Foreground Activation Maps for Weakly Supervised Object Localization

Meng Meng[1], Tianzhu Zhang[1,*] Qi Tian[2], Yongdong Zhang[1], Feng Wu[1]
[1] University of Science and Technology of China
[2] Cloud BU, Huawei Technologies
meng18@mail.ustc.edu.cn, {tzzhang, zhyd73, fengwu}@ustc.edu.cn, tian.qi1@huawei.com

Abstract

Weakly supervised object localization (WSOL) aims to localize objects with only image-level labels, which has better scalability and practicability than fully supervised methods in the actual deployment. However, with only image-level labels, learning object classification models tends to activate object parts and ignore the whole object, while expanding object parts into the whole object may deteriorate classification performance. To alleviate this problem, we propose foreground activation maps (FAM), whose aim is to optimize object localization and classification jointly via an object-aware attention module and a part-aware attention module in a unified model, where the two tasks can complement and enhance each other. To the best of our knowledge, this is the first work that can achieve remarkable performance for both tasks by optimizing them jointly via FAM for WSOL. Besides, the designed two modules can effectively highlight foreground objects for localization and discover discriminative parts for classification. Extensive experiments with four backbones on two standard benchmarks demonstrate that our FAM performs favorably against state-of-the-art WSOL methods.

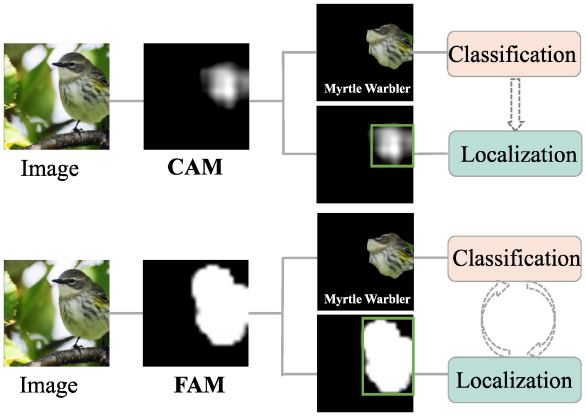

Figure 1. The motivation of our method. In CAM-based methods, object localization is learned as a by-product of the classification, and the model has to rely on class-specific image regions for localization. In this situation, optimizing object classification tends to activate object parts instead of the whole object, while expanding object parts into the whole object could deteriorate classification performance. Unlike the CAM, our FAM aims to discover foreground maps of all classes for localization and select discriminative parts for classification. Meanwhile, the two tasks can complement and enhance each other in a collaborative way.

1. Introduction

Object localization aims to recognize objects and identify their locations in the given images [38]. Because of its broad applications such as autonomous driving [4, 5], face recognition [52, 30], and person re-identification [57, 28, 21], object localization has attracted increasing attention in the research community. However, most existing methods tackle this task in a fully supervised setting [13, 12, 40] by using precise bounding box annotations. Thus, their scalability and practicability are limited in real-world application scenarios because it is expensive and time-consuming to gather massive labeled fine-grained data.

To overcome the above limitations, several recent methods have been proposed by using weakly supervised learning models [63, 7, 2, 41, 43] that require only image-level category labels. However, without box-level annotations, it is challenging to localize objects accurately [60]. Recently, class activation maps (CAM) based methods [63, 41] have been proposed to handle this challenge. These methods use a global average pooling layer and a final fully connected layer (weights of the classifier) to obtain localization maps, which identify discriminative regions for specific object classes [63]. Thus, they perform object localization by using class-specific image regions. In other words, object localization is learned as a by-product of the classification, as shown in Figure 1. Unfortunately, this idea tends to be biased on the most discriminative object part to increase the classification accuracy while ignoring less discriminative object regions, leading to decreased localization accu-

*Corresponding Author

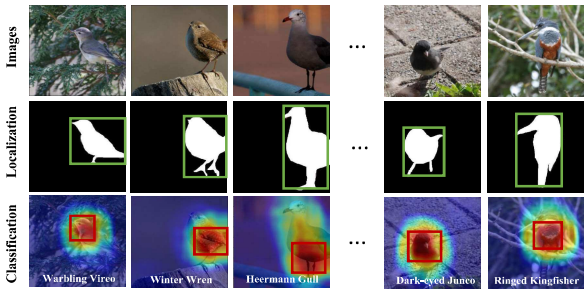

Figure 2. Examples of birds on the CUB-200-2011 dataset. Object localization is to highlight the whole foreground object, and object classification is to select the most discriminative parts.

racy. In pursuit of highlighting the whole object, several techniques have been proposed, which can be mainly categorized into two categories. The first category [51, 61, 26] aims to expand the range of the most discriminative part by exploring context information. However, since there may be large differences among object parts, it is hard to expand the range of the most discriminative part to other object parts, resulting in incomplete object localization. The other category [7, 43, 60, 22, 50] is to erase the most discriminative part and then force the model to discover other relevant parts. Although these approaches can expand class activation maps, they often highlight background regions [54].

To the best of our knowledge, most previous methods utilize the CAM for object classification and localization. Here, object classification is to select the most discriminative object parts, and object localization is to highlight the whole foreground object [41, 47, 23], as shown in Figure 2. However, the CAM achieves object localization as a by-product of object classification, and the model has to rely on class-specific image regions for object localization. In this situation, optimizing object classification tends to activate object parts instead of the whole object, while expanding object parts into the whole object could deteriorate classification performance. To alleviate this problem, we argue that it is better to use **foreground activation maps (FAM)**, which are not class-specific and aim to discover foreground maps of all classes from the background for object localization, as illustrated in Figure 1. The FAM proposes a new perspective for WSOL by optimizing object localization and classification jointly in a unified model, where the two tasks can complement and enhance each other. (1) Object localization can help object classification. Given a dataset, a well-learned object localization model is designed to identify foreground and background regions. As shown in Figure 2, the CUB-200-2011 dataset includes 200 species of birds, which are foreground objects with similar foreground patterns and are different from background regions. As a result, the localization model of FAM can highlight foreground regions. Guided by the learned foreground regions, the classification model can directly choose the most discriminative parts without background interference [24, 31]. (2) Object classification can help object localization. The classifier of FAM is to find the discriminative parts for each class. As shown in Figure 2, different kinds of birds have different discriminative regions [63]. With the guidance of the well-learned classifier, the localization model of FAM is to cover all the discriminative parts. Therefore, the FAM is designed to distinguish foreground regions for localization and select the discriminative parts for classification to achieve remarkable performance for both tasks.

Motivated by the above discussions, we propose foreground activation maps (FAM) to optimize object localization and classification jointly in a unified model via an object-aware attention module and a part-aware attention module. Here, the first module is designed to distinguish foreground from background for localization, and the second module is proposed to exploit discriminative object parts for classification. In the **object-aware attention module**, we design a foreground memory mechanism to identify foreground and background regions in a given dataset, which can deal with large appearance variations of different objects. In specific, we store multiple foreground appearance templates as memory keys, and save multiple foreground classifiers as memory values. Each foreground classifier is designed to determine the likelihood that one specific appearance pattern belongs to a foreground object. Given the feature map of an input image, we can read from the memory and get a set of pixel-wise classifiers by treating each pixel as a query, then a foreground map is obtained based on the pixel-wise classifiers. And based on the fact that foreground object usually occupies a small portion of the image, we add a sparsity constrain on the foreground map, which serves as a prior to guide the learning of the memory. Together with the subsequent classification module, the memory keys and values can be learned through the whole dataset during training. In the **part-aware attention module**, we design a part discovery module to generate several part-aware activation maps. Each part-aware activation map denotes the spatial distribution of one specific part; that is to say, the part-aware activation map has high response values at the pixels belonging to that part. The part-aware features are produced by the attention weighted pooling from the feature map. Because different parts may have different importance for object classification, we exploit part diversity and part importance to constrain the part-aware feature learning in the proposed part discovery module. By jointly optimizing the object-aware and part-aware attention modules, we can achieve robust object localization and classification simultaneously.

The major contributions of this work can be summarized as follows. (1) We propose foreground activation maps (FAM) for weakly supervised object localization (WSOL) to optimize object localization and classification jointly via an object-aware attention module and a part-aware attention

module in a unified model, where the two tasks can complement and enhance each other. To the best of our knowledge, this is the first work that can achieve remarkable performance for both tasks by optimizing them jointly via FAM for WSOL. (2) The designed two modules (object-aware and part-aware attention modules) can effectively highlight foreground objects for localization and discover discriminative parts for classification. (3) Extensive experimental results with four different backbones on two challenging benchmarks show that our FAM performs favorably against state-of-the-art WSOL methods.

2. Related Work

Weakly Supervised Object Localization (WSOL). Given images only with class labels, the WSOL task is to predict both object positions and categories [58, 6]. In [38], it is the first end-to-end approach for weakly supervised object localization. However, the localization is limited to a point rather than the full extent of the object. Later, Zhou et al. [63] generate Class Activation Maps (CAM) with a global average pooling layer and a final fully connected layer (weights of the classifier) to obtain localization maps. To remove the reliance on specific network architectures, Gradient-weighted Class Activation Mapping (Grad-CAM) [41] is proposed. While simple and effective, the CAM-based methods tend to be biased on the most discriminative part. To mitigate this issue, several methods explore object context information to expand the range of the most discriminative part, and the context information can come from different spatial positions [51, 26] or different layers [61]. For example, Wei et al. [51] employ a dilated convolution to consider spatial contexts at various ratios. Several other methods adopt an erasing strategy [43, 22, 49, 50, 27, 18], whose aim is to erase the most discriminative part so that the model needs to seek the relevant object parts from what remains. In [43], the model is designed to hide grid-like patches during training randomly. To erase the most discriminative part effectively, Zhang et al. [60] learn parallel adversarial classifiers to find complementary parts for target objects, and more sophisticated erasing strategies are designed in later works [7, 34]. Besides from the above works, some other approaches explore divergent activations [53], class-agnostic localization maps [58], geometry constrained network [32] and inter-image information [62] to improve localization performance. Most existing methods achieve object localization as a by-product of object classification. Unlike these methods, we propose foreground activation maps to achieve object localization and classification in a collaborative manner.

Memory Networks. Memory networks refer to the architectures that have access to an addressable memory repository for prediction [14, 35]. Different from LSTM [17] and GRU [8], which involve an internal memory implicitly updated in a recurrent process, memory networks explore an explicit memory that can be read or written with an addressing procedure [11, 14, 25, 45, 48]. The addressing methods can be classified into content-based addressing and location-based addressing. The content-based addressing [15, 35, 36] measures the similarity between the query and memory keys to find relevant memory cells. The localization-based addressing [14], on the other hand, enables a simple operation on the query to find out the addresses, regardless of the content of memory keys. Graves et al. [14] first propose Neural Turing Machine (NTM), which can interact with a memory matrix using selective read and write operations. Later, the work of [35, 59] proposes a key-value memory to store information in the form of key-value pairs, which can directly learn the correlation between the input and underlying concepts in memory. To make the read/write operations scalable with a large amount of memory, Chandar et al. [3] propose to organize memory hierarchically and Rae et al. [39] make read and write operations sparse to reduce the cost of operations. Thanks to the addressing design, memory networks typically update query-related memory cells instead of the whole memory. This happens to help to learn the training data structure, such as some common semantic representations [20, 9] or visual patterns [55, 37] sharing among words or images. For weakly supervised object localization, this is the first work to explore multiple foreground patterns in a given dataset with a memory network, which helps to deal with large appearance variations of different objects. As a result, the FAM can better identify foreground and background regions for object localization and classification jointly.

3. Approach

In WSOL, given an image $\boldsymbol{I}$, let $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ denote the feature map extracted from a backbone network. where H, W, and C denote the height, width and channel number of the feature map, respectively. During training, each image is associated with a ground truth label $\boldsymbol{y} \in \mathbb{R}^{L}$, and L refers to the number of categories. During testing, given an image, the outputs are a predicted category label $\tilde{\boldsymbol{y}}$ and the localized object bounding box $\mathcal{B} = (x^*, y^*, h^*, w^*)$, where (x^*, y^*) denotes the center coordinate, and (h^*, w^*) denotes the height and width.

As illustrated in Figure 3, the FAM consists of two modules including an object-aware attention module and a part-aware attention module. The first module is designed to distinguish foreground from background for localization, and the second module is proposed to exploit discriminative object parts for classification. The details are as follows.

3.1. Object-aware Attention Module

Given a specific dataset with several object categories, foreground objects have similar foreground patterns, which

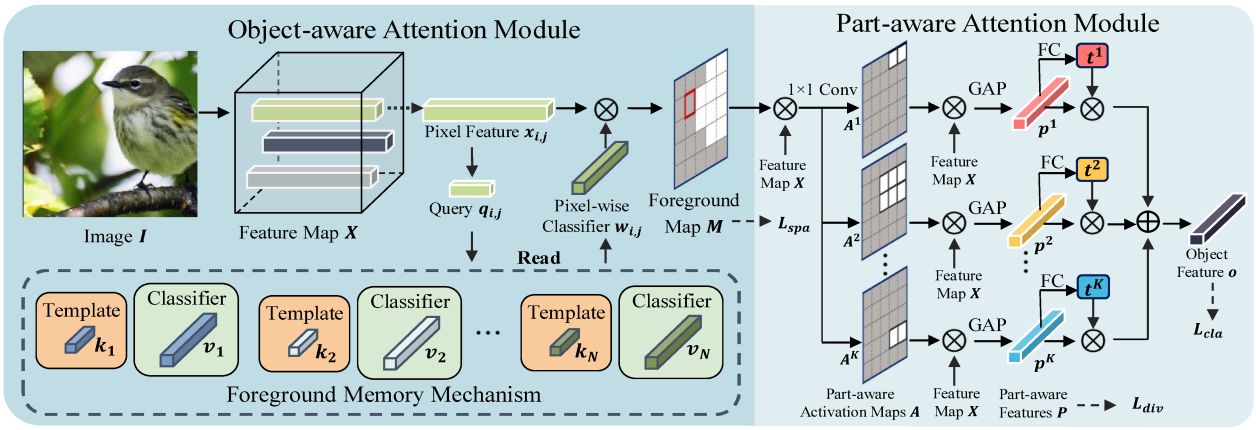

Figure 3. The architecture of our FAM including an object-aware attention module and a part-aware attention module. By optimizing the object-aware attention module and the part-aware attention module jointly, the FAM can achieve robust object localization and classification in a collaborative manner. In this figure, "1 × 1 Conv" denotes a convolutional layer with a 1 × 1 kernel size and "FC" denotes the fully connected layer. Besides, "GAP" represents a global average pooling layer [29]. For more details, please refer to the paper.

are different from the background and can be memorized to identify foreground and background regions by using a foreground memory mechanism. Generally, in a given dataset, different objects may have large appearance variations. Therefore, we need to generate appearance-adaptive foreground classifiers with dynamic weights to handle appearance variations. In specific, we design N keys $\{\boldsymbol{k_n}\}_{n=1}^{N}$ and values $\{\boldsymbol{v_n}\}_{n=1}^{N}$ in the memory. Each key denotes a specific appearance template, and each value represents a key-related foreground classifier. Given a feature map, we first read from the memory and get a set of pixel-related foreground classifiers. These foreground classifiers are adaptive to appearance variations and can be used to generate a foreground map. In practice, the memory keys are embedded as $C/16$ dimensional vectors to improve the efficiency of memory reading, and values are embedded as 1×1 convolutional kernels whose dimension is C.

To read from the memory, we feed the feature map $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ into an encoder to acquire a set of queries $\boldsymbol{Q} \in \mathbb{R}^{H\times W\times C/16}$. The set contains $H \times W$ queries, each of which is represented as $\boldsymbol{q_{i,j}} \in \mathbb{R}^{C/16}$, where $i = 1, 2, \ldots, H$ and $j = 1, 2, \ldots, W$. The similarity $s_{i,j}^{n}$ between each query $\boldsymbol{q_{i,j}}$ and the n-th key $\boldsymbol{k_n}$ is given as

$$s_{i,j}^{n} = \frac{\beta_{i,j}^{n}}{\sum_{n=1}^{N}\beta_{i,j}^{n}},\ \beta_{i,j}^{n} = \frac{\boldsymbol{q_{i,j}^{\mathsf{T}}}\,\boldsymbol{k_n}}{\sqrt{C/16}}, \tag{1}$$

where $n = 1, 2, \ldots, N$ and T refers to the transpose operation. With the similarity score $s_{i,j}^{n}$, we can get the pixel-wise foreground classifier $\boldsymbol{w_{i,j}} \in \mathbb{R}^{C}$ for the query $\boldsymbol{q_{i,j}}$ adaptively by blending memory values $\{\boldsymbol{v_n}\}_{n=1}^{N}$ as

$$\boldsymbol{w_{i,j}} = \sum_{n=1}^{N} s_{i,j}^{n} \cdot \boldsymbol{v_n}. \tag{2}$$

Thus, the generated foreground classifiers are adaptive to appearance variations. Given the feature map $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$, $\boldsymbol{x_{i,j}} \in \mathbb{R}^{C}$ indicates the pixel feature located at (i, j) on the feature map. To generate the foreground map $\boldsymbol{M} \in \mathbb{R}^{H\times W}$, the i-th row and j-th column of $\boldsymbol{M}$ is first calculated by

$$M_{i,j} = \boldsymbol{w_{i,j}^{\mathsf{T}}}\,\boldsymbol{x_{i,j}}. \tag{3}$$

The overall foreground map $\boldsymbol{M}$ is obtained by performing the same operation for all pixels on the feature map, and then is normalized by a sigmoid function.

Based on the fact that the foreground object usually occupies a small portion of the image, we add a sparsity constrain on the foreground map for background suppression, which serves as a prior to guide the learning of the memory.

$$L_{spa} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}|M_{i,j}|. \tag{4}$$

Together with the subsequent classification module, as defined in (10), even without box-level annotations, we can learn foreground maps that highlight nearly the entire object. The intuition behind this idea is simple; the classification loss requires that object-related regions are activated to classify the image correctly, but the sparsity loss requires the foreground map to focus on as few pixels as possible. As a result, these two loss terms together can suppress background regions and highlight foreground regions only.

3.2. Part-aware Attention Module

While the object-aware attention module can effectively highlight foreground objects for localization, the part-aware attention module is designed to exploit discriminative object parts for classification. We first multiply the foreground map $\boldsymbol{M} \in \mathbb{R}^{H\times W}$ and the feature map $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ to generate the foreground feature map $\tilde{\boldsymbol{X}} \in \mathbb{R}^{H\times W\times C}$.

$$\tilde{X}_{i,j,c} = X_{i,j,c} \cdot M_{i,j}, \tag{5}$$

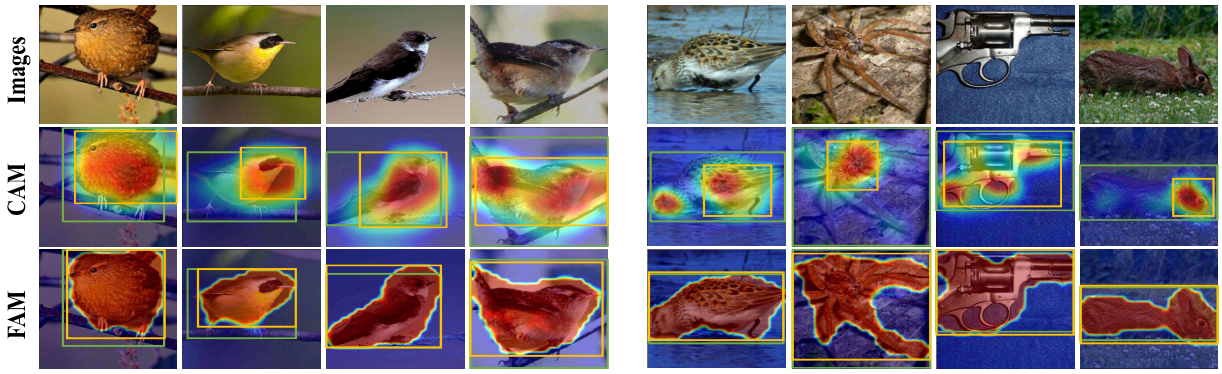

Figure 4. Visualization comparison with CAM [63]. The predicted bounding boxes are in yellow and the ground truth boxes are in green. Our method can highlight nearly the entire object and produce precise bounding boxes for images on CUB-200-2011 and ILSVRC 2016.

where i, j, and c index the height, width and channel number of the foreground feature map.

Then, based on the foreground feature map $\tilde{\boldsymbol{X}}$, we generate part-aware activation maps $\boldsymbol{A} \in \mathbb{R}^{H\times W\times K}$ to mine K object parts through a part discovery module $f(\cdot|\theta)$, parameterized by θ. To discover object parts in a simple and effective way, this module is implemented as a convolution layer followed by a sigmoid function to change the channel size of the foreground feature map to K.

$$\boldsymbol{A} = f(\tilde{\boldsymbol{X}}|\theta), \tag{6}$$

where $\boldsymbol{A} = \{\boldsymbol{A^k}\}_{k=1}^{K}$ denotes a set of part-aware activation maps, and the $\boldsymbol{A^k} \in R^{H\times W}$ corresponds to the k-th part-aware activation map. Each part-aware activation map denotes the spatial distribution of one specific part. That is to say, the part-aware activation map has high response values at the pixels belonging to that part. Based on the feature map $\boldsymbol{X}$, we generate a set of part-aware features $\boldsymbol{P} = \{\boldsymbol{p^k}\}_{k=1}^{K}$ by the attention weighted pooling, and the k-th part-aware feature $\boldsymbol{p^k} = [p_1^k, p_2^k, \ldots, p_C^k]$ is given as

$$p_c^k = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j,c}\cdot A_{i,j}^k, \tag{7}$$

where $c = 1, 2, \ldots, C$.

To discover different object parts with only image-level labels, we impose a diversity loss [33, 56] to expand the discrepancy among part-aware features $\{\boldsymbol{p^k}\}_{k=1}^{K}$ as

$$L_{div} = \frac{1}{K(K-1)}\sum_{k=1}^{K}\sum_{m=1,m\neq k}^{K}\frac{\langle \boldsymbol{p^k}, \boldsymbol{p^m}\rangle}{\|\boldsymbol{p^k}\|_2\|\boldsymbol{p^m}\|_2}. \tag{8}$$

Because different parts have different importance for object classification, we feed the part-aware features into an importance prediction module $g(\cdot|\phi)$, parameterized by ϕ, to evaluate their importance and generate importance weights $\{t^k\}_{k=1}^{K}$. The importance prediction module $g(\cdot|\phi)$ is a linear layer followed by a softmax operation to output probabilities between 0 and 1. The final object feature $\boldsymbol{o} \in R^C$ is obtained by a weighted sum of the part-aware features.

$$t^k = g(\boldsymbol{p^k}|\phi), \quad \boldsymbol{o} = \sum_{k=1}^{K} t^k\cdot\boldsymbol{p^k}. \tag{9}$$

We obtain the category prediction $\tilde{\boldsymbol{y}} = h(\boldsymbol{o}|\sigma)$ through a classification module $h(\cdot|\sigma)$, parameterized by σ and implemented as a fully connected layer. Finally, the class-balanced cross entropy loss is employed between the category prediction and the ground truth label for classification.

$$L_{cla}(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = -\sum_{l=1}^{L} y_l\cdot log\tilde{y}_l. \tag{10}$$

Where L denotes the number of categories, and $\tilde{y}_l$ and y_l are the l-th element of $\tilde{\boldsymbol{y}}$ and $\boldsymbol{y}$, respectively.

3.3. Joint Training

By optimizing the object-aware and part-aware attention modules jointly, the FAM can achieve robust object localization and classification in a collaborative manner. For WSOL, with only image-level category labels, our FAM is trained by minimizing the overall objective as follows

$$L_{final} = L_{cla} + \lambda_{spa}L_{spa} + \lambda_{div}L_{div}, \tag{11}$$

where λ_{spa} and λ_{div} are the balance parameters. When L_{spa} is jointly learned with L_{cla}, L_{cla} of all categories constrains L_{spa} only suppress the background that not related to any class labels, since filtering out the foreground leads to a large L_{cla}. Thus, our FAM can generate foreground maps of all classes to cover nearly the entire object, while CAM uses activation maps of the highest probability class that only activate the most discriminative parts [7].

To perform object localization, the bilinear interpolation is used for upsampling the foreground map to the original image size. We identify the discriminative regions by a hard threshold as in [63, 61]. The detection bounding box B is the coverage of the largest connected area obtained by using the threshold truncation on the foreground map [63].

Table 1. Comparison of the proposed method with other state-of-the-art algorithms.

Method	Backbone	CUB-200-2011			ILSVRC 2016		
		Top-1 Loc	Top-1 Cls	GT-known	Top-1 Loc	Top-1 Cls	GT-known
CAM *(cvpr,2016)*	VGG16	34.41	67.55	-	42.80	66.60	59.00
ACoL *(cvpr,2018)*	VGG16	45.92	71.90	45.90	45.83	67.50	62.96
ADL *(cvpr,2019)*	VGG16	52.36	65.27	-	44.92	69.48	-
DANet *(iccv,2019)*	VGG16	52.52	75.40	67.70	-	-	-
EIL *(cvpr,2020)*	VGG16	57.46	74.77	-	46.81	70.27	-
PSOL *(cvpr,2020)*	VGG16	66.30	-	-	50.89	-	64.03
GCNet *(eccv,2020)*	VGG16	63.24	76.80	81.10	-	-	-
RCAM *(eccv,2020)*	VGG16	58.96	75.01	76.30	44.62	68.67	60.73
FAM *(ours)*	**VGG16**	**69.26**	**77.26**	**89.26**	**51.96**	**70.90**	**71.73**
ADL *(cvpr,2019)*	ResNet50-SE	62.29	80.34	-	48.53	75.85	-
PSOL *(cvpr,2020)*	ResNet50	70.68	-	-	53.98	-	**65.44**
RCAM *(eccv,2020)*	ResNet50	59.53	75.03	77.58	49.42	75.82	62.20
FAM *(ours)*	**ResNet50**	**73.74**	**82.72**	**85.73**	**54.46**	**76.48**	64.56
CAM *(cvpr,2016)*	MobileNetV1	43.70	71.94	-	41.66	68.38	-
HaS *(iccv,2017)*	MobileNetV1	44.67	66.64	-	41.87	67.48	-
ADL *(cvpr,2019)*	MobileNetV1	47.74	70.43	-	43.01	67.77	-
RCAM *(eccv,2020)*	MobileNetV1	59.41	73.51	78.60	44.78	67.15	61.69
FAM *(ours)*	**MobileNetV1**	**65.67**	**76.38**	**85.71**	**46.24**	**70.28**	**62.05**
SPG *(eccv,2018)*	InceptionV3	46.64	-	-	48.60	-	64.69
ADL *(cvpr,2019)*	InceptionV3	53.03	74.55	-	48.71	72.83	-
DANet *(iccv,2019)*	InceptionV3	49.45	71.20	-	48.71	72.83	-
PSOL *(cvpr,2020)*	InceptionV3	65.51	-	-	54.82	-	65.21
I^2C *(eccv,2020)*	InceptionV3	55.59	-	-	53.11	73.30	68.50
GCNet *(eccv,2020)*	InceptionV3	-	-	-	49.06	77.40	68.50
FAM *(ours)*	**InceptionV3**	**70.67**	**81.25**	**87.25**	**55.24**	**77.63**	**68.62**

3.4. Discussions

In this paper, we propose a new perspective for the WSOL task by using FAM. Since CAM-based methods achieve object localization as a by-product of object classification, optimizing classification tends to activate object parts not the whole object, while expanding object parts into the whole object could deteriorate classification performance. Unlike these methods, our FAM utilizes object-aware and part-aware attention modules to perform object localization and classification jointly. Meanwhile, these two tasks can complement each other in a collaborative manner.

4. Experiments

4.1. Experimental Settings

Datasets. We conduct experiments on the most popular benchmarks including CUB-200-2011 [46] and ILSVRC 2016 [10]. The CUB-200-2011 includes 200 categories of birds and contains 11,768 images. The ILSVRC 2016 is a large-scale dataset with 1,000 different classes, consisting of over 1.2 million images of 1,000 categories. Each dataset is divided into three subsets: train-weaksup, train-fullsup and test [62, 6]. Following the protocol in previous methods [63, 7, 41, 43, 60, 61], for both datasets, we train the model with the train-weaksup set and evaluate the performance with the test set. We use the train-fullsup set for the hyperparameter search, since checking the test results with different hyperparameters violates the WSOL protocol.

Evaluation Metrics. Following standard evaluation metrics [7], we use three metrics to evaluate our model. The first metric is **Localization Accuracy**, which measures the ratio of the images with the right class and the bounding box of IoU greater than 50%. The second metric is **GT-known Localization Accuracy**. Unlike the Localization Accuracy, the ground truth class label is given to eliminate the influence caused by classification accuracy when evaluating the localization accuracy. The third metric is **Classification Accuracy**, which represents the ratio of correct classification.

Implementation Details. We implement the proposed algorithm based on four popular backbone networks including VGG16 [42], MobileNetV1 [19], ResNet50 [16] and InceptionV3 [44]. The model is fine-tuned on the pre-trained weights of ILSVRC [10]. The input images are randomly cropped to 224×224 pixels after being resized to 256×256 pixels. Empirically, the weight λ_{spa} for the sparsity loss and the λ_{div} for the diversity loss are set to be 0.04 and 0.01.

4.2. Comparison with State-of-the-art Methods

We compare our method with various recent WSOL methods including CAM [63], HaS [43], ACoL [60], SPG [61], ADL [7], DANet [53], EIL [34], PSOL [58], GCNet [32], RCAM [1] and I^2C [62]. We report the accuracy from the original papers or reproduced works [7, 58, 34]. Table 1 shows the comparison with state-of-the-art methods on CUB-200-2011 test set and ILSVRC 2016 validation set.

Localization Performance: We consistently observe that our FAM outperforms all baseline methods upon all the backbones on both datasets for localization accuracy, especially on CUB-200-2011 dataset. The CUB-200-2011 is a fine-grained dataset which contains many categories of

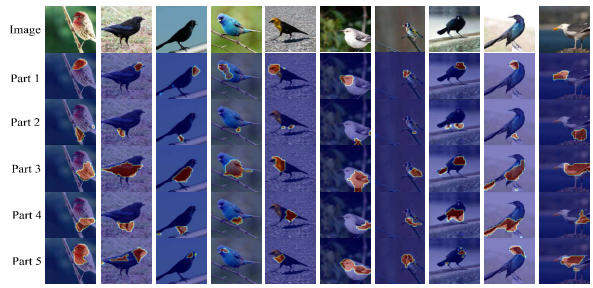

Figure 5. Visualization of the learned part-aware activation maps, which mainly focus on different discriminative object parts.

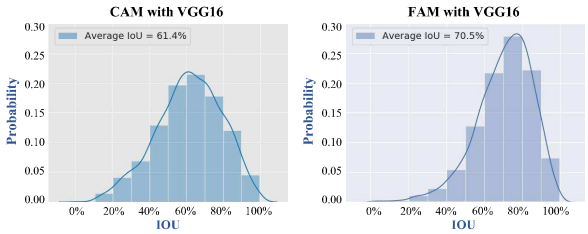

Figure 6. Statistical analysis of the predicted bounding boxes on CUB-200-2011 dataset.

birds, where the intra-class variation is much larger than the inter-class variation. In this case, the image region used to distinguish a certain class may be quite small [58]. Thus, exploiting class-specific image regions for localization would lead to sub-optimal performance. The FAM aims to discover foreground maps of all classes from the background, which can boost localization performance. In specific, FAM-ResNet50 achieves 73.74% and 85.73% accuracy in Top-1 localization and GT-known localization, which exceeds the performance of all other methods by a large margin. Compared to the state-of-the-art method PSOL, FAM-InceptionV3 boosts the Top-1 localization accuracy by 5.16%. On ILSVRC 2016 dataset, which includes a wide variety of classes, FAM-InceptionV3 reports 55.24% and 68.62% accuracy in Top-1 and GT-known localization and sets a new state-of-the-art performance. Besides, FAM-MobilenetV1 obtains 1.46% performance gain over the recent RCAM method. The results show that our method performs well on both fine-grained and large scale datasets, which substantially verifies the effectiveness of our model.

Classification Performance: While some other methods compromise classification accuracy for improving localization, our method achieves the best localization accuracy without damaging the classification accuracy. For example, compared with CAM, HaS-MobileNetV1 reports 0.97% higher Top-1 localization accuracy at the cost of 5.30% classification performance on CUB-200-2011 set, since the random dropout of informative regions would lead to classification degradation. In comparison, FAM-MobileNetV1 obtains the best localization performance, and improves the classification accuracy by 4.44% over the baseline approach CAM, since our FAM avoids the dropout of the important information and exploit discriminative parts for classification. In conclusion, by learning object localization and classification in a collaborative manner, our model achieves significant improvement for object localization as well as maintains remarkable classification accuracy.

Besides, according to the results in Table 1, We also find that the localization maps should not be class-specific. PSOL aims to generate localization maps that are not related to classification labels, which achieves promising localization performance compared with CAM-based methods. Meanwhile, the FAM can also distinguish foreground regions for object localization and consistently performs better than baseline methods on two datasets.

Visualization: Figure 4 visualizes the localization results generated on CUB-200-2011 and ILSVRC 2016 datasets for qualitative evaluation. From the results, we observe that our FAM captures the whole object better than CAM [63]. For example, as shown in Figure 4, the heatmap and bounding box of the right-most sample extracted from CAM only highlight the head of the rabbit, while our method covers nearly the entire area of the rabbit. Besides, in CAM-based methods, the threshold value needs to be tuned manually and carefully so as to extract suitable bounding boxes from the activation maps. Our FAM can distinguish foreground from background with high confidence, so that we can simply set the threshold value to be 0.5 for all datasets. Figure 5 visualizes the learned part-aware activation maps, which are successful in discovering diverse object parts. For example, the 1^{st} part-aware activation map mainly focuses on the head region while the 2^{nd} part-aware activation map mainly focuses on the leg region. This also shows the effectiveness of our proposed part diversity mechanism.

Statistical Analysis: In Figure 6, we show the distribution of the IoU between the predicted bounding boxes and the ground-truth bounding boxes on CUB-200-2011 dataset. Note that the average IoU of CAM-VGG16 is 61.4%. The average IoU of FAM-VGG16 is boosted to 70.5% with a 9.1% performance gain. The comparison of the IoU distribution between CAM-VGG16 and FAM-VGG16 shows that the FAM improves the IoU rates and enhances the quality of the predicted bounding boxes.

4.3. Ablation Studies

To look deeper into the proposed method, we perform a series of ablation studies with VGG16 as the backbone on CUB-200-2011 set, and detailed analyses are as follows.

Effectiveness of the Object-aware Attention Module: This module is designed to discover foreground maps for object localization, and has two components: the sparsity loss and the foreground memory design. Results indicate each design is necessary. (1) Without the sparsity loss, the FAM cannot guarantee that the foreground map is tight and compact to eliminate irrelevant background regions. As

Table 2. Ablation studies about the proposed object-aware attention module on CUB-200-2011 test set.

Sparsity Loss	Foreground Memrory	GT-known	Top-1 Cls
✗	✗	51.81	71.97
✓	✗	73.95	72.36
✓	✓	82.53	72.52

Table 3. Ablation studies about the part-aware attention module on CUB-200-2011 test set.

Diversity Loss	Importance Prediction	Top-1 Cls	GT-known
✗	✗	72.52	82.53
✗	✓	75.63	86.82
✓	✓	77.26	89.26

shown in Table 2, the GT-known localization accuracy increases from 51.81% to 73.95% when the sparsity loss is introduced. As discussed in [54], when optimizing image classification, the model would identify background regions as a class other than the correct class instead of suppressing background regions. In this case, the model may highlight both foreground and background regions for object localization. The sparsity loss is imposed to deal with this problem and boosts the GT-known localization accuracy by 22.14%. (2) As shown in Table 2, with the foreground memory design, the GT-known localization accuracy is further increased to 82.53%. This is because this design can memorize multiple foreground appearances in a given dataset, which can deal with large appearance variations of different objects. Thus, the model has a stronger ability to distinguish foreground regions from background for object localization. Note that we use three 1×1 convolution layers to generate foreground maps with similar parameters when the foreground memory design is removed.

Effectiveness of the Part-aware Attention Module: In this module, whose aim is to select discriminative parts for object classification, there are two designs including the diversity loss and the importance prediction module. (1) As shown in Table 3, with the diversity loss, the classification performance is promoted from 72.52% to 74.13%. The results indicate that the diversity loss can help the model discover different object parts for better classification. (2) Since different parts may have different importance for classification, it is necessary to design an importance prediction module. With the importance prediction module, the classification accuracy is increased by 3.13%.

Relationship between Object Localization and Classification: The results also verify that the two tasks can help each other in a collaborative manner. (1) Object localization can help object classification. With the object-aware attention module, the classification accuracy is boosted from 71.97% to 72.52%, as shown in Table 2. This is because the foreground maps learned by the localization model can help the classifier to avoid background interferences and achieve better performance. (2) Object classification can help object localization. As shown in Figure 4, with only image-level

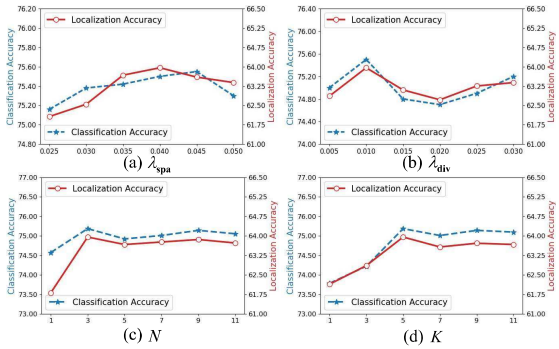

Figure 7. Evaluation of the hyperparameters λ_{spa}, λ_{div} , the number of foreground templates N, and the number of object parts K.

labels, the classifier of FAM can help the localization model to discover foreground maps that highlight nearly the entire object. Besides, better classification results can further improve the localization accuracy. As shown in Table 3, with the part-aware attention module, the localization performance can be improved by 6.73% (82.53% vs. 89.26%).

Hyperparameter Evaluations: We evaluate how λ_{spa} and λ_{div} affect our model learning. Here, λ_{spa} and λ_{div} control the relative importance of the sparsity loss and the diversity loss. As shown in Figure 7, our model achieves much better performance when $\lambda_{spa} = 0.04$, $\lambda_{div} = 0.01$. We then explore the influence of the template number in Figure 7. The best performance is achieved when $N = 3$. Similarly, we explore the influence of the part number in Figure 7. The performance continues to grow until $K = 5$, which means that it is sufficient for classification by mining five parts.

5. Conclusion

In this paper, we propose a new perspective for weakly supervised object localization to optimize object localization and classification jointly by using foreground activation maps. Here, we design an object-aware attention module to effectively highlight foreground objects for object localization and a part-aware attention module to mine discriminative parts for object classification. By jointly learning the two modules in a unified model, the two tasks can help each other. Experiments show the effectiveness.

6. Acknowledgment

This work was partially supported by the National Key Research and Development Program under Grant No. 2018YFB0804204, National Defense Basic Scientific Research Program (JCKY2020903B002), Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050500), National Nature Science Foundation of China (Grant 62022078, 62021001, 62071122), Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant 202000019, and Youth Innovation Promotion Association CAS 2018166.

References

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

[2] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2956–2964, 2015.

[3] Sarath Chandar, Sungjin Ahn, Hugo Larochelle, Pascal Vincent, Gerald Tesauro, and Yoshua Bengio. Hierarchical memory networks. *arXiv preprint arXiv:1605.07427*, 2016.

[4] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2722–2730, 2015.

[5] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[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 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] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[9] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 895–903, 2017.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[14] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[15] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

[16] 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.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[18] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 549–559, 2018.

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

[20] Yan Huang and Liang Wang. Acmm: Aligned cross-modal memory for few-shot image and sentence matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5774–5783, 2019.

[21] Kongzhu Jiang, Tianzhu Zhang, Yongdong Zhang, Feng Wu, and Yong Rui. Self-supervised agent learning for unsupervised cross-domain person re-identification. *IEEE Transactions on Image Processing*, 29:8549–8560, 2020.

[22] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3534–3543, 2017.

[23] Jung Uk Kim, Seong Tae Kim, Eun Sung Kim, Sang-Keun Moon, and Yong Man Ro. Towards high-performance object detection: Task-specific design considering classification and localization separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4317–4321, 2020.

[24] Jung Uk Kim and Yong Man Ro. Attentive layer separation for object classification and object localization in object detection. In *2019 IEEE International Conference on Image Processing*, pages 3995–3999, 2019.

[25] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, pages 1378–1387, 2016.

[26] 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.

[27] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018.

[28] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021.

[29] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[30] Xuxin Lin, Yanyan Liang, Jun Wan, Chi Lin, and Stan Z Li. Region-based context enhanced network for robust multiple face alignment. *IEEE Transactions on Multimedia*, 21(12):3053–3067, 2019.

[31] Yiheng Liu, Wengang Zhou, Jianzhuang Liu, Guo-Jun Qi, Qi Tian, and Houqiang Li. An end-to-end foreground-aware network for person re-identification. *IEEE Transactions on Image Processing*, 30:2060–2071, 2021.

[32] Weizeng Lu, Xi Jia, Weicheng Xie, Linlin Shen, Yicong Zhou, and Jinming Duan. Geometry constrained weakly supervised object localization. *arXiv preprint arXiv:2007.09727*, 2020.

[33] Wang Luo, Tianzhu Zhang, Wenfei Yang, Jingen Liu, Tao Mei, Feng Wu, and Yongdong Zhang. Action unit memory network for weakly supervised temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9969–9979, 2021.

[34] 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 Conference on Computer Vision and Pattern Recognition*, pages 8766–8775, 2020.

[35] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*, 2016.

[36] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 677–685, 2017.

[37] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9226–9235, 2019.

[38] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015.

[39] Jack W Rae, Jonathan J Hunt, Tim Harley, Ivo Danihelka, Andrew Senior, Greg Wayne, Alex Graves, and Timothy P Lillicrap. Scaling memory-augmented neural networks with sparse reads and writes. *arXiv preprint arXiv:1610.09027*, 2016.

[40] 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.

[41] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

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

[43] 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. IEEE, 2017.

[44] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[45] 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.

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

[47] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[48] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[49] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2606–2615, 2017.

[50] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1568–1576, 2017.

[51] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018.

[52] Renliang Weng, Jiwen Lu, and Yap-Peng Tan. Robust point set matching for partial face recognition. *IEEE Transactions on Image Processing*, 25(3):1163–1176, 2016.

[53] 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.

[54] 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.

[55] Tianyu Yang and Antoni B Chan. Learning dynamic memory networks for object tracking. In *Proceedings of the European Conference on Computer Vision*, pages 152–167, 2018.

[56] Wenfei Yang, Tianzhu Zhang, Zhendong Mao, Yongdong Zhanga, Qi Tian, and Feng Wu. Multi-scale structure-aware network for weakly supervised temporal action detection. *IEEE Transactions on Image Processing*, 2021.

[57] Yang Yang, Tianzhu Zhang, Jian Cheng, Zengguang Hou, Prayag Tiwari, Hari Mohan Pandey, et al. Cross-modality paired-images generation and augmentation for rgb-infrared person re-identification. *Neural Networks*, 128:294–304, 2020.

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

[59] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web*, pages 765–774, 2017.

[60] 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.

[61] 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.

[62] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. *arXiv preprint arXiv:2008.05096*, 2020.

[63] 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.