Background Activation Suppression for Weakly Supervised Object Localization

Pingyu Wu[1,†] Wei Zhai[1,†] Yang Cao[1,2,*]
[1] University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{wpy364755620@mail., wzhai056@mail., forrest@}ustc.edu.cn

Abstract

Weakly supervised object localization (WSOL) aims to localize objects using only image-level labels. Recently a new paradigm has emerged by generating a foreground prediction map (FPM) to achieve localization task. Existing FPM-based methods use cross-entropy (CE) to evaluate the foreground prediction map and to guide the learning of generator. We argue for using activation value to achieve more efficient learning. It is based on the experimental observation that, for a trained network, CE converges to zero when the foreground mask covers only part of the object region. While activation value increases until the mask expands to the object boundary, which indicates that more object areas can be learned by using activation value. In this paper, we propose a Background Activation Suppression (BAS) method. Specifically, an Activation Map Constraint module (AMC) is designed to facilitate the learning of generator by suppressing the background activation value. Meanwhile, by using the foreground region guidance and the 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. Code and models are available at github.com/wpy1999/BAS.

1. Introduction

Weakly supervised object localization (WSOL) aims to identify the object's localization in a scene using only image-level labels, no bounding box annotations. WSOL is gaining more and more attention in the research community because it can visualize classification networks [22, 28, 33] and reduce the cost of manual labeling [3, 5, 9, 24, 27].

As an important previous work, Class Activation Map (CAM) [33] is widely used to implement weakly supervised localization. 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. To alleviate this problem, some methods [4, 11, 16, 25, 29] erase the most discriminative regions during the training, forcing the network to learn more object features relevant to localization. [13, 30, 31] are also based on CAM, which improves localization performance by establishing pixel-level spatial correlation. Additionally, some methods [6, 10, 20, 26] suggest adopting a divide-and-conquer strategy to accomplish classification and localization tasks separately to avoid conflicts.

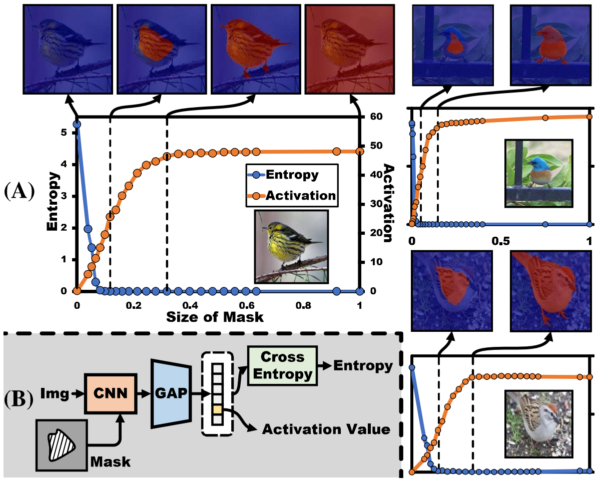


Figure 1. (A) The entropy value of CE loss $w.r.t$ foreground mask and foreground activation value $w.r.t$ foreground mask. To illustrate the generality of this phenomenon, more examples are shown in the subfigure on the right. (B) Experimental procedure and related definitions. Implementation details of the experiment and further results are available in the Supplementary Material.

Very recently, a new paradigm [12, 21] is devised for WSOL by learning a foreground prediction map (FPM) after the feature extraction network to achieve localization without relying on CAM. Typically, ORNet [21] is a two-stage approach, which first trains a classification network as

[*] Corresponding author. [†] Equal contributions.

an evaluator, and then utilizes CE loss to guide the learning of generator by masking the original image with foreground prediction map. In contrast to ORNet, the foreground prediction map in the FAM [12] masks high-level information and is optimized by CE loss through two modules. In this paper, we also follow this FPM-based paradigm.

To better understand how the FPM-based paradigm works, we design the following experiments where we focus on exploring the entropy value of CE loss with respect to ($w.r.t$) foreground mask and activation value $w.r.t$ foreground mask. As shown in Fig. 1 (A), we plot the curves of the two relationships. By observation we can find two important phenomena: 1) There is a "mismatch" between entropy and ground-truth mask, i.e., entropy converges to zero quickly when foreground mask retains only part of the object region. 2) There is a higher correlation between foreground activation value and foreground mask, i.e., the activation value tends to "saturate" when the mask expands to the object boundary. These phenomena suggest that better localization results can be achieved by using activation value compared to entropy. Moreover, from Fig. 1 (B), it can be analyzed that CE actually facilitates the learning of the generator indirectly by influencing the activation value. Based on the above observations, a straightforward manner to obtain a complete foreground prediction map is to maximize the foreground activation value. However, considering that the maximum optimization problem is not friendly to deep neural networks, we propose to promote the learning of generation by suppressing background activation value.

In this paper, we propose a simple but effective Background Activation Suppression (BAS) method. As shown in Figure 2, BAS includes three modules: an extractor, a generator, and an Activation Map Constraint module (AMC). First, an extractor is used to extract the image features for subsequent localization and classification. The generator aims to generate a set of class-specific foreground prediction maps for localization. Then the coupled background prediction map is obtained by inversion and fed into AMC together for 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 and overall activation (the activation generated by the entire image). In the inference phase, we select the Top-k prediction maps to take the mean value as the final localization result based on the predicted category probabilities. Evaluations on CUB-200-2011 [19] and ILSVRC [14] are performed with four different types of backbones, and the experimental results show that our method achieves stable and excellent results with significant improvement over the SOTA methods. The contributions of this paper include:

1) This paper finds that, the essential reason why minimizing CE loss facilitates the generation of foreground maps is that it indirectly increases the foreground activation value, and accordingly proposes to facilitate the generation of foreground prediction maps by suppressing the background activation value.

2) This paper proposes a simple but effective Background Activation Suppression (BAS) approach to facilitate the generation of foreground maps 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 CUB-200-2011 [19] and ILSVRC [14] benchmarks demonstrate that our proposed method outperforms previous methods by a significant margin in terms of GT-known/Top-1/Top-5 localization.

2. Related work

Weakly supervised object localization (WSOL) is a challenging task that requires localizing objects using only image-level labels. To obtain the localization results from the classification network, Zhou et al. [33] proposes to replace top layers with global average pooling, and apply the fully connected weights on depth feature maps to generate the class activation map (CAM) as the localization map. Unfortunately, CAM usually focuses on the most discriminative regions. To alleviate this problem, a type of approach proposes to use erasing strategies. HaS [16] randomly splits the original image into different patches, forcing the classification network to learn more features of objects. ACoL [29] and EIL [11] erase the region with high response in the feature map and utilize two parallel branches for adversarial erasing. Differently, ADL [4] erases the most significant regions or highlighting regions of each layer during forward propagation, to achieve a balance between classification and localization. CutMix [25] uses a data enhancement strategy that blends two different images for training to force the network to learn the relevant regions of different objects.

In addition, another type of approach uses the thought of spreading confidence regions to mine relevant features. SPG [30] uses thresholds to filter foreground and background regions with confidence from CAM to guide shallow network learning. Further, SPOL [20] generates more reliable confidence regions by multiplicative feature fusion strategy, and then feeds the confidence regions into a semantic segmentation network. I2C [31] proposes to increase the robustness and reliability of localization by considering

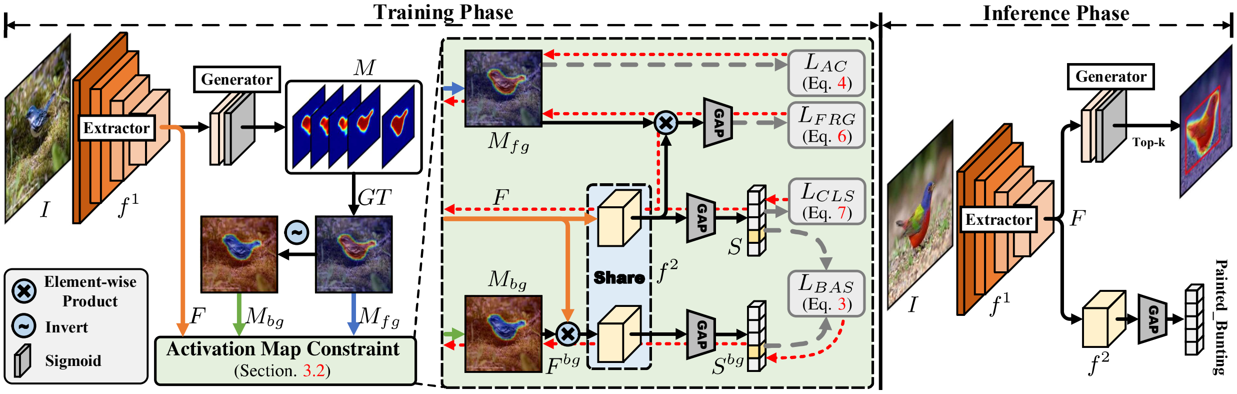


Figure 2. The architecture of the proposed BAS. In the training phase, the class-specific foreground prediction map F^{fg} and the coupled background prediction map F^{bg} are obtained by the generator according to the ground-truth class (GT), and then fed into the Activation Map Constraint module together with the feature map F. In the inference phase, we utilize Top-k to generate the final localization map.

the correlation of different pictures of the same class. Besides, SPA [13] uses post-processing to extract feature maps with structure-preserving. SLT [6] 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 increasing tolerance.

Most recently, two Foreground-Prediction-Map-based works [12, 21], both achieve the localization task by generating a foreground prediction map. ORNet [21] 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, and then fed into the evaluator for classification training, so that the foreground prediction map can learn the object region. FAM [12] utilizes a Foreground Memory Mechanism structure to store different foreground classifiers and generate foreground prediction maps. The foreground prediction map is split into different part regions, and a class-agnostic foreground prediction map is learned by classification learning of each part region in the feature map.

It can be noticed that both ORNet [21] and FAM [12] only consider foreground regions and use cross-entropy to facilitate the learning of generator. Different from these methods, we propose a background activation suppression strategy to learn foreground prediction maps through a simple but effective approach.

3. Methodology

3.1. Overview

Based on the background activation suppression, we obtain more complete object localization maps for WSOL by proposing the BAS approach. As shown in the left subgraph of Fig. 2, BAS consists of three modules: an extractor, a generator, and an Activation Map Constraint module (AMC). The extractor is used to extract features related to classification and localization. The generator is to produce the predictions of foreground maps. The AMC module is to promote the learning of extractor and generator.

We divide the original backbone network into two subnetworks f^1 and f^2 according to the location of the generator, and denote the network parameter by θ. The subnetwork $f^1(I, \theta^1)$ before the generator is used as a feature extractor. Given an image I, the feature map $F \in R^{H \times W \times N}$ is generated by extractor in forward propagation, where H, W, and N denote the height, width, and number of channels of the feature map, respectively. Afterward, the feature map F is fed into the generator, which consists of a 3×3 convolution layer and a $Sigmoid$ activation function for generating a set of class-specific foreground prediction maps $M \in R^{H \times W \times C}$, where C is the number of dataset categories. We choose the foreground prediction map $M_{fg} \in R^{H \times W \times 1}$ corresponding to the ground-truth class and invert it to obtain the coupled background prediction map $M_{bg} \in R^{H \times W \times 1}$. Finally, M_{fg}, M_{bg} and F are fed together into AMC module for prediction map learning. We will detail describe the AMC structure and loss function in Sec. 3.2.

In the inference phase, as shown in the right subgraph of Fig. 2. After obtained by the extractor, the feature map F is fed into the generator and sub-network $f^2(F, \theta^2)$ to generate the foreground prediction maps set M and the classification prediction distribution $\hat{y}$, respectively. We select the prediction maps corresponding to the Top-k categories including the ground-truth class according to the predicted category probabilities, and take their average values as the final localization results.

3.2. Activation Map Constraint

The proposed AMC module utilizes foreground map, background map, and feature map 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 M_{bg}, the background feature map $F^{bg} \in R^{H\times W\times C}$ is obtained by dot product with feature map F. Afterwards, the feature maps F^{bg} and F are fed to two sub-networks $f^2(F,\theta^2)$ and $f^2(F^{bg},\theta^2)$ with shared weights, respectively. For the sub-network with F^{bg} as input, the goal is to generate the background activation value by the same function, so that this sub-network parameter is frozen in back propagation. Following the sub-network $f^2(F,\theta^2)$ and the global average pooling (GAP) [33], F and F^{bg} produce the class probability distributions $\hat{y}$ and $\hat{y}^{bg}$, respectively, which can be expressed as follows:

$$\hat{y} = GAP(f^2(F,\theta^2)), \tag{1}$$

$$\hat{y}^{bg} = GAP(f^2(F^{bg},\theta^2)). \tag{2}$$

We select the values in the $\hat{y}$ and $\hat{y}^{bg}$ according to the ground-truth class, denoted as activation value S and background activation value S^{bg}, respectively. S represents the activation value generated by the unmasked feature map, containing both foreground and background information, and S^{bg} 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 L_{BAS} is defined in the following form:

$$\mathcal{L}_{BAS} = \frac{S^{bg}}{S+\varepsilon}, \tag{3}$$

where ε is a very small value (e^{-8}), to ensure that the equation is meaningful.

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 area well, the $\mathcal{L}_{BAS}$ it produced has to be minimal while the background area should be as large as possible, i.e., the foreground area should be as small as possible. So we use the foreground prediction map area as constraints:

$$\mathcal{L}_{AC} = \frac{1}{H\times W}\sum_{i}^{H}\sum_{j}^{W} M_{fg}(i,j). \tag{4}$$

Foreground Region Guidance ($\mathcal{L}_{FRG}$). Meanwhile, we retain the FPM architecture's form of using classification tasks to drive the learning of foreground prediction map, that is, using high-level semantic information to guide the foreground prediction map to the approximate correct region of the object. Therefore a foreground loss based on cross-entropy is utilized. After F is fed into $f^2(F,\theta^2)$, it is dotted with M_{fg} to produce $\mathcal{L}_{FRG}$:

$$\hat{y}^{fg} = GAP(M_{fg}\cdot f^2(F,\theta^2)), \tag{5}$$

$$\mathcal{L}_{FRG} = -\sum_{i=0}^{C} y_i \log\frac{e^{\hat{y}_i^{fg}}}{\sum_j^C e^{\hat{y}_j^{fg}}}, \tag{6}$$

where 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 $\hat{y}$, which is used for classification learning of the entire image:

$$\mathcal{L}_{CLS} = -\sum_{i=0}^{C} y_i \log\frac{e^{\hat{y}_i}}{\sum_j^C e^{\hat{y}_j}}. \tag{7}$$

3.3. Total loss

By optimizing the foreground and background losses, as well as the area loss in the AMC module, can jointly guide the learning of foreground prediction map to the overall area of the object. The total loss of the BAS training process is defined in the following form:

$$\mathcal{L} = \mathcal{L}_{CLS} + \alpha\mathcal{L}_{FRG} + \beta\mathcal{L}_{AC} + \lambda\mathcal{L}_{BAS}, \tag{8}$$

where α, β, and λ are hyper-parameters, $\mathcal{L}_{CLS}$ and $\mathcal{L}_{FRG}$ are both cross-entropy losses. For all backbones and datasets, we set $\lambda = 1$. The ablation experiments on λ are described in Sec. 4.3, and the ablation experiments on α, β are in Supplementary Materials.

4. Experiment

4.1. Experimental Setup

Datasets. We evaluate the proposed algorithm on the most popular benchmarks including CUB-200-2011 [19] and ILSVRC [14]. CUB-200-2011 contains 200 species of birds with $5,994$ training images and $5,794$ testing images. ILSVRC is divided into $1,000$ classes and contains about 1.2 million training images, $50,000$ validation images. Except for class labels, CUB-200-2011 also provides mask labels, which are only used to evaluate the prediction mask.

Metrics. Following SPA [13], we apply both bounding box and mask metrics to evaluate the performance of our BAS. For bounding box, following [13, 20, 26], we use three metrics for evaluation, including GT-known localization accuracy (**GT-known Loc**), Top-1 localization accuracy (**Top-1 Loc**), and Top-5 localization accuracy (**Top-5 Loc**). Specifically, GT-known Loc is correct when the intersection over union(IoU) between the ground-truth bounding

Methods	Venue	Backbone	CUB-200-2011 [19] Loc. Acc.			ILSVRC [14] Loc. Acc.		
			Top-1	Top-5	GT-known	Top-1	Top-5	GT-known
CAM [33]	CVPR16	VGG16	41.06	50.66	55.10	42.80	54.86	59.00
ACoL [29]	CVPR18	VGG16	45.92	56.51	62.96	45.83	59.43	62.96
ADL [4]	CVPR19	VGG16	52.36	-	75.41	44.92	–	–
DANet [23]	ICCV19	VGG16	52.52	61.96	67.70	–	–	–
I2C [31]	ECCV20	VGG16	55.99	68.34	–	47.41	58.51	63.90
MEIL [11]	CVPR20	VGG16	57.46	–	73.84	46.81	–	–
GCNet [10]	ECCV20	VGG16	63.24	75.54	81.10	–	–	–
PSOL [26]	CVPR20	VGG16	66.30	84.05	89.11	50.89	60.90	64.03
SPA [13]	CVPR21	VGG16	60.27	72.50	77.29	49.56	61.32	65.05
SLT [6]	CVPR21	VGG16	67.80	–	87.60	51.20	62.40	67.20
FAM [12]	ICCV21	VGG16	69.26	–	89.26	51.96	–	**71.73**
ORNet [21]	ICCV21	VGG16	67.73	80.77	86.20	52.05	63.94	68.27
BAS (Ours)	This Work	VGG16	**71.33**	**85.33**	**91.07**	**52.96**	**65.41**	69.64
CAM [33]	CVPR16	MobileNetV1	48.07	59.20	63.30	43.35	54.44	58.97
HaS [16]	ICCV17	MobileNetV1	46.70	–	67.31	42.73	–	60.12
ADL [4]	CVPR19	MobileNetV1	47.74	–	–	43.01	–	–
RCAM [2]	ECCV20	MobileNetV1	59.41	–	78.60	44.78	–	61.69
FAM [12]	ICCV21	MobileNetV1	65.67	–	85.71	46.24	–	62.05
BAS (Ours)	This Work	MobileNetV1	**69.77**	**86.00**	**92.35**	**52.97**	**66.59**	**72.00**
CAM [33]	CVPR16	ResNet50	46.71	54.44	57.35	38.99	49.47	51.86
ADL [4]	CVPR19	ResNet50-SE	62.29	–	–	48.53	–	–
I2C [31]	ECCV20	ResNet50	–	–	–	51.83	64.60	68.50
PSOL [26]	CVPR20	ResNet50	70.68	86.64	90.00	53.98	63.08	65.44
WTL [1]	WACV21	ResNet50	64.70	–	77.35	52.36	–	67.89
FAM [12]	ICCV21	ResNet50	73.74	–	85.73	54.46	–	64.56
SPOL [20]	CVPR21	ResNet50	**80.12**	**93.44**	**96.46**	**59.14**	67.15	69.02
BAS (Ours)	This Work	ResNet50	77.25	90.08	95.13	57.18	**68.44**	**71.77**
CAM [33]	CVPR16	InceptionV3	41.06	50.66	55.10	46.29	58.19	62.68
SPG [30]	ECCV18	InceptionV3	46.64	57.72	–	48.60	60.00	64.69
DANet [23]	ICCV19	InceptionV3	49.45	60.46	67.03	47.53	58.28	–
I2C [31]	ECCV20	InceptionV3	55.99	68.34	72.60	53.11	64.13	68.50
GCNet [10]	ECCV20	InceptionV3	58.58	71.00	75.30	49.06	58.09	–
PSOL [26]	CVPR20	InceptionV3	65.51	83.44	–	54.82	63.25	65.21
SPA [13]	CVPR21	InceptionV3	53.59	66.50	72.14	52.73	64.27	68.33
SLT [6]	CVPR21	InceptionV3	66.10	–	86.50	55.70	65.40	67.60
FAM [12]	ICCV21	InceptionV3	70.67	–	87.25	55.24	–	68.62
BAS (Ours)	This Work	InceptionV3	**73.29**	**86.31**	**92.24**	**58.51**	**69.00**	**71.93**

Table 1. Comparison with state-of-the-art methods. Best results are highlighted in **bold**, second are underlined.

box and the predicted bounding box is greater than 0.5. Top-1/Top-5 Loc is correct when the Top-1/Top-5 predict categories contain the ground-truth class and the GT-known Loc is correct. For mask, we adopt both **Peak_T** and **Peak_IoU** as metrics, which are defined in SEM [32], to compare the prediction mask with the pixel-level ground-truth label. Peak_IoU $\in [0, 1]$ and Peak_T $\in [0, 255]$ denote the maximum intersection and its corresponding threshold, respectively. The larger Peak_T indicates the higher pixel brightness value of the object area in the localization map, which can be better visualized. And a larger Peak_IoU indicates that the localization result is closer to the target object at a specific threshold.

Implementation Details. We evaluate the proposed method on the most popular backbones, including VGG16 [15], InceptionV3 [17], ResNet50 [7], and MobileNetV1 [8]. All networks are fine-tuned on the pre-trained weights of ILSVRC [14]. We train 100 epochs on the CUB-200-2011 [19] and 9 epochs on ILSVRC [14]. In the training phase, the input images are resized to 256×256 and then randomly cropped to 224×224. 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 augmentations to get the final classification results following the settings in [6, 13, 30]. For localization, we replace the random crop with the center crop, as in previous work [4, 20, 25, 26].

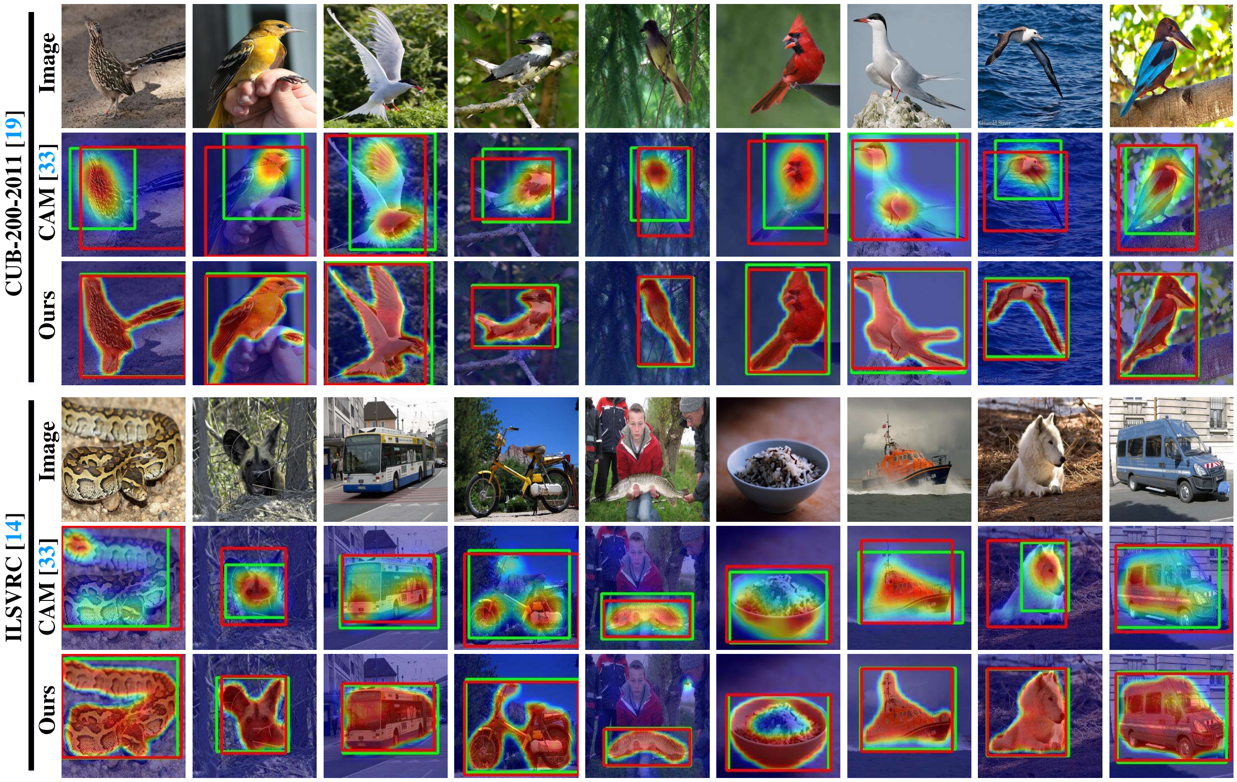


Figure 3. Visualization comparison with the baseline CAM [33] method on CUB-200-2011 [19] and ILSVRC [14]. The ground-truth bounding boxes are in red, and the predictions are in green.

4.2. Comparison with State-Of-The-Arts

We compare the proposed BAS with state-of-the-art methods on CUB-200-2011 [19] and ILSVRC [14] datasets. As shown in Table 1, BAS achieves stable and excellent performance on various backbones. On CUB-200-2011 [19], BAS surpasses all existing methods by a large margin in terms of GT-known/Top-1/Top-5 Loc when the backbone is VGG16, MobileNetV1 and InceptionV3. Compared with the current Foreground-Prediction-Map-based method FAM [12], BAS achieves 6.64% and 4.99% GT-known Loc improvement on MobileNetV1 and InceptionV3, respectively. In addition, ResNet-BAS achieves 95.13% GT-known Loc, which is a significant improvement of 9.40% compared to ResNet-FAM [12]. But compared to ResNet-SPOL [20], BAS is lower than it by 1.33%. SPOL utilizes three separated networks to achieve WSOL, first using a ResNet50 to generate class activation map, then a separate ResNet50 for segmentation, and finally an additional EfficientNet-B7 [18] for classification, while BAS uses only one network, which has significant advantages in efficiency.

On ILSVRC [14], 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 72.00% GT-known Loc, surpassing FAM [12] by 9.95%. Moreover, InceptionV3-BAS and ResNet50-BAS obtain 71.93% and 71.77% 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 compare the localization map of the proposed BAS and CAM [33] on CUB-200-2011 and ILSVRC in Fig. 3. Compared to CAM, BAS can consistently cover the entire area of the object, 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 VGG16 [15] as the backbone. Above all, we conduct ablation experiments on various components of BAS on CUB-200-2011 [19]. 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 Fig. 4, the addition of $\mathcal{L}_{BAS}$ based on baseline can enable the localization map to cover the object region more completely, so as to significantly improve the localization accuracy, achieving 17.43% and 15.60% improvement in terms of GT-known Loc and Top-1 Loc, respectively. Moreover, using Top-k strategy to integrate the final localization results, though

	Baseline	$\mathcal{L}_{BAS}$	Top-k	Top-1	Top-5	GT-known
(a)	√			53.45	65.46	70.39
(b)	√	√		69.05	82.43	87.82
(c)	√	√	√	**71.33**	**85.33**	**91.07**

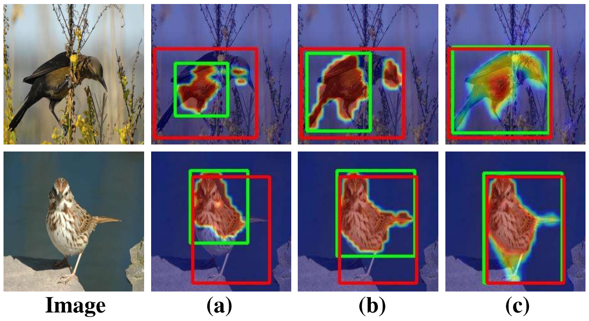


Figure 4. **Ablation study on our method.** (a) the baseline method. (b) add $\mathcal{L}_{BAS}$ to the baseline. (c) synthesize the final result with Top-k strategy.

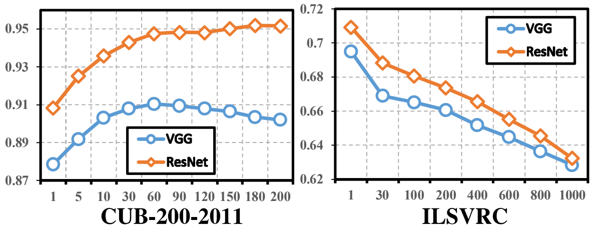


Figure 5. **GT-known Loc.** (%) $w.r.t$ **k.**

making the localization result not as sharp as before, it can further improve the GT-known Loc (from 87.85% to 91.07%) by increasing the connectivity of the localization map and alleviating the problem of the classification network focusing on the distinguish parts.

Hyperparameter k in Top-k strategy. We evaluate the effect of the hyperparameter k in our BAS. As shown in Fig. 5, 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 Top-k strategy can further improve the localization results by integrating the localization results of similar categories on the CUB-200-2011. In contrast, for both VGG16 and ResNet50, the best localization results are obtained for $k = 1$ on ILSVRC.

Hyperparameter λ in total loss. λ denotes the factor of $\mathcal{L}_{BAS}$. A larger λ indicates that more regions in the prediction map are activated. As shown in Table 2, the localization accuracy continues to grow when λ increases from 0.1 to 1, which indicates that the proposed background activation suppression strategy can significantly improve the localization accuracy. The best performance is achieved when $\lambda = 1$ on CUB-200-2011.

$\boldsymbol{\lambda}$	**0.1**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	**1.2**	**1.4**
Top-1	57.99	68.16	68.72	69.99	70.77	**71.33**	69.66	69.81
Top-5	69.24	80.64	82.77	83.16	84.62	**85.33**	83.95	83.62
GT-k.	73.66	85.87	88.36	89.27	90.11	**91.07**	89.57	89.47

Table 2. **Performance** $w.r.t$ $\boldsymbol{\lambda}$**.** λ denotes the factor of $\mathcal{L}_{BAS}$.

Location	Resolution	Top-1 Loc	Top-5 Loc	GT-k. Loc
conv 3-3	56×56	65.35	77.68	83.38
conv 4-3	28×28	**71.33**	**85.33**	**91.07**
conv 5-3	14×14	66.11	79.68	85.36

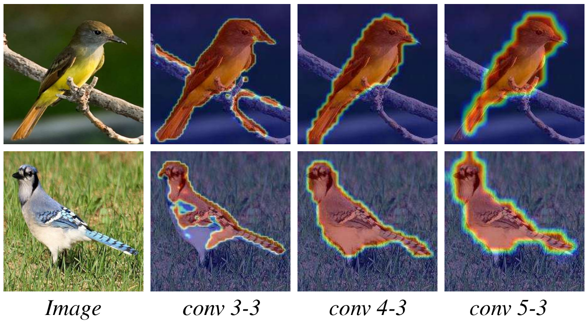


Figure 6. **The results of generator in different layer.**

Generator in different layer. We report the result of inserting the generator at different layers of VGG16. As shown in Fig. 6, we achieve the best results by inserting the generator after the *conv 4-3* layer of VGG16. When the generator learns localization information from shallow feature maps (*conv3-3*), the localization map performs better at the edges of objects, but it is insufficient to resist background distractions. Generator learns localization information from the high-level feature cause imprecise localization due to the limitation of feature map resolution.

Original image vs feature map. We conduct experiments on the intervention position of the background prediction map (original image vs feature map). As shown in Fig. 7, we note that the masked feature map approach achieves higher accuracy and better coverage of the localization results on the object, while the results generated in the original map focus more on the edge texture of the object. It may be due to the fact that the learning process in shallow layers usually focuses on common basic features (e.g., edges, textures) and ignores high-level features.

4.4. Performance Analysis

Localization Quality. In Fig.8, we show the statistical analysis of the IoU between the bounding boxes and the ground-truth boxes when localized correctly, following DANet [23]. On CUB-200-2011, we achieve 77.53% IoU median when localized correctly, exceeding CAM [33] by 17.57%, and correspondingly by 10.04% 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 datasets.

	VGG16 Loc			ResNet50 Loc		
Loc.	**Top-1**	**Top-5**	**GT-k.**	**Top-1**	**Top-5**	**GT-k.**
Img.	70.09	83.25	88.72	73.11	86.98	92.58
Feat.	**71.33**	**85.33**	**91.07**	**77.25**	**90.08**	**95.13**

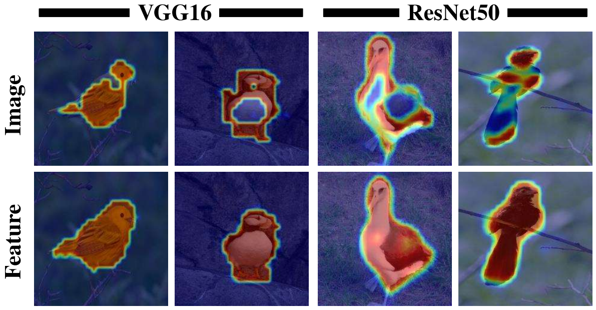


Figure 7. **Comparison of background prediction maps learned from the original image *vs* the feature map.**

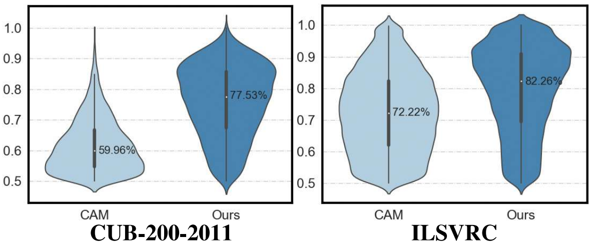
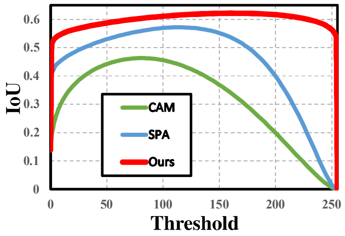


Figure 8. **Statistical analysis of correct bounding boxes.**

Segmentation Quality. We compare the localization map with the ground-truth mask label using two metrics, Peak_T and Peak_IoU, following SPA [13]. As shown in Fig. 9, we evaluate the performance of the proposed BAS with CAM [33] and SPA [13] on VGG16. Compared to SPA, BAS achieves significant and consistent improvement on both Peak_T and Peak_IoU, with a 4.97% improvement in Peak_IoU and 30 in Peak_T, respectively. And it can be seen from the left subgraph of Fig. 9, our IoU-Threshold curve covers a larger area, which indicates that the localization map produced by BAS has fewer low confidence regions and is closer to the original object region.

4.5. Limitation

In this section, we discuss the limitation of BAS. We split the dataset according to the ground-truth box size, as shown in Fig. 10. We note that BAS is inconsistent for localizing objects of different sizes, with poorer localization ability for small objects, especially on ILSVRC. Although it is a great improvement over CAM [33], it is still a challenge to locate small objects better, mainly due to unbalanced distribution between features of the foreground and background. To overcome this limitation, we believe that in further work, object location can be achieved in two stages. 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 objects into a larger one, thereby performing localization in the second stage.



Methods	P.-T	P.-IoU
CAM [33]	80	46.31
SPA [13]	120	57.13
BAS	**150**	**62.10**

Figure 9. **Segmentation Quality.** IoU-Threshold curves for different baseline methods and evaluation results of Peak-T, Peak-IoU on CUB-200-2011.

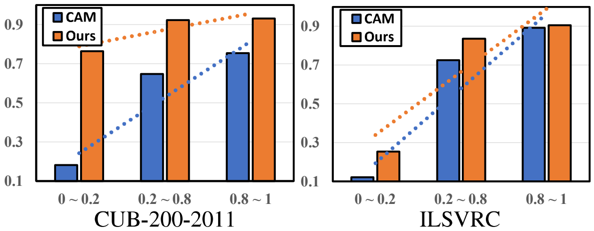


Figure 10. **Limitation.** The performance *w.r.t* different scale. The test set is divided into three intervals of 0~0.2, 0.2~0.8, 0.8~1 according to the percentage of ground-truth box area compared to the image area, and statistics on the localization accuracy of BAS and CAM for various sizes of objects.

5. 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.

Societal Implications. This work may have the following societal Implications. Achieving object localization without the need for location annotations, which will largely reduce manual labeling costs. This is especially valuable for industry or in the medical field since labeling is costly.

Acknowledgments. Supported by National Key R&D Program of China under Grant 2020AAA0105701, National Natural Science Foundation of China (NSFC) under Grants 61872327 and Major Special Science and Technology Project of Anhui (No. 012223665049).

References

[1] Sadbhavana Babar and Sukhendu Das. Where to look?: Mining complementary image regions for weakly supervised object localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1010–1019, 2021. 5

[2] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *European Conference on Computer Vision*, pages 618–634. Springer, 2020. 5

[3] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3133–3142, 2020. 1

[4] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019. 1, 2, 5

[5] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. *arXiv preprint arXiv:2103.14862*, 2021. 1

[6] Guangyu Guo, Junwei Han, Fang Wan, and Dingwen Zhang. Strengthen learning tolerance for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7403–7412, 2021. 1, 3, 5

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[8] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 5

[9] Jeesoo Kim, Junsuk Choe, Sangdoo Yun, and Nojun Kwak. Normalization matters in weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3427–3436, 2021. 1

[10] Weizeng Lu, Xi Jia, Weicheng Xie, Linlin Shen, Yicong Zhou, and Jinming Duan. Geometry constrained weakly supervised object localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 481–496. Springer, 2020. 1, 5

[11] Jinjie Mai, Meng Yang, and Wenfeng Luo. 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*, pages 8766–8775, 2020. 1, 2, 5

[12] Meng Meng, Tianzhu Zhang, Qi Tian, Yongdong Zhang, and Feng Wu. Foreground activation maps for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3385–3395, 2021. 1, 2, 3, 5, 6

[13] Xingjia Pan, Yingguo Gao, Zhiwen Lin, Fan Tang, Weiming Dong, Haolei Yuan, Feiyue Huang, and Changsheng Xu. Unveiling the potential of structure preserving for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11642–11651, 2021. 1, 3, 4, 5, 8

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2, 4, 5, 6

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 6

[16] Krishna Kumar Singh and Yong Jae Lee. 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)*, pages 3544–3553. IEEE, 2017. 1, 2, 5

[17] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[18] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 6

[19] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 4, 5, 6

[20] Jun Wei, Qin Wang, Zhen Li, Sheng Wang, S Kevin Zhou, and Shuguang Cui. Shallow feature matters for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5993–6001, 2021. 1, 2, 4, 5, 6

[21] Jinheng Xie, Cheng Luo, Xiangping Zhu, Ziqi Jin, Weizeng Lu, and Linlin Shen. 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*, pages 132–141, 2021. 1, 3, 5

[22] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[23] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6589–6598, 2019. 5, 7

[24] Seunghan Yang, Yoonhyung Kim, Youngeun Kim, and Changick Kim. Combinational class activation maps for weakly supervised object localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2941–2949, 2020. 1

[25] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regular-

ization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 1, 2, 5

[26] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13460–13469, 2020. 1, 4, 5

[27] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1

[28] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *arXiv preprint arXiv:2202.10108*, 2022. 1

[29] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 1, 2, 5

[30] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 597–613, 2018. 1, 2, 5

[31] Xiaolin Zhang, Yunchao Wei, and Yi Yang. 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*, pages 271–287. Springer, 2020. 1, 2, 5

[32] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Fei Wu. Rethinking localization map: Towards accurate object perception with self-enhancement maps. *arXiv preprint arXiv:2006.05220*, 2020. 5

[33] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1, 2, 4, 5, 6, 7, 8