2623–2632 (2021)
42. Zhang, B., Xiao, J., Wei, Y., Sun, M., Huang, K.: Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 12765–12772 (2020)
43. Zhang, D., Zhang, H., Tang, J., Hua, X.S., Sun, Q.: Causal intervention for weakly-supervised semantic segmentation. Advances in Neural Information Processing Systems pp. 655–666 (2020)
44. 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)