



# Background Activation Suppression for Weakly Supervised Object Localization and Semantic Segmentation

Wei Zhai<sup>1</sup> · Pingyu Wu<sup>1</sup> · Kai Zhu<sup>1</sup> · Yang Cao<sup>1,2</sup>  · Feng Wu<sup>1,2</sup> · Zheng-Jun Zha<sup>1</sup>

Received: 2 October 2022 / Accepted: 19 September 2023 / Published online: 17 October 2023  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

Weakly supervised object localization and semantic segmentation aim to localize objects using only image-level labels. Recently, a new paradigm has emerged by generating a foreground prediction map (FPM) to achieve pixel-level localization. While existing FPM-based methods use cross-entropy to evaluate the foreground prediction map and to guide the learning of the generator, this paper presents two astonishing experimental observations on the object localization learning process: For a trained network, as the foreground mask expands, (1) the cross-entropy converges to zero when the foreground mask covers only part of the object region. (2) The activation value continuously increases until the foreground mask expands to the object boundary. Therefore, to achieve a more effective localization performance, we argue for the usage of activation value to learn more object regions. In this paper, we propose a background activation suppression (BAS) method. Specifically, an activation map constraint module is designed to facilitate the learning of generator by suppressing the background activation value. Meanwhile, by using foreground region guidance and area constraint, BAS can learn the whole region of the object. In the inference phase, we consider the prediction maps of different categories together to obtain the final localization results. Extensive experiments show that BAS achieves significant and consistent improvement over the baseline methods on the CUB-200-2011 and ILSVRC datasets. In addition, our method also achieves state-of-the-art weakly supervised semantic segmentation performance on the PASCAL VOC 2012 and MS COCO 2014 datasets. Code and models are available at <https://github.com/wpy1999/BAS-Extension>.

**Keywords** Weakly supervised · Object localization · Background activation suppression · Semantic segmentation

---

Communicated by SUHA KWAK.

---

Wei Zhai and Pingyu Wu have contributed equally to this work.

---

✉ Yang Cao  
forrest@ustc.edu.cn

Wei Zhai  
wzhai056@ustc.edu.cn

Pingyu Wu  
wpy364755620@mail.ustc.edu.cn

Kai Zhu  
zkzy@mail.ustc.edu.cn

Feng Wu  
fengwu@ustc.edu.cn

Zheng-Jun Zha  
zhazj@ustc.edu.cn

<sup>1</sup> University of Science and Technology of China, Hefei, China

<sup>2</sup> Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China

## 1 Introduction

Weakly supervised object localization (WSOL) aims to identify the object's localization in a scene, where only image-level labels instead of bounding box annotations are available during training. Due to the reduction in the cost of manual labeling, and the potential to use the vast weakly-annotated images on many public datasets and the Web, WSOL is gaining more and more attention in the research community (Selvaraju et al., 2020; Zhai et al., 2022; Luo et al., 2022; Zhang et al., 2021b). Moreover, it can serve various downstream tasks, such as weakly supervised object detection (WSOD) (Song et al., 2021; Zhang et al., 2020b, 2019) and weakly supervised semantic segmentation (WSSS) (Ru et al., 2022a; Chan et al., 2021; Pan et al., 2022).

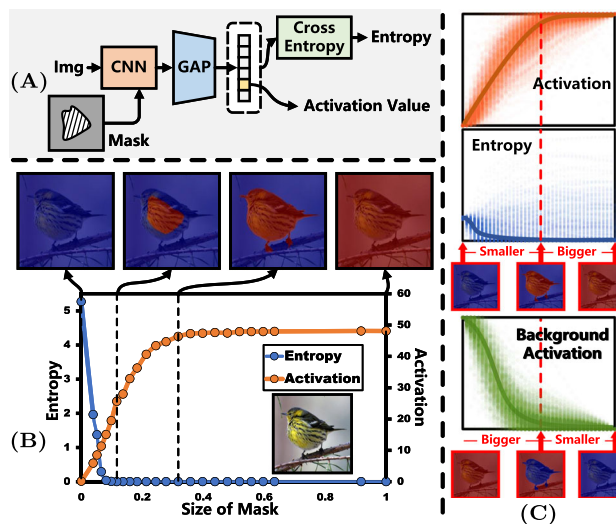
This paper aims to propose an effective approach for WSOL and its downstream task WSSS, since WSOL and WSSS tasks have similarities in that they both use image-level labels as supervision and need to obtain a high-quality

pixel-level localization map from the classification network. Actually, WSSS task can be implemented by directly training a fully supervised semantic segmentation with the localization maps generated from WSOL as pseudo labels. Due to these reasons, they face a similar challenge of establishing supervision between image-level labels and pixel-level localization maps in an effective way.

Previously, most WSOL and WSSS methods utilize class activation map (CAM) (Zhou et al., 2016) to extract localization map from classifier. While CAM can localize approximate object regions, it always prefers to capture the most discriminative regions rather than the overall area of the object, resulting in limited localization performance. Therefore, numerous CAM-based approaches have been proposed to alleviate this problem. Adversarial erasing methods (Singh & Lee, 2017; Zhang et al., 2018a; Choe & Shim, 2019; Mai et al., 2020; Yun et al., 2019) erase the most discriminative regions during the training, forcing the network to learn more object features that facilitate complete localization. Some methods (Zhang et al., 2020d; Pan et al., 2021; Lee et al., 2022a) improve the localization performance of CAM by establishing pixel-level spatial and semantic correlation. Additionally, some other methods (Zhang et al., 2018b; Wei et al., 2021; Kolesnikov & Lampert, 2016) suggest using the thought of region growing to spread confidence regions and mine relevant features.

Although CAM-based method can conveniently extract the localization map from the classifier, this approach will lead to limitations and conflicts in optimization since the classifier needs to implement both localization and classification tasks. Very recently, a CAM-independent paradigm (Meng et al., 2021; Xie et al., 2021) is devised for WSOL to achieve localization with a foreground prediction map (FPM) obtained directly through a generator, which allows the two tasks to be accomplished separately in a unified model. Typically, ORNet (Xie et al., 2021) is a two-stage approach, which first trains a classification network as an evaluator, and then utilizes CE loss to guide the learning of generator by masking the original image with a foreground prediction map. Orthogonally, the foreground prediction map in the FAM (Meng et al., 2021) is split into several parts and separately masks high-level feature maps to achieve learning of different regions through CE loss. Despite FPM-based methods achieving promising performance, they still suffer from incomplete object localization.

To better understand FPM-based methods, we focus on exploring the entropy value of CE loss (entropy) with respect to (*w.r.t.*) foreground mask. As shown in Fig. 1A, by changing the area of the foreground mask and masking the feature map, the relationship between the entropy and foreground mask area is plotted in Fig. 1B. An important phenomenon can be observed that there is a “mismatch” between entropy and ground-truth mask, i.e., entropy is already close to zero

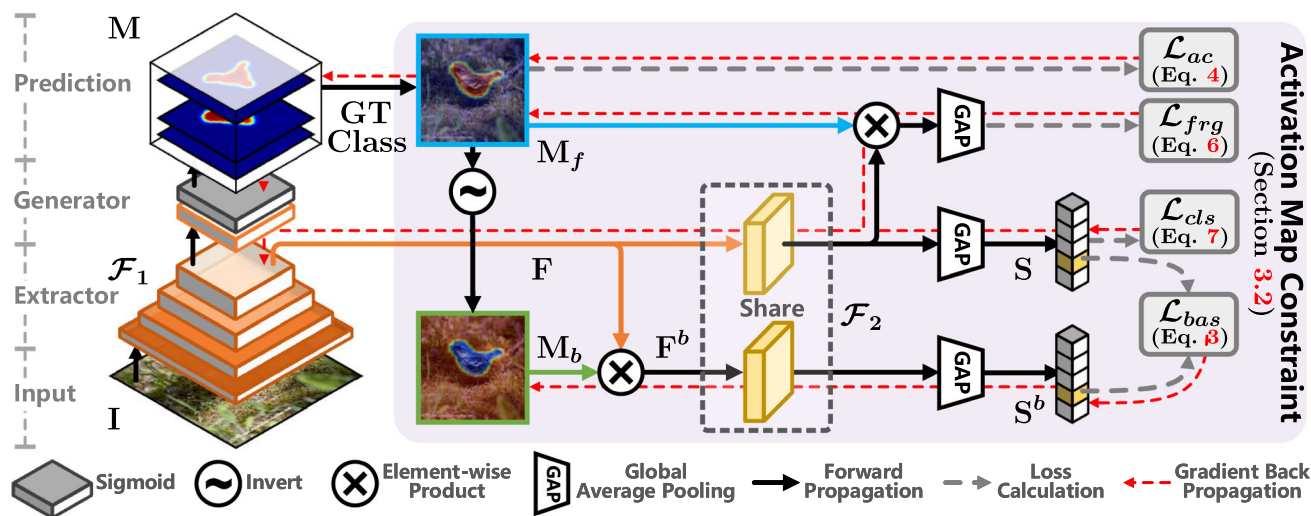


**Fig. 1** **A** Experimental procedure and related definitions. **B** The entropy value of CE loss *w.r.t.* foreground mask and foreground activation value *w.r.t.* foreground mask. **C** The results with statistical significance. Implementation details of the experiment and further results are available in Sect. 3.5

when foreground mask retains only part of the object region, which indicates that entropy cannot force the foreground map to learn the complete object area. The reason is that the exponential form of softmax amplifies the discrepancy in activation values and drives premature convergence of entropy. To find a better factor to facilitate localization learning, we further explore the activation value (before softmax calculation) *w.r.t.* foreground mask. As shown in Fig. 1B, there is a higher “correlation” between activation value and foreground mask, i.e., activation value tends to saturate when the mask expands to the object boundary. This suggests that better localization ability can be learned by optimizing activation value. Figure 1C also confirms the generality of these phenomena in a statistical sense.

Based on the inspiration of the above exploratory analysis, a straightforward manner to obtain a complete foreground prediction map is to maximize the activation value. However, considering that the minimization optimization problem is more conducive to the stability of training and loss convergence than the maximization optimization problem, this paper proposes a novel way to learn a background prediction map by minimizing background activation value, and further obtain the accurate foreground prediction map by inversion. Actually, the statistics on background activation values in Fig. 1C show “symmetry” with the statistics on activation values, both converging at the ground-truth mask area, which further supports the feasibility of background activation value suppression.

In this paper, we propose a simple but effective Background Activation Suppression (BAS) method. As shown in Fig. 2, our method includes three modules: an extractor, a genera-



**Fig. 2** The architecture of the proposed background activation suppression (BAS) in the training phase. The class-specific foreground prediction map  $M_f$  and the coupled background prediction map  $M_b$

are obtained by the generator according to the ground-truth (GT) class, and then fed into the Activation Map Constraint module together with the feature maps  $F$

tor, and an **Activation Map Constraint (AMC)** module. First, an extractor is used to extract the image features for subsequent localization and classification. The generator aims to generate a class-specific foreground prediction map for localization. Then the coupled background prediction map is obtained by inverting the foreground prediction map and fed into AMC together for localization training. The AMC is supervised by four kinds of losses, which are background activation suppression loss, area constraint loss, foreground region guidance loss, and classification loss. The most important one is background activation suppression loss, which is devised to promote the learning of generator by minimizing the ratio of background activation value and overall activation value (the activation value generated by the entire image). In the inference phase, the Top- $k$  prediction maps are selected based on the predicted category probabilities and their average prediction map is adopted as the final localization result. The main contributions of this paper can be summarized as follows:

- (1) This paper identifies that the essential reason why minimizing CE loss facilitates the generation of foreground map is that it indirectly increases the foreground activation value, and accordingly proposes to promote the generation of foreground prediction map by suppressing background activation value.
- (2) This paper proposes a simple but effective Background Activation Suppression (BAS) approach to facilitate the generation of foreground map by an Activation Map Constraint (AMC) in a weakly supervised manner, which is composed of four losses including background activation suppression loss and together contribute to

the generation of the foreground prediction map for localization.

- (3) Extensive experiments on both CUB-200-2011 (Wah et al., 2011) and ILSVRC (Russakovsky et al., 2015) benchmarks demonstrate that our method achieves consistent and significant improvement in terms of GT-known/Top-1/Top-5 Loc. In addition, the proposed BAS approach can be extended to Weakly Supervised Semantic Segmentation (WSSS) task, which also achieves new state-of-the-art results on PASCAL VOC 2012 (Everingham et al., 2010) and MS COCO 2014 (Lin et al., 2014) datasets.

This paper builds upon our conference version (Wu et al., 2021), which has been extended in four distinct aspects. (1) We explain the advantages of Background Activation Suppression and its generalizability (on more complex datasets) in more detail and comprehensively (in a statistical sense), see Fig. 6 and Sect. 3.5. (2) To alleviate the problem of inadequate convergence of BAS loss (Fig. 12), we focus on the location of the ReLU function, which is closely related to the activation value, and further improve the previous BAS after exploration, see Fig. 4 and Sect. 3.2. (3) To verify the extensibility of the BAS approach, we develop a Weakly Supervised Semantic Segmentation (WSSS) framework with proposed BAS in Sect. 5. The framework aims to enhance the quality of the seed generation process in the popular WSSS framework through BAS, resulting in better performance on WSSS task, as shown in Tables 9, 11 and 12. (4) To exploit the advantages of BAS on WSSS in obtaining localization maps through a generator, we propose to produce a class-agnostic foreground map using BAS and further combine it with the class-specific

maps to improve the quality of the initial seed, see Fig. 20 and Table 13. (5) To further improve the segmentation quality, we propose to apply the losses of BAS as evaluation scores in the inference phase to assess each threshold and find the image-specific threshold on WSSS, see Fig. 22 and Table 15. (6) We have made a lot of efforts to improve the presentations (e.g., motivation, related illustrative diagrams, formulation, experimental analysis, key results), and organizations of our paper. Besides, several sections have been refined to improve the readability and provide more detailed explanations about the motivation, quantitative/qualitative comparisons, and discussions.

The rest of this paper is organized as follows. Section 2 describes existing works related to WSOL and WSSS. The detailed method is described in Sect. 3. Sections 4 and 5 present the experimental results of WSOL and WSSS, respectively. Limitation and future work are discussed in Sect. 6. Finally, we conclude our work in Sect. 7.

## 2 Related Work

### 2.1 Weakly Supervised Object Localization

Weakly supervised object localization (WSOL) is a challenging task that requires localizing objects using only image-level labels. To obtain localization results from the classification network, CAM (Zhou et al., 2016) proposes to replace top layers with a global average pooling, and multiply the fully connected weights on depth feature maps to generate class activation map (CAM) as the localization map. Unfortunately, CAM usually focuses on the most discriminative regions. To alleviate this problem, a series of methods propose to use erasing strategies. HaS (Singh & Lee, 2017) splits the original image into different patches and randomly masks part of them, forcing the classification network to learn more features of objects. ACoL (Zhang et al., 2018a) and EIL (Mai et al., 2020) erase areas with high response in the feature map and use two parallel branches for adversarial erasing. Differently, ADL (Choe & Shim, 2019) erases the most significant regions of each layer during forward propagation, to achieve a balance between classification and localization. CutMix (Yun et al., 2019) adopts a data enhancement strategy that mixes two different images to force network to learn relevant regions of different objects.

In addition, another class of approaches adopt the thought of spreading confidence regions to mine relevant features. SPG (Zhang et al., 2018b) uses thresholds to filter foreground and background regions with high confidence from CAM to guide shallow network learning. Further, SPOL (Wei et al., 2021) generates more reliable confidence regions by multiplicative feature fusion strategy and trains a full segmentation network with confidence regions as pseudo labels.

I2C (Zhang et al., 2020d) proposes to increase the robustness and reliability of localization by considering the correlation of different pictures from the same class. Besides, SPA (Pan et al., 2021) uses a post-processing approach to extract feature maps with structure-preserving. SLT (Guo et al., 2021) considers several similar classes as one class when generating classification loss and localization maps, which alleviates the problem of focusing on the most discriminative regions by strengthening learning tolerance. DA-WSOL (Zhu et al., 2022) aligns the feature distributions between the image and pixel domains with the thought of domain adaptation.

Most recently, two Foreground-Prediction-Map-based works (Xie et al., 2021; Meng et al., 2021), both achieve the localization task by generating a foreground prediction map. ORNet (Xie et al., 2021) uses a two-stage approach, where an encode-decode layer is inserted in the shallow layer of the network as a generator and trained by the classification task in the first stage. In the second stage, the parameters of the classification network are fixed as an evaluator, and the foreground prediction map output by the generator is used to mask the image. Then the masked image is fed into the evaluator for classification training, so that the foreground prediction map can learn the object region. FAM (Meng et al., 2021) utilizes a Foreground Memory Mechanism structure to store different foreground classifiers and generate a class-agnostic foreground prediction map. The foreground prediction map is split into several specific parts which are used to mask the feature map to obtain different part-aware feature maps. After classification training with the corresponding foreground classifiers, the class-agnostic foreground map is forced to learn different object regions. It can be noticed that both ORNet (Xie et al., 2021) and FAM (Meng et al., 2021) only consider foreground regions and use cross-entropy to facilitate the learning of generator. Different from these methods, this paper proposes a background activation suppression strategy to learn foreground prediction maps through a simple but effective approach.

### 2.2 Weakly Supervised Semantic Segmentation

Weakly supervised semantic segmentation (WSSS) purposes to alleviate the reliance on pixel-level ground-truth labels by using weak labels instead. Existing WSSS methods usually include the following three stages: (1) Obtaining a high-quality initial seed. (2) Seed refinement and generating pseudo labels. (3) Training a full segmentation network with pseudo labels. It can be seen that generating a high-quality pixel-level localization map is also crucial for WSSS, similar to WSOL.

*Seed Generation.* Extraction of CAM is arguably the most common and convenient approach to generate the initial seed, despite the problem that only the discriminative regions can

be highlighted. To alleviate this issue, some methods propose to improve the quality of CAM by iterative manipulation. AE-PSL (Wei et al., 2017) performs iterative training steps to mine more object-related regions with adversarial erasure. RIB (Lee et al., 2021a) applies a post-processing method to fine-tune the classification model and obtain CAMs by iteration. AdvCAM (Lee et al., 2022a) proposes an anti-adversarial approach to continuously identify more object areas. Besides, a category of methods try to improve the classification learning process. CONTA (Zhang et al., 2020c) aims to avoid contextual confusion by proposing a structural causal model to analyze the causalities among images, contexts, and class labels. SEAM (Wang et al., 2020b) applies consistency regularization on CAMs through various sized images to mitigate the supervision gap issue. ReCAM (Chen et al., 2022b) proposes to use softmax cross-entropy loss to suppress the response of different categories to the same receptive field. CLIMS (Xie et al., 2022a) utilizes the CLIP (Radford et al., 2021) model to assist the network in activating more complete object regions. GAIN (Li et al., 2018) uses Grad-CAM to obtain localization maps and improve them by exploiting the prediction scores of the network as supervision. In contrast, BAS is based on the FPM-based paradigm and proposes a more essential and effective background activation suppression loss compared to the cross-entropy used in the FPM-based methods from the experimental observations.

**Mask Generation.** The initial seed is usually coarse and needs to be refined. Some researchers adopt the thought of region growing to spread the initial seed. SEC (Kolesnikov & Lampert, 2016) proposes three principles: seed, expand and constrain. The initial seed is expanded during the training of segmentation and constrained to the object boundaries. PSA (Ahn & Kwak, 2018) trains a deep network to predict semantic affinity between a pair of adjacent image coordinates and propagate the semantics by random walk (Lovász, 1993). IRN (Ahn et al., 2019) predicts a transition probability matrix from the boundary activation map and generates pseudo masks in a similar way to PSA.

### 3 Methodology

In this section, we first introduce the main architecture of the network and the definition of the symbols in Sect. 3.1. Then we describe the structure of the AMC module, including the form of the four loss functions, and the improvement of BAS compared to the previous conference version in Sect. 3.2. The total loss functions for WSOL and WSSS are listed in Sects. 3.3 and 3.4, respectively. Finally, we provide specific details of the exploratory experiments and statistical results on three different datasets in Sect. 3.5.

### 3.1 Overview

Based on the experimental observation, we enhance the completeness of the localization map for WSOL by proposing a background activation suppression (BAS) approach. As shown in Fig. 2, BAS consists of three modules: an extractor, a generator, and an activation map constraint (AMC) module. The extractor is used to extract features related to classification and localization. The generator is to produce the foreground prediction maps. The AMC module is to promote the learning of extractor and generator through four kinds of losses.

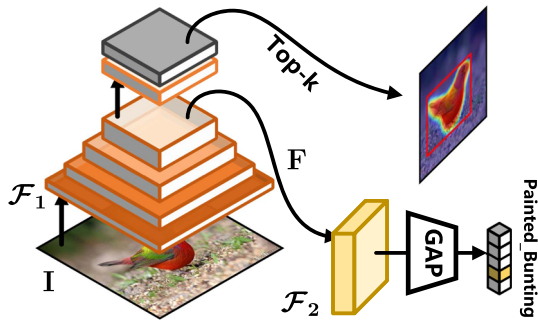
Specifically, we divide the original backbone network into two sub-networks  $\mathcal{F}_1$  and  $\mathcal{F}_2$  according to the location of the generator, and denote the network parameter by  $\Theta$ . The sub-network  $\mathcal{F}_1$  before the generator is used as a feature extractor. Given an image  $\mathbf{I}$ , the feature maps  $\mathbf{F} \in \mathbb{R}^{H \times W \times N}$  are generated by extractor  $\mathcal{F}_1(\mathbf{I}, \Theta_1)$  in the forward propagation, where  $H$ ,  $W$ , and  $N$  denote the height, width, and number of channels of the feature maps, respectively. Afterward, the feature maps  $\mathbf{F}$  are fed into the generator, which consists of a  $3 \times 3$  convolution layer and a Sigmoid activation function for generating a set of foreground prediction maps  $\mathbf{M} \in \mathbb{R}^{H \times W \times C}$  with 0–1 distribution, where  $C$  is the number of categories. We choose the class-specific foreground prediction map  $\mathbf{M}_f \in \mathbb{R}^{H \times W \times 1}$  corresponding to the ground-truth class and invert it to obtain the coupled background prediction map  $\mathbf{M}_b \in \mathbb{R}^{H \times W \times 1}$ , where  $\mathbf{M}_b = 1 - \mathbf{M}_f$ . Finally,  $\mathbf{M}_f$ ,  $\mathbf{M}_b$ , and  $\mathbf{F}$  are fed together into AMC module for prediction map learning. We will detail describe the AMC structure and loss functions in Sect. 3.2.

In the inference phase, as illustrated in Fig. 3, the feature maps  $\mathbf{F}$  obtained by the extractor are input into the generator and sub-network  $\mathcal{F}_2(\mathbf{F}, \Theta_2)$  to generate the foreground prediction maps set  $\mathbf{M}$  and the classification prediction logits  $\tilde{\mathbf{y}}$ , respectively. We select the prediction maps corresponding to the Top- $k$  predicted categories including the ground-truth class, and take their average values as the final localization result. Notably, the Top- $k$  strategy is only used in WSOL and not in WSSS.

### 3.2 Activation Map Constraint

The proposed AMC module utilizes foreground map, background map, and feature maps as input to jointly promote the learning of extractor and generator, which is consisted of four different kinds of losses, including  $\mathcal{L}_{bas}$ ,  $\mathcal{L}_{ac}$ ,  $\mathcal{L}_{frg}$ , and  $\mathcal{L}_{cls}$ .

**Background Activation Suppression ( $\mathcal{L}_{bas}$ ).** For the input background prediction map  $\mathbf{M}_b$ , we multiply it by the feature maps  $\mathbf{F}$  to obtain the background feature maps  $(\mathbf{F} \cdot \mathbf{M}_b)$ , denoted as  $\mathbf{F}^b \in \mathbb{R}^{H \times W \times N}$ . Subsequently, the feature



**Fig. 3** The architecture of the proposed BAS in inference phase. We utilize Top-k to generate final localization map

maps  $\mathbf{F}$  and  $\mathbf{F}^b$  are fed to two sub-networks  $\mathcal{F}_2(\mathbf{F}, \Theta_2)$  and  $\mathcal{F}_2(\mathbf{F}^b, \Theta_2)$  with shared weights, respectively. For the sub-network with  $\mathbf{F}^b$  as input, the goal is to generate the background activation value by the same function, and the parameters of this sub-network are frozen in the back propagation. Following the sub-network  $\mathcal{F}_2(\mathbf{F}, \Theta_2)$  and the global average pooling (GAP) (Zhou et al., 2016),  $\mathbf{F}$  and  $\mathbf{F}^b$  produce the prediction logits  $\tilde{\mathbf{y}} \in \mathbb{R}^C$  and  $\tilde{\mathbf{y}}^b \in \mathbb{R}^C$ , respectively, which can be expressed as follows:

$$\tilde{\mathbf{y}} = \text{GAP}(\mathcal{F}_2(\mathbf{F}, \Theta_2)), \tag{1}$$

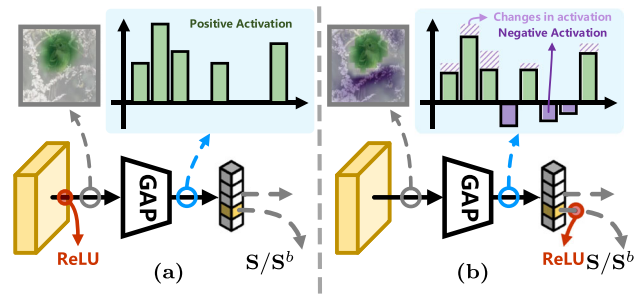
$$\tilde{\mathbf{y}}^b = \text{GAP}(\mathcal{F}_2(\mathbf{F}^b, \Theta_2)). \tag{2}$$

We select the values in the  $\tilde{\mathbf{y}}$  and  $\tilde{\mathbf{y}}^b$  according to the ground-truth class. After applying a ReLU activation function, these values are represented as the activation value  $\mathbf{S} \in \mathbb{R}^1$  and the background activation value  $\mathbf{S}^b \in \mathbb{R}^1$ , respectively.  $\mathbf{S}$  represents the activation value generated by the unmasked feature map, containing both foreground and background information, and  $\mathbf{S}^b$  is the activation value generated by the background feature map, retaining only the background information. Here, we measure the difference between background activation value and activation value in a ratio form as a way to achieve background activation value suppression, and  $\mathcal{L}_{bas}$  is defined as follows:

$$\mathcal{L}_{bas} = \frac{\mathbf{S}^b}{\mathbf{S} + \varepsilon}, \tag{3}$$

where  $\varepsilon$  is a very small value ( $e^{-8}$ ), to ensure that the equation is meaningful. This ratio form not only avoids the addition of more hyperparameters, but also acts as a normalization, so that the range of loss value is maintained under an order of magnitude.

Generating a non-negative  $\mathbf{S}$  and  $\mathbf{S}^b$  is necessary for  $\mathcal{L}_{bas}$ . In the previous conference version, we use a ReLU as the activation function at the end of the network to ensure the non-negativity of the outputs, as shown in Fig. 4. This approach causes pixels with negative values are marked as 0



**Fig. 4** The improvement of BAS. Partial structure of **a** the previous conference version and **b** this work. The green pixels in the localization map indicate positive values and the purple ones indicate negative values (Color figure online)

after ReLU and their gradients will not take part in the back propagation. While pixels with negative values are usually associated with background areas, which are also important for the learning of classification and prediction maps. As shown in Fig. 12, the neglect of negative activation values in the classification loss indirectly causes the BAS loss to become inadequate (the loss value becomes larger instead) later in the training process. To solve this problem, we remove this ReLU layer to make negative pixels also participate in the gradient back propagation. To ensure the non-negativity of  $\mathbf{S}$  and  $\mathbf{S}^b$ , we use the ReLU activation function separately before generating them.

**Area Constraint ( $\mathcal{L}_{ac}$ ).** The background prediction map can be guided by  $\mathcal{L}_{bas}$  in a suppressed way, and a smaller  $\mathcal{L}_{bas}$  means that the region covered by the background prediction map is less discriminative. When the background prediction map can cover the background region well, the  $\mathcal{L}_{bas}$  it produced has to be minimal while the background area should be as large as possible, accordingly, the foreground area should be as small as possible. So we use the foreground prediction map area as constraints:

$$\mathcal{L}_{ac} = \frac{1}{H \times W} \sum_{h=1}^H \sum_{w=1}^W \mathbf{M}_f(h, w). \tag{4}$$

**Foreground Region Guidance ( $\mathcal{L}_{frg}$ ).** Meanwhile, we maintain the FPM’s approach of employing the classification task to drive the learning of foreground prediction maps, which uses high-level semantic information to guide the foreground prediction map to the approximate correct region of the object. Consequently, a foreground region guidance loss based on cross-entropy is utilized. After  $\mathbf{F}$  is fed into  $\mathcal{F}_2(\mathbf{F}, \Theta_2)$ , it is dotted with  $\mathbf{M}_f$  to produce  $\mathcal{L}_{frg}$ :

$$\tilde{\mathbf{y}}^f = \text{GAP}(\mathbf{M}_f \cdot \mathcal{F}_2(\mathbf{F}, \Theta_2)), \tag{5}$$

$$\mathcal{L}_{frg} = - \sum_{i=1}^C \mathbf{y}_i \log \frac{e^{\tilde{\mathbf{y}}_i^f}}{\sum_j^C e^{\tilde{\mathbf{y}}_j^f}}, \tag{6}$$

where  $\mathbf{y}$  denotes the image-level one-hot encoding label.

**Classification ( $\mathcal{L}_{cls}$ ).** Besides, we obtain the classification loss  $\mathcal{L}_{cls}$  by applying cross-entropy to  $\tilde{\mathbf{y}}$ , which is used for classification learning of the entire image:

$$\mathcal{L}_{cls} = - \sum_{i=1}^C \mathbf{y}_i \log \frac{e^{\tilde{y}_i}}{\sum_j^C e^{\tilde{y}_j}}. \quad (7)$$

### 3.3 Weakly Supervised Object Localization

By jointly optimizing background activation suppression loss, area constraint loss, foreground region guidance loss, and classification loss in the AMC module, the foreground prediction map can be guided to the overall area of the object. The total loss of the BAS training process is defined in the following form:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \mathcal{L}_{frg} + \beta \mathcal{L}_{ac} + \lambda \mathcal{L}_{bas}, \quad (8)$$

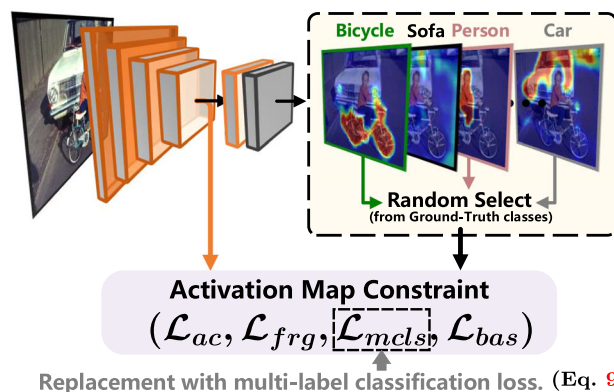
where  $\alpha$ ,  $\beta$ , and  $\lambda$  are hyperparameters,  $\mathcal{L}_{cls}$  and  $\mathcal{L}_{frg}$  are both cross-entropy losses. For all backbones and datasets, we set  $\lambda = 1$ . The ablation experiments of the hyperparameters  $\alpha$ ,  $\beta$ , and  $\lambda$  on WSOL are described in Sect. 4.3.

### 3.4 Weakly Supervised Semantic Segmentation

BAS can also be applied to weakly supervised semantic segmentation to verify the generality of our method. Different from weakly supervised object localization, weakly supervised semantic segmentation no longer assumes that there is only one ground-truth class in an image, which is more challenging. In addition, it is more direct to reflect the segmentation quality of the prediction map by comparing with the weakly supervised semantic segmentation SOTA methods.

Based on the network structure in Fig. 2, we apply BAS to weakly supervised semantic segmentation with minor changes. As shown in Fig. 5, we maintain the learning process for a single prediction map in the AMC module by randomly selecting a foreground category in the image and denoting its corresponding prediction map as  $\mathbf{M}_f$ . In addition, to make the network achieve multi-label classification, we adopt softmax cross-entropy loss and simply modify the form of it instead of using Sigmoid-based loss (binary cross-entropy loss). It mainly due to the activation value  $\mathbf{S}^b$  obtained from the background localization map has to be less than 0 to ensure that the probability generated by  $1/(1 + e^{-\mathbf{S}^b})$  is close to 0, which conflicts with the non-negativity of  $\mathbf{S}^b$ .

**Multi-Label-Classification ( $\mathcal{L}_{mcls}$ ).** For weakly supervised semantic segmentation task, we adopt the multi-label classification loss  $\mathcal{L}_{mcls}$  instead of  $\mathcal{L}_{cls}$  to deal with the multi-label



Replacement with multi-label classification loss. (Eq. 9)

Fig. 5 Applying BAS to weakly supervised semantic segmentation task

case. To avoid the problems of class imbalance and training instability when there are multi-label in the softmax formulation, we only consider the differentiation between foreground and background classes and ignore the interrelationship among foreground categories. It can be expressed as follows:

$$\mathcal{L}_{mcls} = - \sum_{i=1}^L \mathbf{y}_i \log \left( \frac{e^{\tilde{y}_i}}{\sum_j^K e^{\tilde{y}_j} + e^{\tilde{y}_i}} \right), \quad (9)$$

where  $L$  is the set of ground-truth classes in the image, and the remaining set of categories is denoted as  $K$ . The total loss function in weakly supervised semantic segmentation is of the following form:

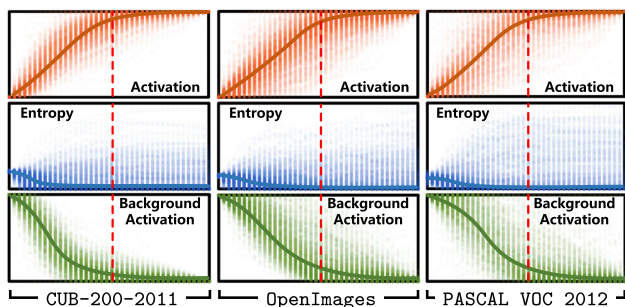
$$\mathcal{L} = \mathcal{L}_{mcls} + \alpha \mathcal{L}_{frg} + \beta \mathcal{L}_{ac} + \lambda \mathcal{L}_{bas}. \quad (10)$$

The  $\lambda$  is set to 1 for all datasets. For PASCAL VOC 2012, we set  $\alpha = 0.2$  and  $\beta = 1.2$ . For MS COCO 2014, we adopt  $\alpha = 0.5$  and  $\beta = 1.5$ . The ablation experiments of the hyperparameters  $\alpha$ ,  $\beta$ , and  $\lambda$  on WSSS, and the results of different combinations of hyperparameters on five datasets are presented in Sect. 5.2.

### 3.5 Empirical Justification

In this part, we empirically justify the advantage of introducing background activation suppression and its generalizability.

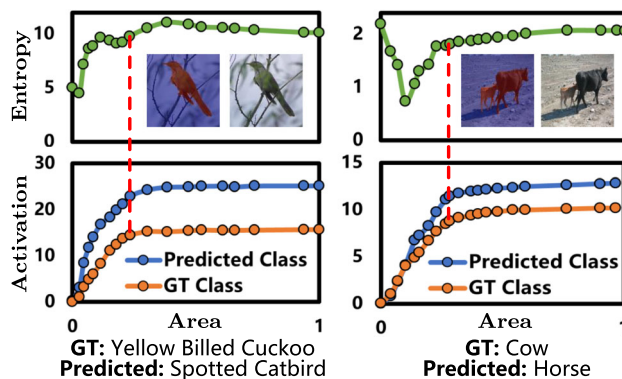
The purpose of the exploratory experiment is to investigate the relationship between activation value (**Activation**), cross-entropy (**Entropy**) and background activation value (**Background Activation**) with the mask area. Specifically, we first train a VGG16 classification network on CUB-200-2011 using  $\mathcal{L}_{cls}$  (Eq. 7) as supervision. Then, for a given pixel-level mask, the activation and entropy corresponding to this mask are generated by masking the feature



**Fig. 6** Motivation. Statistical analysis about exploratory experiments on different datasets

map. We erode and dilate the ground-truth mask with a convolution of kernel size  $5n \times 5n$ , obtain masks with different areas by changing the value of  $n$ , and plot the activation versus entropy with the mask area as the horizontal axis. As shown in Fig. 1A, we display the curve for a single image through the above process.

Due to each image having a different activation value distribution and a different ground-truth mask area, we normalize the activation curve for each image by dividing the activation value generated by the entire image to obtain a more statistically significant result, the same as in Eq. 3. In addition, the area representing the horizontal axis is also normalized based on the ground-truth mask area, which is marked by a red line. As shown in Fig. 6, we present the curves of foreground activation value, cross-entropy, and background activation value with respect to the mask area, which are counted on the CUB-200-2011 test set. It can be noted that the samples on the whole present the following phenomena: When the mask expands near the ground-truth mask, the activation value starts to saturate and the corresponding background activation value tends to converge, while cross-entropy converges to zero early or even diverges with the expansion of the mask. This suggests that the object region learned by activation values is larger and closer to the real object region than that learned by cross-entropy. We further explore why the cross-entropy occasionally diverges and visualize some results as shown in Fig. 7. It can be noted that when the network classifies objects incorrectly, such as identifying cows as horses, the calculated cross-entropy maintains a high value as the mask area increases. In this case, adopting cross-entropy values to supervise the localization map is less feasible and appropriate than using activation values which are not influenced by other categories. Besides, to verify the generality of this observation, we perform the same experiments on the more complex OpenImages and PASCAL VOC 2012 datasets. For PASCAL VOC 2012, we select one ground-truth category and its corresponding mask at a time, convert the multi-label into single-label, and then plot the curve in the same way. As shown in Fig. 6, the



**Fig. 7** Cross-entropy presents a divergence trend as the area of the mask increases when the model classifies the object incorrectly. The dashed line represents the position of ground-truth mask. Entropy: cross-entropy. GT: Ground-Truth

statistical analysis demonstrates similar phenomena, therefore, we believe it is general that better localization ability can be learned through activation values compared to cross-entropy.

## 4 Experiments on Weakly Supervised Object Localization

### 4.1 Experimental Setup

**Datasets.** We evaluate the proposed method on the popular benchmarks including **CUB-200-2011** (Wah et al., 2011), **ILSVRC** (Russakovsky et al., 2015), and **OpenImages** (Choe et al., 2020b). CUB-200-2011 contains 200 fine-grain classes of birds with 5994 training images and 5794 testing images. ILSVRC contains about 1.2 million training images and 50,000 validation images, which are divided into 1000 categories. OpenImages consists of 29,819, 2500 and 5000 samples from 100 classes for training, validation and test, respectively. Except for class labels, CUB-200-2011 and OpenImages also provide pixel-level mask annotations for the evaluation of the prediction mask.

**Metrics.** Following DA-WSOL (Zhu et al., 2022), we apply both bounding box and mask metrics to evaluate the performance of our BAS. For bounding box, following Xu et al. (2022); Zhu et al. (2022); Lee et al. (2022a), four metrics are used for evaluation, including GT-known localization accuracy (**GT-known Loc**), Top-1 localization accuracy (**Top-1 Loc**), Top-5 localization accuracy (**Top-5 Loc**), and maximal box accuracy (**MaxBoxAccV2**). Specifically, GT-known Loc is correct when the intersection over union (IoU) between the ground-truth bounding box and the predicted bounding box is greater than a fixed IoU threshold ( $\delta = 0.5$ ). Top-1/Top-5 Loc is correct when the Top-1/Top-5 predicted categories con-



**Table 1** Comparison with state-of-the-art methods

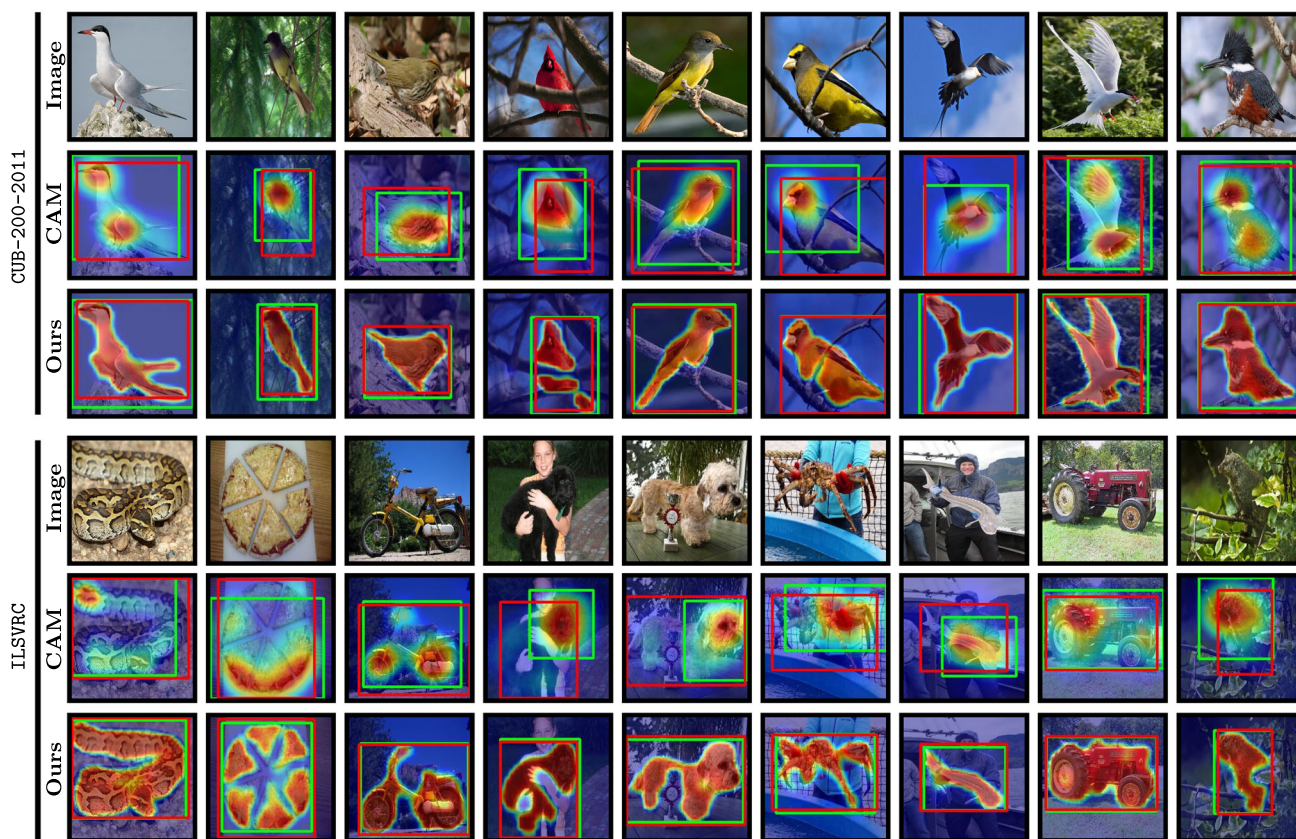
Methods	Venue	Backbone	CUB-200-2011 Loc Acc.			ILSVRC Loc Acc.		
			Top-1 Loc	Top-5 Loc	GT-k.	Top-1 Loc	Top-5 Loc	GT-k.
CAM (Zhou et al., 2016)	CVPR16	VGG16	41.06	50.66	55.10	42.80	54.86	59.00
ACoL (Zhang et al., 2018a)	CVPR18	VGG16	45.92	56.51	62.96	45.83	59.43	62.96
ADL (Choe et al., 2020a)	TPAMI20	VGG16	52.36	–	75.41	44.92	–	–
I2C (Zhang et al., 2020d)	ECCV20	VGG16	55.99	68.34	–	47.41	58.51	63.90
MEIL (Mai et al., 2020)	CVPR20	VGG16	57.46	–	73.84	46.81	–	–
PSOL (Zhang et al., 2020a) $\diamond$	CVPR20	VGG16 $\times$ 2	66.30	84.05	89.11	50.89	60.90	64.03
SPA (Pan et al., 2021)	CVPR21	VGG16	60.27	72.50	77.29	49.56	61.32	65.05
SLT (Guo et al., 2021) $\diamond$	CVPR21	VGG16 $\times$ 3	67.80	–	87.60	51.20	62.40	67.20
FAM (Meng et al., 2021)	ICCV21	VGG16	69.26	–	89.26	51.96	–	<b>71.73</b>
ORNet (Xie et al., 2021) $\diamond$	ICCV21	VGG16 $\times$ 2	67.73	80.77	86.20	52.05	63.94	68.27
Kim et al. (Kim et al., 2022)	CVPR22	VGG16	<u>70.83</u>	<b>88.07</b>	<b>93.17</b>	49.94	63.25	68.92
CREAM (Xu et al., 2022) $\diamond$	CVPR22	VGG16 $\times$ 2	70.44	<u>85.67</u>	90.98	<u>52.37</u>	<u>64.20</u>	68.32
<b>BAS (ours)</b>	This Work	VGG16	<b>70.90</b>	85.36	<u>91.04</u>	<b>52.94</b>	<b>65.38</b>	<u>69.66</u>
CAM (Zhou et al., 2016)	CVPR16	MobileNetV1	48.07	<u>59.20</u>	63.30	43.35	<u>54.44</u>	58.97
HaS (Singh & Lee, 2017)	ICCV17	MobileNetV1	46.70	–	67.31	42.73	–	60.12
ADL (Choe et al., 2020a)	TPAMI20	MobileNetV1	47.74	–	–	43.01	–	–
RCAM (Bae et al., 2020)	ECCV20	MobileNetV1	59.41	–	78.60	44.78	–	61.69
FAM (Meng et al., 2021)	ICCV21	MobileNetV1	<u>65.67</u>	–	<u>85.71</u>	<u>46.24</u>	–	<u>62.05</u>
<b>BAS (ours)</b>	This Work	MobileNetV1	<b>70.54</b>	<b>86.71</b>	<b>93.04</b>	<b>53.05</b>	<b>66.68</b>	<b>72.03</b>
CAM (Zhou et al., 2016)	CVPR16	ResNet50	46.71	54.44	57.35	48.69	58.00	60.58
ADL (Choe et al., 2020a)	TPAMI20	ResNet50	62.29	–	–	48.53	–	–
PSOL (Zhang et al., 2020a) $\diamond$	CVPR20	ResNet50 $\times$ 2	70.68	86.64	90.00	53.98	63.08	65.44
FAM (Meng et al., 2021)	ICCV21	ResNet50	73.74	–	85.73	54.46	–	64.56
DA-WSOL (Zhu et al., 2022)	CVPR22	ResNet50 $\times$ 2	66.65	–	81.83	<u>55.84</u>	–	<u>70.27</u>
Kim et al. (Kim et al., 2022)	CVPR22	ResNet50	73.16	<u>86.68</u>	<u>91.60</u>	53.76	<u>65.75</u>	69.89
CREAM (Xu et al., 2022) $\diamond$	CVPR22	ResNet50 $\times$ 2	<u>76.03</u>	–	89.88	55.66	–	69.31
<b>BAS (ours)</b>	This Work	ResNet50	<b>76.75</b>	<b>90.04</b>	<b>95.41</b>	<b>57.46</b>	<b>68.57</b>	<b>72.00</b>
CAM (Zhou et al., 2016)	CVPR16	InceptionV3	41.06	50.66	55.10	46.29	58.19	62.68
DANet (Xue et al., 2019)	ICCV19	InceptionV3	49.45	60.46	67.03	47.53	58.28	–
I2C (Zhang et al., 2020d)	ECCV20	InceptionV3	55.99	68.34	72.60	53.11	64.13	68.50
GCNet (Lu et al., 2020)	ECCV20	InceptionV3	58.58	71.00	75.30	49.06	58.09	–
PSOL (Zhang et al., 2020a) $\diamond$	CVPR20	InceptionV3 $\times$ 2	65.51	83.44	–	54.82	63.25	65.21
SPA (Pan et al., 2021)	CVPR21	InceptionV3	53.59	66.50	72.14	52.73	64.27	68.33
SLT (Guo et al., 2021) $\diamond$	CVPR21	InceptionV3 $\times$ 3	66.10	–	86.50	55.70	65.40	67.60
FAM (Meng et al., 2021)	ICCV21	InceptionV3	70.67	–	87.25	55.24	–	68.62
CREAM (Xu et al., 2022) $\diamond$	CVPR22	InceptionV3 $\times$ 2	<u>71.76</u>	<u>86.37</u>	<u>90.43</u>	<u>56.07</u>	<u>66.19</u>	<u>69.03</u>
<b>BAS (ours)</b>	This Work	InceptionV3	<b>72.09</b>	<b>88.11</b>	<b>94.63</b>	<b>58.50</b>	<b>69.03</b>	<b>72.07</b>

Best results are highlighted in bold, second are underlined.  $\diamond$  means multi-stage model.  $\times n$  means that there are  $n$  different networks used

tain the ground-truth class and the GT-known Loc is correct. MaxBoxAccV2 compared to GT-known ( $\delta = 0.5$ ) considers multiple IoU thresholds ( $\delta \in \{0.3, 0.5, 0.7\}$ ) and takes the average localization performance as the result. For mask, we adopt both the peak intersection over union (**PIoU**) (Zhang et al., 2020a) and the pixel average precision (**PxAP**) (Choe

et al., 2020b) as metrics when the pixel-level ground-truth label is available.

*Implementation Details.* We evaluate the proposed method on the most popular backbones, including VGG16 (Simonyan & Zisserman, 2014), InceptionV3 (Szegedy et al., 2016), ResNet50 (He et al., 2016), and MobileNetV1 (Howard et al., 2017). All networks are fine-tuned on the pre-trained



**Fig. 8** Visualization comparison with the baseline CAM (Zhou et al., 2016) method on **CUB-200-2011** (Wah et al., 2011) and **ILSVRC** (Russakovsky et al., 2015). The ground-truth bounding boxes are in Red, and the predictions are in Green (Color figure online)

weights of ILSVRC (Russakovsky et al., 2015). We train 120 epochs on the CUB-200-2011 (Wah et al., 2011) and 9 epochs on ILSVRC (Russakovsky et al., 2015). In the training phase, the input images are resized to  $256 \times 256$  and then randomly cropped to  $224 \times 224$ . When  $\mathcal{L}_{bas}$  is larger than 1, we mark it as 1, to ensure the stability of the initial training. In the inference phase, we use ten crop augmentation to get the final classification results following the settings in Pan et al. (2021), Guo et al. (2021), Zhang et al. (2018b). For localization, we replace the random crop with the center crop, as in previous works (Wei et al., 2021; Zhang et al., 2020a; Yun et al., 2019; Choe & Shim, 2019).

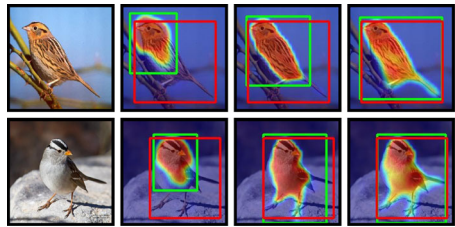
## 4.2 Comparison with State-of-the-Arts

We compare the proposed BAS with state-of-the-art methods on CUB-200-2011 (Wah et al., 2011) and ILSVRC (Russakovsky et al., 2015) datasets. As shown in Table 1, BAS achieves stable and excellent performance on various backbones. On CUB-200-2011 (Wah et al., 2011), BAS surpasses all existing methods by a large margin in terms of GT-known/Top-1/Top-5 Loc when the backbones are MobileNetV1, ResNet50 and InceptionV3. Compared

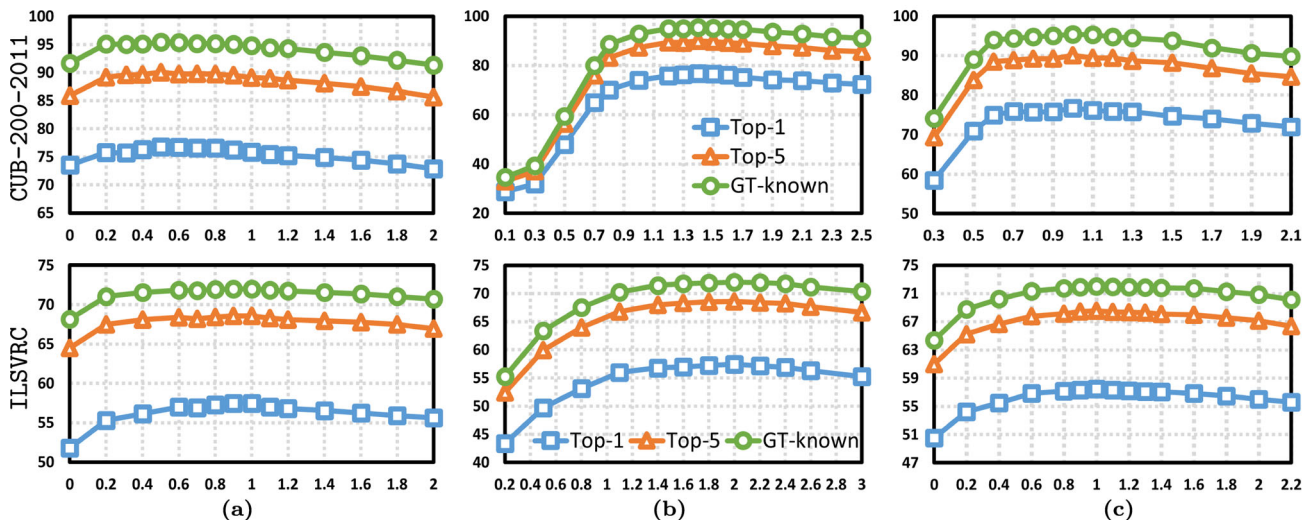
with the current Foreground-Prediction-Map-based method FAM (Meng et al., 2021), BAS achieves **1.78%**, **7.33%**, **9.68%** and **7.38%** improvement on VGG16, MobileNetV1, ResNet50 and InceptionV3 in terms of GT-known Loc, respectively. On ResNet50, BAS achieves **95.41%** GT-known Loc, which is a significant increase of **3.81%** compared to the best performing counterpart Kim et al. (Kim et al., 2022). In addition, our method improves **5.53%** and **4.20%** GT-known Loc compared to the latest multi-stage model CREAM (Xu et al., 2022) on ResNet50 and InceptionV3, respectively.

On ILSVRC (Russakovsky et al., 2015), BAS overall exceeds all baseline methods in terms of GT-known/Top-1/Top-5 Loc on all backbones. When MobileNetV1 is used as the backbone, our BAS achieves **72.03%** GT-known Loc, surpassing FAM (Meng et al., 2021) by a large margin with a **9.98%** improvement. Moreover, InceptionV3-BAS and ResNet50-BAS obtain **72.07%** and **72.00%** GT-known Loc, respectively, establishing a novel state-of-the-art. It shows that BAS performs well on both fine-grained dataset and large universal dataset. Furthermore, we visualize the localization maps of the proposed BAS and CAM (Zhou et al., 2016) on CUB-200-2011 and ILSVRC in Fig. 8. Compared to

**Table 2** Ablation study

	(a)	(b)	(c)	
Baseline	✓	✓	✓	
$\mathcal{L}_{bas}$		✓	✓	
Top- $k$			✓	
Top-1 Loc	57.89	74.15	<b>76.75</b>	
Top-5 Loc	67.48	86.88	<b>90.04</b>	
GT-known	71.14	92.15	<b>95.41</b>	

Bold values indicate the best results among all methods  
 (a) the baseline method. (b) add  $\mathcal{L}_{bas}$  to the baseline. (c) synthesize the prediction maps with Top- $k$  strategy



**Fig. 9** Hyperparameters. **a**  $\alpha$  for foreground region guidance loss  $\mathcal{L}_{frg}$ . **b**  $\beta$  for area constraint loss  $\mathcal{L}_{ac}$ . **c**  $\lambda$  for background activation suppression loss  $\mathcal{L}_{bas}$

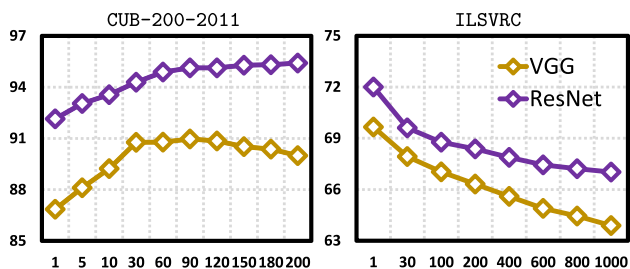
CAM, BAS can robustly cover the entire area of the object even in noisy environments and is sharper and more compact at the edges of the object.

**4.3 Ablation Study**

In this section, we perform a series of ablation experiments using ResNet50 (Simonyan & Zisserman, 2014) as the backbone. Above all, we conduct ablation experiments on various components of BAS on CUB-200-2011 (Wah et al., 2011). We take  $\mathcal{L}_{cls}$ ,  $\mathcal{L}_{frg}$  and  $\mathcal{L}_{ac}$  together as the baseline method for the Foreground-Prediction-Map-based architecture. As shown in Table 2, the addition of  $\mathcal{L}_{bas}$  to the baseline can enable the localization map to cover the object region more completely, thus significantly increasing the localization accuracy, with **21.01%** and **16.26%** improvement in terms of GT-known Loc and Top-1 Loc, respectively. Moreover, using Top- $k$  strategy to integrate the final localization result, though making the localization result not as sharp as before, it can further improve the GT-known Loc (from

**92.15% to 95.41%**) by alleviating the problem of the classification network focusing on the distinguish parts.

*Hyperparameter  $\alpha$ ,  $\beta$ , and  $\lambda$  in total loss* There are three hyperparameters in Eq. 8. Their effectiveness and sensitivity analyses for localization quality are performed on CUB-200-2011 and ILSVRC in Fig. 9. The  $\alpha$  denotes the factor of  $\mathcal{L}_{frg}$ , and it can be noticed from Fig. 9a that the presence of foreground region guidance loss ( $\alpha \geq 0.2$ ) can significantly improve the localization accuracy by ensuring stable learning of foreground activation maps on both datasets. The  $\beta$  reflects the degree of constraint between foreground area and background suppression. When  $\beta$  is small, more areas in the foreground activation map are activated, while when  $\beta$  is too large, it will suppress the learning of the activation map. As shown in Fig. 9b, our method performs stably with high accuracy when  $\beta$  varies from 1.2 to 1.7 on CUB-200-2011 and from 1.6 to 2.4 on ILSVRC. The  $\lambda$  denotes the factor of  $\mathcal{L}_{bas}$ . A larger  $\lambda$  indicates that more regions in the prediction map are activated by background activation suppression. As shown in Fig. 9c, the localization



**Fig. 10** GT-known Loc (%) *w.r.t* *k*. Evaluation results of combining the Top-*k* prediction maps when the backbone is VGG16 and ResNet50 respectively

accuracy continues to grow on CUB-200-2011 when  $\lambda$  increases from 0.3 to 0.6 and remains stable from 0.6 to 1.3 with less than 1% change in GT-known Loc, which shows that the proposed BAS approach can significantly improve the localization accuracy. In summary, although we have three hyperparameters in the loss function, it is easy to choose suitable values for the hyperparameters  $\alpha$ ,  $\beta$ , and  $\lambda$ . In addition, we also provide the results of different combinations of hyperparameters on CUB-200-2011 and ILSVRC in Sect. 5.2.

*Hyperparameter k in Top-k strategy.* We evaluate the effect of the hyperparameter *k* in our BAS. As shown in Fig. 10, the accuracy of GT-known Loc is improved on CUB-200-2011 when  $k > 1$ , comparing  $k = 1$ . For VGG16 and ResNet50, the highest localization accuracy is achieved at *k* of 80 and 200, respectively. It suggests that the Top-*k* strategy can be used to obtain more complete localization results and further improve the localization performance by integrating the localization maps of similar categories on CUB-200-2011. In contrast, for both VGG16 and ResNet50, the best localization results are obtained for  $k = 1$  on ILSVRC dataset, which shows a high variability of classes and few localization features of similarity between categories on ILSVRC.

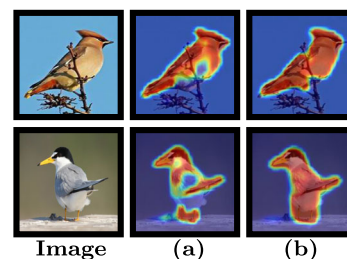
*Generator after different layers.* We report the results of inserting the generator after different layers of ResNet50. As shown in Table 3 (left table), quantitative experiment indicates that inserting the generator after **layer 3** achieves the

**Table 3** Localization accuracy and visualization results about inserting the generator after different layers on ResNet50

	Top-1 Loc	Top-5 Loc	GT-known	
Layer 1	42.47	49.65	52.87	
Layer 2	69.24	81.24	86.16	
Layer 3	<b>76.75</b>	<b>90.04</b>	<b>95.41</b>	
Layer 4	71.63	84.81	90.94	

Bold values indicate the best results among all methods

Loc Acc	(a)	(b)
Top-1 Loc	74.97	<b>76.75</b>
Top-5 Loc	87.78	<b>90.04</b>
GT-known	93.75	<b>95.41</b>



**Fig. 11** Comparison of background prediction maps learned from a original image or b feature maps

best results and is significantly better than other positions. The prediction maps learned from different layers are visualized in Table 3 (right figure). When the generator learns localization information from shallow feature maps (**layer 1** and **layer 2**), the prediction map performs better at the edges of objects, but it is insufficient to resist background distractions and has poor semantic learning ability. In addition, the generator learns localization information from the high-level feature (**layer 4**) resulting in imprecise localization due to the limitation of resolution.

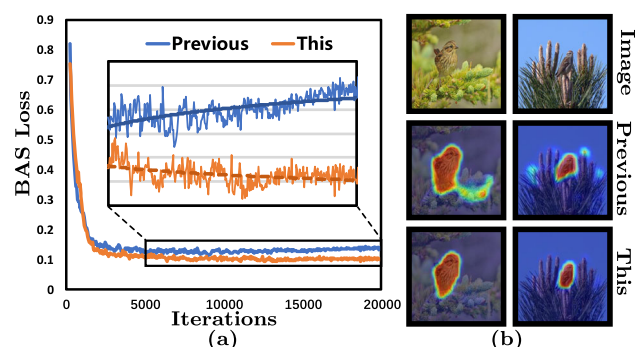
*Original image vs feature maps.* We fix the generator after layer 3 and conduct experiments on the masking position (original image *vs* feature maps) of the background prediction map. As illustrated in Fig. 11, it can be noted that the masking feature maps approach achieves higher accuracy and better coverage of the localization results on the object, while the results generated by the approach of masking the original image focus more on the edge or texture of the object and have less ability to locate smooth regions. It may be because the learning process in shallow layers usually focuses on common basic features (e.g., edges, textures) and ignores high-level semantic features (Table 4).

*Comparison with previous conference version.* We compare the improvement over the conference version in both quantitative and qualitative aspects. Benefiting from the adjustment to the position of the ReLU activation function, BAS can learn the feature map more adequately and efficiently. As shown

**Table 4** Evaluation results in terms of MaxBoxAccV2 on the CUB-200-2011 and ILSVRC datasets using various backbones

Methods	Venue	CUB-200-2011 (MaxBoxAccV2)				ILSVRC (MaxBoxAccV2)			
		VGG	Inception	ResNet	Mean	VGG	Inception	ResNet	Mean
CAM (Zhou et al., 2016)	CVPR16	63.7	56.7	63.0	61.1	60.0	63.4	63.7	62.4
HaS (Singh & Lee, 2017)	ICCV17	63.7	53.4	64.7	60.6	60.6	63.7	63.4	62.6
ACoL (Zhang et al., 2018a)	CVPR18	57.4	56.2	66.5	60.0	57.4	63.7	62.3	61.1
SPG (Zhang et al., 2018b)	ECCV18	56.3	55.9	60.4	57.5	59.9	63.3	63.3	62.2
CutMix (Yun et al., 2019)	ICCV19	62.3	57.5	62.8	60.8	59.4	63.9	63.3	62.2
ADL (Choe et al., 2020a)	TPAMI20	66.3	58.8	58.3	61.1	59.8	61.4	63.7	61.7
IVR (Kim et al., 2021)	ICCV21	65.2	60.8	66.9	64.3	61.5	65.5	65.6	64.2
DA-WSOL (Zhu et al., 2022)	CVPR22	–	68.0	69.9	–	–	64.8	68.2	–
CREAM (Xu et al., 2022)	CVPR22	71.5	64.2	73.5	69.7	66.2	<b>68.9</b>	67.4	67.5
Kim et al. (Kim et al., 2022)	CVPR22	80.1	–	75.9	–	66.6	–	68.7	–
C2AM (Xie et al., 2022b)	CVPR22	81.4	82.4	83.8	82.5	66.3	65.8	66.8	66.3
<b>BAS (ours)</b>	This Work	<b>83.5</b>	<b>82.7</b>	<b>89.4</b>	<b>85.2</b>	<b>68.2</b>	<b>68.9</b>	<b>68.8</b>	<b>68.6</b>

Bold values indicate the best results among all methods



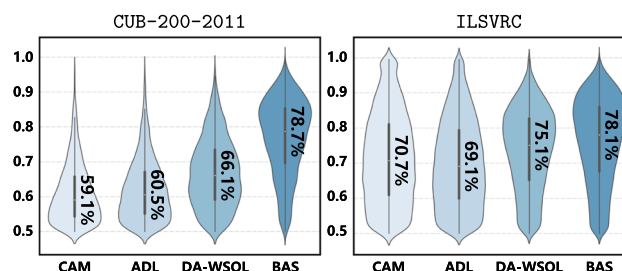
**Fig. 12** Experimental comparison between the previous conference version and this work. **a** The  $\mathcal{L}_{bas}$  training loss curves. **b** Visualization of the localization results

in Fig. 12a, we display the curve of  $\mathcal{L}_{bas}$  (Eq. 3) training loss with the training iterations for both previous conference version and this work. It can be observed that the loss curve (this work) converges to a lower point and shows a more stable convergence trend, while the loss curve in the previous conference version even presents an increasing trend during the iterations. It indicates that the ReLU in the last layer (Fig. 4) makes the classification network learn the background region insufficiently, hence resulting in the inadequate convergence of BAS loss. Figure 12b illustrates some localization maps to support this analysis. Compared with the previous conference version, BAS (this work) demonstrates more robustness in the learning of the background region and consequently improves the localization accuracy in Table 5. We achieve an average of **0.83%** and **0.11%** GT-known Loc gains on the four backbone networks on CUB-200-2011 and ILSVRC, respectively, without additional parameters and computations.

**Table 5** Improvement in GT-known Loc compared to the previous conference version

Backbone	CUB-200-2011	ILSVRC
VGG16	91.04 (−0.03)	69.66 (+0.02)
MobileNetV1	93.04 (+0.69)	72.03 (+0.03)
ResNet50	95.41 (+0.28)	72.00 (+0.23)
InceptionV3	94.63 (+2.39)	72.07 (+0.14)
<b>Mean</b>	93.53 (+0.83)	71.44 (+0.11)

Bold values indicate the best results among all methods



**Fig. 13** Statistical analysis of correct bounding boxes, based on ResNet50 (CAM (Zhou et al., 2016), ADL (Choe et al., 2020a), and DA-WSOL (Zhu et al., 2022))

#### 4.4 Performance Analysis

In this section, we evaluate and analyze in detail the localization quality and segmentation quality of BAS.

**Localization Quality.** Table 4 shows the MaxBoxAccv2 scores compared with other methods on CUB-200-2011 (Wah et al., 2011) and ILSVRC (Russakovsky et al., 2015). Quantitative experiments indicate that our method achieves the best results for different backbone networks and datasets under the MaxBoxAccv2 criterion, which proves the high

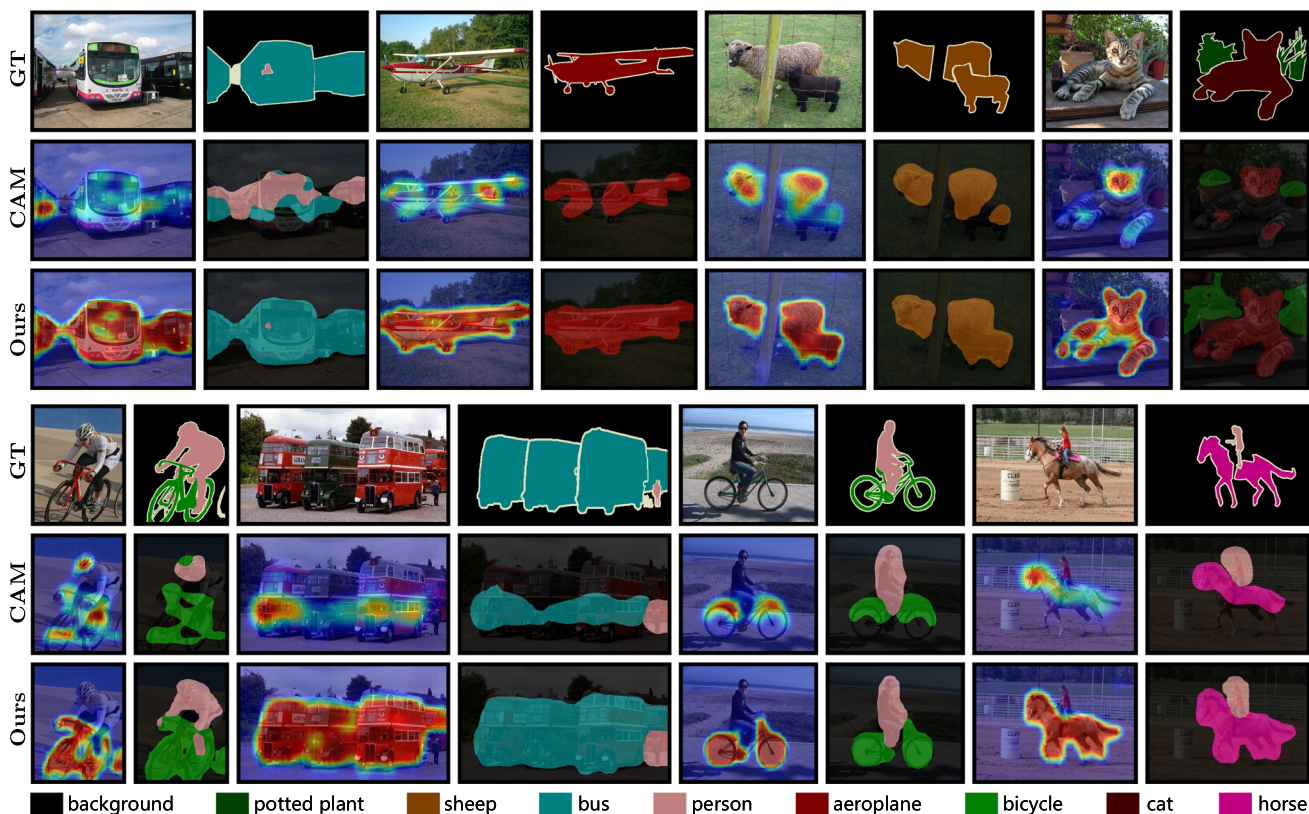


Fig. 14 Visualization of the initial seed generated by CAM and the proposed BAS on the PASCAL VOC 2012 dataset

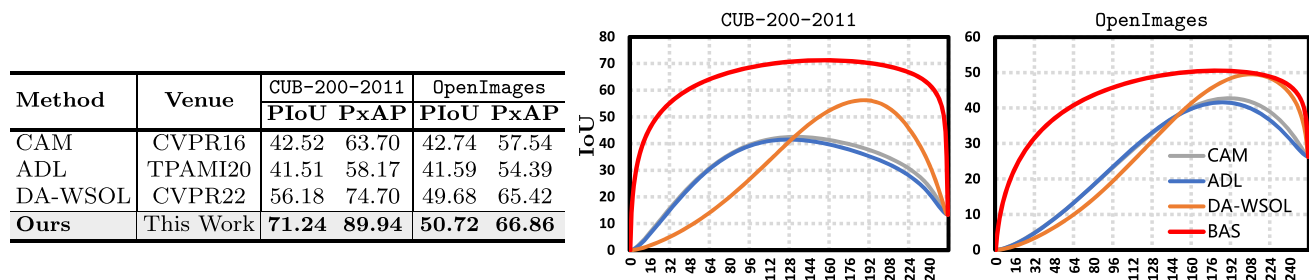


Fig. 15 Segmentation Quality. IoU-Threshold curves for different baseline methods and evaluation results of PIoU, PxAP on CUB-200-2011 (Wah et al., 2011) and OpenImages (Choe et

al., 2020b) datasets, based on ResNet50 (CAM (Zhou et al., 2016), ADL (Choe et al., 2020a), and DA-WSOL (Zhu et al., 2022))

quality of the bounding box generated by BAS and verifies the effectiveness and generalizability of the proposed method. In particular, on CUB-200-2011, we exceed the previous best methods by 2.1% and 5.6% when the backbone networks are VGG16 and ResNet50, respectively. Besides, in Fig. 13, we demonstrate the statistical analysis of IoU based on ResNet50, which plots the IoU distribution curves between the bounding boxes and the ground-truth boxes when localized correctly, following DANet (Xue et al., 2019). On CUB-200-2011, we achieve 78.7% IoU median, exceeding the latest state-of-the-art method DA-WSOL (Zhu et al., 2022) by 12.6%, and correspondingly

by 3.0% on ILSVRC. From the median IoU and the IoU distribution, it can be seen that the proposed BAS significantly improves the localization quality on both CUB-200-2011 and ILSVRC datasets (Fig. 14).

*Segmentation Quality.* We compare the localization map with the ground-truth mask label using two metrics, PIoU and PxAP, following DA-WSOL (Zhu et al., 2022). As shown in Fig. 15 (left table), we evaluate the performance of the proposed BAS with CAM (Zhou et al., 2016), ADL (Choe et al., 2020a) and DA-WSOL (Zhu et al., 2022) on ResNet50. Compared to DA-WSOL, BAS achieves significant and consistent improvement, with a 15.06% increase

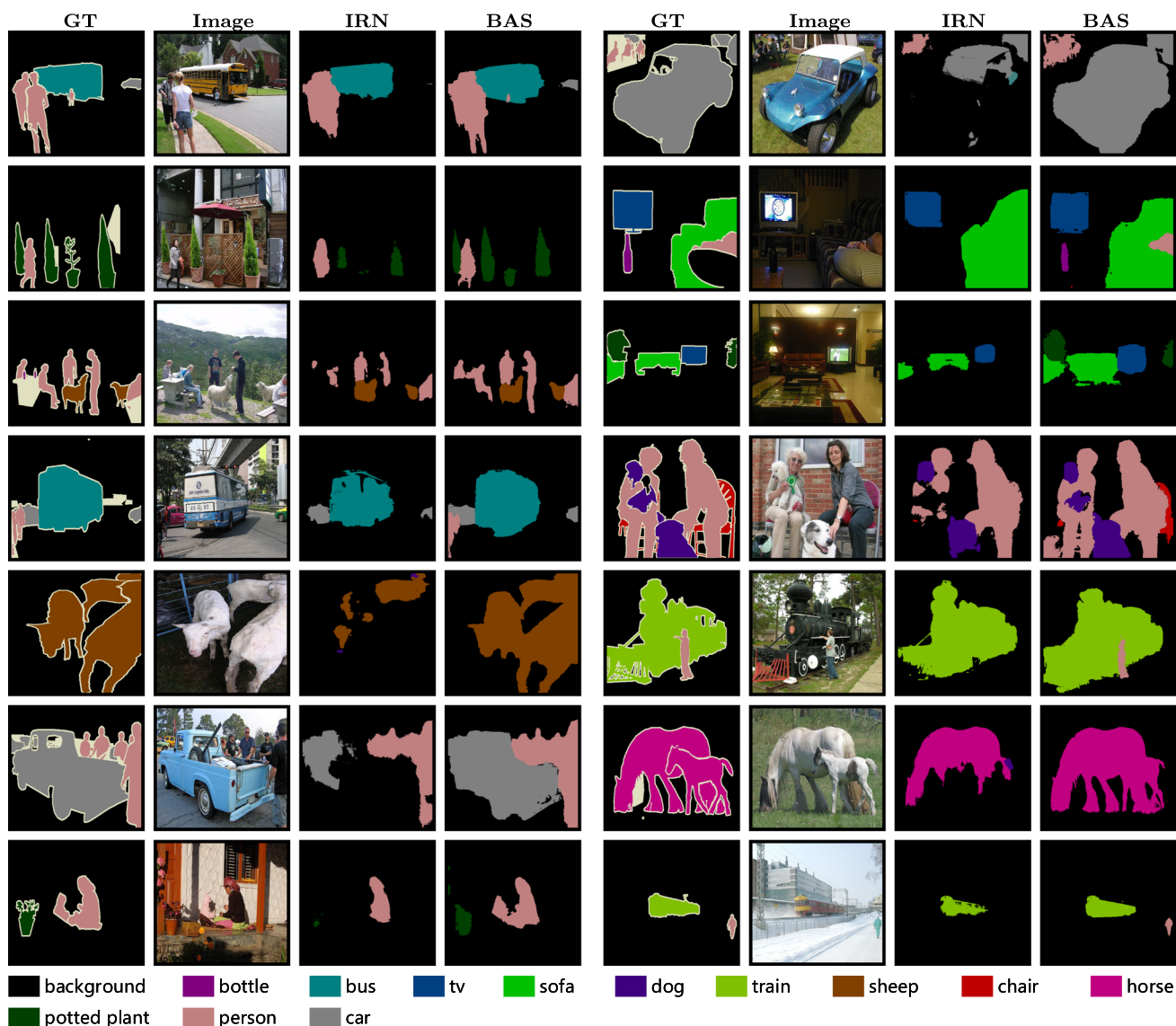


Fig. 16 Examples of semantic segmentation results on PASCAL VOC 2012 for IRN and BAS (with IRN)

in PIoU and **15.24%** in PxAP on CUB-200-2011. The proposed method also surpasses all methods on OpenImages, although OpenImages is a more challenging dataset due to a large number of small objects and complex backgrounds. In addition, we present the IoU-Threshold curves in the right graph of Fig. 15, which represent the IoU values at varying thresholds within the range of  $[0, 255]$ . As observed from the IoU-Threshold curves on both datasets, our method demonstrates a lower sensitivity to the thresholds and achieves better results at arbitrary threshold compared to other methods, which indicates that the localization map produced by BAS has fewer low confidence regions and is closer to the ground-truth object region.

## 5 Experiments on Weakly Supervised Semantic Segmentation

### 5.1 Experimental Setup

*Datasets and Evaluation Metric.* To evaluate the performance of BAS on weakly supervised semantic segmentation task, we conduct experiments on the commonly used **PASCAL VOC 2012** (Everingham et al., 2010) and **MS COCO 2014** (Lin et al., 2014) datasets. PASCAL VOC 2012 contains 21 categories (including one background class). It has 1464, 1449, and 1456 samples in training, val, and test sets, respectively. Following the common experimental protocol (Chen et al., 2014), the training set is augmented with 10,582 weakly annotated images provided

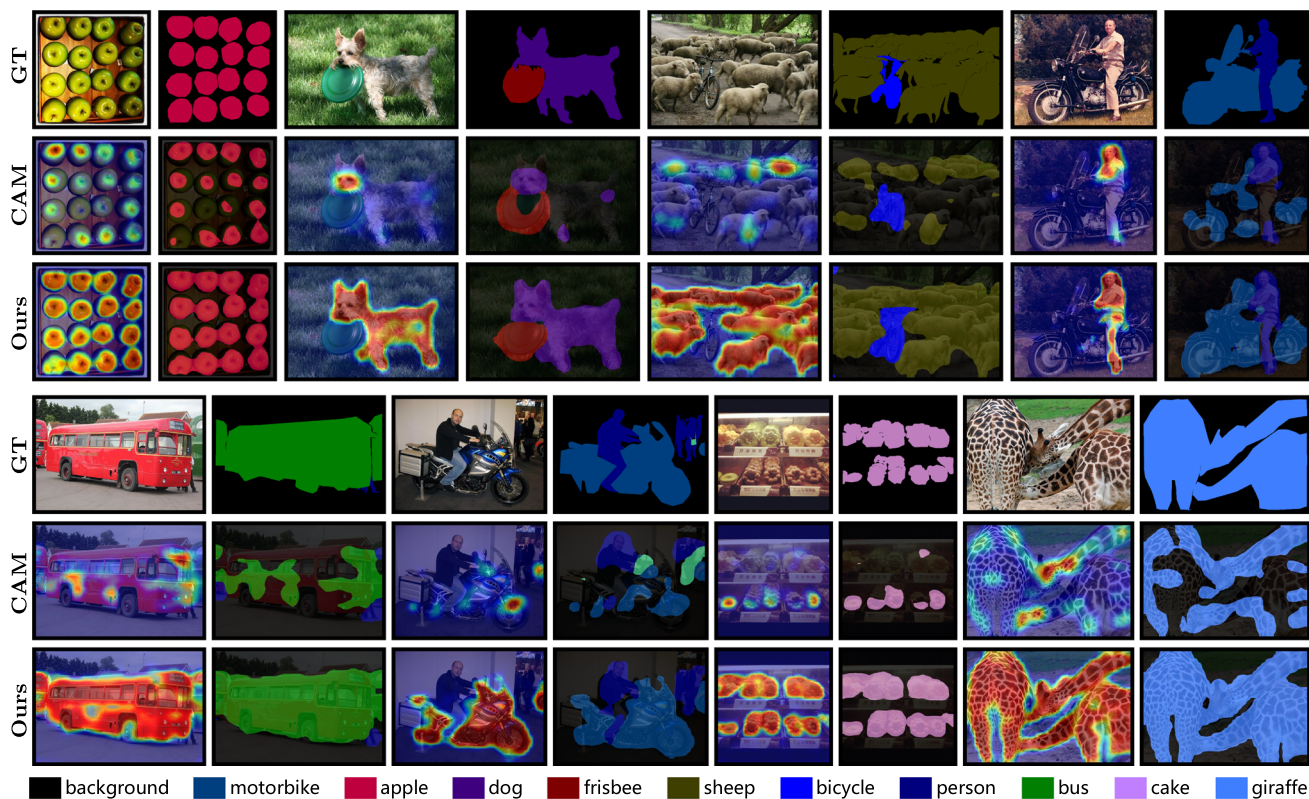


Fig. 17 Visualization of the initial seed generated by CAM and the proposed BAS on the MC COCO 2012 dataset

by SBD dataset (Hariharan et al., 2011). MS COCO 2014 dataset has 81 semantic classes (including one background class). Following Lee et al. (2022a), Jiang et al. (2022), images without the target categories are moved off the dataset, remaining 82,081 training images and 40,137 validation images. We use the mean Intersection-over-Union (mIoU) as the evaluation metric for all experiments (Figs. 16, 17).

**Implementation Details.** For seed generation, the input image is resized to  $512 \times 512$ , then augmented by horizontal flipping and random cropping to  $448 \times 448$ . We train the network for 10 epochs. Batch size is set to 16 and 64 on PASCAL VOC 2012 and MS COCO 2014 respectively. To optimize the network, SGD optimizer is adopted with momentum mechanism and the momentum coefficient is set to 0.9. The initial learning rate is set as 0.005 and decayed following the poly policy  $lr_{init} = lr_{init}(1 - itr/max\_itr)^\rho$  with  $\rho = 0.9$ . Following Lee et al. (2022a), Xie et al. (2022a), we use ResNet50 as the backbone network to generate the initial seed for both PASCAL VOC 2012 and MS COCO 2014 datasets.

**Seed Refinement and Segmentation.** For seed refinement, to make a fair comparison, we follow Lee et al. (2022a), Lee et al. (2021a), Chen et al. (2022b) using IRN (Ahn et al., 2019) to improve the quality of the initial seed. After

Table 6 Ablation study for the components of BAS on PASCAL VOC 2012 and MS COCO 2014

Baseline	$\mathcal{L}_{bas}$	PASCAL	COCO
✓		50.1	32.5
✓	✓	<b>57.7</b>	<b>36.9</b>

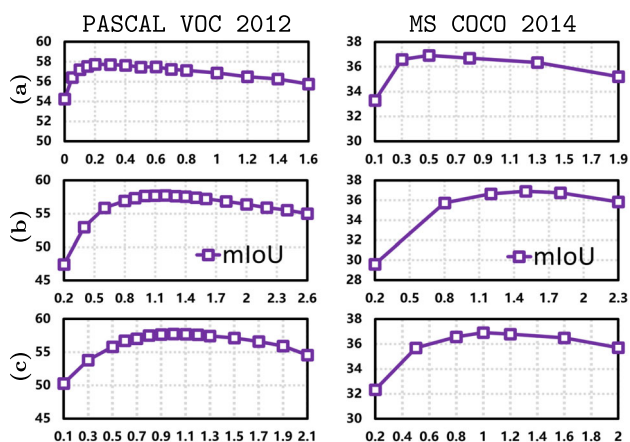
Bold values indicate the best results among all methods

generating pseudo masks, we select DeepLabV2 (Chen et al., 2017) with ResNet-101 (He et al., 2016) as the segmentation network, following Xie et al. (2022a), Jo and Yu (2021). We adopt the default setting to train DeepLabV2 as in Lee et al. (2022a) with weights pretrained on MS COCO 2014.

### 5.2 Ablation Study

In this section, we perform a series of ablation experiments with ResNet50 as the backbone on PASCAL VOC 2012 and MS COCO 2014. We first execute an ablation study regarding the loss composition of the BAS, and as in Sect. 4.3, we take  $\mathcal{L}_{cls}$ ,  $\mathcal{L}_{frg}$ , and  $\mathcal{L}_{ac}$  together as the baseline for the Foreground-Prediction-Map-based architecture. It can be seen from Table 6 that the addition of  $\mathcal{L}_{bas}$  can significantly improve the segmentation quality of baseline with 7.6% and





**Fig. 18** Hyperparameters. **a**  $\alpha$  for foreground region guidance loss  $\mathcal{L}_{frg}$ . **b**  $\beta$  for area constraint loss  $\mathcal{L}_{ac}$ . **c**  $\lambda$  for background activation suppression loss  $\mathcal{L}_{bas}$

**4.4% mIoU gains** on PASCAL VOC 2012 and MS COCO 2014, respectively, which verifies the effectiveness of the proposed  $\mathcal{L}_{bas}$  in capturing object regions relevant to classification.

*Hyperparameter  $\alpha$ ,  $\beta$ , and  $\lambda$  in total loss.* Figure 18 illustrates the sensitivity of the segmentation quality to the hyperparameters  $\alpha$ ,  $\beta$ ,  $\lambda$  on PASCAL VOC 2012 and MS COCO 2014. Among them,  $\alpha$  is the coefficient of  $\mathcal{L}_{frg}$  and a small  $\alpha$  can enable  $\mathcal{L}_{frg}$  to work well. As shown in Fig. 18a, the mIoU result is significantly improved on PASCAL VOC 2012 when the  $\alpha$  is greater than 0.1 and varies very little in the interval 0.15 to 0.5 with less than 0.3% mIoU change.  $\mathcal{L}_{ac}$  aims to constrain the foreground area to avoid unlimited expansion of the foreground area. Therefore, if the coefficient  $\beta$  of  $\mathcal{L}_{ac}$  is too small, it will lead to too many regions to be activated hence drastically reducing the segmentation performance as shown in Fig. 18b. The purpose of  $\mathcal{L}_{bas}$  is to allow the localization map to learn regions contributing to the classification in a background activation suppression manner. As shown in Fig. 18c, the mIoU result remains stable on both datasets when the factor  $\lambda$  of  $\mathcal{L}_{bas}$  is in the range of 0.8 to 1.2.

Although three hyperparameters are included in the total loss, in practice, we simply follow a principle of  $\beta = \alpha + \lambda$ , so that  $\mathcal{L}_{ac}$  is balanced with the losses  $\mathcal{L}_{frg}$  and  $\mathcal{L}_{bas}$ . Meanwhile, in finding the most suitable ratio between  $\alpha$  and  $\lambda$ , for simplicity,  $\lambda$  is fixed at 1 on both WSOL and WSSS. Therefore, when  $\alpha$  is determined,  $\beta$  and  $\lambda$  are also determined. In Table 7, we provide the results of different combinations of hyperparameters on five datasets. When the  $\alpha$  changes from 0.2 to 1.5, the effect on the results is limited with less than 0.6% change. In fact, following the settings of  $\alpha = 0.5$ ,  $\beta = 1.5$ ,  $\lambda = 1.0$  is feasible for all datasets, with very little change compared to the results reported in the

**Table 8** The mIoU results of inserting the generator after different layers with ResNet50 backbone

Dataset	Different layers			
	Layer 1	Layer 2	Layer 3	Layer 4
<b>PASCAL VOC 2012</b>	28.6	41.2	<b>57.7</b>	55.3
<b>MS COCO 2014</b>	17.2	25.3	<b>36.9</b>	34.8

Bold values indicate the best results among all methods

**Table 9** Effects of applying BAS on different baseline methods, including mIoU of the initial seed (**Seed**) and the pseudo ground-truth mask (**Mask**) on the PASCAL VOC 2012 training set

Method	Venue	Seed	Mask
IRN (Ahn et al., 2019)	CVPR19	48.8	66.3
SC-CAM (Chang et al., 2020)	CVPR20	50.9	63.4
SEAM (Wang et al., 2020b)	CVPR20	55.4	63.6
CONTA (Zhang et al., 2020c)	NeurIPS20	48.8	67.9
CDA (Su et al., 2021)	ICCV21	50.8	67.7
CSE (Kweon et al., 2021)	ICCV21	56.0	–
RIB (Lee et al., 2021a)	NeurIPS21	56.5	68.6
ReCAM (Chen et al., 2022b)	CVPR22	54.8	70.9
CLIMS (Xie et al., 2022a)	CVPR22	56.6	70.5
AdvCAM (Lee et al., 2022a)	TPAMI22	55.6	69.9
<b>Ours</b>	This work	<b>57.7</b>	–
<b>Ours + IRN</b>	This work	<b>58.2</b>	<b>71.1</b>
AdvCAM + CDA	–	55.5	69.3
ReCAM + CDA	–	54.5	70.5
<b>Ours + CDA</b>	This work	<b>58.8</b>	<b>71.0</b>
CDA + AdvCAM	–	55.5	69.3
ReCAM + AdvCAM	–	56.6	70.9
<b>Ours + AdvCAM</b>	This work	<b>59.8</b>	<b>71.5</b>

Bold values indicate the best results among all methods

paper. The above experiments illustrate that it is easy to find a suitable set of hyperparameters on different datasets.

*Generator after different layers.* In Table 8, we report the mIoU results of inserting the generator after different layers of ResNet50. Since the generator contains only one convolution layer, the semantic representation of the generated localization map depends mainly on the reused backbone part. Therefore, inserting the generator after **layer 1** or **layer 2** will result in insufficient semantic representation and poor segmentation performance, as presented in Table 8. In addition, inserting the generator after **4** does not perform better than **layer 3**, reducing 2.4% and 2.1% mIoU on PASCAL VOC 2012 and MS COCO 2014, respectively. It is mainly because the feature maps of **layer 4** are usually coarser than the feature maps of **layer 3**, hindering the acquisition of fine segmentation results.

### 5.3 Results on PASCAL VOC 2012 Dataset

*Quality of Initial Seed and Pseudo Labels.* Table 9 compares the quality of the initial seed and the pseudo ground-truth masks on the PASCAL VOC 2012 training set. For the initial seed, we achieve a mIoU of **57.7%**, exceeding the previous method by a large margin. Compared with the state-of-the-art method CLIMS (Xie et al., 2022a), which uses both ResNet50 and CLIP (Radford et al., 2021) networks in the seed generation phase, while BAS uses only ResNet50 network and achieves a gain of **1.1%**. Further, after normalizing the seeds generated by our method and by other methods and adding them together, BAS can combine with various baseline methods and significantly improve their segmentation quality by providing high quality foreground prediction maps. As shown in Table 9, the proposed BAS improves the IRN (Ahn et al., 2019) by **9.4%** mIoU, which is a remarkable boost. In addition, we achieve the best results with **59.8%** mIoU when applying BAS to AdvCAM (Lee et al., 2022a). We also add the initial seeds of the different methods for a fair comparison in Table 9. It is obvious that combining with BAS brings more remarkable improvement than combining with other methods. This is because BAS can produce high and balanced responses on the object, which benefits other methods significantly. We report the per-class mean IoU in Table 10. Although our method achieves consistent improvement on the above baseline methods, it does not perform well in some categories. This is because the classification network has difficulty distinguishing between objects and class-related contexts, especially in some categories, e.g., boats and water, TV and programs on TV, which in turn limits the localization ability of BAS. Figure 14 shows the visual comparison of the initial seed generated by BAS and IRN. It can be clearly noticed that our method has better performance in capturing the whole object area with a high confidence score. For the pseudo ground-truth mask, after refinement by IRN (Ahn et al., 2019), we achieve **4.8%**, **3.3%**, and **1.6%** gains when BAS is deployed on IRN, CDA, and AdvCAM, respectively, which illustrates the effectiveness of the

proposed method. BAS allows to obtain a better foreground-background segmentation and thus provides a strong support for the seed generation stage of the WSSS task.

*Quality of Segmentation.* To further validate the effectiveness of our method, we employ the pseudo segmentation labels to directly train a semantic segmentation network. Table 11 presents the segmentation results of the proposed BAS (with IRN) and other methods on the PASCAL VOC 2012 dataset. It is observed that our BAS exceeds previous methods under the same level of supervision, with **69.6%** and **69.9%** mIoU on the val and test sets. Compared to the latest method ReCAM (Chen et al., 2022b), with the same backbone network, we achieve a **1.1%** mIoU improvement on val set and **1.5%** on test set. We also show some qualitative segmentation results in Fig. 16. Compared with IRN, BAS demonstrates more robustness to various challenging scenarios, such as various sized objects, complex environments, and multi-instance situations.

### 5.4 Results on MS COCO 2014 Dataset

The accuracy of the proposed method and other state-of-the-art approaches on the MS COCO 2014 validation set is compared in Table 12. Our BAS based on IRN achieves a mIoU value of **45.1%**, exceeding all previous methods. Compared to the previous best model AdvCAM (ResNet101 is adopted as the backbone network), we use a smaller ResNet50 as the backbone, but achieve better results. In particular, we surpass our baseline method IRN (Ahn et al., 2019) by **3.7%** mIoU. Figure 17 demonstrates the visual comparison of the initial seed obtained by CAM (Zhou et al., 2016) and our method. Qualitative experiments show that the proposed BAS can capture more object areas compared to CAM, especially for large objects and multiple instances. In addition, BAS can achieve balanced and comprehensive responses on the target regions across various categories. Figure 19 shows some examples of semantic segmentation masks on MS COCO 2014 produced by IRN and by BAS (with IRN). It is observed that our method employed on the

**Table 7** Effect of different combinations of hyperparameters on WSOL and WSSS with ResNet50 backbone

Hyperparameters	GT-k. Loc		PxAP	mIoU	
	<b>CUB-200</b>	<b>ILSVRC</b>	<b>OpenImg</b>	<b>PASCAL</b>	<b>COCO</b>
$\alpha=0.2, \beta=1.2, \lambda=1.0$	95.29	71.75	66.67	<b>57.73</b>	36.79
$\alpha=0.5, \beta=1.5, \lambda=1.0$	<b>95.41</b>	71.87	<b>66.86</b>	57.68	<b>36.91</b>
$\alpha=0.7, \beta=1.7, \lambda=1.0$	95.35	71.94	66.74	57.55	36.84
$\alpha=1.0, \beta=2.0, \lambda=1.0$	95.26	<b>72.00</b>	66.52	57.41	36.67
$\alpha=1.5, \beta=2.5, \lambda=1.0$	95.08	71.89	66.39	57.16	36.43
Report	<b>95.41</b>	<b>72.00</b>	<b>66.86</b>	<b>57.73</b>	<b>36.91</b>

Bold values indicate the best results among all methods

Report: the result reported in the paper. OpenImg: OpenImages. GT-k. Loc: GT-known Loc

**Table 11** Performance comparison of WSSS methods in terms of mIoU (%) on the PASCAL VOC 2012 val and test sets

Method	Venue	Sup.	Val	Test
<i>Full supervision</i>				
DeepLabV2 (Chen et al., 2017)	TPAMI18	$\mathcal{F}$	77.6	79.7
WideResNet38 (Wu et al., 2019)	PR19	$\mathcal{F}$	80.8	82.5
<i>Image-level supervision + Saliency maps</i>				
OAA (Jiang et al., 2021)	TPAMI21	$\mathcal{I} + \mathcal{S}$	66.1	67.2
AuxSegNet (Xu et al., 2021)	ICCV21	$\mathcal{I} + \mathcal{S}$	69.0	68.6
AdvCAM (Lee et al., 2022a)	TPAMI22	$\mathcal{I} + \mathcal{S}$	71.3	71.2
<i>Image-level supervision only</i>				
IRN (Ahn et al., 2019)	CVPR19	$\mathcal{I}$	63.5	64.8
BES (Chen et al., 2020)	ECCV20	$\mathcal{I}$	65.7	66.6
CONTA (Zhang et al., 2020c)	NeurIPS20	$\mathcal{I}$	65.3	66.1
IAL (Wang et al., 2020a)	IJCV20	$\mathcal{I}$	62.0	62.4
ADL (Choe et al., 2020a)	TPAMI20	$\mathcal{I}$	53.7	54.7
LIID (Liu et al., 2020)	TPAMI20	$\mathcal{I}$	66.5	67.5
RIB (Lee et al., 2021a)	NeurIPS21	$\mathcal{I}$	68.3	68.6
CDA (Su et al., 2021)	ICCV21	$\mathcal{I}$	65.8	66.4
ECS (Sun et al., 2021)	ICCV21	$\mathcal{I}$	66.6	67.6
PMM (Li et al., 2021)	ICCV21	$\mathcal{I}$	68.5	69.0
CSE (Kweon et al., 2021)	ICCV21	$\mathcal{I}$	68.4	68.2
CPN (Zhang et al., 2021c)	ICCV21	$\mathcal{I}$	67.8	68.5
A <sup>2</sup> GNN (Zhang et al., 2021a)	TPAMI21	$\mathcal{I}$	66.8	67.4
AFA (Ru et al., 2022b)	CVPR22	$\mathcal{I}$	66.0	66.3
Du et al. (Du et al., 2022)	CVPR22	$\mathcal{I}$	67.7	67.4
ReCAM (Chen et al., 2022b)	CVPR22	$\mathcal{I}$	68.5	68.4
SIPE (Chen et al., 2022a)	CVPR22	$\mathcal{I}$	68.8	69.7
MCIS (Wang et al., 2022)	TPAMI22	$\mathcal{I}$	66.2	66.9
AdvCAM (Lee et al., 2022a)	TPAMI22	$\mathcal{I}$	68.1	68.0
<b>Ours</b>	This Work	$\mathcal{I}$	<b>69.6</b>	<b>69.9</b>

Bold values indicate the best results among all methods

Sup.: supervision.  $\mathcal{F}$ : full supervision.  $\mathcal{I}$ : image-level supervision.  $\mathcal{S}$ : saliency map supervision

**Table 12** Evaluation results on MS COCO 2014 validation set

Method	Venue	Bac.	Sal.	mIoU
IRN (Ahn et al., 2019)	CVPR19	R101		41.4
IAL (Wang et al., 2020a)	IJCV20	VGG16		27.7
ADL (Choe et al., 2020a)	TPAMI20	VGG16	✓	30.8
CONTA (Zhang et al., 2020c)	NeurIPS20	R50		33.4
EPS (Lee et al., 2021b)	ICCV21	WR38	✓	35.7
CSE (Kweon et al., 2021)	ICCV21	WR38		36.4
PMM (Li et al., 2021)	ICCV21	WR38		36.7
RIB (Lee et al., 2021a)	NeurIPS21	R101		43.8
ReCAM (Chen et al., 2022b)	CVPR22	R50		44.1
L2G (Jiang et al., 2022)	CVPR22	R101	✓	44.2
AdvCAM (Lee et al., 2022a)	TPAMI22	R101		44.4
<b>Ours</b>	This Work	R50		<b>45.1</b>

Bold values indicate the best results among all methods

Sal: Saliency. Bac: Backbone. WR38: WideResNet38. R50/101: ResNet50/101



Fig. 19 Examples of semantic segmentation results on MS COCO 2014 for IRN and BAS (with IRN)

IRN can achieve more accurate segmentation and show a better demarcation between different objects, because the proposed BAS can provide a more complete and accurate seed region compared to IRN.

### 5.5 Analysis

In this section, we will explore how to fully leverage BAS, especially focusing on its crucial background activation suppression loss. Furthermore, we aim to enhance the segmentation capability of BAS by integrating it with other methods.

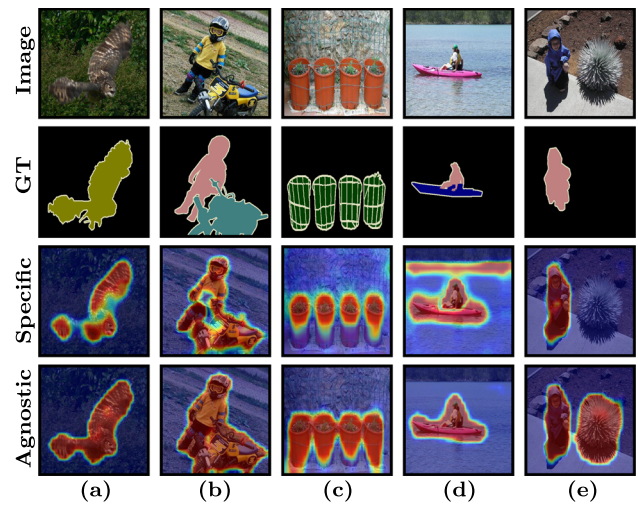
*Class-agnostic foreground map.* Different from the CAM-based approaches to extract class activation maps from the classifier, the proposed BAS obtains localization maps

through an extra generator. In addition to generating class-specific localization maps, BAS can also produce a class-agnostic foreground map by providing suitable objective functions. To this end, we consider all the classes existing in the image as a foreground class and sum the  $\mathcal{L}_{bas}$  of existing classes to supervise the foreground map. In this way, the foreground map can be fully trained from the entire dataset. As shown in Fig. 20, the class-agnostic foreground map localizes objects more completely and robustly than the class-specific localization map, and generates less noise. However, the foreground map is unable to distinguish objects of different categories and often identifies objects that are not in the target classes as shown in Fig. 20e. To utilize the foreground map to improve the performance of class-specific localization maps, we follow an intuitive idea that the foreground

**Table 10** Semantic segmentation performance gains for per-class on PASCAL VOC 2012

Method	bkg	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	mbk	Person	Plant	Sheep	Sofa	Train	Tv	mIoU
IRN	80.0	45.8	30.1	41.5	<b>38.8</b>	45.6	61.4	52.6	43.3	<b>29.0</b>	56.8	40.7	44.2	53.3	62.7	51.1	45.2	63.3	44.6	50.5	<b>44.2</b>	48.8
+ Ours	<b>82.3</b>	<b>50.1</b>	<b>39.1</b>	<b>54.9</b>	30.4	<b>56.4</b>	<b>76.7</b>	<b>58.5</b>	<b>73.5</b>	27.6	<b>72.7</b>	<b>50.3</b>	<b>68.9</b>	<b>70.3</b>	<b>73.7</b>	<b>59.7</b>	<b>48.7</b>	<b>79.2</b>	<b>52.0</b>	<b>55.1</b>	42.1	<b>58.2</b>
CDA	80.2	44.6	28.9	45.8	<b>36.9</b>	52.9	65.2	54.4	55.7	<b>28.4</b>	57.5	42.1	54.1	54.2	60.2	54.3	47.7	65.7	46.6	49.3	<b>43.9</b>	50.8
+ Ours	<b>82.6</b>	<b>50.1</b>	<b>39.0</b>	<b>56.1</b>	30.6	<b>57.5</b>	<b>77.1</b>	<b>59.8</b>	<b>75.3</b>	27.6	<b>73.8</b>	<b>49.6</b>	<b>70.5</b>	<b>71.4</b>	<b>73.4</b>	<b>60.1</b>	<b>51.3</b>	<b>80.6</b>	<b>51.1</b>	<b>54.6</b>	41.4	<b>58.8</b>
AdvCAM	81.3	50.6	33.5	57.3	<b>36.9</b>	53.1	67.9	54.8	64.8	<b>35.0</b>	68.4	42.0	58.4	67.9	67.1	56.0	42.6	76.1	48.5	56.8	<b>45.9</b>	55.5
+ Ours	<b>83.1</b>	<b>52.1</b>	<b>39.4</b>	<b>59.8</b>	31.4	<b>58.3</b>	<b>77.5</b>	<b>60.2</b>	<b>75.9</b>	30.3	<b>76.2</b>	<b>49.5</b>	<b>70.4</b>	<b>73.5</b>	<b>74.8</b>	<b>62.3</b>	<b>48.3</b>	<b>82.0</b>	<b>51.6</b>	<b>56.9</b>	43.0	<b>59.8</b>

Bold values indicate the best results among all methods



**Fig. 20** Class-agnostic foreground map vs class-specific localization map in the following five aspects: **a** Completeness. **b** Connectivity. **c** Less noise. **d** Identifying class-related background. **e** Class-aware

map usually covers all class-specific localization maps and has higher segmentation quality. If the class-specific localization map has a higher response in some regions than the foreground map, it may be caused by noise or confusing background, as shown in Fig. 20c, d. Therefore, we should weaken the response in these regions by directly replacing them with the response in the foreground map or averaging the response of both maps. The experimental results in Table 13 show that both strategies can improve the quality of the initial seed and hence increase the accuracy of the pseudo ground-truth mask. The best results are achieved by the average approach which not only reduces the response of the uncertain region but also combines the prediction probabilities of both class-agnostic and class-specific maps. It improves the initial seed and pseudo ground-truth mask by **0.6%** and **0.5%** mIoU results.

*Improve the quality of BAS.* As analyzed in Sect. 5.3, It can be noted that BAS does not perform well in some categories, which is usually due to the co-occurring context providing support to the classification discrimination, causing the localization map to learn the context. To alleviate this problem, we apply the proposed BAS to the W-OoD (Lee et al., 2022b) method, which uses additional out-of-distribution data to address the spurious relevance of the background, such as boat-water and aeroplane-sky/runway. As presented in Table 14, benefiting from the strong discriminative ability of the classification network in W-OoD method, BAS can achieve better performance, with a **1.8%** mIoU improvement on the initial seed, including **16.0%** and **7.1%** mIoU gains on the boat and aeroplane categories, respectively. After applying IRN and DeepLabV2, BAS w/ W-OoD obtains **71.3%** and **71.1%** mIoU on PASCAL VOC 2012 val and test sets. In addition, we apply CLIMS (Xie et al., 2022a) to BAS

**Table 13** Applying the class-agnostic foreground map to the class-specific localization maps with different strategies on the PASCAL VOC 2012

Method	Strategy	Seed	Mask	Val	Test
Ours	—	57.7	71.1	69.6	69.9
w/ Foreground	Replace	57.9	71.4	69.9	70.0
w/ Foreground	Average	<b>58.3</b>	<b>71.6</b>	<b>70.3</b>	<b>70.1</b>

Bold values indicate the best results among all methods

**Table 14** Applying BAS to CLISM and W-OoD on the PASCAL VOC 2012

Method	Seed	Mask	Val	Test
Ours	57.7	71.1	69.6	69.9
w/ CLIMS (Xie et al., 2022a)	<b>59.0</b>	<b>72.3</b>	<b>70.6</b>	<b>70.9</b>
w/ W-OoD (Lee et al., 2022b)	<b>59.5</b>	<b>72.7</b>	<b>71.3</b>	<b>71.1</b>

Bold values indicate the best results among all methods

to suppress the co-occurring background by using natural language supervision in CLIP (Radford et al., 2021), which also significantly improves the quality of the initial seed and brings a **8.9%** boost in the boat category. Consequently, BAS w/ CLIMS obtains **70.6%** and **70.9%** mIoU on the val set and test set, substantially improving the segmentation ability of BAS.

*Finding image-specific threshold by BAS.* Unlike CAM, the proposed BAS designs a set of loss functions to evaluate the quality of the localization map and uses them for training, similarly, they are also suitable for the testing phase. As shown in Fig. 22a, it can be noted that the unbalanced response of CAM causes the segmentation performance heavily dependent on the threshold, while the optimal threshold value even varies significantly across images. It is obviously not appropriate to use a global threshold for the whole dataset. Therefore, we propose to find the image-specific threshold by employing background activation suppression loss  $\mathcal{L}_{bas}$  and area constraint loss  $\mathcal{L}_{ac}$  as the evaluation of the threshold values. As illustrated in Fig. 22b, we obtain a series of binary masks by changing the threshold values and input them into the AMC module to generate  $\mathcal{L}_{bas}$  and  $\mathcal{L}_{ac}$ , the same process as in Fig. 2. Then, we simply add  $\mathcal{L}_{bas}$  and  $\mathcal{L}_{ac}$  together as the evaluation score and select the binary mask with the smallest evaluation score as the final result. Table 15 compares the effect of this image-specific threshold post-processing with global threshold on different methods on the PASCAL VOC 2012 training set. Experimental results show that the proposed post-processing approach is helpful to improve the segmentation quality by providing feedback on different thresholds to select the best threshold value specific to the image, especially for CAM (Zhou et al., 2016) and CDA (Su et al., 2021),

**Table 15** Effect of applying image-specific threshold on different methods compared to global threshold on the PASCAL VOC 2012 training set

Threshold	Method				
	CAM	CDA	ReCAM	AdvCAM	Ours
Global	48.8	50.8	54.8	55.5	57.7
Image-specific	<b>49.5</b>	<b>51.3</b>	<b>55.0</b>	<b>55.7</b>	<b>57.8</b>

Bold values indicate the best results among all methods

bringing **0.7%** and **0.5%** mIoU improvement, respectively. However, the enhancement is limited when applying it to the proposed BAS, mainly because BAS produces few uncertainty regions and is not very sensitive to the threshold.

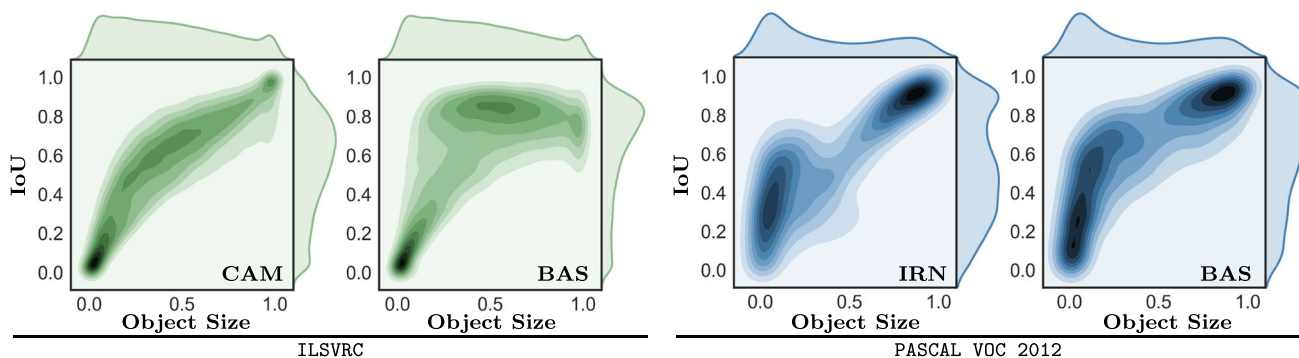
## 6 Discussion

*Limitation.* In this section, we discuss the localization ability of BAS for different size objects. We first visualize the density distribution of the IoU about BAS and CAM (Zhou et al., 2016) in Fig. 21. It can be noted that BAS performs better on medium and large objects, but not enough on small objects. We believe the main reason is the following two aspects: the localization of small objects is an inherent problem of computer vision, on the other hand, the area constraint loss penalizes different size objects unequally and will penalize small objects less, which causes BAS cannot balance both large and small objects with only a single hyperparameter to adjust the area constraint loss (Fig. 22).

*Future Works.* In the future, there are two main aspects of work, (1) improving the performance of the localization capability at different object sizes and (2) further extending the application of BAS.

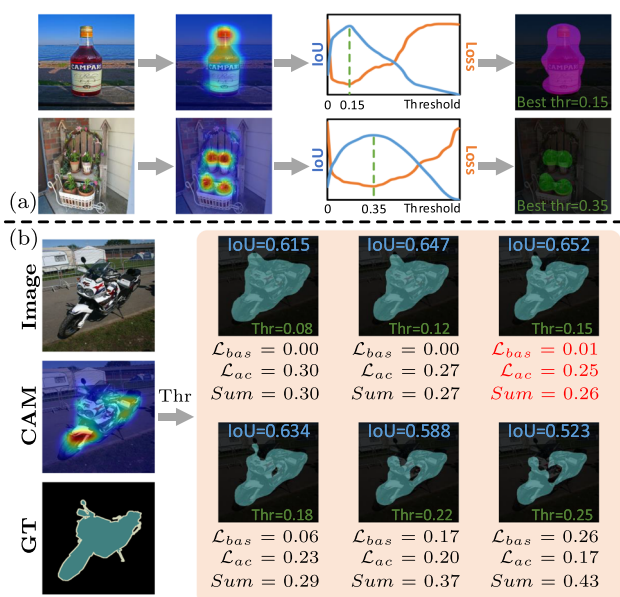
To solve the issue of inconsistent localization ability for different size objects, we would like to explore the following promising researches: (1) The area constraint loss can be improved to allow different tolerance for objects of various sizes. (2) Based on the fact that WSOL works better for localizing large objects, we can determine the approximate region of the objects in the first stage, and then crop and resize the corresponding region to convert the original small object into a larger one, thereby performing localization in the second stage.

Apart from the above possible improvements, BAS can also be extended to weakly supervised instance segmentation (WSIS), since obtaining a high-quality locality map is also essential for WSIS.



**Fig. 21** Limitation. The density distribution map about IoU and object size. For WSOL, the experiment is implemented on ILSVRC and bounding boxes are used to measure IoU and object size. For WSSS,

experimental results are calculated by pixel-level masks on the PASCAL VOC 2012 training set at the seed phase



**Fig. 22** Finding image-specific threshold by BAS. **a** IoU-threshold curve and Loss-threshold curve. Loss indicates the summation of  $L_{bas}$  and  $L_{ac}$ . **b** Process of finding the image-specific threshold by using  $L_{bas}$  and  $L_{ac}$  as evaluation

## 7 Conclusion

In this paper, we find previous FPM-based work using cross-entropy to facilitate the learning of foreground prediction maps, essentially by changing the activation value, and the activation value shows a higher correlation with the foreground mask. Thus, we propose a background activation suppression (BAS) approach to promote the generation of foreground maps by an activation map constraint (AMC) module, which facilitates the learning of foreground prediction maps mainly through the suppression of background activation. Extensive experiments on CUB-200-2011 and ILSVRC verify the effectiveness of the proposed BAS, which

surpasses previous methods by a large margin. In addition, BAS can also be extended on WSSS to enhance the seed quality of other methods by providing high quality foreground maps, and achieves the state-of-the-art performance on PASCAL VOC 2012 and MS COCO 2014.

## References

- Ahn, J., & Kwak, S. (2018). 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).
- Ahn, J., Cho, S., & Kwak, S. (2019). Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2209–2218).
- Bae, W., Noh, J., & Kim, G. (2020). Rethinking class activation mapping for weakly supervised object localization. In *European conference on computer vision* (pp. 618–634). Springer.
- Chan, L., Hosseini, M. S., & Platanotis, K. N. (2021). A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. *International Journal of Computer Vision*, 129(2), 361–384.
- Chang, Y. T., Wang, Q., Hung, W. C., Piramuthu, R., Tsai, Y. H., & Yang, M. H. (2020). Weakly-supervised semantic segmentation via subcategory exploration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8991–9000).
- Chen, L., Wu, W., Fu, C., Han, X., & Zhang, Y. (2020). Weakly supervised semantic segmentation with boundary exploration. In *European conference on computer vision* (pp. 347–362). Springer.
- Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. [arXiv:1412.7062](https://arxiv.org/abs/1412.7062)
- Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). 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.
- Chen, Q., Yang, L., Lai, J. H., & Xie, X. (2022a). Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4288–4298).

- Chen, Z., Wang, T., Wu, X., Hua, X. S., Zhang, H., & Sun, Q. (2022b). Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 969–978).
- Choe, J., Lee, S., & Shim, H. (2020a). Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12), 4256–4271.
- Choe, J., Oh, S. J., Lee, S., Chun, S., Akata, Z., & Shim, H. (2020b). Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3133–3142).
- Choe, J., & Shim, H. (2019). Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2219–2228).
- Du, Y., Fu, Z., Liu, Q., & Wang, Y. (2022). Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4320–4329).
- Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2), 303–338.
- Guo, G., Han, J., Wan, F., & Zhang, D. (2021). Strengthen learning tolerance for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7403–7412).
- Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., & Malik, J. (2011). Semantic contours from inverse detectors. In *2011 International conference on computer vision* (pp. 991–998). IEEE.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. [arXiv:1704.04861](https://arxiv.org/abs/1704.04861)
- Jiang, P. T., Han, L. H., Hou, Q., Cheng, M. M., & Wei, Y. (2021). Online attention accumulation for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 7062–7077.
- Jiang, P. T., Yang, Y., Hou, Q., & Wei, Y. (2022). L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16886–16896).
- Jo, S., & Yu, I. J. (2021). Puzzle-cam: Improved localization via matching partial and full features. In *2021 IEEE international conference on image processing (ICIP)* (pp. 639–643). IEEE.
- Kim, E., Kim, S., Lee, J., Kim, H., & Yoon, S. (2022). Bridging the gap between classification and localization for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14258–14267).
- Kim, J., Choe, J., Yun, S., & Kwak, N. (2021). Normalization matters in weakly supervised object localization. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3427–3436).
- Kolesnikov, A., & Lampert, C. H. (2016). Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision* (pp. 695–711). Springer.
- Kweon, H., Yoon, S. H., Kim, H., Park, D., & Yoon, K. J. (2021). Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6994–7003).
- Lee, J., Choi, J., Mok, J., & Yoon, S. (2021). Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 27408–27421.
- Lee, J., Kim, E., Mok, J., & Yoon, S. (2022a). Anti-adversarially manipulated attributions for weakly supervised semantic segmentation and object localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Lee, J., Oh, S. J., Yun, S., Choe, J., Kim, E., & Yoon, S. (2022b). Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16897–16906).
- Lee, S., Lee, M., Lee, J., & Shim, H. (2021b). Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5495–5505).
- Li, K., Wu, Z., Peng, K. C., Ernst, J., & Fu, Y. (2018). Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9215–9223).
- Li, Y., Kuang, Z., Liu, L., Chen, Y., & Zhang, W. (2021). Pseudo-mask matters in weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6964–6973).
- Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer.
- Liu, Y., Wu, Y. H., Wen, P. S., Shi, Y. J., Qiu, Y., & Cheng, M. M. (2020). Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Lovász, L. (1993). Random walks on graphs. *Comb. Paul Erdos Eighty*, 2(1–46), 4.
- Lu, W., Jia, X., Xie, W., Shen, L., Zhou, Y., & Duan, J. (2020). Geometry constrained weakly supervised object localization. In *Computer vision—ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16* (pp. 481–496). Springer.
- Luo, H., Zhai, W., Zhang, J., Cao, Y., & Tao, D. (2022). Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2252–2261).
- Mai, J., Yang, M., & Luo, W. (2020). Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8766–8775).
- Meng, M., Zhang, T., Tian, Q., Zhang, Y., & Wu, F. (2021). Foreground activation maps for weakly supervised object localization. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3385–3395).
- Pan, J., Zhu, P., Zhang, K., Cao, B., Wang, Y., Zhang, D., Han, J., & Hu, Q. (2022). Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation. *International Journal of Computer Vision*, 130(5), 1181–1195.
- Pan, X., Gao, Y., Lin, Z., Tang, F., Dong, W., Yuan, H., Huang, F., & Xu, C. (2021). Unveiling the potential of structure preserving for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11642–11651).
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.



- Ru, L., Du, B., Zhan, Y., & Wu, C. (2022a). Weakly-supervised semantic segmentation with visual words learning and hybrid pooling. *International Journal of Computer Vision*, 130(4), 1127–1144.
- Ru, L., Zhan, Y., Yu, B., & Du, B. (2022b). Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16846–16855).
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2), 336–359.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
- Singh, K. K., & Lee, Y. J. (2017). Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 3544–3553). IEEE.
- Song, L., Liu, J., Sun, M., & Shang, X. (2021). Weakly supervised group mask network for object detection. *International Journal of Computer Vision*, 129(3), 681–702.
- Su, Y., Sun, R., Lin, G., & Wu, Q. (2021). Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7004–7014).
- Sun, K., Shi, H., Zhang, Z., & Huang, Y. (2021). ECS-Net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7283–7292).
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).
- Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The caltech-ucsd birds-200–2011 dataset.
- Wang, W., Sun, G., & Van Gool, L. (2022). Looking beyond single images for weakly supervised semantic segmentation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Wang, X., Liu, S., Ma, H., & Yang, M. H. (2020). Weakly-supervised semantic segmentation by iterative affinity learning. *International Journal of Computer Vision*, 128(6), 1736–1749.
- Wang, Y., Zhang, J., Kan, M., Shan, S., & Chen, X. (2020b). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12275–12284).
- Wei, J., Wang, Q., Li, Z., Wang, S., Zhou, S. K., & Cui, S. (2021). Shallow feature matters for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5993–6001).
- Wei, Y., Feng, J., Liang, X., Cheng, M. M., Zhao, Y., & Yan, S. (2017). 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).
- Wu, P., Zhai, W., & Cao, Y. (2021). Background activation suppression for weakly supervised object localization. [arXiv:2112.00580](https://arxiv.org/abs/2112.00580)
- Wu, Z., Shen, C., & Van Den Hengel, A. (2019). Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90, 119–133.
- Xie, J., Hou, X., Ye, K., & Shen, L. (2022a). Cross language image matching for weakly supervised semantic segmentation. [arXiv:2203.02668](https://arxiv.org/abs/2203.02668)
- Xie, J., Luo, C., Zhu, X., Jin, Z., Lu, W., & Shen, L. (2021). Online refinement of low-level feature based activation map for weakly supervised object localization. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 132–141).
- Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., & Shen, L. (2022b). Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. [arXiv:2203.13505](https://arxiv.org/abs/2203.13505)
- Xu, J., Hou, J., Zhang, Y., Feng, R., Zhao, R. W., Zhang, T., Lu, X., & Gao, S. (2022). Cream: Weakly supervised object localization via class re-activation mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9437–9446).
- Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Sohel, F., & Xu, D. (2021). Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6984–6993).
- Xue, H., Liu, C., Wan, F., Jiao, J., Ji, X., & Ye, Q. (2019). Danet: Divergent activation for weakly supervised object localization. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6589–6598).
- Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6023–6032).
- Zhai, W., Luo, H., Zhang, J., Cao, Y., & Tao, D. (2022). One-shot object affordance detection in the wild. *International Journal of Computer Vision*, 130, 1–29.
- Zhang, B., Xiao, J., Jiao, J., Wei, Y., & Zhao, Y. (2021a). Affinity attention graph neural network for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 8082–8096.
- Zhang, C. L., Cao, Y. H., & Wu, J. (2020a). Rethinking the route towards weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13460–13469).
- Zhang, D., Han, J., Cheng, G., & Yang, M. H. (2021b). Weakly supervised object localization and detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 5866–5885.
- Zhang, D., Han, J., Zhao, L., & Meng, D. (2019). Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. *International Journal of Computer Vision*, 127(4), 363–380.
- Zhang, D., Han, J., Zhao, L., & Zhao, T. (2020b). From discriminant to complete: Reinforcement searching-agent learning for weakly supervised object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12), 5549–5560.
- Zhang, D., Zhang, H., Tang, J., Hua, X. S., & Sun, Q. (2020c). Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33, 655–666.
- Zhang, F., Gu, C., Zhang, C., & Dai, Y. (2021c). Complementary patch for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7242–7251).
- Zhang, X., Wei, Y., Feng, J., Yang, Y., & Huang, T. S. (2018a). Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1325–1334).
- Zhang, X., Wei, Y., Kang, G., Yang, Y., & Huang, T. (2018b). Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 597–613).

- Zhang, X., Wei, Y., & Yang, Y. (2020d). Inter-image communication for weakly supervised localization. In *Computer vision—ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16* (pp. 271–287). Springer.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921–2929).
- Zhu, L., She, Q., Chen, Q., You, Y., Wang, B., & Lu, Y. (2022). Weakly supervised object localization as domain adaption. [arXiv:2203.01714](https://arxiv.org/abs/2203.01714)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.