Where to Look?: Mining Complementary Image Regions for Weakly Supervised Object Localization

Sadbhavana Babar, Sukhendu Das
Visualization and Perception Lab,
Dept. of Computer Science and Engineering, IIT Madras, India
sadbhav@cse.iitm.ac.in, sdas@iitm.ac.in

Abstract

Humans possess an innate capability of recognizing objects and their corresponding parts and confine their attention to that location in a visual scene where the object is spatially present. Recently, efforts to train machines to mimic this ability of humans in the form of weakly supervised object localization, using training labels only at the image-level, have garnered a lot of attention. Nonetheless, one of the well-known problems that most of the existing methods suffer from is localizing only the most discriminative part of an object. Such methods provide very little or no focus on other pertinent parts of the object. In this paper, we propose a novel way of scrupulously localizing objects using training with labels as for the entire image by mining information from complementary regions in an image. Primarily, we adapt to regional dropout at complementary spatial locations to create two intermediate images. With the help of a novel Channel-wise Assisted Attention Module (CAAM) coupled with a Spatial Self-Attention Module (SSAM), we parallely train our model to leverage the information from complementary image regions for excellent localization. Finally, we fuse the attention maps generated by the two classifiers using our Attention-based Fusion Loss. Several experimental studies manifest the superior performance of our proposed approach. Our method demonstrates a significant increase in localization performance over the existing state-of-the-art methods on CUB-200-2011 and ILSVRC 2016 datasets.

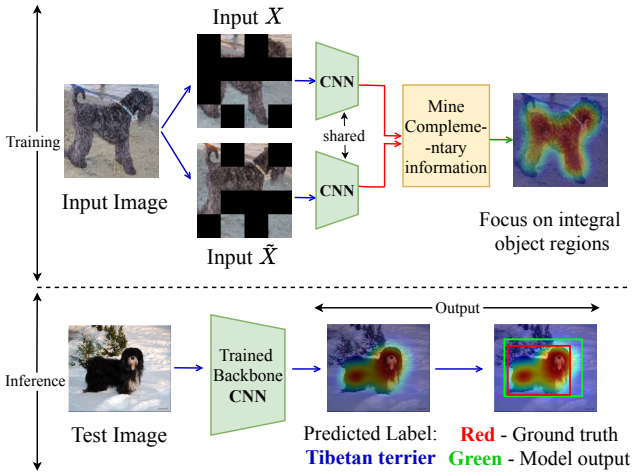


Figure 1. **Overview of our proposed approach.** From a single input image, we create two complementary images during training, denoted by X and $\tilde{X}$. For hiding patches in our complementary inputs, we adapt from [28]. We then extract features from X and $\tilde{X}$ using a shared CNN. Finally, we fuse information from the complementary inputs and perform parallel training of two classifiers to discover integral object regions. Our parallel classifiers aid in mining all relevant parts of the object (*e.g.*, for a dog - its face, forelegs, hindlegs) along with its most-discriminative part (head). In this way, our model learns to focus attention on "where to look" for the specified object in the given input image as well as localize objects in a weakly-supervised manner. During inference, we do not hide any patches of the input image. The test image as a whole is provided to our trained CNN model.

1. Introduction

Given a visual scene, humans have an inherent ability to recognize and localize objects of interest with minimal effort. With the advent of deep convolutional neural networks [16, 17], there has been a remarkable improvement in image recognition [15, 26, 29] and object detection [10, 20, 21, 23, 24, 27, 30, 35]. However, these methods rely on full supervision during training. Recently, there has been an increasing focus on Weakly Supervised Learning (WSL) techniques that require minimal supervision or coarse annotation during training, which reduces the effort of using costly pixel-level annotations. One of the fundamental computer vision tasks like semantic segmentation that require fine pixel-level annotations, can now be trained using only bounding box annotations or image-based labels using the

WSL approach [1, 18, 36, 44].

Weakly Supervised Object Localization (WSOL) aims to classify as well as localize objects without using expensive bounding box annotations during training. Recently, a lot of approaches [7, 22, 28, 39, 42, 48, 49, 51] have been proposed to tackle this challenging problem. Zhou *et al*. [51] put forward the idea of appending a Global Average Pooling (GAP) [19] layer at the end of convolutional neural networks (CNNs) followed by a fully-connected layer to generate a class activation map (CAM). CAM highlights the discriminative image region used to recognize that object category. However, a crucial limitation of this approach is that it only localizes the most discriminative class-specific region instead of the entire object. For *e.g*., given an image of a dog, it only tries to generate implicit attention on its face, without paying any heed to its remaining body parts. Hence, it often leads to sub-optimal localization performance.

To overcome this problem, a few recent methods [8, 28, 42, 46] have come up with making changes to *input image* rather than modifying the algorithm. In the paper, Hide-and-Seek (HaS) [28], Singh and Lee attempt to randomly hide patches of an input image during training so that their model tries to seek other visible relevant parts of the object. Even though this approach focuses on non-discriminative object parts, it loses information during training when the patches are hidden, leading to a limited localization performance. This gives rise to an interesting question: Is there any way to optimize the localization performance by maximally utilizing the information lost in regional dropout?

We propose to solve the above problem by introducing to strategically mine information from complementary image regions. Regional dropout methods [8, 50] have significantly demonstrated the ability to generalize well on image classification and object localization. We also venture to leverage this generalization ability and create two complementary images, each possessing regional dropout at complementary spatial locations in the respective images. To create these input images, we adapt to randomly hide patches in the input image similar to Hide-and-Seek [28], as illustrated in figure 1. We perform joint training of these complementary image regions as two input channels, using two parallel classifiers. Further, we try to fuse the information captured in both these input channels by incorporating a novel Channel-wise Assisted Attention Module (CAAM) along with a Spatial Self-Attention Module (SSAM). Both these modules take input features extracted from pre-trained CNNs. CAAM takes inspiration from [11, 40, 47], and tries to model interactions in the channel dimension between features extracted from two complementary images. SSAM is inspired by [11, 36, 47] to capture feature dependencies in the spatial dimension. We finally aggregate the interdependencies modeled by these two modules: CAAM and SSAM, for better localization ability. We also propose an Attention-based Fusion Loss, inspired by [43], to fuse the two attention maps obtained using the complementary images. Almost all the previous works rely only on the classification objective to learn the implicit attention maps, which serve as a testimony of visual explanations learned by the model to localize objects. However, we feel that relying only on the classification objective for localizing objects limits the overall localization performance. The use of our proposed Attention-based Fusion Loss, along with the usual cross-entropy loss to train our localization model, to the best of our knowledge, is the first of its kind.

Our key contributions are summarized as follows: 1) We propose a novel way of training a network for weakly supervised object localization that mines information from complementary regions in an image, individually as well as when fused. 2) We propose a novel Channel-wise Assisted Attention Module (CAAM). Along with a Spatial Self-Attention Module (SSAM), CAAM jointly aids in localizing integral object regions. 3) We also propose an Attention-based Fusion Loss criteria to fuse the attention maps generated by the two parallel classifiers. Our proposed loss function diligently captures all relevant parts of the concerned object of interest, thereby suppressing background regions. 4) Our method achieves state-of-the-art object localization performances on two benchmark datasets: CUB-200-2011 [34] and ILSVRC 2016 [25]. We achieve a Top-1 localization of 64.70% on CUB-200-2011 and 52.36% on ILSVRC 2016 datasets.

2. Related Work

Correspondence with human visual perception: The two-stream theory of the human visual system proposed by Goodale *et al*. [13] highlights the two distinct visual pathways in the human visual system viz., the ventral pathway or the "what pathway" and the dorsal pathway or the "where pathway" that jointly aid in recognizing and localizing objects respectively. We take motivation from the human vision system to jointly model the "what" and "where" pathways and efficiently perform object localization in a weakly supervised setting.

Weakly Supervised Learning: Learning strong predictive models using imprecise labels is becoming a trend since it involves cheaper annotations and reduced human effort for manual labeling of data. Also, the huge availability of weakly labeled data in the form of videos and images over the internet makes it possible to explore various real-world problems in deep learning in a weakly supervised paradigm. Although supervised object detection methods [21, 23, 24, 27] have made tremendous progress, the fact that they require costly bounding box annotations has led to the exploration of weakly supervised object detection methods [9, 31, 32, 45] using only image-level labels.

Regional Dropout: Randomly masking certain regions

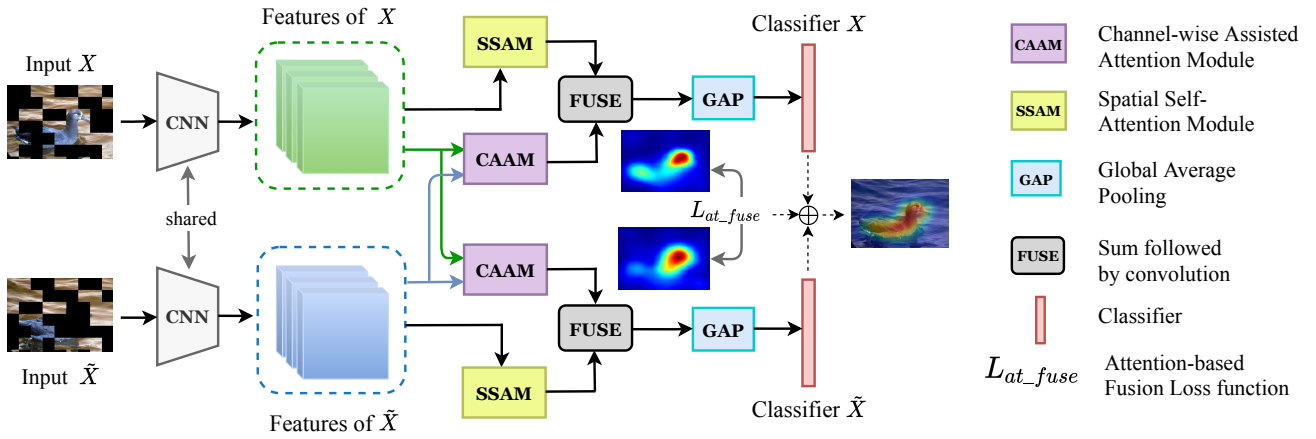


Figure 2. **Our proposed architecture.** After extracting CNN features for our complementary inputs, X and $\tilde{X}$, from a shared backbone, we feed them to our proposed Channel-wise Assisted Attention Module (CAAM) along with a Spatial Self-Attention Module (SSAM) to generate better feature representations. The aggregated outputs of CAAM and SSAM are then fed to the Global Average Pooling (GAP) [19] layer. We perform parallel training for both the input branches corresponding to our inputs X and $\tilde{X}$ and obtain two localization maps. We combine the attention between these localization maps using our proposed Attention-based Fusion Loss function. The dual training branches with Classifiers X and $\tilde{X}$ focus attention on all regions of the object present in an image, thereby effectively localizing it.

in an input image have found to be effective in capturing richer object context and better generalization performance. Bazzani *et al*. [4] proposed to mask out certain regions in an image that lead to a drop in the image recognition performance, finally feeding the regions to an agglomerative clustering algorithm which indicate higher objectness of such merged regions in the input image. In Hide-and-Seek [28], the crux was to randomly hide patches in an input image forcing the network to focus on other relevant object parts. Cutout [8] is yet another successful generalizable approach that drops a certain amount of input region from the input image. However, these methods lose information while training the network using regional dropout. We make use of information lost in regional dropout while training the network, by generating two images to mask complementary spatial locations.

Attention in Deep Neural Networks: Attention mechanism was first proposed in the pioneering work [33] by Vaswani *et al*. in machine translation to model long-range dependencies that the recurrent neural networks failed to handle. Since then, attention has been used in a wide variety of applications including image captioning [37], visual grounding [12], visual question answering [2], sound source localization [3]. Self-attention has also made its way in visual question answering [41], in which it was used for question embedding, and in [40], it was used along with a guided-attention module to model interactions for different feature modalities. In one of the most recent works [47], self-attention was used in image generation. In all these works, self-attention has proved to generate better feature representations. In our proposed framework, we adapt to attention mechanism coupled with regional dropout for more meaningful representations of our complementary images.

Other methods for weakly supervised object localization: The work in [49] generates self-produced guidance masks, which in turn are used in the form of pixel-level supervision for localizing objects. Zhang *et al*. [48] proposed adversarial erasing in feature space that mine information from two adversarial parallel classifiers for superior localization performance. Choe and Shim, in their work [7], proposed to use self-attention mechanism to generate a drop mask and an importance map from the input feature map and randomly select either of them along with the input feature map for localizing objects. Yang *et al*. [39] uses a linear combination of activation maps from the highest probability score of a class to the lowest probability score, thereby assisting in suppressing the background regions and focusing more on the foreground object of interest. The most recent work of EIL [22] by Mai *et al*. attempts to jointly perform adversarial erasing and mining discriminative regions to localize objects efficiently.

3. Proposed Approach

Looking at complementary image regions helps the network in paying attention to the concise details of the object. Our detailed network architecture is illustrated in figure 2.

3.1. Notations

Given an input image I, with its image-level label, y_i, the goal of weakly supervised object localization is to learn a model that is capable of classifying the input image I into one of C object categories in the dataset and localizing the

object in that image using a bounding box B. From I, we create two input images, X and $\tilde{X}$, with regional dropout at complementary spatial locations in an image, as shown in figure 1. We adapt to hide patches in the input image as in [28]. We form our input X by randomly hiding patches of an image. However, we regain the information lost in input X by forming another input $\tilde{X}$, which we call the X complement. Input $\tilde{X}$ reveals the information in the hidden patches of input X, whereas hides the information in visible patches of input X. We extract features from inputs X and $\tilde{X}$ using a shared CNN having parameters $\hat{\theta}$. We denote these features as F_x and $F_{\tilde{x}}$, where $F_x, F_{\tilde{x}} \in \mathbb{R}^{C \times H \times W}$.

3.2. Mining Information from Complementary Image Regions

The information captured from CNN features (F_x and $F_{\tilde{x}}$) of both inputs X and $\tilde{X}$ are used individually as well as combined (fused) using Spatial Self Attention Module (SSAM) and Channel-wise Assisted Attention Module (CAAM) modules respectively (shown in figure 2). Finally, we aggregate the features captured both by SSAM and CAAM for better feature representations of F_x and $F_{\tilde{x}}$, denoted by F_x^t and $F_{\tilde{x}}^t$ respectively. F_x^t and $F_{\tilde{x}}^t$ are then passed to a global average pooling layer [19], followed by their respective classifiers. By jointly training the two branches concerning the inputs X and $\tilde{X}$, viz., classifier X and classifier $\tilde{X}$, our model precisely gets an idea regarding "where to look" in the input image while classifying it correctly.

3.3. Channel-wise Assisted Attention Module

To compute CAMs [51], Zhou *et al*. proposed to multiply the weights of the last fully connected layer of the classifier to the feature maps of the preceding convolution layer. Towards the last layers of the CNN, the feature maps tend to capture the class-specific responses. Hence, CAM highlights the most discriminative region of the object belonging to that category. In the work [11], Fu *et al*. put forth the idea of a Channel Attention Module to capture long-range inter-dependencies between channels of feature maps in a fully supervised setting for the task of semantic segmentation. Adapting the idea from [11], we attempt to leverage the class-specific inter-dependencies between channels of input features from both the branches, F_x and $F_{\tilde{x}}$. So, our CAAM module takes as input the CNN features, F_x and $F_{\tilde{x}}$ and outputs features with more meaningful representation, denoted by F_x^c and $F_{\tilde{x}}^c$ respectively. $F_x^c, F_{\tilde{x}}^c \in \mathbb{R}^{C \times H \times W}$. A similar approach has been studied in [52] recently using a cross-correlated attention network in the spatial dimension. However, our CAAM module tries to capture inter-dependencies in the channel dimension of two feature maps.

To compute F_x^c, we take the input features F_x, $F_{\tilde{x}}$ and reshape them to $\mathbb{R}^{C \times N}$, where $N = H \times W$ corresponds to the number of pixels in the feature map. We then generate channel attention matrix Q_x as follows:

$$Q_x = Softmax(F_x \otimes F_{\tilde{x}}^T) \tag{1}$$

where, $Q_x \in \mathbb{R}^{C \times C}$ and $\otimes$ denotes matrix multiplication. Q_x consists of the learnable attention weights denoted by $\lambda_x^{ij} \in Q_x, i, j \in \{1...C\}$, which capture inter-dependencies between the channels of F_x and $F_{\tilde{x}}$. Further, we multiply transpose of Q_x with $F_{\tilde{x}}$ to get Y_x, as:

$$Y_x = Q_x^T \otimes F_{\tilde{x}} \tag{2}$$

We then reshape Y_x as $\mathbb{R}^{C \times H \times W}$ and generate F_x^c as :

$$F_x^c = F_x + \delta_x Y_x \tag{3}$$

Here, δ_x is used to scale the features of Y_x. It is initially set to 0 and iteratively trained similar to that in [11, 47]. The detailed expression for F_x^c is as follows :

$$F_x^{c^{ij}} = F_x^{ij} + \delta_x \sum_{k=1}^{C} \lambda_x^{ki} F_{\tilde{x}}^{kj} \tag{4}$$

F_x^c refers to channel-wise attended features of input F_x assisted by input $F_{\tilde{x}}$. We can see in equation (4) that the final features F_x^c are a weighted sum of features of all locations of input feature $F_{\tilde{x}}$ and the original features F_x.

Similarly, to compute $F_{\tilde{x}}^c$ we follow the same set of steps as followed for F_x^c. However, we save computations as well

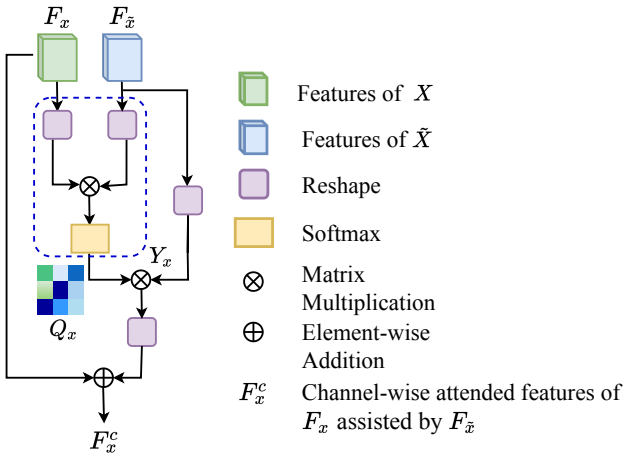


Figure 3. **Channel-wise Assisted Attention Module (CAAM).** The input to this module is the CNN features, F_x and $F_{\tilde{x}}$ of inputs X and $\tilde{X}$ respectively, and it outputs the Channel-wise attended features of input F_x assisted by input $F_{\tilde{x}}$. We interchange F_x and $F_{\tilde{x}}$ in figure 3 to obtain the Channel-wise attended features of input $F_{\tilde{x}}$ assisted by input F_x. The part indicated in dotted blue rounded rectangle indicates shared computation while computing both F_x^c and $F_{\tilde{x}}^c$.

as parameters in generating the channel attention matrix $Q_{\tilde{x}}$ (shown in figure 3), as it is the transpose of Q_x.

$$Q_{\tilde{x}} = Softmax(F_{\tilde{x}} \otimes F_x^T) \tag{5}$$

From equations (1) and (5), it is evident that $Q_{\tilde{x}}$ is actually transpose of Q_x. But for the purpose of simplicity, we denote it as $Q_{\tilde{x}}$ itself. Also, we denote its learnable attention weights as $\lambda_{\tilde{x}}^{ij} \in Q_{\tilde{x}}$ and $i, j \in \{1...C\}$. Similarly, for $F_{\tilde{x}}^c$:

$$F_{\tilde{x}}^{c^{ij}} = F_{\tilde{x}}^{ij} + \delta_{\tilde{x}} \sum_{k=1}^{C} \lambda_{\tilde{x}}^{ki} F_x^{kj} \tag{6}$$

Similar to equation (4), $\delta_{\tilde{x}}$ is also used as a scaling factor for $Y_{\tilde{x}}$. It is initially set to 0 and learns weight as training progresses. $F_{\tilde{x}}^c$ refers to channel-wise attended features of input $F_{\tilde{x}}$ assisted by input F_x. As shown in equation (6), the features of $F_{\tilde{x}}^c$ are a weighted sum of features of all locations of input feature F_x and the original features $F_{\tilde{x}}$.

3.4. Spatial Self-Attention Module

Apart from the inter-dependencies between the features of channels F_x and $F_{\tilde{x}}$ modeled by CAAM, it is significant to consider their individual contribution as well for efficient feature representation. We also hypothesize that to get the object's correct spatial location, it is important to have an overall view of the visual scene and give the corresponding weightage to the entire scene as per the objectness. In the work [47], self-attention was used in GANs [14]. Taking motivation from [47], we propose to use spatial self-attention for localizing objects. So the input to our SSAM module is the features F_x and $F_{\tilde{x}}$ and its output is spatially attended features F_x^s and $F_{\tilde{x}}^s$ respectively. Both F_x^s and $F_{\tilde{x}}^s$ are of dimension $\mathbb{R}^{C \times H \times W}$. As illustrated in figure 4, given features F_x, we compute matrices M, L and P using 1x1 convolution where, $\{M, L\} \in \mathbb{R}^{\tilde{C} \times H \times W}$, where $\tilde{C} = C/8$, and $P \in \mathbb{R}^{C \times H \times W}$. Mathematically, we compute F_x^s as:

$$\begin{aligned} K_x &= Softmax(M^T \otimes L) \\ R_x &= P \otimes K_x^T \\ F_x^s &= F_x + \alpha_x R_x \end{aligned} \tag{7}$$

where, α_x is a weight factor for R_x. The parameter α_x and the weight matrices M, L, P, K_x and R_x are learnt during training. Similarly, $F_{\tilde{x}}^s$ can be formulated as:

$$F_{\tilde{x}}^s = F_{\tilde{x}} + \alpha_{\tilde{x}} R_{\tilde{x}} \tag{8}$$

3.5. Aggregation

For features F_x and $F_{\tilde{x}}$ coming from each of the input branches, we have two enhanced feature representations, $\{F_x^c, F_x^s\}$ and $\{F_{\tilde{x}}^c, F_{\tilde{x}}^s\}$: the channel assisted features and the spatially attended features respectively. To take advantage of complementary information in both these features,

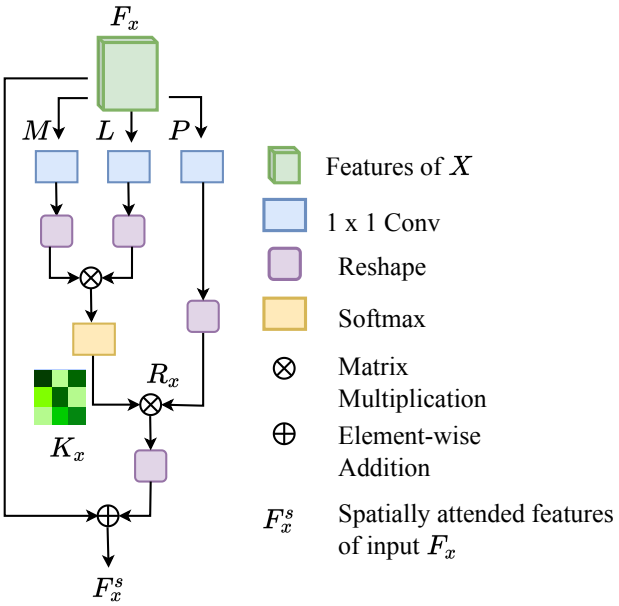


Figure 4. **Spatial Self-Attention Module (SSAM).** This module helps in spatially localizing the object as it takes help from all feature locations of the input feature map F_x and outputs the corresponding spatially boosted features F_x^s. Similarly, we also get $F_{\tilde{x}}^s$ from the corresponding complementary features $F_{\tilde{x}}$.

we fuse them using an element-wise sum. Finally, a convolution layer is used to bind them together as follows:

$$F_x^t = conv(F_x^c + F_x^s); \;\; F_{\tilde{x}}^t = conv(F_{\tilde{x}}^c + F_{\tilde{x}}^s) \tag{9}$$

Here, F_x^t and $F_{\tilde{x}}^t$ denote the feature maps from final convolution layer in our proposed framework. We use outputs of these final convolution layers to generate localization maps.

3.6. Attention-based Fusion Loss

We train our model in an end-to-end way to obtain two localization maps, in a manner similar to CAM [51]. We use cross-entropy loss for training both the classifier branches. However, both the classifiers discover complementary object parts during training. Thus, it is necessary to fuse the pair of localization maps. This is done by our Attention-based Fusion Loss, such that our model learns to focus on the entire object during training and generalizes well during testing (as we do not use two branches during testing).

Calculating localization maps: The features from our last convolution layer, F_x^t and $F_{\tilde{x}}^t$ having parameters θ^x and $\theta^{\tilde{x}}$ consist of C feature maps each having spatial dimension $H \times W$. These features are fed to the global average pooling (GAP) [19] layer. Let the value of k^{th} feature map at spatial location (m, n) of F_x^t and $F_{\tilde{x}}^t$ be denoted as $F_x^{t^k}(m, n)$ and $F_{\tilde{x}}^{t^k}(m, n)$ respectively. After performing GAP on the k^{th} feature maps, we get the activation units G_x^k and $G_{\tilde{x}}^k$ respectively. We pass the outputs from GAP layer to the respective

Algorithm 1: Our training algorithm

Input: N training images along with their image-level labels, $\{(I_i, y_i)\}_{i=1}^{N}$, hyperparameter β.

while *convergence condition not met* **do**

- Create two images X and $\tilde{X}$ with spatial dropout at complementary image locations, for an input image;
- Compute CNN features as: $F_x \leftarrow f(X, \hat{\theta})$ and $F_{\tilde{x}} \leftarrow f(\tilde{X}, \hat{\theta})$;
- Use CAAM to compute: $F_x^c \leftarrow f^{CAAM}(F_x)$, $F_{\tilde{x}}^c \leftarrow f^{CAAM}(F_{\tilde{x}})$;
- Use SSAM to compute: $F_x^s \leftarrow f^{SSAM}(F_x)$, $F_{\tilde{x}}^s \leftarrow f^{SSAM}(F_{\tilde{x}})$;
- Aggregate CAAM and SSAM outputs: $F_x^t \leftarrow conv(F_x^c + F_x^s)$, $F_{\tilde{x}}^t \leftarrow conv(F_{\tilde{x}}^c + F_x^s)$;
- Compute predicted labels as: $p_x = g_x(X, \theta^x, F_x^t)$, $p_{\tilde{x}} = g_{\tilde{x}}(\tilde{X}, \theta^{\tilde{x}}, F_{\tilde{x}}^t)$;
- Compute cross entropy loss for classifiers X and $\tilde{X}$: $L_{CE_x} = -\sum_i y_i \log p_{x,i}$, $L_{CE_{\tilde{x}}} = -\sum_i y_i \log p_{\tilde{x},i}$;
- Compute Attention-based Fusion Loss as in eq. (14);
- Obtain total loss as: $L_{tot} = L_{CE_x} + L_{CE_{\tilde{x}}} + \beta * L_{at\_fuse}$;
- Backpropagate loss and update parameters $\hat{\theta}$, θ^x, $\theta^{\tilde{x}}$;

end

classifiers. Let the weights for a given class c coming from the k^{th} activation unit be denoted as W_c^k. The softmax outputs of the classifiers for a particular class c are denoted by H_{x_c} and $H_{\tilde{x}_c}$. Mathematically, we denote this process as:

$$G_x^k = \sum_{m,n} F_x^{t^k}(m,n); \qquad G_{\tilde{x}}^k = \sum_{m,n} F_{\tilde{x}}^{t^k}(m,n) \tag{10}$$

$$H_{x_c} = \sum_k W_c^k G_x^k; \qquad H_{\tilde{x}_c} = \sum_k W_c^k G_{\tilde{x}}^k \tag{11}$$

From equations (10) and (11),

$$H_{x_c} = \sum_{m,n} \sum_k W_c^k F_x^{t^k}(m,n); \tag{12}$$

Similar to equation (12), we express $H_{\tilde{x}_c}$ in terms of W_c^k, $F_{\tilde{x}}^{t^k}$. For a particular class c, we denote the localization maps for both the input features F_x^t and F_y^t, as follows:

$$\begin{aligned} A_{x_c}(m,n) &= \sum_k W_c^k F_x^{t^k}(m,n); \\ A_{\tilde{x}_c}(m,n) &= \sum_k W_c^k F_{\tilde{x}}^{t^k}(m,n) \end{aligned} \tag{13}$$

We finally combine these localization maps A_{x_c} and $A_{\tilde{x}_c}$ using our proposed Attention-based Fusion Loss function (as illustrated in figure 5).

Fusing the localization maps: Unlike in [48], which relies on non-differentiable max function for fusing localization maps from two classifiers, we propose to combine the localization maps using an Attention-based Fusion Loss inspired from [43]. We first convert the obtained localization maps into their respective vectorized forms, i.e., $V_{x_c} = vec(A_{x_c})$ and $V_{\tilde{x}_c} = vec(A_{\tilde{x}_c})$ and perform l_2-normalization of V_{x_c} and $V_{\tilde{x}_c}$. Our proposed Attention-based Fusion Loss is formulated as follows:

$$L_{at\_fuse} = \left(\frac{V_{x_c}}{||V_{x_c}||_2} - \frac{V_{\tilde{x}_c}}{||V_{\tilde{x}_c}||_2} \right)^2 \tag{14}$$

We simply train our network with the proposed Attention-based Fusion Loss coupled with the categorical cross-entropy loss for efficient and integral object localization. The total loss function for training our model is:

$$L_{tot} = L_{CE}(y, p_x) + L_{CE}(y, p_{\tilde{x}}) + \beta * L_{at\_fuse} \tag{15}$$

where, L_{CE} denotes the categorical cross-entropy loss function, β is a hyperparameter used to scale our Attention-based Fusion Loss. Empirically, we choose $\beta = 50$ in our experiments. y denotes the true labels, p_x and $p_{\tilde{x}}$ denote the predictions made by our complementary classifiers.

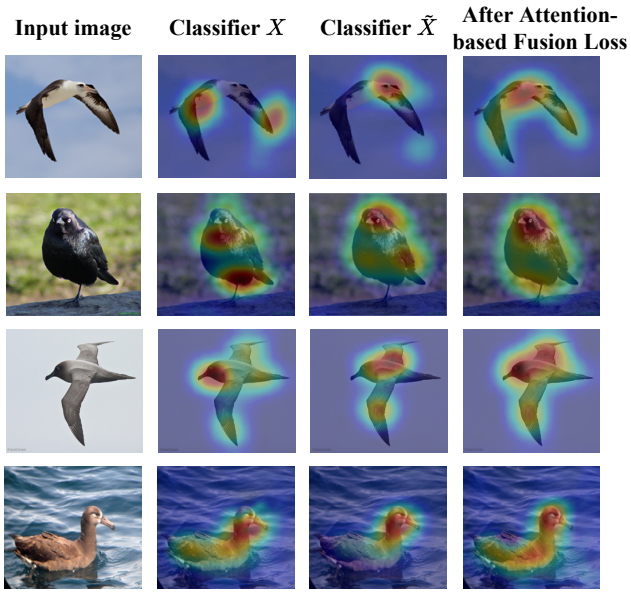
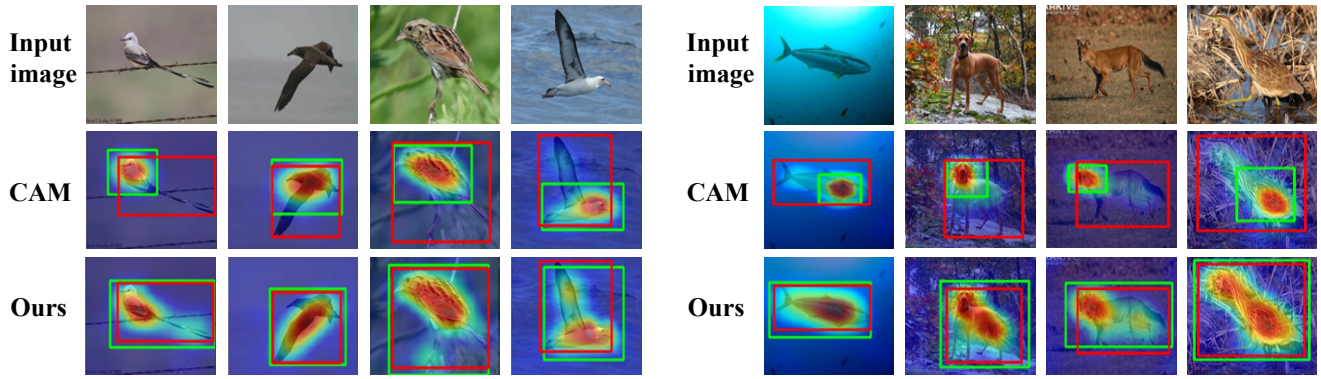


Figure 5. **Visualizing the effect of the proposed Attention-based Fusion Loss.** During training, we visualize the effect of our proposed loss function. The left column denotes the input image, the second and third columns denote the localization maps of our two classifiers and the right column denotes the localization map after applying our Attention-based Fusion Loss.

4. Experiments

4.1. Experimental Setup

Datasets: We perform our experiments on two benchmark datasets used for object localization, CUB-200-2011 [34] and ILSVRC 2016 [25]. CUB-200-2011 has a total



(a) CUB-200-2011 (b) ILSVRC 2016

Figure 6. **Qualitative Results.** We compare our results qualitatively with the baseline CAM [51] model. Ground truth bounding boxes are denoted in Red, whereas predicted bounding boxes are denoted in Green. Visually, we observe that our attention maps are much precise and our model tries to localize non-discriminative object parts (like the wings, legs, tail of the bird) as well.

of 11,788 images spanning across 200 bird categories, of which 5,994 images are used for training and 5,794 for testing. ILSVRC 2016 has approximately 1.2 million images in the training set across 1000 different categories, and 50,000 images in the validation set. We compare our results across different methods on the ILSVRC 2016 validation set.

Evaluation Metrics: We evaluate our method using the following metrics: 1) Top-1 localization (Top-1 Loc) accuracy [25] calculates the fraction of images that are correctly classified and the predicted bounding box has 50% IoU with the ground truth bounding box. 2) Top-1 classification (Top-1 Clas) accuracy determines the fraction of images that are correctly classified. 3) GT-known localization (*GT-Loc*) accuracy [28] only considers the fraction of images for which the predicted bounding box has 50% IoU with the ground truth bounding box, independent of the Top-1 classification accuracy. 4) Apart from the above three standard metrics, we also evaluate our method on the recently proposed *MaxBoxAccv2* [6] metric (as shown in table 4).

4.2. Implementation Details

We experiment with VGG16 [26] and ResNet50 [15] as the backbone CNN architectures for our proposed approach. As in [51], we remove the layers after conv5-3 in the VGG16 network. We insert our CAAM and SSAM modules after conv5-3 layer of the original VGG16 network. The aggregated outputs from both CAAM and SSAM modules are then fed to a global average pooling (GAP) layer [19], followed by a fully-connected layer for classification. We follow similar steps for ResNet50 backbone as well. Both VGG16 and ResNet50 architectures are initialized with weights pre-trained on ImageNet [25] dataset. We extract our localization maps followed by bounding boxes, in a similar way to [51]. During testing, we do not hide patches in the input image, similar to [28]. Also, we deactivate CAAM and SSAM modules during testing, similar to vanilla CAM [51] model for fair comparison with other existing state-of-the-art methods (shown in tables 1, 2 & 3).

Method	Top-1 Loc	Top-1 Clas
InceptionV3-CAM [51]	43.67	**73.80**
InceptionV3-SPG [49]	46.64	–
InceptionV3-DANet [38]	**49.45**	71.20
VGG-CAM [51]	34.41	67.55
VGG-ACoL [48]	45.92	71.90
VGG-ADL [7]	52.36	65.27
VGG-CCAM [39]	50.07	73.20
VGG-EIL [22]	56.21	72.26
Ours-VGG	**58.12**	**72.59**
ResNet50-CAM [51]	49.41	75.68
ResNet50-CutMix [42]	54.81	–
Ours-ResNet50	**64.70**	**77.28**

Table 1. Quantitative Results on CUB-200-2011 dataset.

4.3. Ablation Studies

Hyperparameters: For our complementary input images, we use a hide probability of 0.5 to hide and render complementary image locations, in the corresponding input images. Similar to [28], we also experiment with different patch sizes, $\{16, 32, 44, 56\}$ for regional dropout during training. As illustrated in Algorithm 1, we use a hyperparameter β to scale the proposed Attention-based Fusion Loss during training. We set β to 50 in all our experiments.

Importance of each module in the architecture: Our proposed architecture has three components, CAAM, SSAM, and an Attention-based Fusion Loss to fuse localization maps from two distinct branches during training. We study the effect of each of these modules on localization ac-

Method	Top-1 Loc	GT-Loc
InceptionV3-CAM [51]	46.29	–
GoogLeNet-HaS-32 [28]	45.21	60.29
InceptionV3-SPG [49]	48.60	64.69
InceptionV3-DANet [38]	47.53	–
InceptionV3-MEIL [22]	**49.48**	–
VGG-CAM [51]	42.80	57.72
VGG-ACoL [48]	45.80	62.96
VGG-ADL [7]	44.92	–
VGG-CCAM [39]	48.22	63.58
VGG-EIL [22]	46.27	–
Ours-VGG	**51.64**	**66.32**
ResNet50-CAM [51]	38.99	51.86
ResNet50-SE-ADL [7]	48.53	–
ResNet50-CutMix [42]	47.25	–
Ours-ResNet50	**52.36**	**67.89**

Table 2. Localization Results on ILSVRC 2016 dataset.

Method	Top-1 Clas (in %)
InceptionV3-CAM [51]	68.10
GoogLeNet-HaS-32 [28]	**70.70**
VGG-CAM [51]	66.60
VGG-ACoL [48]	67.50
VGG-ADL [7]	69.48
VGG-CCAM [39]	66.60
VGG-EIL [22]	70.48
Ours-VGG	**71.24**

Table 3. Classification performance on ILSVRC 2016 dataset.

Method	CUB-200-2011	ILSVRC 2016
CAM [51]	71.1	61.1
HaS-32 [28]	76.3	61.8
ACoL [48]	72.3	60.3
SPG [49]	63.7	61.6
ADL [7]	75.7	60.8
CutMix [42]	71.9	62.1
Ours	**77.5**	**63.4**

Table 4. **Evaluating our method on MaxBoxAccv2.** We evaluate our model on the recently proposed $MaxBoxAccv2$ metric [6] on VGG16 as the backbone. Experiments for ResNet50 are provided in the Supplementary Section A.

curacies. Table 5 shows how accuracies vary with different modules in our architecture on ILSVRC dataset. Overall, we observe that all our proposed modules are crucial to significantly boost the localization accuracy.

Effect of Patch Size on localization accuracy: We perform experiments with different patch sizes as in [28]. The patch sizes we use during training are either one of $\{16, 32, 44, 56\}$. We also come up with a *Mixed* model

CAAM	SSAM	L_{at_fuse}	Top-1 Loc	GT-Loc
✓	✓	✗	50.49	64.87
✓	✗	✓	50.95	65.58
✗	✓	✓	49.46	65.10
✓	✓	✓	**51.64**	**66.32**

Table 5. **Effect of each module in the architecture.** For the above experiment, we have used VGG16 as the backbone CNN.

Patch Size	CUB-200-2011		ILSVRC 2016	
	Top-1 Loc	GT-Loc	Top-1 Loc	GT-Loc
16	61.89	74.16	50.97	65.31
32	60.49	72.14	51.37	66.24
44	61.51	73.35	51.89	67.10
56	62.58	75.22	51.64	67.31
Mixed	**64.70**	**77.35**	**52.36**	**67.89**

Table 6. **Localization accuracies with different patch sizes.** For the above experiment, we have used ResNet50 as the backbone CNN architecture.

wherein the patch size is randomly sampled among the patch sizes $\{16, 32, 44, 56\}$, with uniform probability, for every image in every epoch during training. Unlike [28], we do not show full image during training in our *Mixed* approach. Our *Mixed* approach outperforms all the existing state-of-the-art models on localization accuracy (as shown in table 6). We do lose on some classification accuracy, as our model never encounters full image during training. Still, our method achieves comparable *Top-1 Clas* and best *Top-1 Loc* performance on both CUB-200-2011 and ILSVRC 2016 datasets. Qualitative results (shown in figure 6) ensure that our model looks at all object parts to make correct predictions. In future, we plan to evaluate our method on the recently proposed OpenImages30K [5, 6] dataset.

5. Discussion and Conclusion

We propose a novel way of mining information from complementary image regions to tackle the problem of Weakly-Supervised Object Localization. We show that our novel Channel-wise Assisted Attention Module (CAAM), when combined with a Spatial Self-Attention Module (SSAM), boosts existing feature representations for localizing integral object regions. We also propose a novel Attention-based Fusion Loss function to fuse the localization maps coming from two different input branches during training. In this way, our method is able to focus on discriminative as well as non-discriminative object parts for precise localization. Even though we study the task of single-object detection in a weakly-supervised manner, it will be interesting to explore the case of detecting multiple objects in a scene, laying the foundation for significantly bridging the gap between supervised and weakly-supervised methods.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

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

[5] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11700–11709, 2019.

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

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

[8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[9] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 914–922, 2017.

[10] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019.

[11] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[12] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[13] Melvyn A Goodale, A David Milner, et al. Separate visual pathways for perception and action. 1992.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[17] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[18] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019.

[19] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *International Conference on Learning Representations*, 2014.

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

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

[22] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8766–8775, 2020.

[23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

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

[27] Bharat Singh, Mahyar Najibi, and Larry S Davis. Sniper: Efficient multi-scale training. In *Advances in Neural Information Processing Systems*, pages 9310–9320, 2018.

[28] Krishna Kumar Singh and Yong 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*, pages 3544–3553, 2017.

[29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[30] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020.

[31] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):176–191, 2018.

[32] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2843–2851, 2017.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[34] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[35] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[36] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.

[37] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

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

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

[40] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.

[41] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):5947–5959, 2018.

[42] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019.

[43] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[44] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7223–7233, 2019.

[45] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8292–8300, 2019.

[46] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[47] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363, 2019.

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

[49] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European Conference on Computer Vision*, pages 597–613, 2018.

[50] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020.

[51] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[52] Jieming Zhou, Soumava Kumar Roy, Pengfei Fang, Mehrtash Harandi, and Lars Petersson. Cross-correlated attention networks for person re-identification. *Image and Vision Computing*, page 103931, 2020.