Half Yearly Progress Report for Jan-May 2020 & Jul-Nov 2020

Data Sheet for M.S Scholars

Name: Sadbhavana Babar

Registration Number: CS18S029

Department: Department of Computer Science and Engineering

Date of Joining: 09/07/2018

Specialization / Stream: Computer Vision

Area of Research work: Weakly Supervised Object Localization

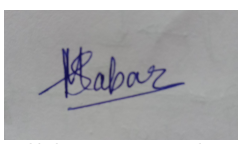
Category of Admission: Project

Guide(s): Prof. Sukhendu Das

Details of Course work:

S.No.	Course No.	Course Title	Sem/Year	Credits	Grade
Core Courses					
1	CS6015	Linear Algebra and Random Processes	01 / 01	12	B
2	CS6730	Probabilistic Graphical Models	02 / 01	12	A
3	CS7015	Deep Learning	02 / 01	12	A
Elective Courses					
1	CS6350	Computer Vision	02 / 01	12	B
2	CS6777	Optimization Methods for Computer Vision Applications	03 / 02	12	S
Optional Courses					
1	CS5691	Pattern Recognition and Machine Learning	01 / 01	15	C
2	CS5020	Non-Linear Optimization: Theory and Algorithms	01 / 01	12	C
Compulsory Courses					
1	ID6020	Introduction to Research (Institute Module)	01 / 01	0	P
2	CS6021	Introduction to Research	01 / 01	0	P

Signature of the Scholar



(Sadbhavana Babar)

Signature of the Guide

(Prof. Sukhendu Das)

Contents

(i) Title of Research Work:
Weakly Supervised Object Localization

(ii) Problem Definition / Research Objectives:
Object Localization is one of the fundamental problems in Computer Vision which helps to understand visual scenes in a better way. 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. E.g. Given an image of a dog, existing methods try to generate implicit attention map only on the face of the dog, leaving its other body parts like legs, tail unattended. This often leads to sub-optimal localization performace. Thus, the focus of this work has been to design an architecture that can cover the entire extent of integral objects. This work proposes 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. We validate our method on two benchmark datasets for object localization: CUB-200-2011 and ILSVRC 2016 datasets.

(iii) Summary of Work Done before Review (From the date of admission till now):

- Course work : Completed seven courses, three in the first semester (Jul - Nov 2018), three in the second semester (Jan - May 2019) and one in the third semester (Jul - Nov 2019). In the first three semesters I took courses which would help me build my fundamentals to work in the field of machine learning and computer vision for my research work.
- Literature Review : Read research papers about recent advances in Object Detection and Recognition. Found out some unexplored areas in the field of object detection.
- Attended various seminars, workshops and talks in and outside the department related to my field of research.

(iv) Work Done During Review:
Considering the limitations of existing methods in weakly supervised object localization, a new approach has been developed by us to tackle the problem of localizing only the most-discriminative part of the object. And the fact that it does not require full supervision during training makes it an interesting problem and a quite challenging one too. Our approach is discussed in detail in the following sections.

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

A.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 [11, 12], there has been a remarkable improvement in image recognition [13–15] and object detection [18–20, 22–26]. 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 [30–33].

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 [34–39, 47, 48] have been proposed to tackle this challenging problem. Zhou et al. [34] put forward the idea of appending a Global Average Pooling (GAP) [10] 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.

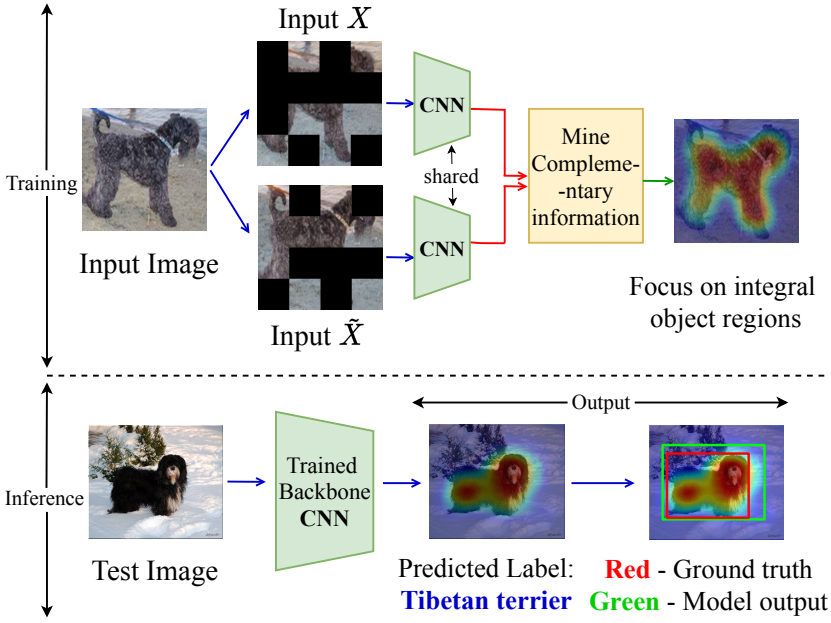


Figure 1: An overview of our proposed approach

To overcome this problem, a few recent methods [35, 39, 45, 46] have come up with making

changes to *input image* rather than modifying the algorithm. In the paper, Hide-and-Seek (HaS) [35], 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 [42, 46] 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 [35], 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 [43, 44, 50], and tries to model interactions in the channel dimension between features extracted from two complementary images. SSAM is inspired by [30, 43, 50] to capture feature dependencies in the spatial dimension. We finally aggregate the inter-dependencies modeled by these two modules: CAAM and SSAM, for better localization ability. We also propose an Attention-based Fusion Loss, inspired by [29], 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.

A.2 Related Work

Zhou et al. [34] put forward the idea of appending a Global Average Pooling (GAP) [10] 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. Randomly masking certain regions in an input image have found to be effective in capturing richer object context and better generalization performance. Bazzani et al. [51] 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 [35], the crux was to randomly hide patches in an input image forcing the network to focus on other relevant object parts. Cutout [46] 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. The work in [36] generates self-produced guidance masks, which in turn are used in the form of pixel-level supervision for localizing objects. Zhang et al. [37] proposed adversarial erasing in feature space that mine information from two adversarial parallel classifiers for superior localization performance. Choe and Shim, in their work [38], 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. [47] 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 [48] by Mai et al. attempts to jointly perform adversarial erasing and mining discriminative regions to localize objects efficiently.

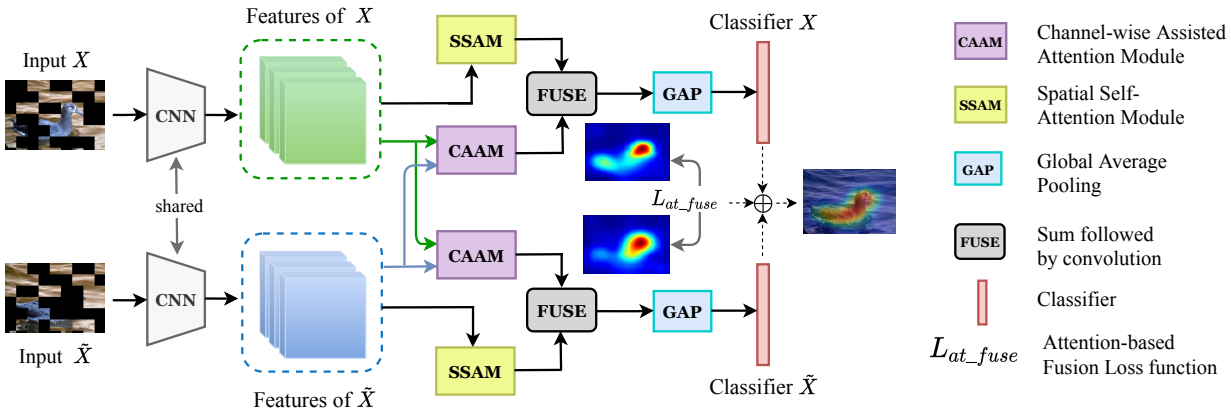


Figure 2: Our proposed architecture

A.3 Proposed Approach

A.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 [35]. 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, $F_{\tilde{x}} \in R^{C \times H \times W}$.

A.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 [10], 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.

A.3.3 Channel-wise Assisted Attention Module

To compute CAMs [34], Zhou 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 [50], Fu 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 [50], 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, $F_{\tilde{x}}^c$ $\in R^{C\times H\times W}$. A similar approach has been studied in [49] 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.

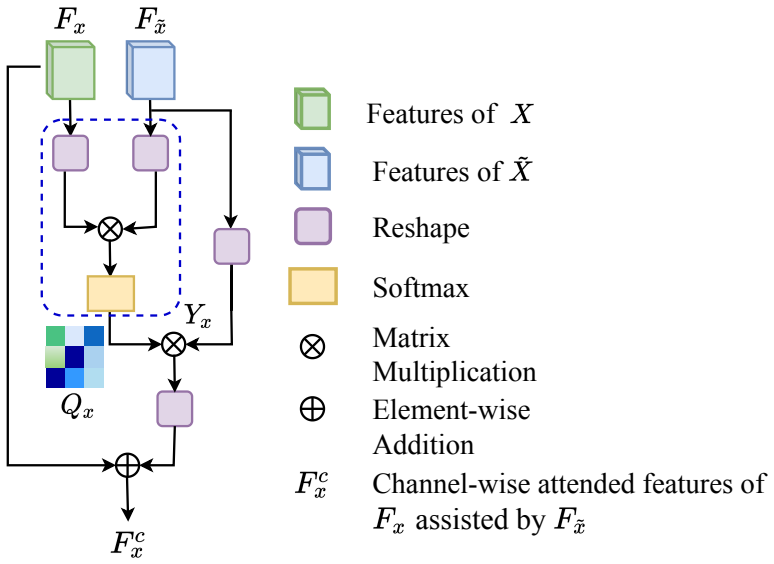


Figure 3: Channel-wise Assisted Attention Module (CAAM)

To compute F_x^c, we take the input features F_x, $F_{\tilde{x}}$ and reshape them to $R^{C\times N}$, 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, $Q_x \in R^{C\times C}$ 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 $R^{C\times H\times W}$ 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 [43, 50]. 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 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 $F_{\tilde{x}}$.

A.3.4 Spatial Self-Attention Module

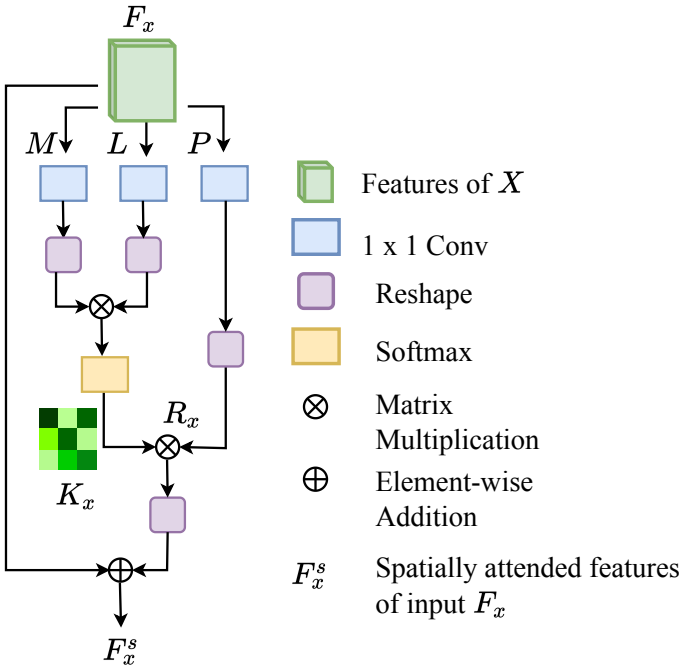


Figure 4: Spatial Self-Attention Module (SSAM)

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 [43], self-attention was used in GANs [1]. Taking motivation from [43], 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 $R^{C\times H\times W}$. As illustrated in figure 4, given features F_x, we compute matrices M, L and P using 1x1 convolution where, $\{M, L\} \in R^{\tilde{C}\times H\times W}$, where $\tilde{C} = C/8$, and $P \in R^{C\times H\times W}$. 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:

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

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

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

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.

A.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 [34]. 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).

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

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) [10] 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 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 [37], 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 [29]. 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.

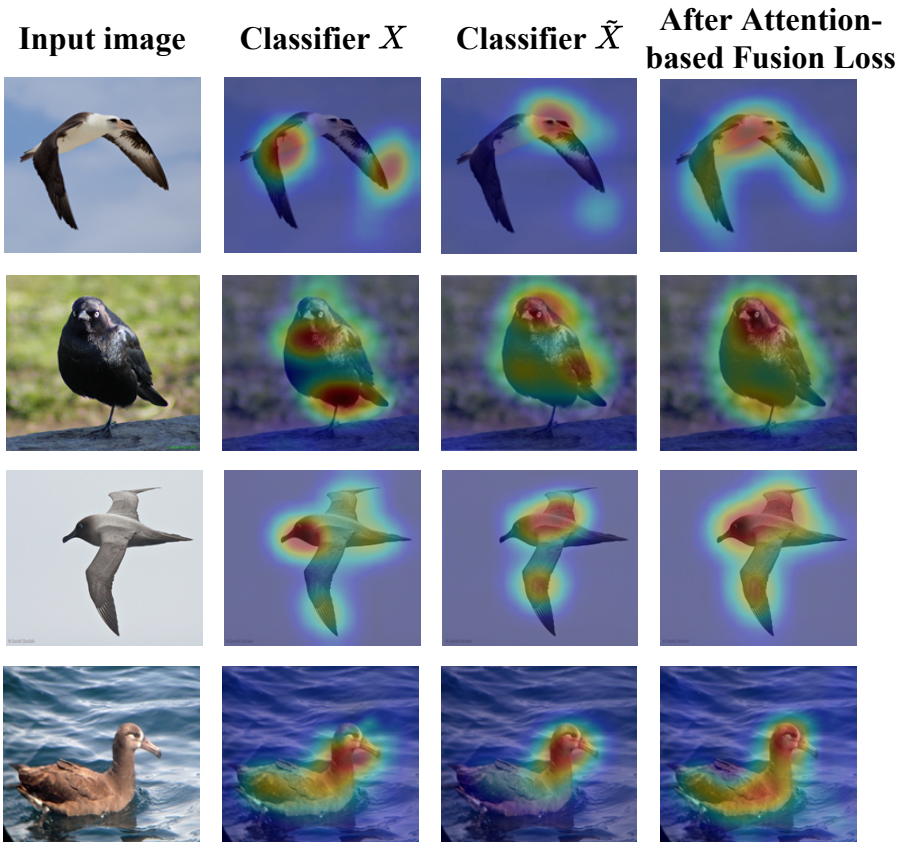


Figure 5: Visualization of the proposed Attention-based Fusion Loss During training, we visualize the effect of our proposed loss function. The left column denotes the input image, the second and third columns denote the localization maps of our two classifiers and the right column denotes the localization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

A.4 Experiments and Results

A.4.1 Experimental Setup

Datasets: We perform our experiments on two benchmark datasets used for object localization, CUB-200-2011 [3] and ILSVRC 2016 [4]. CUB-200-2011 has a total 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 [4] 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 [35] 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* [2] metric (as shown in table 4).

Method	Top-1 Loc (%)	Top-1 Clas (%)
InceptionV3-CAM [34]	43.67	**73.80**
InceptionV3-SPG [36]	46.64	–
InceptionV3-DANet [40]	**49.45**	71.20
VGG-CAM [34]	34.41	67.55
VGG-ACoL [37]	45.92	71.90
VGG-ADL [38]	52.36	65.27
VGG-CCAM [47]	50.07	73.20
VGG-EIL [48]	56.21	72.26
Ours-VGG	**58.12**	**72.59**
ResNet50-CAM [34]	49.41	75.68
ResNet50-CutMix [39]	54.81	–
Ours-ResNet50	**64.70**	**77.28**

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

A.4.2 Implementation Details

We experiment with VGG16 [13] and ResNet50 [15] as the backbone CNN architectures for our proposed approach. As in [34], 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 [10], 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 [4] dataset. We extract our localization maps followed by bounding boxes, in a similar way to [34]. During testing, we do not hide patches in the input image, similar to [35]. Also, we deactivate CAAM and SSAM modules during testing, similar to vanilla CAM [34] model for fair comparison with other existing state-of-the-art methods (shown in tables 1, 2 & 3).

Method	Top-1 Clas (in %)
InceptionV3-CAM [34]	68.10
GoogLeNet-HaS-32 [35]	**70.70**
VGG-CAM [34]	66.60
VGG-ACoL [37]	67.50
VGG-ADL [38]	69.48
VGG-CCAM [47]	66.60
VGG-EIL [48]	70.48
Ours-VGG	**71.24**

Table 2: Classification performance on ILSVRC 2016 dataset.

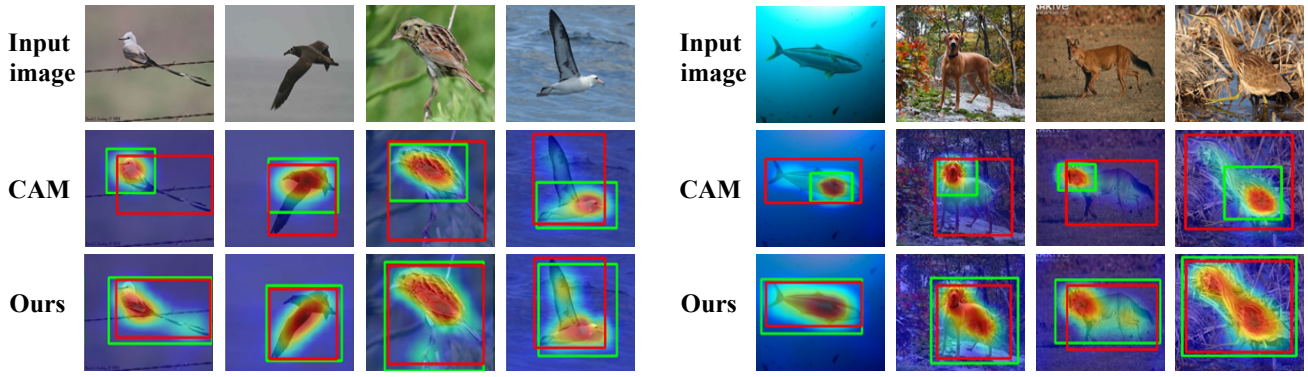
Method	Top-1 Loc (%)	GT-Loc (%)
InceptionV3-CAM [34]	46.29	–
GoogLeNet-HaS-32 [35]	45.21	60.29
InceptionV3-SPG [36]	48.60	64.69
InceptionV3-DANet [40]	47.53	–
InceptionV3-MEIL [48]	**49.48**	–
VGG-CAM [34]	42.80	57.72
VGG-ACoL [37]	45.80	62.96
VGG-ADL [38]	44.92	–
VGG-CCAM [47]	48.22	63.58
VGG-EIL [48]	46.27	–
Ours-VGG	**51.64**	**66.32**
ResNet50-CAM [34]	38.99	51.86
ResNet50-SE-ADL [38]	48.53	–
ResNet50-CutMix [39]	47.25	–
Ours-ResNet50	**52.36**	**67.89**

Table 3: Localization Results on ILSVRC 2016 dataset.

Method	CUB-200-2011	ILSVRC 2016
CAM [34]	71.1	61.1
HaS-32 [35]	76.3	61.8
ACoL [37]	72.3	60.3
SPG [36]	63.7	61.6
ADL [38]	75.7	60.8
CutMix [39]	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 [2] on VGG16 as the backbone.

In figure 6, we compare our results qualitatively with the baseline CAM [34] 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.



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

Figure 6: Qualitative Results

(v) Future Plans: Extend the work of weakly-supervised object localization for the case of multiple objects in an image, more like weakly-supervised object detection, laying the foundation for significantly bridging the gap between supervised and weakly-supervised methods.

(vi) Visible Research Output:

(a) Full Paper(s) Published in Conference Proceedings:

- "Where to Look?: Mining Complementary Image Regions for Weakly Supervised Object Localization", **Sadbhavana Babar** and Sukhendu Das; In the IEEE Winter Conference on Applications of Computer Vision (WACV) [Rank A], Fully Virtual, Jan 5-9, 2021.

(b) Seminars/Workshops/Conferences/Exchange Programmes attended and Papers Presented:

- Attended the Computer Vision Track of the Google Research India AI Summer School from August 20-22, 2020.
- Attended the Workshop on Machine Intelligence and Brain Research organized by the Center for Computational Brain Research (CCBR), IIT Madras from January 2-10, 2020.
- Attended IMPRINT Project Review Meeting II held at IIT Delhi from 31/10/2019 to 01/11/2019.
- Visited DRDO, CAIR in Bangalore to gain insights about the ongoing IMPRINT Project from January 10-11, 2019.
- Attended 11th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP) 2018 held at IIIT Hyderabad from December 18-22, 2018.

(c) Awards/Honours, if any:

- Awarded cash prize for being in the top-20 performers in the CVIT Computer Vision Summer School, 2019 held at IIIT Hyderabad from 01/07/2019 to 07/07/2019.

References

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

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

[3] Wah, Catherine and Branson, Steve and Welinder, Peter and Perona, Pietro and Belongie, Serge *The caltech-ucsd birds-200-2011 dataset*, California Institute of Technology, 2011.

[4] Russakovsky, Olga and Deng, Jia and Su, Hao and Krause, Jonathan and Satheesh, Sanjeev and Ma, Sean and Huang, Zhiheng and Karpathy, Andrej and Khosla, Aditya and Bernstein, Michael and others *Imagenet large scale visual recognition challenge* International Journal of Computer Vision, 2015

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

[6] Tang, Peng and Wang, Xinggang and Bai, Xiang and Liu, Wenyu *Multiple instance detection network with online instance classifier refinement* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

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

[8] Diba, Ali and Sharma, Vivek and Pazandeh, Ali and Pirsiavash, Hamed and Van Gool, Luc *Weakly supervised cascaded convolutional networks* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

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

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

[11] Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E *Imagenet classification with deep convolutional neural networks* Advances in Neural Information Processing Systems (NeurIPS), 2012.

[12] LeCun, Yann and Boser, Bernhard and Denker, John S and Henderson, Donnie and Howard, Richard E and Hubbard, Wayne and Jackel, Lawrence D *Backpropagation applied to handwritten zip code recognition* Neural computation, 1989.

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

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

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

[16] Girshick, Ross and Donahue, Jeff and Darrell, Trevor and Malik, Jitendra *Region-based convolutional networks for accurate object detection and segmentation* IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), 2015.

[17] Girshick, Ross *Fast r-cnn* Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[18] Ren, Shaoqing and He, Kaiming and Girshick, Ross and Sun, Jian *aster r-cnn: Towards real-time object detection with region proposal networks* Advances in Neural Information Processing Systems (NeurIPS), 2015.

[19] Liu, Wei and Anguelov, Dragomir and Erhan, Dumitru and Szegedy, Christian and Reed, Scott and Fu, Cheng-Yang and Berg, Alexander C *Ssd: Single shot multibox detector* European Conference on Computer Vision (ECCV), 2016.

[20] Redmon, Joseph and Divvala, Santosh and Girshick, Ross and Farhadi, Ali *You only look once: Unified, real-time object detection* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[21] Redmon, Joseph and Farhadi, Ali *YOLO9000: better, faster, stronger* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[22] Lin, Tsung-Yi and Dollár, Piotr and Girshick, Ross and He, Kaiming and Hariharan, Bharath and Belongie, Serge *Feature pyramid networks for object detection* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[23] Singh, Bharat and Najibi, Mahyar and Davis, Larry S *Sniper: Efficient multi-scale training* author=Singh, Bharat and Najibi, Mahyar and Davis, Larry S, Advances in Neural Information Processing Systems (NeurIPS), 2018.

[24] Duan, Kaiwen and Bai, Song and Xie, Lingxi and Qi, Honggang and Huang, Qingming and Tian, Qi *Centernet: Keypoint triplets for object detection*, Proceedings of the IEEE International Conference on Computer Vision, (CVPR), 2019.

[25] Tan, Mingxing and Pang, Ruoming and Le, Quoc V *Efficientdet: Scalable and efficient object detection*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

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

[27] Huang, Zilong and Wang, Xinggang and Wang, Jiasi and Liu, Wenyu and Wang, Jingdong *Weakly-supervised semantic segmentation network with deep seeded region growing* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

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

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

[30] Wang, Yude and Zhang, Jie and Kan, Meina and Shan, Shiguang and Chen, Xilin *Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation*, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[31] Ahn, Jiwoon and Cho, Sunghyun and Kwak, Suha *Weakly supervised learning of instance segmentation with inter-pixel relations*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[32] Zeng, Yu and Zhuge, Yunzhi and Lu, Huchuan and Zhang, Lihe *Joint learning of saliency detection and weakly supervised semantic segmentation* In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019.

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

[34] Zhou, Bolei and Khosla, Aditya and Lapedriza, Agata and Oliva, Aude and Torralba, Antonio *Learning deep features for discriminative localization* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[35] Singh, Krishna Kumar and Lee, Yong Jae *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 (ICCV), 2017.

[36] Zhang, Xiaolin and Wei, Yunchao and Kang, Guoliang and Yang, Yi and Huang, Thomas *Self-produced guidance for weakly-supervised object localization* In Proceedings of the European Conference on Computer Vision (ECCV), 2018.

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

[38] Choe, Junsuk and Shim, Hyunjung *Attention-based dropout layer for weakly supervised object localization* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

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

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

[41] Benenson, Rodrigo and Popov, Stefan and Ferrari, Vittorio *Large-scale interactive object segmentation with human annotators* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[42] Zhong, Zhun and Zheng, Liang and Kang, Guoliang and Li, Shaozi and Yang, Yi *Random Erasing Data Augmentation*, AAAI 2020.

[43] Zhang, Han and Goodfellow, Ian and Metaxas, Dimitris and Odena, Augustus *Self-Attention Generative Adversarial Networks* In International Conference on Machine Learning (ICLR), 2019.

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

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

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

[47] Yang, Seunghan and Kim, Yoonhyung and Kim, Youngeun and Kim, Changick *Combinational Class Activation Maps for Weakly Supervised Object Localization*, In the IEEE Winter Conference on Applications of Computer Vision (WACV), 2020.

[48] Mai, Jinjie and Yang, Meng and Luo, Wenfeng *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 (CVPR), 2020.

[49] Zhou, Jieming and Roy, Soumava Kumar and Fang, Pengfei and Harandi, Mehrtash and Petersson, Lars *Cross-Correlated Attention Networks for Person Re-Identification* Image and Vision Computing (IVC), 2020.

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

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