This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3309621

Background-aware Classification Activation Map for Weakly Supervised Object Localization

Lei Zhu, Qi She, Qian Chen, Xiangxi Meng, Mufeng Geng, Lujia Jin, Yibao Zhang, Qiushi Ren, Yanye Lu*

Abstract—Weakly supervised object localization (WSOL) relaxes the requirement of dense annotations for object localization by using image-level annotation to supervise the learning process. However, most WSOL methods only focus on forcing the object classifier to produce high activation score on object parts without considering the influence of background locations, causing excessive background activations and ill-pose background score searching. Based on this point, our work proposes a novel mechanism called the background-aware classification activation map (B-CAM) to add background awareness for WSOL training. Besides aggregating an object image-level feature for supervision, our B-CAM produces an additional background image-level feature to represent the pure-background sample. This additional feature can provide background cues for the object classifier to suppress the background activations on object localization maps. Moreover, our B-CAM also trained a background classifier with image-level annotation to produce adaptive background scores when determining the binary localization mask. Experiments indicate the effectiveness of the proposed B-CAM on four different types of WSOL benchmarks, including CUB-200, ILSVRC, OpenImages, and VOC2012 datasets.

Index Terms—Weakly Supervised Object Localization, Weakly Supervised Learning, Object Localization

1 Introduction

Weakly supervised learning (WSL), using minimal supervision or coarse annotations for model learning, has attracted extensive attention in recent years and has been widely used in computer vision tasks [1]–[5]. Among them, weakly supervised object localization (WSOL) has immensely profited from WSL, where the requirement of location annotations such as pixel-level masks or bounding boxes can be replaced by easily obtained image-level classification labels. It usually adopts the flow of classification activation map (CAM) [4] that utilizes the structure of image classification to generate the localization score via appending a global average pooling (GAP) operation and a fully connected layer after the feature extractor, *i.e.*, the convolutional network.

Unfortunately, CAM usually activates the most discriminative object part rather than the whole object and requires post-processing to generate the localization mask when used for the WSOL tasks. Thus, a series of WSOL methods have been developed to overcome the above issues. These methods can be divided into multi-stage [6]–[9] and one-stage [10]–[17] methods. The former involves additional training stages as pre- or post-processing to enhance the quality of the localization map or generate class-agnostic localization results, which seriously increases the complexity of both the training and the test processes; while the latter usually adopts different data-augmentation strategies [10]–[13] to erase discriminative object parts, or uses the coarse pixel-level mask as additional pixel-level supervision [14]–[17] to enhance the activation of undiscriminating parts of the objects. Though raising the activation of object locations is a straightforward improvement way, the influence of background locations is not considered, causing ill-posed background threshold searching [18], [19] and unexpected excessive background activation [20].

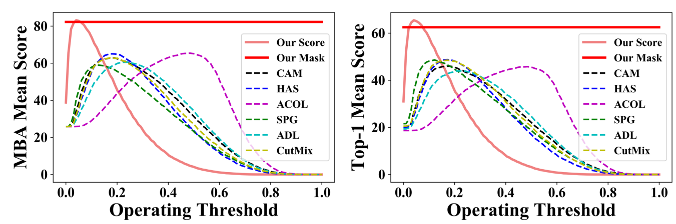


Fig. 1. The performance of WSOL relies much on the background threshold. Our work solves this problem by training an additional background classifier with image label to provide adaptive background scores.

Specifically, the training images of WSOL must contain at least one object, making their image-level label cannot effectively provide background cues. In other words, the pure-background sample remains "unseen" for the image-label-supervised WSOL tasks. Due to this unawareness of background, CAM only can discern different object classes but cannot simultaneously identify whether the location belongs to object parts or background stuff. Thus, current

This work was supported by the Beijing Natural Science Foundation under Grant Z210008, in part by Peking University Medicine Sailing Program for Young Scholars' Scientific & Technological Innovation under Grant BMU2023YFJHMX007, in part by Shenzhen Science and Technology Program under Grant KQTD20180412181221912, and in part by Shenzhen Nanshan Innovation and Business Development Grant.

Lei Zhu, Qian Chen, Mufeng Geng, Lujia Jin, Qiushi Ren are with the Institute of Medical Technology, Peking University Health Science Center, Peking University, Beijing 100191, China, also with the Department of Biomedical Engineering, College of Future Technology, Peking University, Beijing 100871, China, also with the Institute of Biomedical Engineering, Peking University Shenzhen Graduate School, Shenzhen 518055, China, and also with Shenzhen Bay Laboratory, Shenzhen 5181071, China.
Qi She is with the ByteDance AI Lab, ByteDance, Beijing 100086, China.
Xiangxi Meng, Yibao Zhang is with the Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing 100142, China
Yanye Lu is with the Institute of Medical Technology, Peking University Health Science Center, Peking University, Beijing 100191, China, also with the National Biomedical Imaging Center, Peking University, Beijing, 100871, China, also with the Institute of Biomedical Engineering, Peking University Shenzhen Graduate School, Shenzhen 518055, China, email: yanye.lu@pku.edu.cn

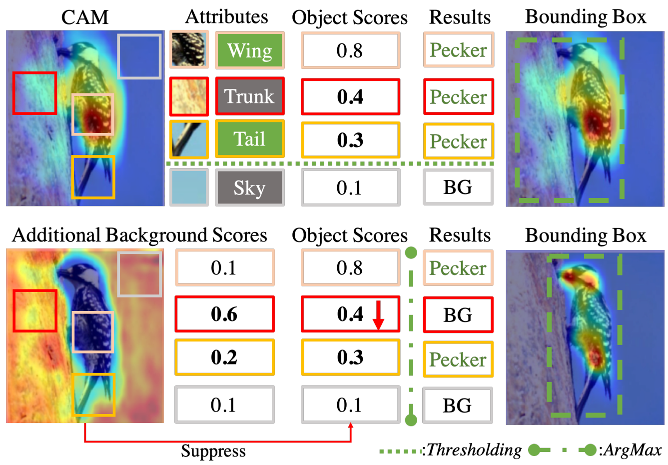


Fig. 2. Activation of object-related background limits the upper bound of WSOL. Our method generates pixel-level background scores to replace the image-level threshold and suppress the background activations.

WSOL methods require additional training stages or post-thresholding to generate the background scores. As indicated in Fig. 1, this fixed background score dramatically influences the functional performance of one-stage WSOL methods.

Beyond that, the absence of pure-background samples also prevents CAM from suppressing the excessive activation of the background locations [20], especially the object-related background that is also discriminative for some objects. For example, in the first row of Fig. 2, the background "trunk" is also informative for discerning "woodpecker", resulting in a higher activation score in the locations of "trunk" relative to "the bird's tail". Even if using the optimal threshold, the bird's tail will still be assigned to the background rather than the foreground woodpecker. Thus, except for the functional performance, the upper bound performance of WSOL methods is also limited by background unawareness.

Compared with raising the activation of object locations upon a fixed threshold, utilizing background cues to generate adaptive background scores and suppress the excessive background activation for WSOL is also a feasible choice to locate objects better, as in the second row of Fig. 2. Inspired by this points, our work focuses on adding background awareness for one-stage WSOL by proposing a novel structure called the background-aware classification activation map (B-CAM). Instead of aggregating a single object image-level feature with GAP, our B-CAM proposes to produce an additional image-level background feature with attention-pooling strategies. This additional background feature acts as the "unseen pure-background samples" for the object classifier to further suppress background activation on the localization maps. Moreover, our B-CAM also learns a background classifier simultaneously with the object classifier by considering background prediction as a multi-label classification task. This background classifier can provide adaptive background scores to replace the threshold searching step when determining the localization mask.

In a nutshell, our contributions are threefold:

- To our knowledge, our paper is the first one-stage WSOL work that simultaneously learns both object and background classifiers with image-level labels.
- A novel structure B-CAM is presented for WSOL to generate pixel-level background scores and suppress the background activation with image-level label.
- Experiments indicate that our method can effectively localize objects with less background activation on four different types of WSOL benchmarks.

2 Related Work

2.1 One-stage Weakly Supervised Object Localization

One-stage WSOL methods follow the pipeline of CAM [4], adopting the classification structure to generate localization score by projecting the classification head (object estimator) back to the pixel-level feature map. However, due to the absence of localization supervision, CAM cannot effectively catch the indiscriminating parts of objects. To solve this problem, some one-stage WSOL methods focused on applying augmentation on input images or feature maps to erase the discriminative object parts. Yun *et al.* [13] proposed a CutMix strategy, which replaces a patch of an image with another image to force the model to capture the indiscriminative features. Singh *et al.* [10] randomly hid the patches of images in the training process to discover different object parts. Zhang *et al.* [11] then simplified this augmentation by proposing an end-to-end network that contains two adversarial classifiers to capture object parts complementarily. Choe *et al.* [12], [21] further adopted the attention mechanism to drop the discriminative parts of the feature map. Chen *et al.* [22] considered the rotation variations of objects and proposed the E^2Net to attend to less discriminative object features. Though these methods can capture more parts of the objects, they inevitably increase the activation of background stuff, especially the object-related background location that also contributes to determining the class of objects.

Apart from adopting augmentation strategies, some one-stage WSOL methods also attempt to use coarse pixel-level supervision to train the object estimator. Zhang *et al.* [14] proposed the self-produced guidance (SPG) approach, which generates an auxiliary pixel-level mask based on the attention map of different extractor stages to perceive background cues. Kou *et al.* [15] further generalized SPG by adding an additional object estimator to adaptively produce the auxiliary pixel-level mask, which is then utilized to design a metric learning loss to better supervise the training process. Ki *et al.* [23] focused on enlarging the distance between features of object locations and background locations in the latent space with the help of the coarse mask generated by non-local attention. Babar *et al.* [16] attempted to enhance the localization map by aligning the localization scores of two complementary images, where these two scores supervise each other at the pixel level. Zhu *et al.* [25] proposed to derive multiple regional localizers based on pixel-level features to reduce the feature discrepancy of the global learned classifier [26].

Recently, to pursue high capabilities for catching long-range dependencies, some methods also explored using self-attention strategies to assist WSOL. Yang *et al.* [27] integrated non-local blocks [28] into the convolutional neural network (CNN) to catch long-range spatial relations for both low-level and high-level features. Gao *et al.* [29] explored utterly replacing the CNN-based baseline with the self-attention-based structure, *i.e.*, the visual transformer [30], for generating better localization maps. Chen *et al.* [31] argued that the visual transformer deteriorates the local feature details and proposed a local continuity transformer to better percept local cues. Bai *et al.* [32] focused on adding spatial coherence for the transformer baseline to enhance the localization performance near object boundaries. Rather than training a transformer for localization, Murtaza *et al.* [33] adopted a frozen-weight transformer to generate class-agnostic bounding boxes, which are used as pseudo-labels to train the CNN-based localization network. Xu *et al.* [34] utilized contrastive language-image pre-training to provide texture tokens for the transformer to assist localization of dense objects. However, though the visual transformer has better representation ability than the CNN, their training process requires large-scale pre-training and careful fine-tuning, limiting its performance on small-scale datasets [35], e.g., in medical image analysis.

In contrast to the one-stage WSOL methods above, our B-CAM only uses image-level labels in the training process to perceive background cues rather than using additional pixel-level supervision. Moreover, our B-CAM also avoids the post-thresholding step required by other one-stage WSOL methods without using any additional training stages.

2.2 Multi-stage Weakly Supervised Object Localization

Multi-stage WSOL methods add additional pre- or post-stages upon the classification structure to pursue better localization performance. Some multi-stage WSOL methods were elaborated to enhance the localization map of the one-stage WSOL by proposing novel post-processing. Zhang *et al.* [17] added an additional learning-free post-stage upon CAM to generate the self-enhanced map, which explores the correlation between each location and the seeds (locations with high localization scores). Pan *et al.* [6] further extended this approach by considering both first- and second-order self-correlation when aggregating the enhanced localization map. Xie *et al.* [36] focused on considering low-level features for localization and proposed a method that included two stages trained for generating and refining the localization map respectively. Belharbi *et al.* [37] adopted an additional training stage to decode the localization map of CAM to pursue higher resolution and boundary adherence. Though these methods enhance the quality of localization maps, they still require post-thresholding to generate background scores.

Some other multi-stage WSOL methods focus on generating class-agnostic localization masks by the additional stages. The most typical work is the pseudo-supervised object localization (PSOL) proposed by Zhang *et al.* [7]. PSOL adds two additional training stages upon the classification stage to generate localization results. In the first stage, the one-stage WSOL method is learned to produce coarse class-agnostic bounding boxes. Then in the second stage, those coarse boxes are used as the ground truth to fully-supervised train bounding boxes regression that generates the region of interest-objects (ROI). Based on this route, Guo *et al.* [9] further proposed SLT-Net that improves PSOL by using a class-tolerance classification model for the localizer to enhance the quality of the coarse bounding boxes. However, these two methods cannot generate pixel-level localization masks as one-stage WSOL methods. As a replacement, another three-stage WSOL method was proposed by Lu *et al.* [8]. This method adopts a generator, implemented by learning- or model-driven approaches, to generate class-agnostic binary masks based on the ROI with different geometry shapes (rectangle or ellipse). In addition, a detector and a classifier are also trained to generate the ROI and class of objects, respectively. More recently, Meng *et al.* [38] improved the multi-stage WSOL methods by jointly optimizing class-agnostic localization and classification to pursue better localization results. Wei *et al.* [24] optimized both inter-class feature similarity and intra-class appearance consistency to reduce the background influence when localizing objects. Though these methods can better generate localization results profited by separating the localization and classification structure or adopting additional localization refining stages, both time and space complexities of the training process are increased. In addition, this type of method only generates class-agnostic localization maps, limiting their application for multi-object localization, where objects with different classes can co-occur in an image.

Compared with these multi-stage WSOL methods, our B-CAM simultaneously learns the background and object classifiers rather than adopting additional training stages for class-agnostic localization. Moreover, both the object and background scores generated by our B-CAM are class-knowable, enhancing flexibility when engaging in multi-object localization and downstream tasks.

2.3 Background Effect in Weakly Supervised Learning

There are also some weakly supervised-learning methods in other scopes designed to capture background cues. Oh *et al.* [39] proposed a background-aware pooling strategy for the weakly supervised semantic segmentation (WSSS) with bounding-boxes annotations, which uses the region out of the ground-truth bounding boxes to catch the inner-boxes background locations. Lee *et al.* [40] utilized the additional saliency map as pixel-level supervision to perceive background cues and reserve rich boundaries for WSSS. Fan *et al.* [41] generated background scores for each class by learning intra-class boundaries, which requires additional superpixel and coarse pixel-level mask during network training. Lee *et al.* [42] proposed two background-aware losses that suppress the localization score of the background frame in the weakly supervised action localization.

Unlike these methods, our B-CAM is designed for WSOL tasks that is harder to locate background cues. Moreover, our B-CAM can perceive the background cues through only image-level labels rather than using the additional pixel-level supervision or off-the-shelf process, for example, the object proposal [43], saliency detection [44], superpixel segmentation [45], or conditional random fields [46].

3 Methodology

In this section, we first analyze the problem of current WSOL methods, *i.e.*, lacking considerations on the background locations, and overview our solution. Then, we illustrate the proposed B-CAM, which adds background awareness

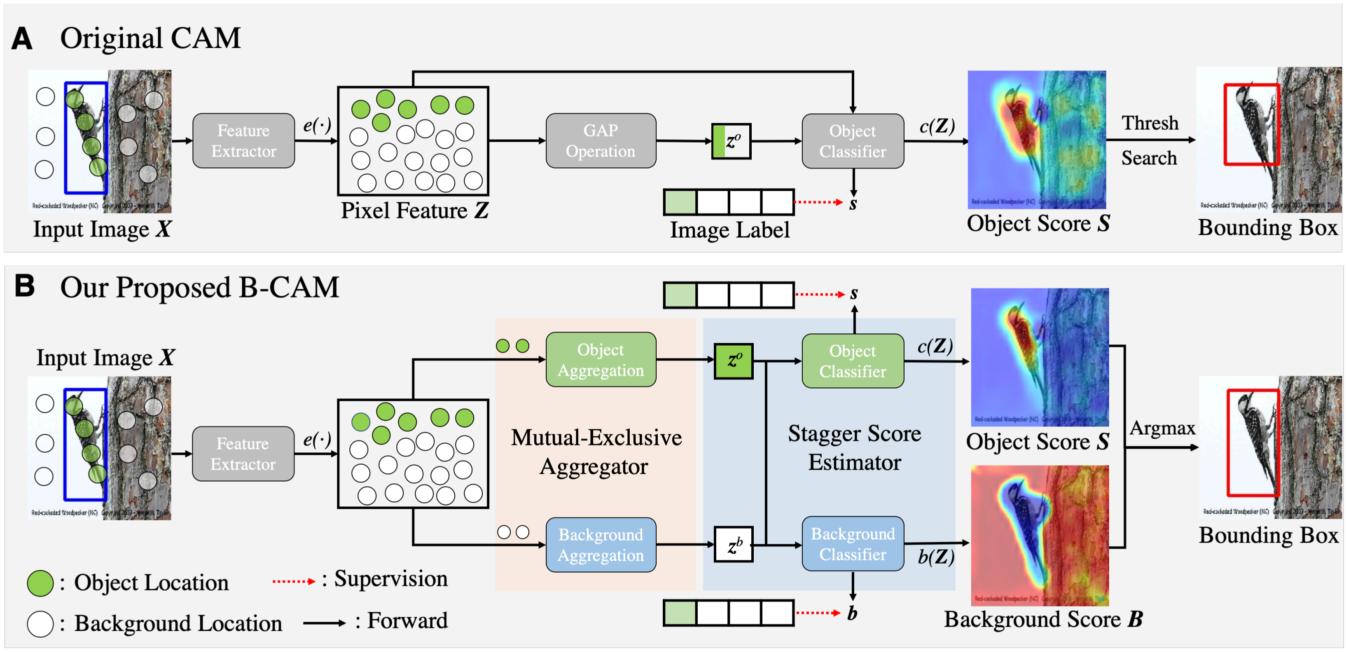


Fig. 3. The comparison of CAM and our B-CAM. A: the structure of CAM. B: Our B-CAM that aggregates two image-level features and produces spatial-specific background scores to produce the localization results with the proposed MEA and SSE.

with only image-level supervision. Finally, we summarize the workflow of our B-CAM for training and inference process.

In this paper, we use **bold** uppercase characters to denote the matrix-valued random variables (the parameter matrices), and ***italic bold*** uppercase characters to represent other matrices (such as feature maps). Vectors are denoted with ***italic bold*** lowercase, and other notations (constants or functions) are represented by normal style. A essential notation list is also provided in our Appendix 1 to clarify the meaning of pivotal symbols used in our paper.

3.1 Problem Definition

Given an input image represented by a matrix $\boldsymbol{X} \in \mathbb{R}^{3\times N}$, the object localization task aims to identify whether the N pixels in $\boldsymbol{X}$ belong to a set of object classes. For this purpose, the localization model adopts a feature extractor $e(\cdot)$ to extract the pixel-level feature $\boldsymbol{Z} \in \mathbb{R}^{C\times N}$, where C represents the dimension of features. Then, an object classifier $c(\cdot)$ further generates the object classification score for each spatial location of $\boldsymbol{Z}$:

$$\boldsymbol{S} = c(\boldsymbol{Z}) = c(e(\boldsymbol{X})) \quad , \tag{1}$$

where $\boldsymbol{S} \in \mathbb{R}^{K\times N}$ represents the localization map of the K target object classes. Finally, the localization map is filtered by a background mask to produce the final localization result $\boldsymbol{Y}^* \in \mathbb{R}^{K\times N}$, whose element $\boldsymbol{Y}^*_{k,i}$ identifies whether or not pixel i belongs to the object of a specific class k.

In contrast to the fully supervised object localization that utilizes the ground truth mask $\mathbf{Y} \in \mathbb{R}^{K\times N}$ to supervise the learning process, WSOL refers to the condition that only the image-level annotation $\boldsymbol{y} \in \mathbb{R}^{K\times 1}$ is available for the whole training process. Thus, an additional GAP layer is required to aggregate $\boldsymbol{Z}$ into the object image-level feature $\boldsymbol{z}^o \in \mathbb{R}^{K\times 1}$ to produce an image-level classification score with the object classifier. Though this aggregation enables WSOL to generate an image-level score for supervision, it also makes the training process pay too much attention to the image-level object classification without concerning the influence of background locations that are also crucial and need to be discerned for the localization task.

Specifically, the GAP-based aggregation contaminates the object image-level feature with the feature of background, causing excessive activation of background locations. As shown in Fig. 3 A, the GAP layer, proposed for the image classification task, treats pixel-level features of the object and the background equally when summarizing the image representations. As a result, $\boldsymbol{z}^o$ is inevitably contaminated by the background locations, where some object-related background cues can also assist the classifier in discerning image classes, as in the case of the background "trunk" *vs.* the object "woodpecker". Although this influence can improve the accuracy and interpretability of image classification, it causes undesirable background activation for WSOL that generates object localization scores by projecting the object classifier back to the pixel-level features, where background locations are also contained.

Moreover, the GAP-based aggregation also disables the training process aware pure-background samples, which are crucial for object localization to percept background locations. In detail, it only aggregates a single object image-level feature, serving as the positive sample of object classification under the supervision of the image-level mask $\boldsymbol{y}$. But, unlike the pixel-level classification supervised by $\boldsymbol{Y}$, this image-level classification does not contain any sample that satisfies $\boldsymbol{y} = \mathbf{0}$, making the pure-background samples unaware during the training process. This absence not only diminishes the capacity of the object classifier to suppress background activation but also disables training a background classifier to generate the pixel-level background scores for filtering the localization map.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3309621

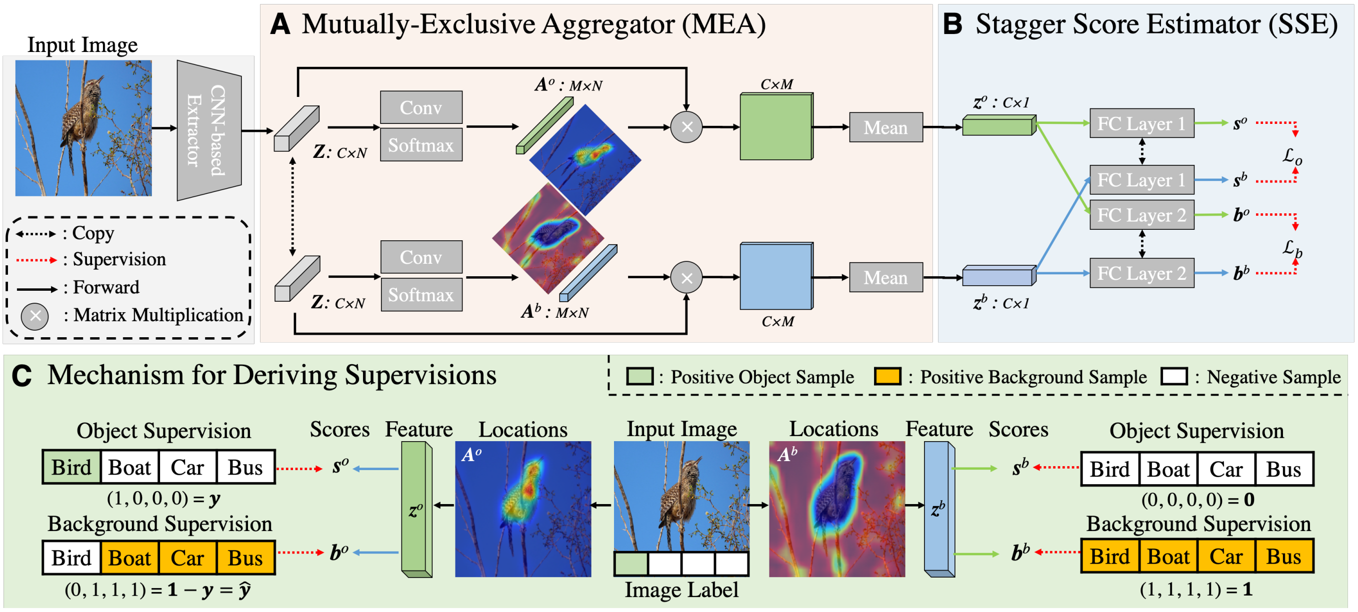


Fig. 4. The structure of the proposed modules and our B-CAM. A: the structure of our MEA that aggregates to image-level features respectively with the object and background locations. B: the structure of our SSE implemented as two fully connected layer with stagger connection to generate four image-level classification scores. C: the mechanism that derives the supervision based on image-level annotations for the scores generated by SSE.

To solve these problems, our B-CAM is proposed as generalized in Fig. 3 B. Instead of generating a single object image-level feature with GAP, the key idea of our B-CAM is to produce an additional background image-level feature $\boldsymbol{z}^b \in \mathbb{R}^{C\times 1}$ to ensure background awareness during the training process. This background image-level feature $\boldsymbol{z}^b$ can simulate the feature aggregated from "the pure-background image" to suppress the background activation on the object classifier $c(\cdot)$. In addition, it also supports training an additional background classifier $b(\cdot)$ with image-level annotation to produce adaptive background scores. Thus, the total target of our B-CAM contains two parts to optimize both the object and background classification tasks with these two image-level features under the supervision of only image-level labels:

$$\mathcal{L} = \mathcal{L}_o(\boldsymbol{z}^b, \boldsymbol{z}^o, \boldsymbol{y}) + \mathcal{L}_b(\boldsymbol{z}^b, \boldsymbol{z}^o, \boldsymbol{y}) \quad , \tag{2}$$

where $\mathcal{L}_o$ and $\mathcal{L}_b$ are the loss function of the object and background classification task, respectively.

3.2 Background-aware Classification Activation Map

For achieving the above purpose, our B-CAM proposes two modules to add background awareness for WSOL: (1) the mutual-exclusive aggregator (MEA) that generates both object and background image-level features by respectively aggregating features on the potential location of the object part and background part; (2) the stagger score estimator (SSE) that adopts a dual classifier structure to predict both the object and background classification scores for the two image-level features as well as derives their supervision. In addition, a stagger classification (SC) loss is also elaborated to train our B-CAM with only image-level annotations effectively.

3.2.1 Mutual-exclusive Aggregator

The proposed MEA aims at purifying the object image-level features to contain more object cues and produce an additional background image-level feature to simulate the pure-background sample. For this purpose, two image-level features $\boldsymbol{z}^o$ and $\boldsymbol{z}^b$ are produced by respectively aggregating the object and background locations.

Firstly, a multi-head spatial attention structure is used to produce two localization priors that coarsely identify whether a spatial position belongs to the object or background. Specifically indicated in Fig. 4 A, two groups of spatial attention maps are utilized as the location priors, which are produced by feeding the pixel-level feature $\boldsymbol{Z}$ into two convolution layers with softmax activation:

$$\begin{cases} \boldsymbol{A}^o_{:,i} = \dfrac{\exp(\mathbf{W}_1 * \boldsymbol{Z}_{:,i})}{\sum_j^N \exp(\mathbf{W}_1 * \boldsymbol{Z}_{:,j})} \\ \boldsymbol{A}^b_{:,i} = \dfrac{\exp(\mathbf{W}_2 * \boldsymbol{Z}_{:,i})}{\sum_j^N \exp(\mathbf{W}_2 * \boldsymbol{Z}_{:,j})} \end{cases}, \tag{3}$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{M\times C}$ are the learnable weight matrixs of convolution layers. $\boldsymbol{A}^o, \boldsymbol{A}^b \in \mathbb{R}^{M\times N}$ represents the object and background location priors, whose accuracy can be guaranteed by the proposed SC loss and detailed in Sec. 3.2.3. M is a hyper-parameters to control the number of spatial attention maps for each group.

Then, these two localization priors are fed into the attention pooling layer [47] to reduce the influence of irrelevant regions when aggregating the two image-level features:

$$\begin{cases} \boldsymbol{z}^o = \dfrac{1}{M}\sum_m^M \sum_i^N \boldsymbol{A}^o_{m,i}\boldsymbol{Z}_{:,i} \\ \boldsymbol{z}^b = \dfrac{1}{M}\sum_m^M \sum_i^N \boldsymbol{A}^b_{m,i}\boldsymbol{Z}_{:,i} \end{cases}. \tag{4}$$

Compared with simply aggregating a single image-level feature with GAP, adopting attention pooling with the localization priors make $\boldsymbol{z}^o$ less contaminated by the feature of background locations. Meanwhile, the additional image-level background feature $\boldsymbol{z}^b$ is also produced to simulate the feature aggregated from "the pure-background image". This sample then supports SSE to learn a background classifier and suppress background activations on localization maps.

3.2.2 Stagger Score Estimator

Benefitting from the proposed MEA, image-level features can be purified and enriched. Thus, SSE is elaborated to better utilize those image-level features for supervising the training process. As shown in Fig. 4, SSE adopts a dual classifier structure to predict both the object and background classification scores for these features and derive the corresponding supervisions with only the image-level label.

Object Classification: Both object and background image-level features are fed into the object classifier, implemented as a fully connected layer, to proceed object classification:

$$\boldsymbol{s}^o = s(\boldsymbol{z}^o) \ \ , \ \ \boldsymbol{s}^b = s(\boldsymbol{z}^b) \ \ , \tag{5}$$

where $\boldsymbol{s}^o \in \mathbb{R}^{K\times 1}$ and $\boldsymbol{s}^b \in \mathbb{R}^{K\times 1}$ are the object classification scores for the object and background image-level features, representing the probability that an object existed in the corresponding aggregated locations. Based on these two classification scores, the supervision of the image-level object classification task can be derived by the following properties:

Property 1. *The image-level feature aggregated mainly by regions of a particular object i.e., $\boldsymbol{z}^o$, is the positive sample for the object classification task on this object. For example in Fig. 4* C *(top-left), the feature aggregated by the locations of "bird" is the positive sample for "bird" classification. Thus, the image-level label $\boldsymbol{y}$ can be used as the supervision for $\boldsymbol{s}^o$ to force the training process of the object classification task.*

Property 2. *The image-level feature aggregated mainly by background locations, i.e., $\boldsymbol{z}^b$, is the negative sample of all objects for the object classification task. For example in Fig. 4* C *(top-right), the feature aggregated by the locations of "trunk" or "sky" does not belong to any objects, i.e., "bird", "boat", "car" and "bus". Thus, zero vector $\boldsymbol{0}$ can be used as the supervision for $\boldsymbol{s}^b$ to force the training process of the object classification task.*

Compared with existing works [4], [10], [12] that only estimate the classification score of the object image-level feature during weakly-supervised training, the additional supervision on the score of background image-level features, *i.e.*, $\boldsymbol{s}^b$, can suppress the activation of background locations to enhance the quality of object localization maps.

Background Classification: Except for engaging background image-level features for training the object classifier, a background classifier, implemented by another fully connected layer, is also utilized by SSE to predict additional background classification scores. Similarly, this background classifier also predicts two scores for the image-level features, representing the probability that their aggregated locations belong to the background of a certain object:

$$\boldsymbol{b}^o = b(\boldsymbol{z}^o) \ \ , \ \ \boldsymbol{b}^b = b(\boldsymbol{z}^b) \ \ , \tag{6}$$

where $\boldsymbol{b}^o \in \mathbb{R}^{K\times 1}$ and $\boldsymbol{b}^b \in \mathbb{R}^{K\times 1}$ represent the class-specific background classification scores for object and background image-level features. With these two scores, the image-level annotation can also be used to train the background classification task based on the following properties:

Property 3. *The feature aggregated mainly on parts of a particular object, i.e., $\boldsymbol{z}^o$, is the negative sample for the background classification task of this object. But it is the positive sample for the background classification task of other objects. For example in Fig. 4* C *(down-left), the feature aggregated by the locations of "bird" is the background of "boat", "car", "bus" and other classes except for "bird". Thus, $\hat{\boldsymbol{y}} = \boldsymbol{1} - \boldsymbol{y}$ can be used as the supervision for $\boldsymbol{b}^o$ to force the training of the background classification task, where $\boldsymbol{1}$ is a vector filled with 1.*

Property 4. *The feature aggregated by some background locations, i.e. $\boldsymbol{z}^b$, is the positive sample for the background classification task of all objects. For example in Fig. 4* C *(down-right), the feature aggregated by the locations of "trunk" or "sky" is the background sample of all objects, including "bird", "boat", "car" and "bus". Thus, $\boldsymbol{1}$ can be used as the supervision for $\boldsymbol{b}^b$ to force the training of the background classification task.*

Profited by engaging the additional background classification task, adaptive background localization scores can be produced for each spatial location by projecting $b(\cdot)$ onto the pixel-level feature $\boldsymbol{Z}$ for the inference process:

$$\boldsymbol{B} = b(\boldsymbol{Z}) = b(e(\boldsymbol{X})) \ \ , \tag{7}$$

where $\boldsymbol{B} \in \mathbb{R}^{K\times N}$ is the background localization maps. Thus, the final localization mask can be produced without using post-processes to search a fixed background threshold [18]:

$$\boldsymbol{Y}^*_{k,i} = \arg\max(\boldsymbol{B}_{k,i}, \boldsymbol{S}_{k,i}) \tag{8}$$

3.2.3 Stagger Classification Loss

Based on the image-level classification scores and their corresponding labels derived by the SSE, an SC loss is further designed to train our B-CAM with only image-level annotations. The proposed SC loss serves as s a multi-task loss that learns both the object classification and background classification task:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_o(\boldsymbol{z}^b, \boldsymbol{z}^o, \boldsymbol{y}) + \mathcal{L}_b(\boldsymbol{z}^b, \boldsymbol{z}^o, \boldsymbol{y}) \\ &= \lambda_1 l_1(\boldsymbol{s}^o, \boldsymbol{y}) + \lambda_2 l_1(\boldsymbol{s}^b, \boldsymbol{0}) + \lambda_3 l_2(\boldsymbol{b}^o, \hat{\boldsymbol{y}}) + \lambda_4 l_2(\boldsymbol{b}^b, \boldsymbol{1})\end{aligned} \ , \tag{9}$$

where $l_1(\cdot)$ is the object classification criterion that is implemented by cross-entropy. $l_2(\cdot)$ is the background classification criterion implemented as multi-label soft margin loss because a location can be the background of multiple classes. In detail, the accuracy of the object classification task is forced by the first two terms. The former ensures the object classification accuracy for the object classifier, and the latter helps suppress its activation on the background locations by the pure-background sample. The other two terms aim at regulating the background scores generated by the background classifier to ensure the accuracy of the background classification.

Moreover, the proposed SC loss can also ensure MEA to aggregate features of pure-object and background locations to form $\boldsymbol{z}^o$ and $\boldsymbol{z}^b$, respectively. To show this effect, we take Eq. 5 and Eq. 6 into Eq. 9 and split it into two parts:

$$\mathcal{L} \begin{array}{l} \nearrow \lambda_1 l_1(s(\boldsymbol{z}^o), \boldsymbol{y}) + \lambda_3 l_2(b(\boldsymbol{z}^o), \boldsymbol{1} - \boldsymbol{y}) \\ \searrow \lambda_2 l_1(s(\boldsymbol{z}^b), \boldsymbol{0}) + \lambda_4 l_2(b(\boldsymbol{z}^b), \boldsymbol{1}) \end{array} . \tag{10}$$

It can be seen that the upper part forces $\boldsymbol{z}^o$ to have a high probability of being discerned as a specific object and a low

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3309621

TABLE 1
Results of WSOL methods on CUB-200 test set

	Top-1 Localization Scores				MBA Localization Scores				Complexity		
	Top-1 70%	Top-1 50%	Top-1 30%	Top-1 Mean	MBA 70%	MBA 50%	MBA 30%	MBA Mean	Flops	Size	T
CAM	$15.38_{\pm 0.22}$	$53.95_{\pm 0.37}$	$69.05_{\pm 0.33}$	$46.12_{\pm 0.27}$	$20.27_{\pm 0.24}$	$72.90_{\pm 0.26}$	$95.61_{\pm 0.12}$	$62.92_{\pm 0.15}$	19.13G	23.92M	✓
HAS	$21.46_{\pm 0.52}$	$56.45_{\pm 0.46}$	$70.16_{\pm 0.40}$	$49.36_{\pm 0.34}$	$27.74_{\pm 0.69}$	$74.33_{\pm 0.61}$	$94.33_{\pm 0.23}$	$65.47_{\pm 0.48}$	19.13G	23.92M	✓
ACOL	$16.19_{\pm 0.54}$	$53.31_{\pm 0.76}$	$66.81_{\pm 0.50}$	$45.43_{\pm 0.41}$	$21.97_{\pm 0.92}$	$74.83_{\pm 1.04}$	$96.53_{\pm 0.32}$	$64.45_{\pm 0.68}$	63.85G	80.55M	✓
ADL	$11.68_{\pm 1.49}$	$48.52_{\pm 2.42}$	$65.53_{\pm 1.71}$	$41.91_{\pm 1.74}$	$16.36_{\pm 1.72}$	$67.50_{\pm 1.79}$	$94.61_{\pm 0.58}$	$59.49_{\pm 1.06}$	19.13G	23.92M	✓
SPG	$13.51_{\pm 0.25}$	$55.20_{\pm 0.49}$	$\mathbf{76.03}_{\pm 0.31}$	$48.25_{\pm 0.29}$	$16.00_{\pm 0.24}$	$65.97_{\pm 0.52}$	$93.93_{\pm 0.20}$	$58.63_{\pm 0.23}$	56.45G	61.67M	✓
CutMix	$17.38_{\pm 0.28}$	$56.18_{\pm 0.24}$	$71.91_{\pm 0.20}$	$48.49_{\pm 0.17}$	$21.86_{\pm 0.36}$	$72.20_{\pm 0.36}$	$94.90_{\pm 0.11}$	$62.99_{\pm 0.22}$	19.13G	23.92M	✓
Ours^m	$\mathbf{43.52}_{\pm 2.84}$	$\mathbf{67.32}_{\pm 1.80}$	$74.15_{\pm 1.23}$	$\mathbf{61.56}_{\pm 1.67}$	$\mathbf{55.20}_{\pm 2.36}$	$\mathbf{87.58}_{\pm 1.60}$	$\mathbf{97.78}_{\pm 0.58}$	$\mathbf{80.20}_{\pm 1.45}$	19.45G	24.74M	×
Ours^p	$\underline{\mathbf{46.20}}_{\pm 1.79}$	$\underline{\mathbf{70.80}}_{\pm 0.69}$	$\underline{\mathbf{77.22}}_{\pm 0.19}$	$\underline{\mathbf{64.74}}_{\pm 0.83}$	$\underline{\mathbf{57.99}}_{\pm 2.32}$	$\underline{\mathbf{90.10}}_{\pm 0.79}$	$\underline{\mathbf{98.87}}_{\pm 0.17}$	$\underline{\mathbf{82.32}}_{\pm 1.00}$	19.45G	24.74M	✓

∗ *"MBA 50%" is also called "GT-Known Loc" [12], considering whether the IoU between the estimated box and the ground-truth box is higher than 50%.*
∗ *"Top-1 50%" is also called "Top-1 Loc" [12], considering whether the classification results and "MBA 50%" are both correct.*

Algorithm 1 Workflow of training the proposed B-CAM

Input: Images set $\{\boldsymbol{X}^i\}$, Labels set $\{\boldsymbol{y}^i\}$

while not reaching stop conditions **do**
  Calculating the pixel-level features $\boldsymbol{Z} \leftarrow e(\boldsymbol{X^i})$
  Producing location priors $\boldsymbol{A}^o, \boldsymbol{A}^b$ by Eq. 3
  Generating image-level features $\boldsymbol{z}^o, \boldsymbol{z}^b$ with Eq. 4
  Extracting image-level classification scores $\boldsymbol{s}^o \leftarrow s(\boldsymbol{z}^o)$ and $\boldsymbol{s}^b \leftarrow s(\boldsymbol{z}^b)$, $\boldsymbol{b}^o \leftarrow b(\boldsymbol{z}^o)$ and $\boldsymbol{b}^b \leftarrow b(\boldsymbol{z}^b)$
  Calculating SC loss $\mathcal{L}$ with Eq. 9
  Backward updating the learning parameters
end while

likelihood of being classified as its background. Likewise, the lower part forces $\boldsymbol{z}^b$ to be indiscriminating for all classes and have a high probability of being the background of all categories. Thus, aggregating pure-object locations for $\boldsymbol{z}^o$ and pure-background locations for $\boldsymbol{z}^b$ will minimize the SC loss, ensuring the accuracy of the localization priors of MEA.

3.3 Workflows

Algorithm 1 summarizes the workflow of training the proposed B-CAM. Specifically, the pixel-level feature $\boldsymbol{Z}$ is firstly computed by the feature extractor, implemented by CNN-based backbone structures [48]–[50]. Then, MEA is utilized to aggregate $\boldsymbol{z}^o$ and $\boldsymbol{z}^b$ with localization priors, representing the object and background image-level features. Next, SSE estimates object and background classification scores for these image-level features and derives their corresponding supervision with only image-level label. Finally, the SC loss is calculated based on the four score/label pairs to guide the update of learning parameters in the training process.

As for the inference process, the pixel-level feature $\boldsymbol{Z}$ is directly fed into SSE to generate the binary localization mask $\boldsymbol{Y}^*$ with Eq. 8. Note that gradient-based approaches [51]–[53] can also use to produce these two localization maps based on the gradient of the classification difference $\frac{\partial(\boldsymbol{s}^o-\boldsymbol{s}^b)}{\partial \boldsymbol{Z}}$ and $\frac{\partial(\boldsymbol{b}^o-\boldsymbol{b}^b)}{\partial \boldsymbol{Z}}$, which improves the localization performance by engaging the whole MEA in the inference process.

4 Experiments

In this section, experiments on different types of datasets are first illustrated to validate our proposed B-CAM, including the single object localization dataset (CUB-200), the single object localization dataset with noisy label (ILSVRC and OpenImages), and the multiple object localization dataset (VOC2012). In addition, the effectiveness and limitation of our B-CAM are further discussed to inspire future works. All experiments in this section were conducted with the help of the Pytorch [54] toolbox on an Intel Core i9 CPU and an Nvidia RTX 3090 GPU. Codes are available at *https://github.com/zh460045050/BCAM*. More experiments are also given in Appendix 3.

4.1 Single Object Localization

Experiments on single object localization were conducted on the CUB-200 dataset [55]. It contains $11,788$ *single-class images* annotated for 200 classes with the corresponding *object bounding box annotations* to benchmark the localization tasks. Following the official setting, $5,994$ images were used as the training set to train the WSOL methods with only image-level labels, and the other $5,794$ images were used to report the performance. Additionally, $1,000$ extra images (5 images per class) annotated by Choe [18] were adopted as the validation set to search the optimal hyper-parameters.

Maximal box accuracy (MBA) [18] was used to evaluate the bounding boxes generated by the localization map. Specifically, for each background threshold, the largest connected component of the predicted binary mask was used as the predicted bounding box. Then, the box accuracy was calculated by counting the number of images where the IoU between the predicted box and the ground truth box was higher than a ratio, *e.g.*, 30%, 50%, and 70%. The maximum scores for all possible thresholds were reported as MBA. Moreover, we also used Top-1 localization accuracy (Top-1) to evaluate both the localization and classification accuracy of the WSOL methods. Note that MBA and Top-1 under 50% IoU are also called "Top-1 Loc" and "GT-Known Loc" in some works [12], respectively.

ImageNet pre-trained ResNet50 [48], [56] was used as the feature extractor. Following Choe [18], its downsample layers before $res4$ and the final fully connected layer were removed to enhance the localization performance. In the training process, input images were resized to 256×256 and then randomly cropped to 224×224, followed by a random horizontal flip operation to form the batches of 32 images. Hyper-parameters were set as $M = 100$, $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$ for our B-CAM. SGD optimizer with weight decay $1e$-4 and momentum 0.9 was used to train our B-CAM for

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3309621

TABLE 2
Comparing with SOTA methods on the CUB-200 test set

	Backbone	Top-1 50%	MBA 50%	MBA Mean	A	M	T
DANet [57]	INC	49.45	67.03	-	✓		✓
I2C [58]	INC	55.99	-	-			✓
MEIL [59]	INC	57.46	-	-			✓
UPSP [6]	INC	53.59	72.14	-			✓
GCNet [8]	VGG	63.24	81.10	-	✓	✓	✓
ORNet [36]	VGG	67.74	86.20	-		✓	✓
ACOL [11]	VGG	45.92	-	-			✓
CCAM [27]	VGG-NL	52.40	-	-			✓
CSOA [15]	VGG	62.31	-	-			✓
TS-CAM [29]	Deit-S	71.30	87.80	-			✓
LCTR [31]	Deit-S	<u>**79.20**</u>	89.90	71.85			✓
PSOL [7]	RES	68.17	-	-	✓	✓	✓
SLT [9]	RES	-	90.70	-	✓	✓	✓
F-CAM [37]	RES	59.10	90.30	79.40	✓	✓	✓
FAM [38]	RES	**73.74**	85.73	-	✓		✓
CutMix [13]	RES	54.81	-	-			✓
ADL [12]	RES-SE	62.29	-	-			✓
PAS [19]	RES	59.53	77.58	-			✓
ICLCA [23]	RES	56.10	72.79	63.20			✓
DGL [60]	RES	61.72	74.65	-			✓
CAAM [16]	RES	64.70	77.35	-			✓
IVR [61]	RES	-	-	71.23			✓
E^2Net [22]	RES	65.10	78.30	-			✓
Ours^m	RES	70.60	89.33	**81.67**			
Ours^p	RES	71.41	<u>**90.83**</u>	<u>**82.90**</u>			✓

∗ "<u>***bold underline***</u>" *indicates the best and* "***bold***" *indicate the second best.*
∗ *"A" indicates the method generates the class-agnostic localization map.*
∗ *"M" indicates the method needs multi training stages.*
∗ *"T" indicates the method needs thresholding to generate localization mask.*

20 epochs. The initial learning rate was set as $1.7e\text{-}4$, divided by 10 every 15 epoch.

Six one-stage WSOL methods were re-implemented with the same backbone structure as ours for fair comparisons, including CAM [4], HAS [10], ACOL [11], SPG [14], ADL [12], and CutMix [13]. Hyper-parameter of those methods were tuned ourselves to guarantee the quality of our re-implementations and given in Appendix 2.1. We also run each method with ten different random seeds and report the mean performance and standard deviation to remove the influence of randomness. For the proposed B-CAM, we evaluated both the object localization score (noted as Ours^p), i.e. $\boldsymbol{S}$, and the final binary mask (noted as Ours^m), i.e. $\boldsymbol{Y}^*$.

Corresponding results are given in Table 1. Our proposed B-CAM significantly improves the quality of the object localization map (Ours^p) and achieves better performance on all evaluation metrics for this fine-grained dataset (16.85% MBA Mean and 15.38% Top-1 Mean scores higher than the best of others) with only a minor complexity increase (0.3 GFlops). This excellent improvement benefits from the trait that our B-CAM can perceive the unseen pure-background samples (images without birds) by the image-level background feature $\boldsymbol{z}^b$ and use it to suppress the localization score of the background area. Moreover, the background localization map $\boldsymbol{B}$ of our B-CAM can also release the background threshold searching process. Directly adopting the background score map $\boldsymbol{B}$ as the binary map (Ours^m) just causes a little reduction in these matrices.

In addition, we also plotted the performance of WSOL methods under different thresholds in Fig. 1. It can be seen that the peak value of our localization map is the highest among all the WSOL methods, indicating the effectiveness of our B-CAM in reducing the activation of background location. Though using the adaptive background score generated by our background classifier will lower the peak performance, it releases the post-threshold searching step, which influences the performance of one-stage WSOL methods. Finally, we also used the recently released localization mask on CUB-200 test set to evaluate the performance of our B-CAM with the peak intersection over union (pIoU) and pixel average precision (PxAP) [18] score. Table 3 shows that the improvement of our B-CAM is still remarkable when evaluated with the fine-grained pixel-level mask, indicating the effectiveness of our B-CAM in suppressing the background activations.

Except for those re-implemented methods, we also compared our B-CAM with some other state-of-the-art WSOL methods on the CUB-200 dataset in Table 2 with their reported localization metrics. It can be seen that our method outperforms all those methods in MBA 50% and MBA Mean localization scores, indicating the satisfactory performance of our B-CAM in localizing objects. Only the Top-1 50% localization score is a bit lower than LCTR [31] and FAM [38], which adopt the visual transformer as the backbone or assist classification by class-agnostic localization map. However, compared with these two methods, our B-CAM is completely based on CNN structure and can generate class-specific localization results, making our method easy to train and can be used for multi-object localization tasks.

To qualitatively represent the performance of the WSOL methods, the localization maps and bounding boxes with optimal thresholds are visualized in Fig. 5. It can be seen that SPG [14] and ACOL [11] seriously suffer from the excessive activation of the background locations, especially for the objects with object-related background (woodpecker/trunk). This is because these two methods both affirm the locations with high activation (may contain object-related background) belong to the object parts. Though the methods that adopt random-erasing augmentation (HAS [10], ADL [12], CutMix [13]) can better catch object parts than CAM [4], they cannot effectively suppress the activation of the background locations, especially near object boundaries. This makes the localization map generated by these methods still larger than the real objects. Compared with those methods, our B-CAM can activate more object parts and avoid excessive background activation, which is beneficial from our awareness of background cues. Thus, the localization boxes generated by our B-CAM have higher IoU than others.

4.2 Single Object Localization with Noisy Label

ILSVRC Dataset: Experiments on object localization with label noise were conducted on the large-scale ILSVRC dataset [56], containing 1.3 million images of 1000 classes. Though images in the ILSVRC dataset may contain objects of multi-classes [62], only the single-class label is provided, where just *the most conspicuous object* is annotated. For example, the image with both "person" and "bird" are only labeled as "person". For the ILSVRC dataset, 50,000 images with bounding box annotations were used to calculate Top-1 50%, MBA 50%, and MBA mean scores for evaluation. The rest images serve as the training set to train WSOL methods with the noise image-level annotations.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3309621

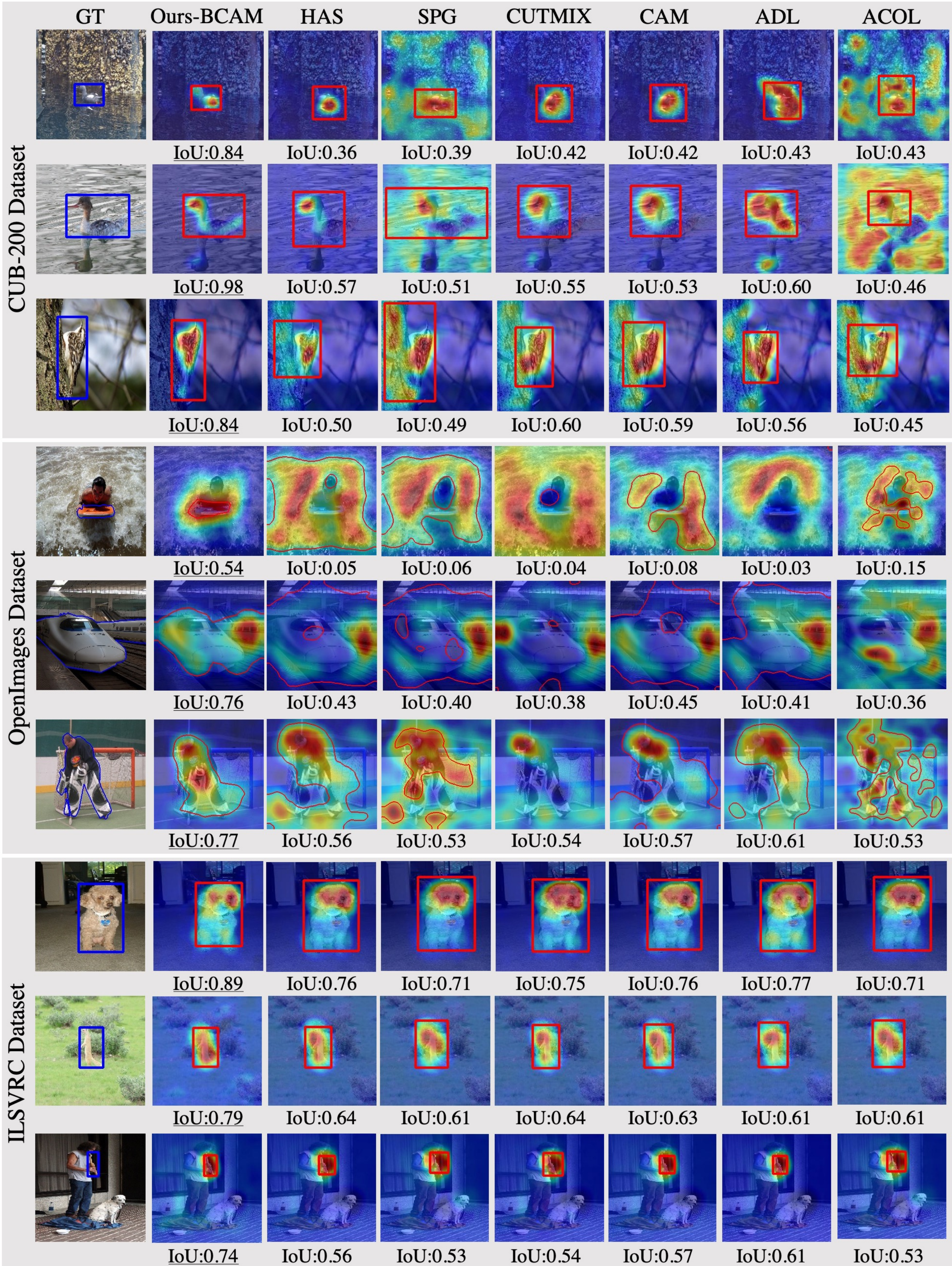


Fig. 5. Visualizations of the object localization scores and predicted bounding boxes of WSOL methods on the CUB-200, ILSVRC and OpenImage datasets. The ground truth bounding boxes/object boundaries are noted in blue color, while the predicted bounding boxes/object boundaries are noted in red. Note that the bounding boxes and localization masks with the highest IoU among all thresholds are visualized for each method in these figures.

TABLE 3
Results of WSOL methods on OpenImages dataset

	CUB-200)		OpenImage			
	Test Set		Test Set		Validation Set	
	pIoU	PxAP	pIoU	PxAP	pIoU	PxAP
CAM	48.60	67.79	42.95	**58.19**	43.42	**58.59**
HAS	49.74	**68.75**	41.92	55.10	42.47	55.84
ACOL	44.17	56.43	41.68	56.37	42.73	57.70
ADL	43.39	56.96	42.05	55.02	42.33	55.26
SPG	43.89	62.01	41.79	55.76	42.17	56.45
CutMix	47.06	65.96	42.73	57.47	43.43	58.18
Ours^m	**53.66**	-	**42.98**	-	**43.70**	-
Ours^p	65.69	85.37	44.31	59.46	44.73	60.27

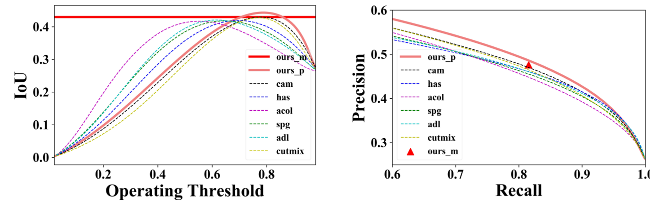
TABLE 4
Comparing with SOTA methods on the ILSVRC validation set

	Backbone	Top-1 50%	MBA 50%	MBA Mean	A	M	T
PSOL [7]	RES	-	65.44	-	✓	✓	✓
SLT [9]	RES	**56.20**	**68.50**	-	✓	✓	✓
FAM [38]	RES	-	64.56	-	✓		✓
CAM* [4]	RES	52.56	65.72	63.78			✓
HAS* [10]	RES	52.33	65.39	63.42			✓
ACOL* [11]	RES	44.90	64.99	62.13			✓
ADL* [12]	RES	50.63	65.85	63.90			✓
SPG* [14]	RES	47.10	64.49	62.17			✓
CutMix* [13]	RES	51.49	64.52	62.73			✓
PAS [19]	RES	-	64.42	63.30			✓
ICLCA [23]	RES	-	65.22	63.40			✓
DGL [60]	RES	-	66.52	-			✓
CAAM [16]	RES	52.36	**67.89**	-			✓
IVR [61]	RES	-	64.93	63.84			✓
E^{2}Net [22]	RES	49.10	63.25	-			✓
Ours^m	RES	**53.29**	66.84	**64.89**			
Ours^p	RES	53.26	66.75	**65.05**			✓

In the training process of ILSVRC, we set $M = 100$, $\lambda_1 = 1$, $\lambda_2 = \lambda_3 = 0.2$, and $\lambda_4 = 0.4$. We also adopted the soft multi-class label on top-5 predictions [63] to reduce the side-effect caused by the label noise when deriving the label of the object image-level feature $\boldsymbol{z}^o$. $1e$-5 was set as the learning rate to train our B-CAM for 3 epochs. The settings of the feature extractor, data pre-processing, and SGD optimizer were the same as the settings of the CUB-200 dataset.

Table 4 shows the performance of our proposed B-CAM and other WSOL methods on ILSVRC datasets. Even though the label noise takes side effects when deriving the label of image-level features with SSE, our B-CAM still outperforms the majority of one-stage methods on this challenged benchmark and effectively solves the dependency of the post-thresholding. In addition, compared with the multi-stage WSSS method such as SLT [9], our B-CAM is lightweight for training and can generate pixel-level localization masks to support downstream weakly supervised semantic segmentation task. Fig. 5 also visualized the quality of localization results of our approach on the ILSVRC dataset. The localization results of our B-CAM are more fining and cover more object locations, which contributes to our higher localization performance.

OpenImages Dataset: Except for the ILSVRC dataset, experiments were also conducted on the OpenImages WSOL dataset [18], [64], whose image-level annotations also contain label noise. This dataset contains 37,319 images of 100 classes, where 2,9819, 2,500, and 5,000 images serve as the training, validation, and test set, respectively. Unlike CUB-200 and



(a) IOU on OpenImages (b) P-R Curve on OpenImages

Fig. 6. Threshold-related metrics on OpenImages dataset. Metrics of our B-CAM are highlighted with solid lines. (a) IoU with different thresholds. (b) P-R curve plotted with different thresholds.

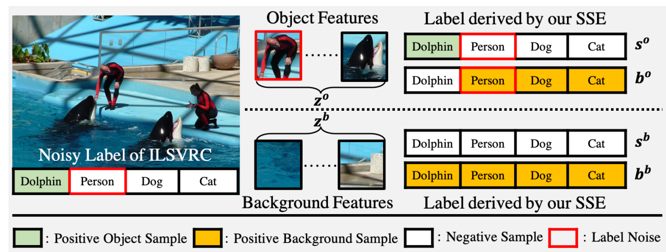


Fig. 7. An Example of the noise-labeled image in the ILSVRC dataset.

ILSVRC datasets, the OpenImages WSOL dataset provides *pixel-level object binary masks* with the *single-class image-level annotation* for validating WSOL in a more fine-grained way.

IoU between the pixel-level ground truth and predicted binary mask was used to quantitatively evaluate the WSOL methods for the OpenImages dataset, where the predicted binary mask can be obtained by thresholding the localization map generated by the WSOL methods with parameter $\tau \in (0, 1)$. The pIoU and PxAP [18] were adopted as the metric to evaluate the performance of WSOL methods based on the pixel-level ground truth.

In the training process, $M = 80$, $\lambda_1 = \lambda_3 = \lambda_4 = 1$ and $\lambda_2 = 0.5$ were set, and our B-CAM was trained for total 10 epochs. The learning rate was set as $1.7e$-4, which was divided by 10 every 3 epoch. The settings of the feature extractor, data pre-processing, and SGD optimizer were the same as the settings of the CUB-200 dataset.

Corresponding results are given in Fig. 6. It shows that the peak of our localization map (Ours^p) is the highest among all the WSOL methods. Though our binary mask (Ours^m) has a relatively lower peak than our localization map (Ours^p), it is still higher than all other WSOL methods and avoids the post-threshold searching step. Moreover, the precision-recall (P-R) curves of the localization maps were plotted based on the precision/recall pairs of different background thresholding scales for evaluation. The P-R curve of our B-CAM is closer to the top right corner, indicating the effectiveness of locating objects. Table 3 also gives the threshold-free metric pIoU and PxAP metrics of the WSOL methods. Our method obtains the maximal improvement over the original CAM among all WSOL methods, achieving 1.36 higher pIoU and 1.27 higher PxAP on the test set. Note that we cannot calculate the PxAP (area under the P-R curve) of our binary masks whose P-R curve degrades into a dot because of its insensitivity to the thresholds. Finally, the qualitative comparisons are also visualized in Fig. 5. The localization results generated

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3309621

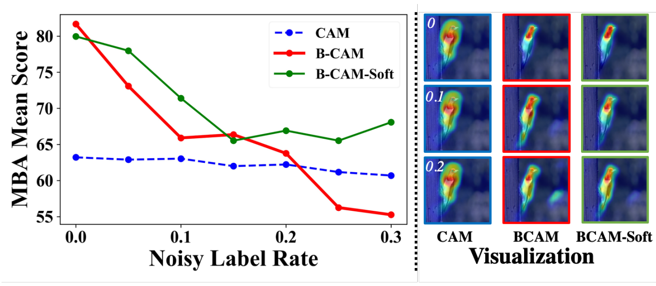


Fig. 8. Results with different noisy label rates on CUB-200 dataset.

TABLE 5
Metric of WSOL methods on VOC2012 dataset

	Official Train Set				Official Validation Set			
	pIoU	SE	PR	SP	pIoU	SE	PR	SP
CAM	45.43	43.53	55.64	33.77	46.60	43.67	56.30	33.15
HAS	45.14	43.50	55.32	33.79	46.32	43.72	56.02	33.26
ACOL	45.28	42.71	55.51	33.97	46.60	42.92	56.08	32.57
SEAM	49.68	51.09	62.81	41.13	51.78	52.01	64.10	40.86
Ours^m	**52.69**	**56.18**	**69.91**	**51.17**	**54.51**	**56.38**	**70.49**	**50.96**
Ours^p	**52.64**	**56.08**	**69.52**	**50.75**	**54.43**	**56.26**	**70.09**	**50.51**

TABLE 6
The mIoU of each classes on VOC2012 official validation dataset

	bg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
CAM	73.0	35.7	24.0	40.1	26.6	41.7	**64.4**	53.0	52.2	24.6	48.5
HAS	73.0	35.7	24.0	39.2	25.6	41.4	63.8	52.9	53.1	24.3	48.2
ACOL	72.0	33.7	23.9	38.4	25.6	45.4	**67.4**	54.3	52.1	23.4	48.9
SEAM	80.0	47.4	25.9	46.3	31.4	48.0	53.5	59.0	55.3	26.8	**49.5**
Ours^m	**82.5**	**54.6**	**29.2**	**55.1**	**39.4**	**48.2**	59.4	**59.3**	**69.1**	**30.6**	49.1
Ours^p	**82.4**	**54.0**	**29.2**	**54.7**	**39.2**	**48.4**	59.1	**59.1**	**69.1**	**30.5**	**50.0**

	table	dog	horse	motor	man	plant	sheep	sofa	train	tv	**avg**
CAM	**44.1**	53.2	49.1	56.4	49.6	32.8	53.5	**46.0**	48.6	37.0	46.6
HAS	43.5	53.3	48.6	56.4	50.5	32.8	53.3	45.7	48.9	33.6	46.3
ACOL	43.9	52.3	48.7	57.1	46.9	33.0	53.0	**46.7**	45.4	**39.1**	46.6
SEAM	**45.9**	58.3	51.0	58.1	58.8	40.0	63.0	50.3	54.3	**40.7**	51.8
Ours^m	36.3	**71.4**	**56.1**	**59.9**	**64.1**	**40.7**	**60.6**	43.1	**61.0**	36.9	**54.5**
Ours^p	36.3	**71.2**	**57.0**	**59.9**	**64.1**	**40.8**	**60.6**	42.9	**60.9**	37.1	**54.4**

by our B-CAM also have better localization performance, less contaminated by object-related background locations (such as "water" for "surfboard") due to our awareness of background cues.

Influence of Label Noise: To better indicate the influence of noise labels for our B-CAM, Fig. 7 gives an example of the noise-labeled image in the ILSVRC dataset, where the image with both "dolphin" and "person" are only labeled as "dolphin". Under such case, our MEA aggregates parts of both "person" and "dolphin" as the object feature z^o. However, due to the label noise, our B-CAM assumes that z^o is the negative sample of "person" for the object classification task based on our Property 1. Correspondingly for Property 3, our B-CAM also assumes that z^o is the positive sample of "person" for the background classification task. Under these supervisions, MEA may tend to catch less part of "person", and SSE will be contaminated in discerning both foreground and background of "person". Thus, in this situation, the improvement of our B-CAM is not as apparent as on the dataset with clean annotations.

For further analyzing the effect of label noise, we artificially added noisy labels into the clean CUB-200 dataset by replacing an image patch with the object part of another image. Under this setting, those images also contain objects that are not annotated by the image-level label. Corresponding results are shown in Fig. 8, indicating our B-CAM (noted by red) is more sensitive to the miss-labeled images than the original CAM (noted by blue). When the noisy label rate reaches 20%, our B-CAM even has lower performance. Fortunately, simply adopting soft multi-class label on top-5 predictions [63] can reduce this side-effect (noted by green), making our B-CAM persistently outperform the baseline even with large label noise rate.

4.3 Multiple Object Localization

The multi-object localization dataset VOC2012 was also used to evaluate the proposed B-CAM, where all the objects with different classes are annotated for a certain image. The VOC2012 dataset [65] contains 14,978 images of 20 classes, where $10,582$ images are annotated by SBD [66]. Unlike the CUB-200, ILSVRC, and OpenImages datasets, the annotation of the VOC2012 dataset gives the *multi-class image annotation*, *i.e.*, annotating all the objects that exist in an image. The pIoU metric and its corresponding sensitivity (SE), precision (PR), and specificity (SP) were used to evaluate the performance.

ResNet38 [67] was used as the feature extractor for this dataset to guarantee fair comparison with the existing method [3]. In the training process, input images were first randomly resized into range (448, 768), and then cropped into 448×448 followed by a color jittering operation to form batches of 8 images. The hyper-parameters were set as $M = 20$ and $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$. SGD optimizer with weight decay $1e$-5 and momentum 0.9 was used to train the WSOL models for a total of 8 epochs. The initial learning rate was set as 0.01, which was delayed by the poly strategy.

Results of our B-CAM and other WSOL methods including CAM [4], ACOL [11], HAS [10], and SEAM [3] are shown in Table 5. It shows that those object localization methods cannot effectively improve the original CAM on the VOC2012 dataset that contains multi-objects in an image. However, our B-CAM can improve the performance to a great extent (7.16% higher mIoU for the validation set), owing to our background awareness. Moreover, compared with the class-agnostic post-thresholding used by other WSOL methods, our background classifier can also generate the background score for each class, which is more reasonable for multi-object localization. So our binary masks (Ours^m) even have a higher mIoU than localization scores (Ours^p).

We also exhibit the performance of the 20 classes on the VOC2012 dataset in Table 6, where our B-CAM obtains better performance nearly on all the categories, especially for the categories with an object-related background (20.90% IoU higher for "plane", 14.4% higher IoU for "train" and 13.8% higher for "boat"). Moreover, for the background class, our B-CAM also has a much larger improvement (10.56% higher IoU), indicating the effectiveness of our B-CAM for suppressing background activations.

Finally, the localization maps of those methods are visualized in Fig. 9, where the masks are selected by the ones with the highest mIoU among all background thresholds. It shows that all other methods face excessive activation on the background locations, especially the object-related

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3309621

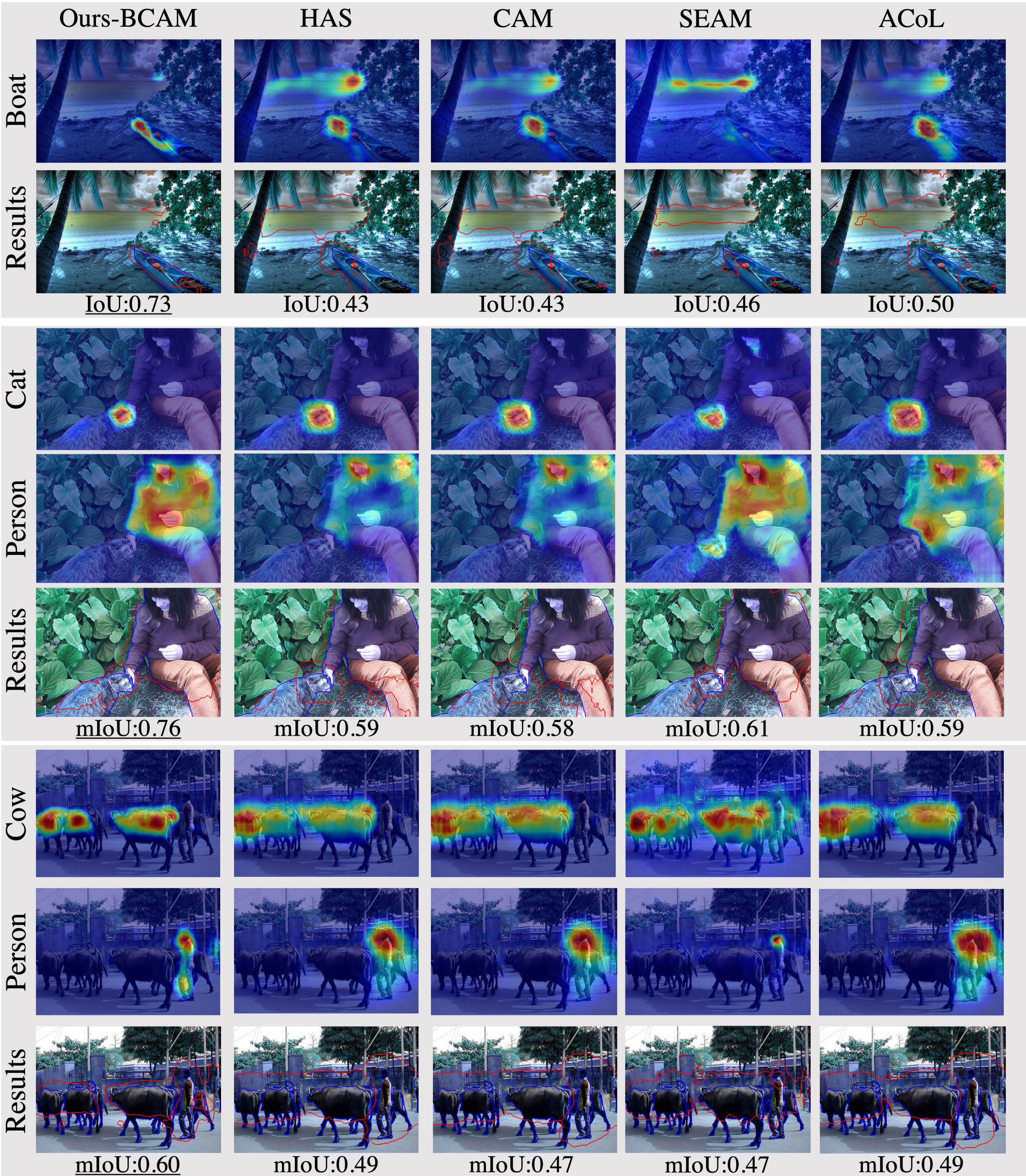


Fig. 9. Visualizations of the localization scores of WSOL methods on the VOC2012 dataset. The ground truth object boundaries are noted in blue color, while the predicted bounding object boundaries are noted in red.

TABLE 7
Ablation studies on the CUB-200 test set

	Top-1 Mean	MBA Mean	OA	BA	BC	SP	Grad	T
CAM	46.71	62.28	×	×	×	×	×	✓
Ours_1	46.40	56.06	✓	×	×	×	×	✓
Ours_2	49.67	59.75	✓	✓	✓	×	×	✓
Ours_3^p	58.43	72.28	✓	✓	✓	✓	×	✓
Ours_3^m	57.25	71.54	✓	✓	✓	✓	×	×
Ours_4^p	65.23	82.90	✓	✓	✓	✓	✓	✓
Ours_4^m	64.65	81.67	✓	✓	✓	✓	✓	×

TABLE 8
MBA for WSOL with different backbones on CUB-200 test set

	VGG16				InceptionV3			
	70%	50%	30%	Mean	70%	50%	30%	Mean
CAM	21.23	73.14	96.77	63.72	13.03	62.31	94.70	56.68
HAS	29.10	69.93	92.10	63.71	12.63	56.21	91.30	53.38
ACOL	15.17	63.20	93.77	57.38	11.10	62.31	95.12	56.17
ADL	23.04	78.06	97.72	66.28	16.86	65.79	93.77	58.81
SPG	17.40	60.98	90.46	56.28	14.26	61.37	92.10	55.91
CutMix	28.58	67.28	91.08	62.31	16.24	62.91	93.13	57.43
Ours	32.94	84.20	98.39	71.85	12.47	75.80	99.34	62.54

TABLE 9
Metrics of the background localization score

	pIoU	PxAP
OpenImage Validation Set	72.75	69.28
OpenImage Test Set	73.71	69.95
CUB-200 Test Set	86.66	81.71

TABLE 10
The metrics in OIS scale of WSOL methods on CUB-200 test set

	Top-1 Localization Scores				MBA Localization Scores			
	70%	50%	30%	Mean	70%	50%	30%	Mean
CAM	34.29	68.05	71.30	57.88	46.51	94.79	99.95	80.42
HAS	42.16	70.68	73.06	61.97	56.52	96.00	99.98	84.17
ACOL	34.04	65.15	68.17	55.79	47.00	94.74	100.00	80.58
SPG	33.36	74.46	79.39	62.40	40.59	92.75	99.93	77.76
ADL	30.53	66.67	69.76	55.66	42.42	94.29	99.98	78.90
CutMix	37.33	71.26	74.97	61.19	48.53	94.15	99.91	80.87
Ours	64.29	77.10	78.08	73.16	81.62	98.62	99.98	93.41

background (water locations for the boat image). Moreover, when facing images with multi-objects, the localization maps of SEAM are also contaminated by those classes. For example, the locations of the cat/person (the second/third images) also have high activation on the localization map of the person/cow. Our B-CAM can avoid this problem because the background cues of each class can be perceived.

4.4 Discussions

Ablation Studies: Ablation studies were also conducted, where the effectiveness of all the proposed parts of our B-CAM are explored with four different settings: 1) Ours_1 only used our object aggregator (OA) to replace the original GAP-based aggregator of CAM; 2) Ours_2 further added the background aggregator (BA) that helps to train an additional background classifier (BC); 3) Ours_3 used the complete SSE that added the staggered path (SP) for generating $\boldsymbol{s}_b$ upon Ours_2 to suppress the background activation. 4) Ours_4 further adopted the gradient-based localization map generation (Grad) to engage the whole MEA in the inference. All models contained the object classifier and adopted the same initialization weights for the common parts.

Table 7 shows the results of these B-CAMs. It illustrates that instead of enhancing the performance, only using OA (Ours_1) even drops the performance compared with the baseline. This is because in such a condition, the object feature is only coarsely formed by OA without any restrictions, which may undesirably contain excessive background or missing object parts. When adding BA and BC (Ours_2), additional restrictions can be added to ensure that the image-level object feature is not classified into the background, which enhances the purity of the object feature. Thus the quality of our localization map raises about 3.27% in Top-1. Next, when adopting the complete SSE, $\boldsymbol{s}^b$ can help to suppress the background activation on the localization map (Ours_3^p) by the second term of SC loss, which brings an 8.76% improvement over Ours_2, when directly evaluating the binary mask (Ours_3^m), the supervised thresholding can be removed with only a 1.25% drop in Top-1. Finally, when engaging our MEA for inference by utilizing the gradient-based map generation, the performance reaches the best, *i.e.*, 64.65 and 81.67 for Top-1 Mean and MBA Mean, respectively.

Generalization for Different Backbones: Besides adopting ResNet50 as the extractor, InceptionV3 [49] and VGG16 [50] structure were also used as the feature extractor. We also compared the performance under these backbones with other WSOL methods to illustrate the generalization of our B-CAM. Corresponding results are given in Table 8, which is in accordance with ResNet50. Specifically, when adopting InceptionV3 as the extractor, our B-CAM achieves 56.68 mean MBA metric, 5.86 higher than the baseline methods. As for VGG16, the improvement is also remarkable, *i.e.*, about 8.13 improvement compared with the baseline for MBA Mean metric. These show the effectiveness of our B-CAM to generalize for different network structures. Note that implementation details and qualitative results of our B-CAM with these backbones are also given in Appendix 2.2.

Effectiveness of the Background Classifier: We evaluated our background localization score on the CUB-200 and Open-Images datasets to verify our background classifier. Specifically, different thresholds are adopted for the background localization score to generate the background localization mask. Then, for an image with class k, we use $1 - \boldsymbol{Y}_{k,:}$ as the ground truth of the background localization task to calculate the pIoU and PxAP metrics that evaluate our background localization score. Corresponding scores are given in Table 9, where the background localization maps of our B-CAM obtain satisfactory scores on these datasets. This indicates the effectiveness of our background classifier.

Upper-bound Performance: To confirm that our better localization map is not attributed to calibration dependency [18], we also explored the upper-bound performance for our B-CAM and other WSOL methods. Specifically, we searched the optimal image-scale (OIS) threshold to generate the binary mask based on the localization map for evaluation. Table 10 shows the scores of our B-CAM and other one-stage WSOL methods. Owing to suppressing the activation on background locations, our B-CAM still outperforms other methods to a great extent. This guarantees

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3309621

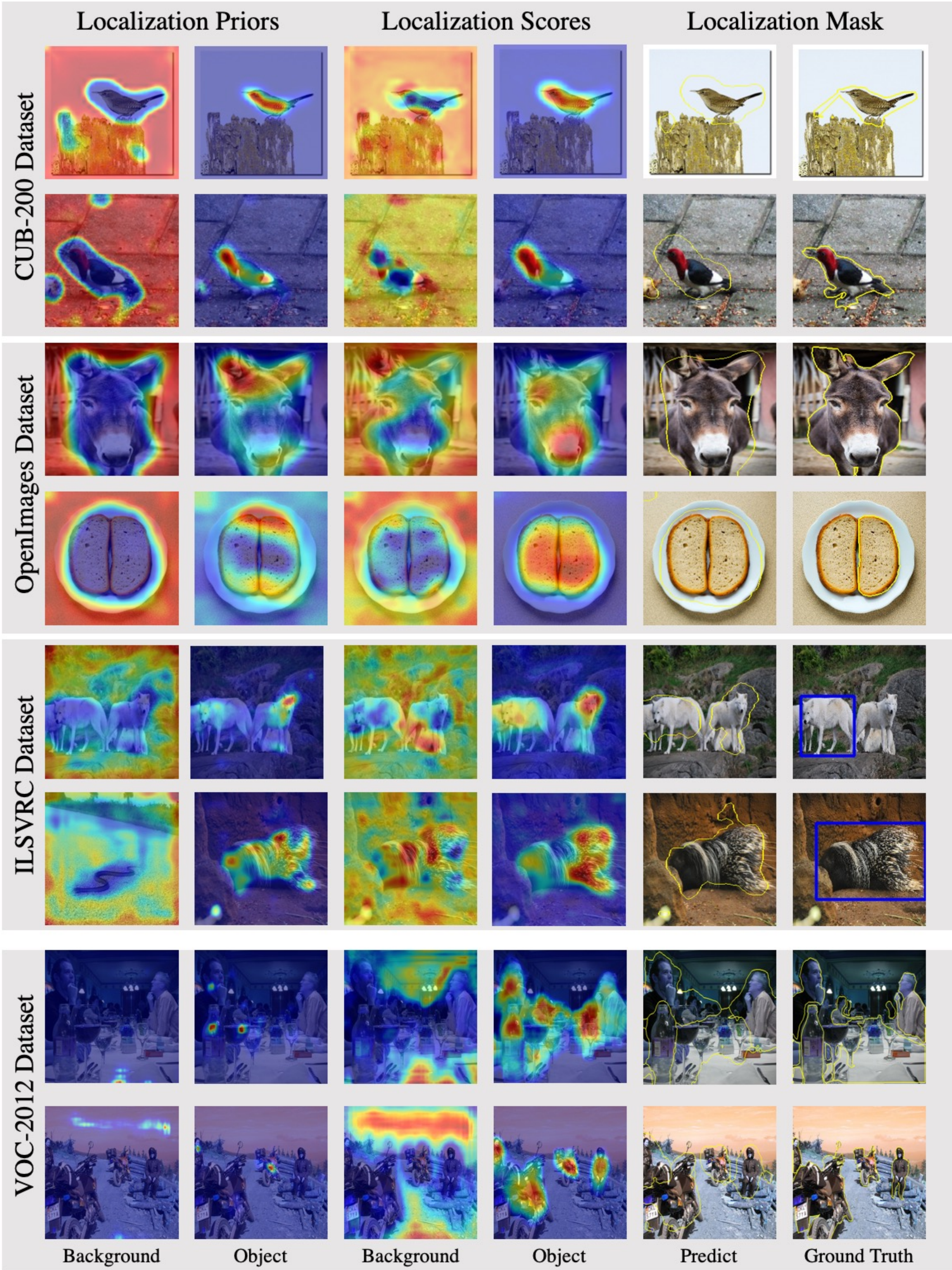


Fig. 10. Visualizations for the intermediate results of our B-CAM, from left to right are the background localization prior $\boldsymbol{A}^b$, the object localization prior $\boldsymbol{A}^o$, the background localization score $\boldsymbol{B}$, the object localization score $\boldsymbol{S}$, the edge map of the predicted mask $\boldsymbol{Y}^*$ and ground truth $\boldsymbol{Y}$.

the effectiveness of our B-CAM in improving the upper-bound quality of the localization map. In addition, it is also worth noting that simply adopting the OIS can obviously improve their performance, e.g., MBA 50% of the CAM with OIS threshold is 94.97, which is already higher than all SOTAs. This indicates that there is still much potential for enhancing WSOL performance by exploring how to generate background scores under image-level supervision better.

Visual Interpretability: Intermediate results are visualized in Fig. 10 to provide visual interpretabilitys of our B-CAM, including the localization priors $\boldsymbol{A}^o$, $\boldsymbol{A}^b$ and the localization scores $\boldsymbol{S}$, $\boldsymbol{B}$. The localization priors are visualized by their mean strength. Specifically, the localization priors efficiently capture some representative background/object locations, which are then used to fuse the two aggregation features to represent pure-background and object samples. Then, the object classifier, trained based on these two aggregation features, can generate better localization maps with less background activation. Moreover, our background classifier can also generate precise background localization, assisting the decision of the final binary masks and bounding boxes. Though the boundary adherence is not good enough due to the weakly supervised manner, our localization map still capture most of the object parts in images.

5 CONCLUSION

This paper proposes the B-CAM to improve WSOL methods by supplementing background awareness, which not only suppresses the excessive activation on background location but eliminates the need for threshold searching step. Experiments on four different types of WSOL benchmarks indicate the effectiveness of our proposed approach. Future works will extend the proposed B-CAM into the downstream localization tasks and some specific fields, such as lesion localization of medical images.

REFERENCES

[1] C.-C. Hsu, K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised instance segmentation using the bounding box tightness prior," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 6586–6597, 2019.

[2] Z. Zhang, Z. Zhao, Z. Lin, X. He *et al.*, "Counterfactual contrastive learning for weakly-supervised vision-language grounding," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 18 123–18 134, 2020.

[3] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 275–12 284.

[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

[5] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Weakly-supervised learning of visual relations," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5179–5188.

[6] X. Pan, Y. Gao, Z. Lin, F. Tang, W. Dong, H. Yuan, F. Huang, and C. Xu, "Unveiling the potential of structure preserving for weakly supervised object localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 642–11 651.

[7] C.-L. Zhang, Y.-H. Cao, and J. Wu, "Rethinking the route towards weakly supervised object localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 460–13 469.

[8] W. Lu, X. Jia, W. Xie, L. Shen, Y. Zhou, and J. Duan, "Geometry constrained weakly supervised object localization," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 481–496.

[9] G. Guo, J. Han, F. Wan, and D. Zhang, "Strengthen learning tolerance for weakly supervised object localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7403–7412.

[10] K. Kumar Singh and Y. Jae Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2017, pp. 3524–3533.

[11] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1325–1334.

[12] J. Choe, S. Lee, and H. Shim, "Attention-based dropout layer for weakly supervised single object localization and semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[13] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6023–6032.

[14] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang, "Self-produced guidance for weakly-supervised object localization," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 597–613.

[15] Z. Kou, G. Cui, S. Wang, W. Zhao, and C. Xu, "Improve cam with auto-adapted segmentation and co-supervised augmentation," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[16] S. Babar and S. Das, "Where to look?: Mining complementary image regions for weakly supervised object localization," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[17] X. Zhang, Y. Wei, Y. Yang, and F. Wu, "Rethinking localization map: Towards accurate object perception with self-enhancement maps," *arXiv preprint arXiv:2006.05220*, 2020.

[18] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim, "Evaluating weakly supervised object localization methods right," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3133–3142.

[19] W. Bae, J. Noh, and G. Kim, "Rethinking class activation mapping for weakly supervised object localization," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 618–634.

[20] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[21] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2219–2228.

[22] Z. Chen, L. Cao, Y. Shen, F. Lian, Y. Wu, and R. Ji, "E2net: Excitative-expansile learning for weakly supervised object localization," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 573–581.

[23] M. Ki, Y. Uh, W. Lee, and H. Byun, "In-sample contrastive learning and consistent attention for weakly supervised object localization," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020.

[24] J. Wei, S. Wang, S. K. Zhou, S. Cui, and Z. Li, "Weakly supervised object localization through inter-class feature similarity and intra-class appearance consistency," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*. Springer, 2022, pp. 195–210.

[25] L. Zhu, Q. Chen, L. Jin, Y. You, and Y. Lu, "Bagging regional classification activation maps for weakly supervised object localization," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*. Springer, 2022, pp. 176–192.

[26] L. Zhu, Q. She, Q. Chen, Y. You, B. Wang, and Y. Lu, "Weakly supervised object localization as domain adaption," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 637–14 646.

[27] S. Yang, Y. Kim, Y. Kim, and C. Kim, "Combinational class activation maps for weakly supervised object localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2941–2949.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3309621

[28] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[29] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, and Q. Ye, "Ts-cam: Token semantic coupled attention map for weakly supervised object localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2886–2895.

[30] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[31] Z. Chen, C. Wang, Y. Wang, G. Jiang, Y. Shen, Y. Tai, C. Wang, W. Zhang, and L. Cao, "Lctr: On awakening the local continuity of transformer for weakly supervised object localization," *arXiv preprint arXiv:2112.05291*, 2021.

[32] H. Bai, R. Zhang, J. Wang, and X. Wan, "Weakly supervised object localization via transformer with implicit spatial calibration," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 2022, pp. 612–628.

[33] S. Murtaza, S. Belharbi, M. Pedersoli, A. Sarraf, and E. Granger, "Discriminative sampling of proposals in self-supervised transformers for weakly supervised object localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 155–165.

[34] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Learning multi-modal class-specific tokens for weakly supervised dense object localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 596–19 605.

[35] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, "Visformer: The vision-friendly transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 589–598.

[36] J. Xie, C. Luo, X. Zhu, Z. Jin, W. Lu, and L. 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*, 2021, pp. 132–141.

[37] S. Belharbi, A. Sarraf, M. Pedersoli, I. Ben Ayed, L. McCaffrey, and E. Granger, "F-cam: Full resolution class activation maps via guided parametric upscaling," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3490–3499.

[38] M. Meng, T. Zhang, Q. Tian, Y. Zhang, and F. Wu, "Foreground activation maps for weakly supervised object localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 3385–3395.

[39] Y. Oh, B. Kim, and B. Ham, "Background-aware pooling and noise-aware loss for weakly supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6913–6922.

[40] S. Lee, M. Lee, J. Lee, and H. Shim, "Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5495–5505.

[41] J. Fan, Z. Zhang, C. Song, and T. Tan, "Learning integral objects with intra-class discriminator for weakly supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4283–4292.

[42] P. Lee, J. Wang, Y. Lu, and H. Byun, "Weakly-supervised temporal action localization by uncertainty modeling," *arXiv preprint arXiv:2006.07006*, 2020.

[43] X.-S. Wei, C.-L. Zhang, J. Wu, C. Shen, and Z.-H. Zhou, "Unsupervised object discovery and co-localization by deep descriptor transformation," *Pattern Recognition*, vol. 88, pp. 113–126, 2019.

[44] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3085–3094.

[45] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 11, pp. 2274–2282, 2012.

[46] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 24, pp. 109–117, 2011.

[47] L. Chen, W. Wu, C. Fu, X. Han, and Y. Zhang, "Weakly supervised semantic segmentation with boundary exploration," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*. Springer, 2020, pp. 347–362.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[51] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE/CVF International Conference on Computer Bision (ICCV)*, 2017, pp. 618–626.

[53] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.

[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.

[55] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, 2009, pp. 248–255.

[57] H. Xue, C. Liu, F. Wan, J. Jiao, and Q. Ye, "Danet: Divergent activation for weakly supervised object localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[58] X. Zhang, Y. Wei, and Y. Yang, "Inter-image communication for weakly supervised localization," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 271–287.

[59] J. Mai, M. Yang, and W. Luo, "Erasing integrated learning: A simple yet effective approach for weakly supervised object localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8766–8775.

[60] C. Tan, G. Gu, T. Ruan, S. Wei, and Y. Zhao, "Dual-gradients localization framework for weakly supervised object localization," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1976–1984.

[61] J. Kim, J. Choe, S. Yun, and N. Kwak, "Normalization matters in weakly supervised object localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3427–3436.

[62] V. Shankar, R. Roelofs, H. Mania, A. Fang, B. Recht, and L. Schmidt, "Evaluating machine accuracy on imagenet," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8634–8644.

[63] S. Yun, S. J. Oh, B. Heo, D. Han, J. Choe, and S. Chun, "Re-labeling imagenet: from single to multi-labels, from global to localized labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2340–2350.

[64] R. Benenson, S. Popov, and V. Ferrari, "Large-scale interactive object segmentation with human annotators," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 700–11 709.

[65] M. Everingham, S. Eslami, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision (IJCV)*, vol. 111, no. 1, pp. 98–136, 2015.

[66] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 991–998.

[67] X. Wang, S. You, X. Li, and H. Ma, "Weakly supervised semantic segmentation by iteratively mining common object features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1354–1362.

Lei Zhu received the master degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2020. He is currently pursuing the Ph.D degree in the Peking University. His current research interests include image segmentation, weakly supervised learning and medical image processing.

Qi She obtained Ph.D. in machine learning and neural computation from the Department of Electrical Engineering at the City University of Hong Kong. He is now working as a Research Scientist at Bytedance. His research interests is statistical machine learning, dynamical systems and theory for representation learning.

Qian Chen is a master student in College of Information Science and Technology, University of Science and Technology of China (USTC). Her research interest is in saliency detection and medical image analysis.

Xiangxi Meng is a research assistant professor in the Department of Nuclear Medicine, Peking University Cancer Hospital. He received Biomedical Engineering PhD degrees from Peking University, Emory University, and Georgia Institute of Technology in 2019. His research interest is in the engineering aspects of molecular imaging, on the interdisciplinary frontiers of nuclear medicine and biomedical photonics.

Mufeng Geng received the bachelor degree from the Beijing Forestry University, Beijing, China, in 2017. He is currently pursuing the Ph.D degree in the Peking University. His current research interests include deep learning and medical image processing.

Lujia Jin received the B.E. degree in Biomedical Engineering from Peking University, Beijing, China, in 2019. He is currently working toward the Ph.D. degree in Biomedical Engineering in Peking University, Beijing, China. His research interests include deep learning and medical image processing.

Yibao Zhang is an associate professor of Medical Physics at Peking University Cancer Hospital. He received his Ph.D. degree from Peking University School of Physics in 2013, during which he did 1.5-year postgraduate fellowship with Prof. Jun Deng at Yale University School of Medicine. His research interests include application of artificial intelligence to the image guided adaptive radiotherapy.

Qiushi Ren is a professor at Peking University. His research interests include multi-modality molecular imaging; and biomedical optics and laser medicine. In 2012, he became the Fellow in the American Institute for Medical and Biological Engineering's (AIMBE) College of Fellows. In December 2014, he was selected as the Fellow in the International Society for Optics and Photonics (SPIE).

Yanye Lu is currently an Assistant Professor at Peking University. He received his Ph.D. degree from Peking University in 2016. Before join Peking University, he was a Postdoc Researcher in Pattern Recognition Lab at Friedrich-Alexander-University Erlangen-Nuremberg, from 2016 to 2021. His research interests include deep learning, computer vision and medical imaging.