Online Refinement of Low-level Feature Based Activation Map for Weakly Supervised Object Localization

Jinheng Xie, Cheng Luo, Xiangping Zhu, Ziqi Jin, Weizeng Lu, Linlin Shen*
Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University
Shenzhen Institute of Artificial Intelligence of Robotics of Society
Guangdong Key Laboratory of Intelligent Information Processing
xiejinheng2020@email.szu.edu.cn, xiangping.zhu2010@gmail.com, llshen@szu.edu.cn

Abstract

We present a two-stage learning framework for weakly supervised object localization (WSOL). While most previous efforts rely on high-level feature based CAMs (Class Activation Maps), this paper proposes to localize objects using the low-level feature based activation maps. In the first stage, an activation map generator produces activation maps based on the low-level feature maps in the classifier, such that rich contextual object information is included in an online manner. In the second stage, we employ an evaluator to evaluate the activation maps predicted by the activation map generator. Based on this, we further propose a weighted entropy loss, an attentive erasing, and an area loss to drive the activation map generator to substantially reduce the uncertainty of activations between object and background, and explore less discriminative regions. Based on the low-level object information preserved in the first stage, the second stage model gradually generates a well-separated, complete, and compact activation map of object in the image, which can be easily thresholded for accurate localization. Extensive experiments on CUB-200-2011 and ImageNet-1K datasets show that our framework surpasses previous methods by a large margin, which sets a new state-of-the-art for WSOL. Code will be available soon.

1. Introduction

Supervised object localization and detection methods based on deep neural networks [22, 14, 21, 13] have achieved great advances. However, these methods usually rely on massive training data with intensive annotations, especially location-level labels. To alleviate the high annotation costs, weakly supervised object localization requiring only image-level annotations has gained lots of attentions.

*Corresponding Author

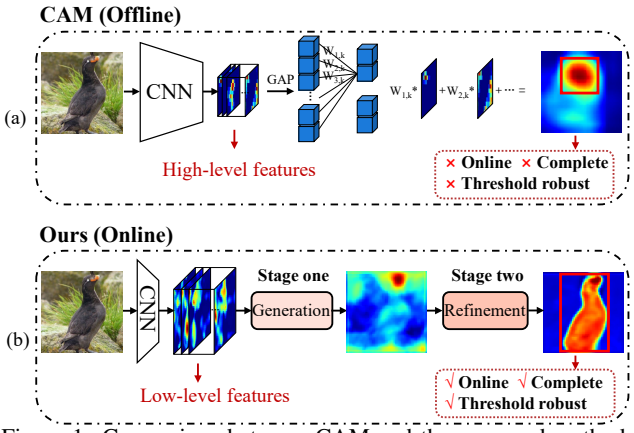


Figure 1: Comparison between CAM and the proposed method. (a) Overview of the CAM pipeline. (b) Overview of the proposed two-stage learning framework. Red bounding boxes illustrate the localization results.

Instead of exploring the entire extent of the objects, classification networks incline to identify patterns from small and sparse regions. Due to this limitation, class activation maps [44] (CAM), a weighted average of high-level feature maps, usually indicate only the discriminative regions of objects. Fig. 1(a) gives the diagram of CAM. As shown in figure, only the head region, which is the most discriminative part to distinguish the bird, is localized. See Fig. 2 for more examples of CAM based localization. However, discriminative region is insufficient for accurate object localization. To address this issue, various solutions [28, 4, 31, 2, 41, 27, 3, 40, 38, 16] have been explored. For example, [11, 40, 3] attempt to erase the most discriminative regions, promoting the network to discover less discriminative regions. As high-level features are used as region guidance, these approaches have limited potentials to derive complete and compact activation maps. Furthermore, when the following thresholding is used to separate foreground and background pixels, the result is very sensitive to the setting of threshold, due to the diverse distribution of

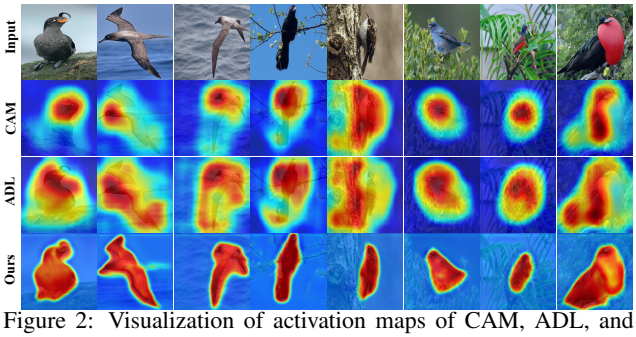


Figure 2: Visualization of activation maps of CAM, ADL, and ours. The images are from the CUB-200-2011 testing set.

activation maps across different objects and background.

The defects of CAM based solutions can be summarized as: (1) The generation of CAM is offline and thus can not be easily refined online. (2) Abstract semantics of high-level features are difficult to produce both complete and compact activation maps. (3) Due to the ambiguity between foreground and background pixels across different activation maps, the localization results are sensitive to the threshold used in post-processing.

In this paper, we develop a two-stage learning framework for weakly supervised object localization, as shown in Fig. 1(b). Compared to high-level features, we argue that low-level features, which contain details or contour information of the object, are more suitable and effective region guidance for object localization. Based on this observation, we propose in the first stage low-level features based activation map generator, to explore those underlying information in the low-level features. It mainly contains an image classifier and an activation map generator (generator for short) with a classification head. The generator is directly incorporated into the shallow layer of the image classifier. Co-supervision of two classification losses leads to an online activation maps generation based on the low-level features. However, in most cases, these activation maps also suffer from two limitations. 1) The pixel value in activation maps intensively locates on around 0.5 (Fig. 4(b)). When thresholding is applied to locate the object, localization results are sensitive to the threshold. 2) While more regions containing contour and context information are included in the low-level feature based activation map, discriminative regions like bird head still dominate the highly activated regions (Fig. 3(b)). To address these issues, the second stage, called entropy-guided refinement, is proposed. The network architecture consists of a generator (initialized from the first stage) and an evaluator (a pre-trained image classifier). The evaluator aims to evaluate and ensure the quality of activation maps produced by the generator through an evaluation loss. Based on this, we further design a weighted entropy loss to alleviate the first limitation, by reducing the uncertainty of each single pixel in activation maps. Moreover, two components, including an attentive erasing and an area loss, are designed to deal with the second limitation. The two components can explore more alternative contents for classification and penalize the background pixels. Designs in the second stage adversarially encourage the generator to further explore the contextual semantics of objects preserved in the low-level features. Fig. 2 shows the comparison results of our method and two other existing popular object localization methods, *i.e.*, CAM [44] and ADL [3]. From Fig. 3(b) to (c) and from left to right in Fig. 4, it reveals that, after the entropy-guided refinement, our method can get well-separated, complete and compact object activation map, to achieve the accurate object localization.

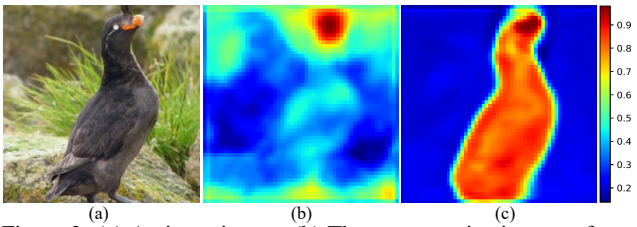
Collectively, the main contributions of this paper can be summarized as:

- We propose to employ low-level features as region guidance to generate activation maps in an online manner, which provides rich contextual information of objects for the following refinement.
- We design the entropy-guided refinement to adversarially drive the network to further explore the low-level features to obtain well-separated, complete, and compact activation maps for accurate object localization. As the activations of objects and background are now well-separated, the localization is not sensitive to the threshold.
- Extensive experiments on CUB-200-2011 [30] and ImageNet-1K [24] datasets show that the proposed method surpasses the previous methods by a large margin, setting a new state-of-the-art performance. Experiments on additional datasets verify the robustness and generalization ability of our method across various scenarios and species.

2. Related Work

Weakly Supervised Object Localization (WSOL) aims to locate objects with less cost of annotation. Among various forms of weak supervision, the image-level label is mostly preferred by researchers. With only image-level labels, diverse solutions [28, 4, 31, 2, 44, 41, 27, 3, 40, 38, 16] have been explored to train deep neural networks for object localization.

CAM based methods. [44, 27, 40, 37, 3] usually employ discriminative regions as guidance to locate the target objects. These class activation maps, a weighted average of high-level feature maps, approximate spatial distribution of discriminative regions. Unfortunately, the active responses in class activation maps only cover the most discriminative regions, instead of the entire object. The random erase of image patches [27] helps to locate less discriminative regions of objects. Furthermore, [11, 12, 3, 17] develop the erasing technique to drop the most discriminative regions.



(a) (b) (c)

Figure 3: (a) An input image. (b) The coarse activation map from stage one. (c) The refined activation map from stage two.

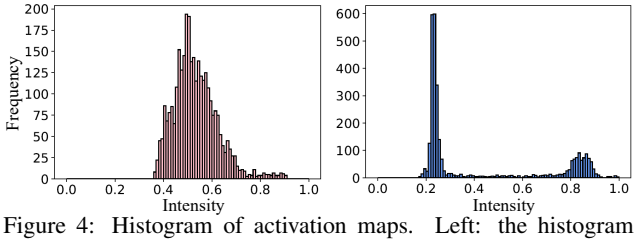


Figure 4: Histogram of activation maps. Left: the histogram calculated from Fig. 3(b). Right: the histogram calculated from Fig. 3(c).

Other than the above solutions, self-produced guidance (SPG) [41] progressively generates object masks in a stage-wise manner. It employs high confident object regions as the seeds of foreground and the supervision of the lower layers. By integrating foreground regions, it enforces classification networks to learn pixel correlations from multiple layers. However, uncertain regions are less explored.

Geometric constraint. [16] designs a novel network architecture called GC-Net, which contains a detector, a generator, and a classifier. The detector predicts a set of location coefficients, which are transformed by the generator to a 2D mask. Then the input image is split into foreground and background regions using the mask. During training, the categorical cross-entropy is adopted to minimize the uncertainty of object classification and the negative entropy loss is used to maximize the uncertainty of background classification. In this way, the mask gradually approaches the target object. However, the training of GC-Net might be unstable, due to the abstract transformation from numerical values to the 2D matrix. Besides, this approach has less potential to be applied to multi-instance localization.

Attention mechanism. Deep networks with attention mechanisms focus on informative semantics. Since the proposal of attention [29, 33, 19, 35, 9, 39], various tasks (*e.g.* classification and detection) have involved attention mechanism for better features learning.

[3] utilizes an attention mechanism to hide the most discriminative parts and stochastically highlight the informative regions. However, a hard drop of the most discriminative part loses pixel correlations and expanding cues. In contrast, in our setting, we add constraints to the areas to be dropped and drop pixels with adaptive probabilities.

Residual Attention Networks (RAN) [32] stacks multiple attention modules to capture mixed attention. Attention Branch Network (ABN) [6] introduces a branch structure with an attention mechanism. This branch structure built on a top layer directly generates an attention map in the training iteration. Likewise, our method designs a sub-network to produce the activation map simultaneously, but on the shallow layer.

3. Methodology

Details of the proposed two-stage learning framework are presented in Fig. 5. **First stage**: As shown in Fig. 5(a), the activation map generator with the classification head is integrated into the shallow layer of an image classifier (*e.g.* [7, 26]). During training, supervised by two classification losses, the generator employs the low-level features to online yield activation maps of objects with rich contextual information. **Second stage**: After the first stage, we employ an evaluator (a pre-trained image classification network, *e.g.* [26, 7]), coupled with the generator (initialized from the first stage), to evaluate the generated activation maps (Fig. 5(b)). During training, three loss functions and an attentive erasing are proposed to supervise the model. In particular, the generator gradually yields well-separated, complete, and compact activation maps based on the contextual object information preserved in the low-level features. During inference, as shown in Fig. 5(c), only shallow layers of the classifier and the generator are needed to predict activation maps for object localization. In the following, we provide more details about the proposed framework.

3.1. Low-level Feature Based Activation Map

First stage. The designed activation map generator $\mathcal{F}^g$ is incorporated into shallow layer of the image classifier $\mathcal{F}^c$, which separates the classifier into two sub-networks $\mathcal{F}_1^c$ and $\mathcal{F}_2^c$ (as shown in Fig. 5(a)). Specifically, given an input image I, the low-level features can be derived as follow:

$$f^c = \mathcal{F}_1^c(I; \mathbf{W}_1^c), \tag{1}$$

where $f^c \in \mathbb{R}^{h \times w \times c}$ is the low-level features. $\mathbf{W}_1^c$ represents the learning parameters of $\mathcal{F}_1^c$. h, w, c denote the height, width, and number of channel of f^c, respectively. To generate the activation map p^a, f^c is fed to the generator, which consists of encoder-decoder layers $\mathcal{F}^d$, a 2D convolution layer ($\mathbf{Conv}$), and a Batch Normalization layer ($\mathbf{BN}$):

$$f^a = \mathcal{F}^d(f^c; \mathbf{W}^d), \quad p^a = \mathbf{BN}(\mathbf{Conv}(f^a; \mathbf{W}_1^p)), \tag{2}$$

where $\mathbf{W}^d$ and $\mathbf{W}_1^p$ represent the learning parameters of $\mathcal{F}^d$ and $\mathbf{Conv}$-$\mathbf{BN}$ layer, respectively, $f^a \in \mathbb{R}^{h \times w \times k}$ is the output features of $\mathcal{F}^d$ with k channels, where k denotes the number of classes. $p^a \in \mathbb{R}^{h \times w \times 1}$ is the generated activation map. Following the sub-network $\mathcal{F}_1^c$, $\mathcal{F}_2^c$ aims to produce

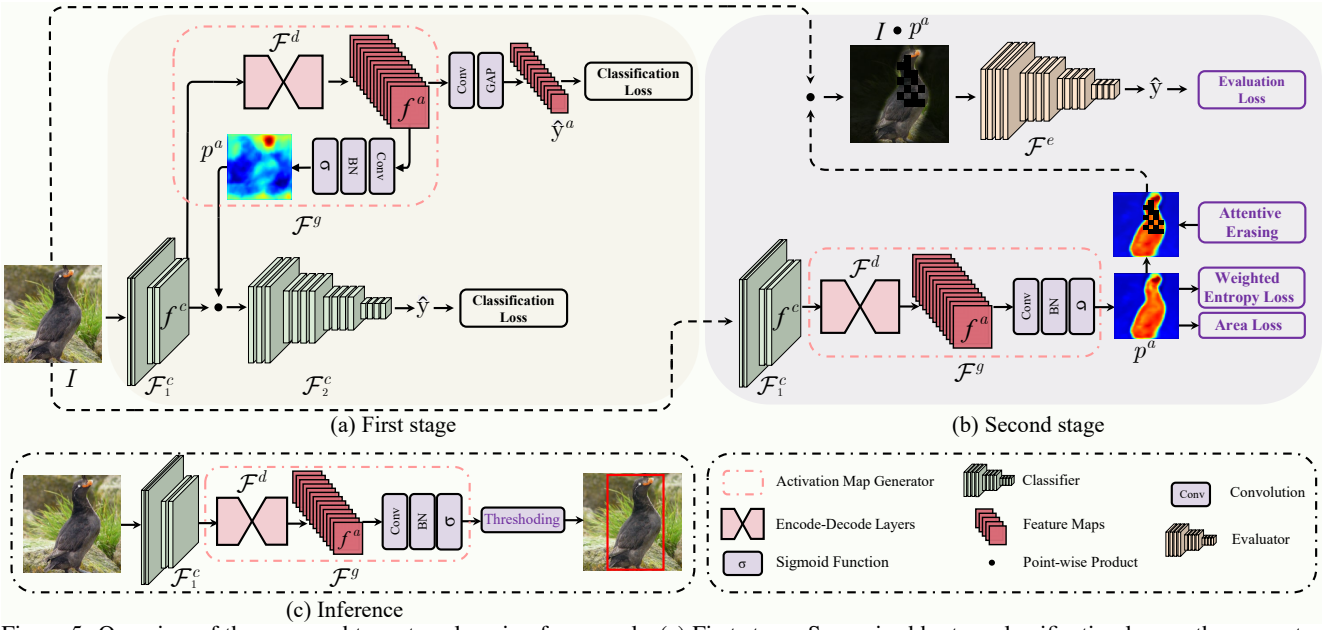


Figure 5: Overview of the proposed two-stage learning framework. (a) First stage. Supervised by two classification losses, the generator $\mathcal{F}^g$, incorporated in the shallow layer of the classifier $\mathcal{F}^c$, turns the low-level features into coarse activation maps p^a. (b) Second stage. The network consists of a generator $\mathcal{F}^g$ (initialized from the first stage) and an evaluator $\mathcal{F}^e$. The evaluator aims to evaluate activation maps predicted by the generator through the evaluation loss. In addition, the weighed entropy loss, attentive erasing, and area loss are proposed to adversarially drive the generator to refine the coarse activation maps. (c) The inference for locating objects.

the class probability distribution $\hat{y}$:

$$\hat{y} = \mathcal{F}_2^c(p^a \cdot f^c; \mathbf{W}_2^c), \tag{3}$$

where $\mathbf{W}_2^c$ represents the learning parameters of $\mathcal{F}_2^c$, $p^a \cdot f^c$ is the point-wise product between p^a and f^c. p^a act as the mask to mask out the background clutters in f^c. To get more accurate activation maps, an auxiliary classification head, including one Global Average Pooling ($\mathbf{GAP}$) layer and one Convolution ($\mathbf{Conv}$) layer, is added following $\mathcal{F}^d$:

$$\hat{y}^a = \mathbf{GAP}(\mathbf{Conv}(f^a; \mathbf{W}_2^p)), \tag{4}$$

in which $\mathbf{W}_2^p$ represents the learning parameters of $\mathbf{Conv}$.

With corresponding image-level one-hot encoding label y, the classification losses $\mathcal{L}_c(y, \hat{y})$ and $\mathcal{L}_c(y, \hat{y}^a)$, respectively corresponding to $\hat{y}$ and $\hat{y}^a$, are formulated as:

$$\mathcal{L}_c(y, \hat{y}) = -\sum_i^k y_i \log\left(\frac{e^{\hat{y}_i}}{\sum_j^k e^{\hat{y}_j}}\right), \tag{5}$$

$$\mathcal{L}_c(y, \hat{y}^a) = -\sum_i^k y_i \log\left(\frac{e^{\hat{y}_i^a}}{\sum_j^k e^{\hat{y}_j^a}}\right). \tag{6}$$

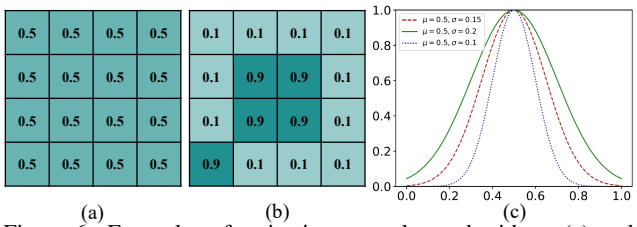
$\mathcal{F}^g$ and $\mathcal{F}_2^c$ can work complementarily during network training. As aforementioned, before feeding f^c into the classification subnetwork $\mathcal{F}_2^c$, p^a is applied to mask out the background clutters in f^c. Thus, with the supervision of the label y, the generator $\mathcal{F}^g$ learns to explore the contextual semantics in the low-level features, *i.e.* excite the object regions in p^a.

In addition, the auxiliary classification head following $\mathcal{F}^g$ can further encourage $\mathcal{F}^g$ to learn more underlying information of the object in the low-level features, such that activation map p^a includes more details and contextual information.

The Attention Branch Network (ABN) [6] has a similar architecture as our model. However, there are several differences: (1) We employ the low-level features as region guidance to generate the coarse activation map p^a. Compared with the high-level features used in ABN, the low-level features, persevering more details and contour information of the object, are more suitable for object localization. (2) The low-level features usually contain high-frequency noise, *e.g.*, twigs, and gravels. Based on this observation, the encoder-decoder layers $\mathcal{F}^d$ is added to alleviate this problem. (3) Our network architecture is designed to preserve more information from low-level features.

3.2. Entropy-guided Refinement

Second stage. As shown in Fig. 5(b), the network architecture consists of the shallow layers of the classifier, a generator, and an evaluator. Both the shallow layers and generator are initialized from the first stage. Supervised by the evaluation loss, the evaluator, following the generator, aims to evaluate the quality of the generated activation maps. Based on this, we further propose the weighed entropy loss, attentive erasing, and area loss to address the limitations aforementioned above.



(a) (b) (c)

Figure 6: Examples of activation maps learned without (a) and with (b) entropy loss. (c) illustrates the weighted function with different parameter settings, *i.e.* μ, σ.

Forward inference of the network can be formulated as:

$$f^c = \mathcal{F}_1^c(I; \mathbf{W}_1^c), \tag{7}$$

$$p^a = \mathcal{F}^g(f^c; \mathbf{W}^p), \tag{8}$$

$$\hat{y} = \mathcal{F}^e(I \cdot p^a; \mathbf{W}^{e*}), \tag{9}$$

where $\mathbf{W}_1^c$ and $\mathbf{W}^p$ are initialized from the first stage, pre-trained parameters $\mathbf{W}^{e*}$ of the evaluator $\mathcal{F}^e$ are fixed.

The overall training loss of the second stage can be formulated as:

$$\mathcal{L} = \mathcal{L}_e(y, \hat{y}) + \alpha \mathcal{L}_w(p^a) + \beta \mathcal{L}_a(p^a), \tag{10}$$

where we adopt Eq. 5 as the evaluation loss $\mathcal{L}_e$. $\mathcal{L}_w$ is the weighed entropy loss, $\mathcal{L}_a$ is the area loss. α and β are the hyper-parameters. In addition, in order to get a more complete activation map of the object, attentive erasing is designed and applied in model training.

3.2.1 Weighed Entropy Loss

Entropy, a measure of uncertainty of random variable X, is defined as:

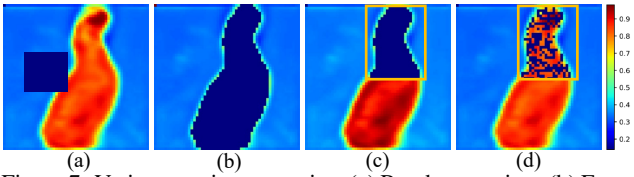
$$H(X) = -\sum_{x \in X} P(x) \log P(x), \tag{11}$$

where $P(x)$ is the occurrence probability of event x. Minimizing the entropy $H(X)$ can reduce the uncertainty of X. Based on this observation, we propose the entropy loss to reduce the uncertainty of activation map p^a. The $(i,j)^{th}$ pixel in the activation map p^a is represented as a random variable $X_{i,j}$, which takes the values 1 (*i.e.* foreground) and 0 (*i.e.* background). Hence, $P(X_{i,j} = 1) = p^a_{i,j}, P(X_{i,j} = 0) = 1 - p^a_{i,j}$. Specifically, the entropy of a single pixel can be formulated as:

$$H(X_{i,j}) = -p^a_{i,j} \log(p^a_{i,j}) - (1 - p^a_{i,j}) \log(1 - p^a_{i,j}), \tag{12}$$

where $p^a_{i,j}$ indicates the $(i,j)^{th}$ element of p^a. Then the entropy loss of p^a is defined as the mean of entropy of $X_{i,j}$:

$$\mathcal{L}_h(p^a) = \frac{1}{hw} \sum_{i=1}^{h} \sum_{j=1}^{w} H(X_{i,j}). \tag{13}$$



(a) (b) (c) (d)

Figure 7: Various erasing strategies. (a) Random erasing. (b) Erasing the regions with activation higher than a threshold. (c) Erasing the regions within a restricted rectangle. (d) The proposed attentive erasing.

While it's difficult to predict foreground/background for pixels with activation around 0.5, they are the main contributions of uncertainty. This motivates us to extend Eq. 13 to our weighted entropy loss, in which the Gaussian distribution shown in Fig. 6(c), is used to assign adaptive weight to pixels with bigger uncertainty. Specifically, the proposed weighted entropy loss is:

$$\mathcal{L}_w(p^a) = \frac{1}{hw} \sum_{i=1}^{w} \sum_{j=1}^{w} \gamma_{i,j} \cdot H(X_{i,j}), \tag{14}$$

where $\gamma_{i,j}$ is defined as:

$$\gamma_{i,j} = e^{-\frac{(p^a_{i,j} - \mu)^2}{2\sigma^2}}, \tag{15}$$

where σ and μ represent the variance and mean of the Gaussian Distribution, respectively. In our setting, $\sigma = 0.1$ and $\mu = 0.5$.

With the co-supervision of evaluation loss and weighted entropy loss, the generator tends to excite the foreground pixels and suppress the background pixels, which generates exact and well-separated activation maps (Fig. 3(c)). Fig. 6 shows the examples of activation map learned without (a) and with (b) the entropy loss, one can observe that the number of uncertain pixels (0.5) has been substantially reduced by the proposed entropy loss.

3.2.2 Attentive Erasing

With image-level labels, classification network mainly emphasizes the most discriminative object regions. However, the completed region of object is needed for accurate localization. A number of erasing strategies have been proposed to encourage the network to explore less discriminative regions of the object, during the generation of activation map. Fig. 7 compares a number of available erasing strategies. As shown in the figure, random region erasing could involve the background pixels (Fig. 7(a)) and erasing the region with activation higher than a threshold sometimes causes the removal of whole object (Fig. 7(b)).

To further improve the completeness of our low-level feature based activation map, we design a so called attentive erasing in this paper, to randomly erase regions at pixel level. Firstly, we choose the coordinate of peak response in

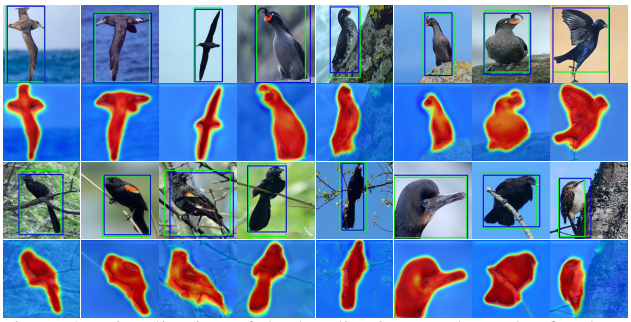


Figure 8: Visualization of the localization results and refined activation maps on CUB-200-2011. Ground-truth and predicted bounding boxes are highlighted in blue and green, respectively.

p^a as the center to produce a rectangle with a random height and width. As shown in Fig. 7(c), the rectangle aims to restrict the erasing region, such that only pixels with value higher than a threshold and fall into the rectangle are considered as the candidates for erasing. The candidate pixels are then dropped with a probability of 0.5 (Fig. 7(d)).

Compared to other strategies in Fig. 7, our approach can encourage the network to explore entire object and retain part of the discriminative region at the same time.

3.2.3 Area Loss

Until now, there is no area constraint on the activation map p^a, which may lead to yield overlarge bounding box for inaccurate object localization. To solve this problem, we propose the area loss:

$$\mathcal{L}_a(p^a) = \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w} p^a_{i,j}. \tag{16}$$

This encourages the generator to reduce the excitation of irrelevant background clutters and ensures the compactness of activation maps for accurate object localization.

In the entropy-guided refinement stage, both the network architecture and components are designed to adversarially explore the object information preserved in the low-level features for accurate object localization. Fig. 3(a) and (b) gives examples of the activation map generated from stage one and stage two. It can be found that a more accurate, well-separated, complete, and compact activation map can be generated after the refinement.

4. Experiments

We evaluate the proposed method on CUB-200-2011 [30] and ImageNet-1K [24] datasets. Extensive experiments show that our approach consistently achieves significant improvements on these benchmarks. Besides, the quantitative analysis is also conducted to verify the effectiveness of each component. We also apply our localization model to person re-identification datasets like Market-1501 [43], Duke-MTMC [23] and MSMT17 [34] and fine-grained classification datasets like Standford Dog [10] and FGVC-Aircraft [18], and show the visual results in supplementary materials. The results clearly justify that our model can achieve robust localization results across a large number of different objects.

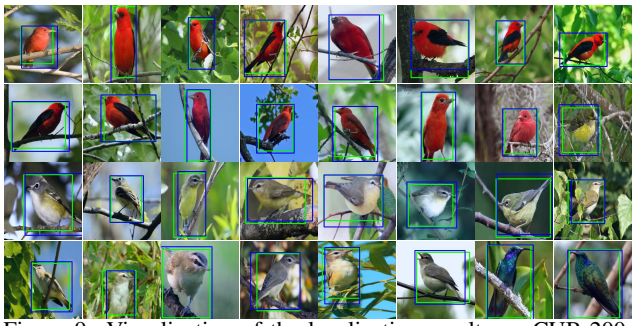


Figure 9: Visualization of the localization results on CUB-200-2011. Ground-truth and predicted bounding boxes are highlighted in blue and green, respectively.

	ClsErr		LocErr		
Methods compared	Top1	Top5	Top1	Top5	CorLoc
CAM-GoogLeNet [44]	26.2	8.5	58.94	49.34	55.1
Friend or Foe-GoogLeNet [36]	-	-	-	-	56.5
SPG-GoogLeNet [41]	-	-	53.36	42.28	-
GC-Net-Elli-GoogLeNet [16]	23.2	6.6	43.46	31.58	72.6
GC-Net-Rect-GoogLeNet [16]	23.2	6.6	41.42	29.00	75.3
DA-Net-Inception-V3 [37]	28.8	9.4	50.55	39.54	67.0
CAM-VGG16 [44]	23.4	7.5	55.85	47.84	56.0
ACoL-VGG16 [40]	28.1	-	54.08	43.49	54.1
TSC-VGG16 [8]	-	-	-	-	65.5
ADL-VGG16 [3]	34.7	-	47.64	-	-
DA-Net-VGG16 [37]	24.6	7.7	47.48	38.04	67.7
RCAM-VGG16 [1]	29.9	-	42.63	-	78.6
I^2C-VGG16 [42]	-	-	44.01	31.66	-
GC-Net-Elli-VGG16 [16]	23.2	7.7	41.15	30.10	74.9
GC-Net-Rect-VGG16 [16]	23.2	7.7	36.76	24.46	81.1
Ours(ORNet)-VGG16	**23.0**	**7.0**	**32.26**	**19.23**	**86.2**

Table 1: Comparison of the performance between the proposed method and the state-of-the-art on CUB-200-2011 test set. Here 'ClsErr' represents the classification error.

	ClsErr		LocErr		
Methods compared	Top1	Top5	Top1	Top5	CorLoc
Backprop-GoogLeNet [25]	-	-	61.31	50.55	-
GMP-GoogLeNet [44]	35.6	13.9	57.78	45.26	-
CAM-InceptionV3 [44]	-	-	53.71	41.81	62.68
HaS-32-GoogLeNet [27]	-	-	54.53	-	
SPG-InceptionV3 [41]	-	-	51.40	40.00	64.69
ACol-GoogLeNet [40]	29.0	11.8	53.28	42.58	-
DA-Net-InceptionV3 [37]	27.5	8.6	52.47	41.72	-
GC-Net-Rect-InceptionV3 [16]	**22.6**	**6.4**	50.94	41.91	-
Backprop-VGG16 [25]	-	-	61.12	51.46	-
CAM-VGG16 [44]	33.4	12.2	57.20	45.14	-
ACol-VGG16 [40]	32.5	12.0	54.17	40.57	62.96
ADL-VGG16 [40]	30.5	-	55.08	-	-
RCAM-VGG16 [1]	34.8	-	43.91	-	61.48
I^2C-VGG16 [42]	30.6	10.7	52.59	41.49	63.90
PSOL-VGG16-Sep [38]	-	-	49.11	39.10	64.03
Ours(ORNet)-VGG16	28.4	9.6	**47.95**	**36.06**	**68.27**

Table 2: Comparison of the performance between the proposed method and the state-of-the-art on the ImageNet-1K validation set. Here 'ClsErr' represents the classification error.

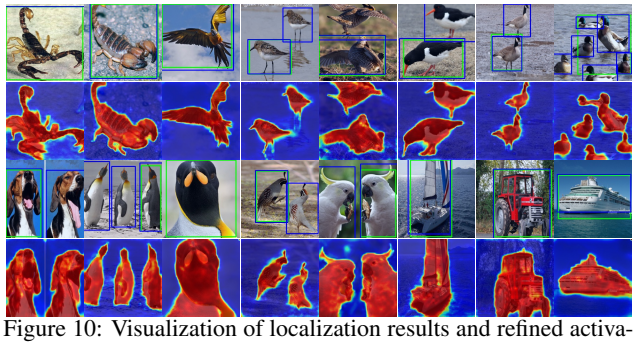


Figure 10: Visualization of localization results and refined activation maps on ImageNet-1K dataset. Ground-truth and predicted bounding boxes are highlighted in blue and green, respectively.

4.1. Experimental Setup

Datasets. CUB-200-2011 [30] is a fine-grained classification dataset with 200 species of birds. It consists of 11788 images, which is divided into 5994 images for training and 5794 images with bounding box annotations for testing. ImageNet-1K [24] contains images from 1000 object categories, which is split into 1.3 million images for training and 50000 images for testing. We only use the bounding boxes annotations of testing images for evaluation.

Implementation details. All the training images are first resized to 256×256 and then augmented by random cropping to 224×224. We adopt Adam as the default optimizer, with a weight decay 0.0. We adopt the cosine annealing policy [15] to schedule the learning rate. Moreover, for CUB-200-2011 dataset, the mini-batch size is 16. The initial learning rate is 0.001 for the first stage and 0.0001 for the second stage. The number of training epoch is 100, including 70 epochs for stage one and 30 epochs for stage two. For ImageNet-1K dataset, we set mini-batch size to 256. The initial learning rate is 0.001 for stage one and 0.0001 for stage two, and the number of training epoch is 8, *i.e.* 5 epochs for stage one and 3 epochs for stage two. Our model is implemented in PyTorch [20] and trained on NVIDIA Tesla P100 GPU with a 16GB memory.

Evaluation metrics. Following [5, 24], we adopt the localization error (LocErr), correct localization (CorLoc), Top1 accuracy, and Top5 accuracy as the metrics for evaluating the performances of the proposed methods. The LocErr is calculated based on the Top1 accuracy and location accuracy, a prediction is correct only if both localization(*i.e.* IoU ≥ 0.5) and classification are correct. For CorLoc metric, the prediction is correct as long as the localization is correct(*i.e.* IoU ≥ 0.5).

4.2. Comparison to State-of-the-Arts

Visual Comparison. As shown in Fig. 2, CAM [44] mostly focuses on the most discriminative parts of objects (*e.g.* bird head). ADL [3] alleviates the reliance on representative features. However, it does not fully address the intrinsic defects of CAM. In contrast, in our two-stage setting, the activation maps are exact, well-separated, and compact. Fig. 8 and Fig. 10 are the visual results of our method on CUB-200-2011 and ImageNet-1K, respectively, which also prove the advantages of our approach. Besides, Fig. 9 reveals that our algorithm is also robust to locate objects even in a noisy environment. More visual results on other datasets are provided in supplementary materials.

Components	LocErr Top1	LocErr Top5	CorLoc
Stage one	-	-	64.52
$\mathcal{L}_e$	44.36	33.02	71.35
$\mathcal{L}_e$+$\alpha\mathcal{L}_w$	36.05	23.40	81.68
$\mathcal{L}_e$+$\beta\mathcal{L}_a$	44.01	32.40	72.11
$\mathcal{L}_e$+$\mathcal{AE}$	45.75	35.02	69.14
$\mathcal{L}_e$+$\alpha\mathcal{L}_w$+$\beta\mathcal{L}_a$	37.31	24.92	80.10
$\mathcal{L}_e$+$\alpha\mathcal{L}_w$+$\mathcal{AE}$	33.21	20.21	85.09
$\mathcal{L}_e$+$\beta\mathcal{L}_a$+$\mathcal{AE}$	39.99	27.91	76.68
$\mathcal{L}_e$+$\alpha\mathcal{L}_w$+$\beta\mathcal{L}_a$+$\mathcal{AE}$	**32.26**	**19.23**	**86.19**

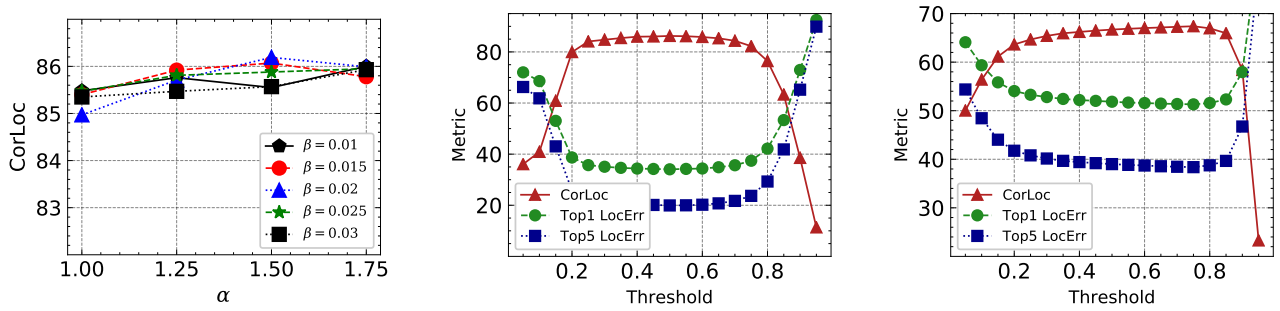
Table 3: Comparison of the object localization performance on CUB-200-2011 dataset using different components, including $\mathcal{L}_e$, $\mathcal{L}_w$, $\mathcal{L}_a$, and $\mathcal{AE}$.

Components	LocErr Top1	LocErr Top5	CorLoc
$\mathcal{L}_e$+$\beta\mathcal{L}_a$+$\mathcal{AE}$ + $\alpha\mathcal{L}_\mathbf{h}$	42.53	31.34	72.89
$\mathcal{L}_e$+$\beta\mathcal{L}_a$+$\mathcal{AE}$ + $\alpha\mathcal{L}_\mathbf{w}$	**32.26**	**19.23**	**86.19**

Table 4: The localization performance of entropy loss $\mathcal{L}_h$ and weighed entropy loss $\mathcal{L}_w$ on CUB-200-2011 dataset.

Quantitative Comparison. Table 1 shows the comparison results with state-of-the-art methods on CUB-200-2011 [30] dataset. As shown in the table, our method has the best localization performance in terms of all evaluation metrics. Specifically, our method achieves significant improvements of 5.1% on CorLoc and 4.5% on Top1 Loc Err over state-of-the-art GCNet [16]. Compared with CAM-VGG, our method has a large improvement of 30.2% and 23.59% on CorLoc and Top1 LocErr, respectively.

Table 2 provides the results on ImageNet-1K [24] dataset. Our method achieves 68.27% CorLoc, outperforming all the considered state-of-the-art methods (*e.g.* PSOL [38], I^{2}C [42]) by a large margin($\geq 4.16\%$). When VGG 16 is considered, our model also achieves the best classification accuracy than other approaches like CAM and ADL. As both GC-Net and our model directly take the publically available backbone trained by ImageNet, *i.e.* Inception-V3 or VGG-16, as the classifier, the ClsErr is independent with the generator of activation map. As a result, the classification accuracy of our VGG-16 trained by ImageNet is lower than that of Inception V3 used in GC-Net. However, our model still achieves much lower Top1 LocErr (47.95%) and Top5 LocErr (36.06%) than those of DA-Net-Inception V3 (52.47% and 41.72%) and GC-Net-Inception



(a) Analyses of hyper-parameters α and β. (b) Thresholding analyses on CUB-200-2011. (c) Thresholding analyses on ImageNet-1K.

Figure 11: Sensitive analyses of parameters α and β and the thresholding procedure.

V3 (50.94% and 41.91%). Top1 and Top5 LocErr consider both location and classification accuracy, *i.e.* a prediction is correct only if both localization and classification are correct. While fewer number of images are correctly classified, the results of LocErr suggest that our model achieves significantly higher location accuracy than DA-Net and GC-Net.

4.3. Ablation Study

We now perform ablation studies using CUB-200-2011 [30] dataset to evaluate the effectiveness of different loss functions and attentive erasing.

We first investigate the effectiveness of different combinations of the proposed components. Table 3 shows the quantitative contribution of each loss function and the attentive erasing (*i.e.* the evaluation loss $\mathcal{L}_e$, the weighed entropy loss $\mathcal{L}_w$, the area loss $\mathcal{L}_a$ and the attentive erasing $\mathcal{AE}$). As shown in the table, the generator trained in the first stage only achieves 64.52% CorLoc score. However, in the second stage, a substantial improvement of CorLoc, *i.e.* 6.83%, is gained by integrating the evaluator $\mathcal{F}^e$ supervised by $\mathcal{L}_e$. This verifies the necessity and effectiveness of $\mathcal{F}^e$. Besides, $\mathcal{L}_w$ solely can also significantly promote the performance of our model, *i.e.* 10.33% improvement of CorLoc, 8.31% and 9.62% decreases in Top1 LocErr and Top5 LocErr. This reveals that well-separated confidence maps essentially reduce ambiguity between foreground and background regions, which benefits the thresholding procedure for accurate localization. As expected, integrating either $\mathcal{L}_a$ or $\mathcal{AE}$ fails to improve the performance, as these two components are complementary to each other. $\mathcal{L}_e$+$\mathcal{L}_a$+$\mathcal{AE}$ could significantly boost CorLoc (from 71.35% to 76.68%) and decrease LocErr score. Moreover, $\mathcal{L}_e$+$\mathcal{L}_w$+$\mathcal{L}_a$, $\mathcal{L}_e$+$\mathcal{L}_w$+$\mathcal{AE}$, or $\mathcal{L}_a$+$\mathcal{L}_a$+$\mathcal{AE}$ also improves the performance due to the effectiveness of weighted entropy loss $\mathcal{L}_w$ or combination of $\mathcal{L}_a$ and $\mathcal{AE}$. The combination of three loss functions $\mathcal{L}_e, \mathcal{L}_w, \mathcal{L}_a$ and the attentive erasing $\mathcal{AE}$, further increase CorLoc from 71.35% to 86.19%, and decrease Top 1 LocErr and Top 5 LocErr from 44.36% to 32.26% and from 33.02% to 19.23%, respectively.

Table 4 lists the performance of activation map refined using weighted entropy loss $\mathcal{L}_w$ (Eq. 14) and mean entropy loss $\mathcal{L}_h$ (Eq. 13). As shown in the table, $\mathcal{L}_w$ increase CorLoc from 72.89% to 86.19%, and decrease Top 1 LocErr and Top 5 LocErr from 42.53% to 32.26% and from 31.34% to 19.23%, respectively.

4.4. Sensitivity Analysis

There are two hyper-parameters α and β in Eq. 10. The sensitivity analyses of these two parameters are performed on CUB-200-2011 test set and the results are presented in Fig. 11(a). As seen, stable CorLoc performance is obtained by a wide range setting of α and β. This suggests the lower sensitivity to hyper-parameters setting and good performance stability of our method. In experiments, the default value of α and β are 1.5 and 0.02.

Fig. 11(b) and (c) show the sensitivity analysis of thresholding on CUB-200-2011 test set and ImageNet-1K validation set. As seen, the performances of our method are stable when the threshold varies from 0.2 to 0.8, which proves that the weighted entropy loss $\mathcal{L}_w$ can substantially reduce the number of uncertain pixels (0.5) and improve the robustness of our method to the setting of threshold.

5. Conclusion and Discussion

This article proposes a two-stage learning framework to explore low-level features based activation maps for weakly supervised object localization. The first stage uses the low-level features to yield activation map of the target object with rich contextual information. The second stage refines the activation map based on the proposed weighted entropy loss, leading to an accurate pixel-level object localization. Experiments on CUB-200-2011 and ImageNet-1K datasets validate the effectiveness of our framework. Our work suggests a promising alternative for weakly supervised semantic and instance segmentation.

Acknowledgments

The work is supported by the National Natural Science Foundation of China under Grant 91959108.

References

[1] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *The European Conference on Computer Vision (ECCV)*, pages 618–634, 2020. 6

[2] Loris Bazzani, Alessandra Bergamo, Dragomir Anguelov, and Lorenzo Torresani. Self-taught object localization with deep networks. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. 1, 2

[3] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2219–2228, 2019. 1, 2, 3, 6, 7

[4] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold mil training for weakly supervised object localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2409–2416, 2014. 1, 2

[5] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision (IJCV)*, 100(3):275–293, 2012. 7

[6] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10705–10714, 2019. 3, 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[8] Xiangteng He and Yuxin Peng. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *The AAAI Conference on Artificial Intelligence (AAAI)*, volume 31, 2017. 6

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 3

[10] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2011. 6

[11] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 3534–3543, 2017. 1, 2

[12] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Guided attention inference network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(12), 2019. 2

[13] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 1

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *The European Conference on Computer Vision (ECCV)*, pages 21–37, 2016. 1

[15] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *The International Conference on Learning Representations (ICLR)*, 2017. 7

[16] Weizeng Lu, Xi Jia, Weicheng Xie, Linlin Shen, Yicong Zhou, and Jinming Duan. Geometry constrained weakly supervised object localization. In *The European Conference on Computer Vision (ECCV)*, pages 481–496, 2020. 1, 2, 3, 6, 7

[17] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8763–8772, 2020. 2

[18] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[19] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. BAM: bottleneck attention module. In *The British Machine Vision Conference (BMVC)*, page 147, 2018. 3

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 7

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 1

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(6):1137–1149, 2017. 1

[23] Ergys Ristani, Francesco Solera, Roger S. Zou, R. Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *The European Conference on Computer Vision Workshop (ECCVW)*, 2016. 6

[24] 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 (IJCV)*, 115(3):211–252, 2015. 2, 6, 7

[25] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 6

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *The International Conference on Learning Representations (ICLR)*, 2015. 3

[27] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised

object and action localization. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 3544–3553. IEEE, 2017. 1, 2, 6

[28] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *The International Conference on Machine Learning (ICML)*, pages 1611–1619, 2014. 1, 2

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *The International Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 3

[30] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2, 6, 7, 8

[31] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *The European Conference on Computer Vision (ECCV)*, pages 431–445, 2014. 1, 2

[32] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6458, 2017. 3

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018. 3

[34] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 79–88, 2018. 6

[35] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *The European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 3

[36] Zhe Xu, Dacheng Tao, Shaoli Huang, and Ya Zhang. Friend or foe: Fine-grained categorization with weak supervision. *IEEE Transactions on Image Processing*, 26(1):135–146, 2016. 6

[37] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 6589–6598, 2019. 2, 6

[38] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13460–13469, 2020. 1, 2, 6, 7

[39] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *The International Conference on Machine Learning (ICML)*, pages 7354–7363, 2018. 3

[40] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1325–1334, 2018. 1, 2, 6

[41] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *The European Conference on Computer Vision (ECCV)*, pages 597–613, 2018. 1, 2, 3, 6

[42] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *The European Conference on Computer Vision (ECCV)*, pages 271–287, 2020. 6, 7

[43] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 6

[44] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 1, 2, 6, 7