Contrastive Attention Maps for Self-supervised Co-localization

Minsong Ki[1] Youngjung Uh[2,3] Junsuk Choe[4]* Hyeran Byun[1,3]

[1]Department of Computer Science, Yonsei University
[2]Department of Applied Information Engineering, Yonsei University
[3]Department of Artificial Intelligence, Yonsei University
[4]Department of Computer Science and Engineering, Sogang University

Abstract

The goal of unsupervised co-localization is to locate the object in a scene under the assumptions that 1) the dataset consists of only one superclass, e.g., *birds, and 2) there are no human-annotated labels in the dataset. The most recent method achieves impressive co-localization performance by employing self-supervised representation learning approaches such as predicting rotation. In this paper, we introduce a new contrastive objective directly on the attention maps to enhance co-localization performance. Our contrastive loss function exploits rich information of location, which induces the model to activate the extent of the object effectively. In addition, we propose a pixel-wise attention pooling that selectively aggregates the feature map regarding their magnitudes across channels. Our methods are simple and shown effective by extensive qualitative and quantitative evaluation, achieving state-of-the-art co-localization performances by large margins on four datasets: CUB-200-2011, Stanford Cars, FGVC-Aircraft, and Stanford Dogs. Our code will be publicly available online for the research community.*

1. Introduction

Object localization aims to capture the location of the target object in a given image. Over the past decade, deep learning approaches have become mainstream in object localization. These methods typically train a convolutional neural network (CNN) with human-annotated locations in the form of bounding boxes [24, 29, 30]. This has shown great performance but has the downside that the location annotations on all images are too expensive.

To alleviate this, object localization with weaker supervision, such as image-level class labels [43] or dataset-level superclass label [2, 37], has drawn a lot of attention recently. In general, the former is called weakly-supervised object localization (WSOL), whereas we refer to the latter as image co-localization. This paper focuses on the problem of image co-localization, which aims to locate common objects in a dataset consisting of only one superclass.

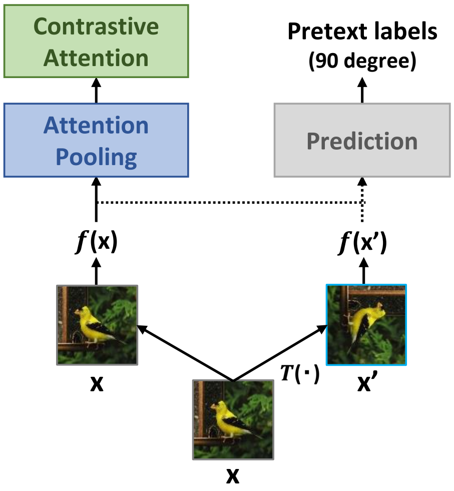

Figure 1: Contrastive learning framework for image co-localization. An encoder embeds two views of one image into feature maps which become attention maps by channel-wise pooling. Then we train the encoder by contrastive objective on the attention maps, which preserve signals from different locations, and by classification objective on the feature maps for the pretext task.

Existing image co-localization methods can be divided into two categories: multiple instance learning (MIL) and self-supervised representation learning (SSL). MIL-based methods [17, 33] firstly generate candidate boxes and then identify if each box contains the target object using handcrafted features. These approaches require high computational costs for inference, making it difficult to operate in real-time.

*Work done as a research scientist at NAVER AI Lab.

On the other hand, the SSL-based methods employ image transformation as a pretext task. If the selected transform is *rotation*, the model learns to predict the amount of rotation applied to the image. Interestingly, the attention map from the learned representation is strongly activated at the location of the target object. The current state-of-the-art method [2] trains the network by cross-entropy for classifying artificial labels from the pretext task.

However, Figure 2 illustrates that the activations smear in the backgrounds, which hamper the co-localization performance. We suppose one of the reasons be the discrepancy between the goals of classification and localization [9, 32]: the classification loss function trains the model to learn the task-relevant information. We believe that the co-localization performance can be further improved by additional location-related information.

Contrastive learning [13] became popular in self-supervised representation learning in recent years. However, it is not straightforward to adapt it into image co-localization because current state-of-the-art methods encode the transformed images into feature vectors by neural networks [3, 14] and the feature vector does not contain spatial information.

To adopt the contrastive learning for image co-localization, we believe that three questions need to be answered: (1) *how do we encode the input image into the embeddings that contain spatial information?* (2) *how do we define positive and negative pairs of embeddings for contrastive learning?* and (3) *which image transformation do we use?*

To this end, we propose a contrastive learning framework (Figure 1) for image co-localization considering these three questions. First, we aggregate the last convolutional feature map of across the channels to generate an attention map which will serve as an embedding for the contrastive framework. It allows the contrastive framework not to lose spatial information. Specifically, we introduce a simple and effective pooling method that chooses the contributed channels separately in each location for computing attention maps. Second, we maximize the similarity between the attention map of the input original image and the inverse transformed attention map of the transformed image, and maximize the dissimilarity between the former and the attention map on the background. It makes the attention map contain the extent of the object more accurately. Last, we explore various image transformations for the positive pairs. Then, we suggest the optimal combination for image co-localization. Overview of our framework is illustrated in Figure 3.

We demonstrate the effectiveness of the proposed method through extensive experiments. Qualitative evaluation results show that our method can localize the full extent of the object and ignore the background. In quantitative evaluation, our method achieves new state-of-the-art localization performances on CUB-200-2011 [34], Stanford Cars [20], FGVC-Aircraft [26], and Stanford Dogs [18].

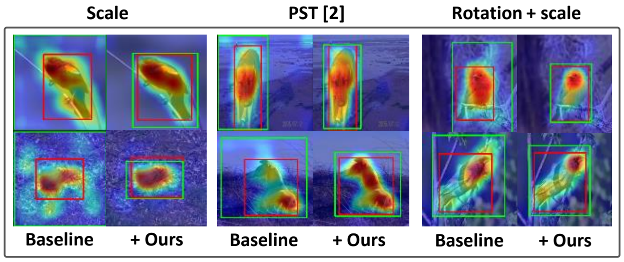

Figure 2: Activation maps extracted from baseline and our model. The baseline method trains the network using only the classification loss. Thus, it activates even on backgrounds to predict the amount of transformations. The red boxes are the ground truth, and the green boxes are the predictions.

In summary, our main contributions are:

- We propose a novel way to adopt a contrastive learning framework for image co-localization. The proposed framework successfully leads the model to learn the full extent of the target object.
- We extensively study how to adopt contrastive learning for image co-localization: (1) the definition of positive and negative pairs, (2) a simple yet effective attention extraction method, and (3) an optimal combination of image transformations.
- Our method achieves new state-of-the-art localization performances with significant margins on four different benchmark datasets. Consistent results are observed through qualitative evaluation.

2. Related Work

Image co-localization aims to discover the common object using only unlabeled positive image sets. Li et al. [22] apply the object detector [12] to generate heat maps by modeling the distribution of confidence scores. However, they require a supervised detector to obtain object candidates, and localization performance depends on the quality of initial proposals.

Several methods reuse CNN pre-trained models to localize target objects. Selective convolutional descriptor aggregation [36] discards the noisy backgrounds and aggregates the remaining deep convolutional descriptors from the pretrained model. Deep descriptor transforming [37, 38] proposes an indicator matrix using principal component analysis that can indicate the correlations of deep descriptors. Co-attention recurrent unit [21] explores all training set to

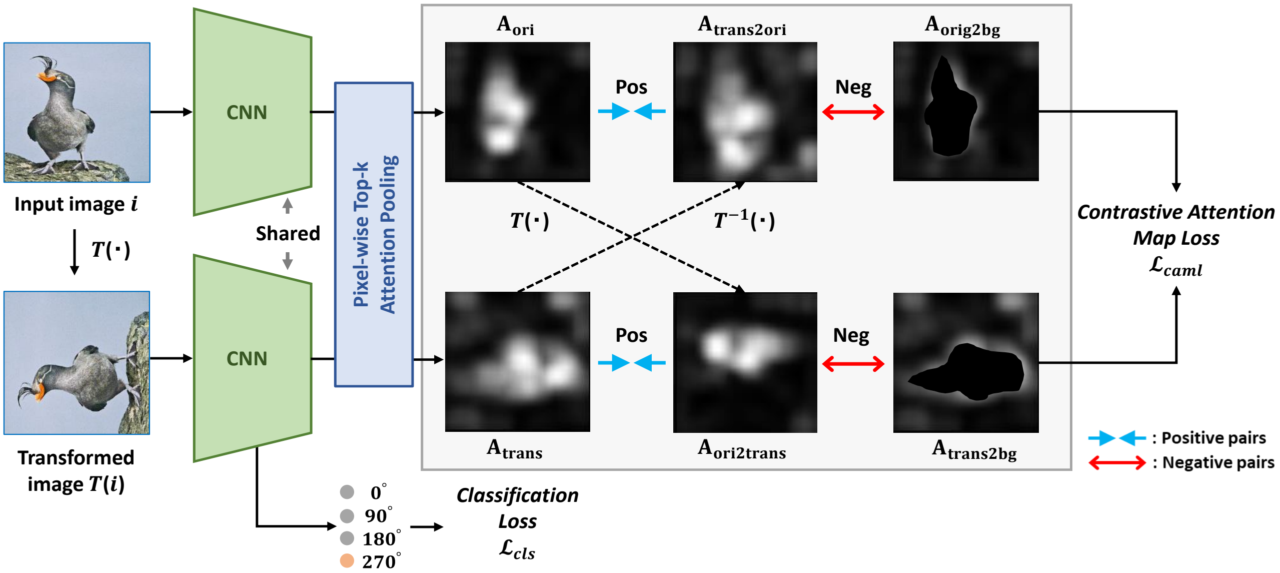

Figure 3: Our proposed self-supervised co-localization network based on contrastive learning. The pixel-wise top-k attention pooling (PTAP) generates the attention maps from original and transformed images. We compute the contrastive attention map loss (CAML) so that the positive pairs are closer to each other and away from the negative pairs. Our model can use various pretext tasks other than the rotation in Figure 3.

learn the valuable group representation. Spatial-semantic modulated deep network [40] trains a mask to coarsely localize the co-object regions that capture the correlations of image features. We also use the CNN features but propose an effective attention pooling that indicates the foreground area of the target object well. Most recently, Baek *et al.* [2] proposed PsyNet that employs self-supervised learning to solve the unsupervised co-localization task. We also focus on self-supervised learning but introduce a novel objective function to localize the entire object more accurately.

Self-supervised learning. Many self-supervised methods have been proposed to learn meaningful feature representation without expensive manual annotations. Doersch et al. [10] extract random pairs of patches in the 3 x 3 grid and predict the relative position between two patches. Context-free network [27] defines a set of jigsaw puzzle permutations to classify the randomly permutated index. Counting [28] exploits image transformations (*e.g.*, scaling and tiling) to estimate the number of visual primitives in the image. RotNet [11] predicts random multiples of 90 degrees from rotated images. We find these methods useful, but we complement their limitations with contrastive learning.

Also, contrastive leaning [13] pulls similar samples closer and pushes different samples away in an embedding feature space. Recent self-supervised approaches [3, 4, 5, 14] consider the different images in the minibatch to minimize the agreement for negative pairs. These methods use the last feature vector of the neural network to calculate the similarity. However, the feature vector does not contain location information; hence it cannot be employed for image co-localization. To address this, we propose contrastive attention maps to compute the similarity.

3. Proposed Method

We first obtain the contrastive attention maps by encoding multiple views of the input image with our pixel-wise top-k attention pooling (§3.1). Then, we compute contrastive attention map loss with the attention maps to train the model (§3.2). The schematic overview of our method is shown in Figure 3.

3.1. Pixel-wise Top-k Attention Pooling

Motivation. A common method to locate objects is to extract class activation maps (CAM) [43] from the last convolutional feature map of a trained classifier [1, 9, 19, 25]. However, it requires class labels to specify the target objects for CAM.

To obtain the attention maps in the image co-localization setting where the target class labels are not given, previous methods use max-pool [39] or average-pool [2]. Especially, [2] finds average-pool effective for extracting attention maps and name it class-agnostic activation mapping (CAAM). We question the choice for max or average and design an alternative that improves co-localization performance.

Our attention pooling. We propose a new attention pooling based on activations' priority which is defined as follows. To obtain the priority, we first employ a channel attention module [35] to assign importance weights for each channel of the feature map. Specifically, we apply a global average pooling (GAP) [23] layer, a 1D convolution, and sigmoid function sequentially on the feature map $\mathbf{F}_x \in R^{C\times H\times W}$. Then, we compute the weighted feature map $\mathbf{F}_w \in R^{C\times H\times W}$ by:

$$\mathbf{F}_w = \mathbf{F}_x \odot \sigma(\text{Conv1D}(\text{GAP}(\mathbf{F}_x))), \tag{1}$$

where Conv1D indicates a 1D convolution and $\odot$ denotes element-wise multiplication with broadcast along spatial dimensions. We treat the activation values of the weighted feature map $\mathbf{F}_w$ as the priority of activations. Finally, with priority, we define the pixel-wise top-k pooled attention map $\mathbf{A}$. Specifically, we collect top-k activations in each location and perform average pooling over the channel dimension:

$$\mathbf{A}(x,y) = \frac{1}{|C|} * \sum_{j\in C} \mathbf{F}_w(x,y,j), \tag{2}$$

where C is a set of selected channel indices regarding their magnitudes across channels. It is worth noting that max and average pooling can be regarded as special cases of our pooling method. This is because, when all the channels are selected, ours becomes average pooling and when $|C| = 1$, ours becomes max pooling.

3.2. Contrastive Attention Map Loss

Contrastive objectives [3, 14] learn representations by maximizing agreement between differently augmented views (positive pair) from the original image. We propose a novel objective function for the co-localization task using this concept of contrastive learning. Figure 3 illustrates the schematic of our method. Predicting rotation is chosen as an example pretext task for brevity here and we will cover other pretext tasks later. We train the model with a classification loss that predicts the degree of rotation and with a contrastive loss function that maximizes the similarity between positive pairs and minimizes that between negative pairs.

To construct the positive and negative pairs, we first obtain two attention maps: $\mathbf{A}_{\text{ori}}$ and $\mathbf{A}_{\text{trans}}$ of original and transformed input images, respectively, by our attention pooling. Then, we generate transformed attention map $\mathbf{A}_{\text{orig2trans}}$ by applying the transformation $\mathbf{T}$ to the attention map $\mathbf{A}_{\text{ori}}$ of the original input image. Next, we apply the inverse transformation $\mathbf{T}^{\text{-1}}$ to the attention map $\mathbf{A}_{\text{trans}}$ of the transformed input image to obtain the inverse transformed attention map $\mathbf{A}_{\text{trans2ori}}$. Finally, the background attention maps ($\mathbf{A}_{\text{ori2bg}}$, $\mathbf{A}_{\text{trans2bg}}$) for negative pairs are computed as below:

$$\begin{aligned} \mathbf{M}_{\text{bg}} &= \mathbb{1}[(1 - \mathbf{A}_{(\cdot)}) > \theta_{\text{bg}}], \\ \mathbf{A}_{(\cdot)\text{2bg}} &= \mathbf{A}_{(\cdot)} \odot \mathbf{M}_{\text{bg}}. \end{aligned} \tag{3}$$

$\mathbb{1}$ denotes a matrix with the same shape with the reverse attention map having ones according to the logical operation. θ_{bg} is a background threshold hyperparameter which set by prefixed ratio of the minimum intensity of the reverse attention map (1-$\mathbf{A}_{(\cdot)}$).

Then, we build two triplets of (anchor, positive sample, negative sample): $(\mathbf{A}_{\text{trans2ori}}, \mathbf{A}_{\text{ori}}, \mathbf{A}_{\text{ori2bg}})$ and $(\mathbf{A}_{\text{ori2trans}}, \mathbf{A}_{\text{trans}}, \mathbf{A}_{\text{trans2bg}})$. Formally, our contrastive attention map loss (CAML) is given by:

$$\begin{aligned} \mathcal{L}_{\text{caml}} = \mathbb{E}_{\mathbf{x}}[[d(\mathbf{A}_{\text{trans2ori}}, \mathbf{A}_{\text{ori}}) - d(\mathbf{A}_{\text{trans2ori}}, \mathbf{A}_{\text{ori2bg}}) + m]_+ + \\ [(d(\mathbf{A}_{\text{ori2trans}}, \mathbf{A}_{\text{trans}}) - d(\mathbf{A}_{\text{ori2trans}}, \mathbf{A}_{\text{trans2bg}}) + m]_+]. \end{aligned} \tag{4}$$

where $[\cdot]_+ = \max(\cdot, 0)$ and $d(\cdot,\cdot)$ denotes L_2 distance in attention map. m indicates the margin.

Our objective encourages consistency between the attention maps before and after the transformation of the input image. Also, it penalizes the attention maps of two anchors being activated in backgrounds.

3.3. Training and Inference

Training. We train our network with the full objective:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{caml} \tag{5}$$

To compute the classification loss, we employ a GAP layer at the end of the network, produce softmax output $\hat{y}$, and then compute the loss given the ground truth label y:

$$\mathcal{L}_{cls} = \text{CrossEntropy}(\hat{y}, y) \tag{6}$$

Inference. We first generate the attention map $\mathbf{A}_{\text{ori}}$ from the original input image by using our PTAP. Next, the final output, the bounding box, is obtained from the attention map. Note that we follow the common practice [2, 8, 43] to extract a bounding box from an attention map.

4. Experiments

Datasets. We evaluate the proposed method on four benchmarks: CUB-200-2011 [34], Stanford Cars [20], FGVC-Aircraft [26], and Stanford Dogs [18]. Each dataset is divided into two subsets: `train` and `test`. The `train` sets include only images without labels for training. The CUB dataset is also commonly used in WSOL, and we perform comparative experiments with recent WSOL state-of-the-art methods.

Evaluation metrics. Following the common practice [2, 36, 38, 41], we evaluate our method in terms of CorLoc$^{\text{IoU=0.5}}$, where the predicted box is considered as correct if an intersection over union (IoU) exceeds 50%. We also compute Mean that averages accuracies at three IoU criterions $\in \{0.3, 0.5, 0.7\}$ to address diverse demands for localization fineness. In addition, MaxBoxAccV2 [8] is also measured for comparison with WSOL methods. MaxBoxAccV2 measures the ratio of the samples with the correct box, while the correctness is defined by an IoU criterion 0.5 at the optimal activation threshold.

Implementation details. We build the proposed method upon two CNN backbones: VGG16 [31], SE-ResNet50 [15, 16]. We insert our PTAP at the last convolution layer for each backbone during the both training and inference phases. We define two hyperparameters: k for PTAP and θ_{bg} for background thresholding. The background threshold is set to minimum intensity of $\mathbf{A}_{\text{PTAP}}$ times pre-defined ratio θ_{bg}. We use 2.3 for θ_{bg} regardless of the dataset. We set the batch size to 64 and margin m to 1. The initial learning rate and the momentum of the SGD optimizer are set to 0.001 and 0.9, respectively. We also begin by loading ImageNet pre-trained weights for comparison with existing co-localization works [2, 36, 38, 41] and then fine-tune the network. Our model is implemented using PyTorch and trained using two NVIDIA GeForce RTX 2080 Ti GPUs for approximately three hours.

4.1. Comparison with state-of-the-art methods

In Table 1 and Table 2, we compare the proposed method with the weakly-supervised methods and unsupervised co-localization in terms of CorLoc$^{\text{IoUs=0.5}}$ and MaxBoxAccV2, respectively. We also show the upper bound accuracy based on few-shot learning [8]. It uses only a few fully labeled samples per class at training. The CUB dataset is commonly used in WSOL tasks, and three datasets (CUB, Cars, Aircraft) are traditionally used in unsupervised co-localization tasks. Additionally, we also compare with recent state-of-the-art works [2, 36, 41] on the Stanford Dogs dataset [18].

Our method achieves state-of-the-art performance both CorLoc$^{\text{IoUs=0.5}}$ on the four benchmarks and MaxBoxAccV2 [8] on the CUB dataset. In Table 1, we observe that our method achieves state-of-the-art localization performance. We believe that this is particularly impressive because our method does not need image-level class labels.

In Table 2, we observe that the proposed method has achieved the best performance on all four datasets. On the CUB and Stanford Dogs datasets, our method increases performance by 2% and 9% compared to PsyNet [2], respectively. Our improvements on Stanford Cars and Aircraft are not significant as those of other datasets, because the performances on Stanford Cars and Aircraft benchmarks have already been saturated. We also measure the performance on the SE-ResNet50 [15, 16] backbone for the comparison with PsyNet [2]. The performances of our method on all datasets outperform those of PsyNet in terms of CorLoc$^{\text{IoUs=0.5}}$.

Table 1: MaxBoxAccV2 [8] comparisons with the WSOL state-of-the-art methods on the CUB dataset. The values are taken from [8] except for MEIL [25]. The result of MEIL is reproduced with the officially provided checkpoint.

	Method	MaxBoxAccV2@IoU (%)			
		0.3	0.5	0.7	Mean
VGG16	Few-shot [8]	-	86.30	-	-
	CAM [43]	96.77	73.14	21.23	63.72
	ACoL [42]	93.77	63.20	15.17	57.38
	ADL [9]	**97.72**	78.06	23.04	66.28
	MEIL [25]	96.19	70.99	18.38	61.85
	InCA [19]	96.20	77.20	26.75	66.72
	Ours	96.42	**84.15**	**50.60**	**77.06**

4.2. Ablation study

The ablation studies for the proposed components are performed with VGG16 [31] on CUB-200-2011 [34]. We validate the necessity of each proposed element over the baseline.

Necessity of the proposed components. We propose two components to locate the correct extent of the target object. Table 3 shows the effectiveness of each proposed element on the baseline. The baseline utilizes a rotation pretext task that estimates four rotation angles by using only the classification loss $\mathcal{L}_{\text{cls}}$.

We observe that each component of our method plays an important role in improving co-localization performance. Specifically, ours without the $\mathcal{L}_{\text{caml}}$ achieves 3.76% lower performance than the full setting. PTAP also improves the performance by 3.70%. Excluding the $\mathcal{L}_{\text{cls}}$ in our full setting, the performance degrades by 2.46%, and the degradation is the smallest compared to the two elements. It means that $\mathcal{L}_{\text{cls}}$ has the lowest contribution to the performance improvement of our method. It is worthy to note that the best performance is achieved when all components are employed. Interestingly, our method especially boosts the accuracy of CorLoc$^{\text{IoU=0.5}}$ and CorLoc$^{\text{IoU=0.7}}$. This indicates that our proposed method induces the model to learn the extent of the object more effectively than the baseline method.

Superiority of the contrastive attention over contrastive feature map. In general, contrastive learning [3, 5, 6, 14] compares similarity by embedding the samples as feature representations. Following the original way of contrastive learning, we compare the performance for two cases using

Table 2: $\texttt{CorLoc}^{\texttt{IoU=0.5}}$ comparisons with the co-localization state-of-the-art methods on four benchmarks. SE: SE-Res50 [15, 16]. The values are taken from their respective papers.

	Method	CUB-200-2011	Stanford Cars	FGVC-Aircraft	Stanford Dogs
VGG16	Part-based [7]	69.37	93.05	42.91	36.23
	SCDA [36]	76.79	90.96	94.91	78.76
	DDT [38]	82.26	71.33	92.53	-
	OLM [41]	80.45	92.51	94.94	80.70
	PsyNet [2]	83.78	96.61	95.59	73.71
	Ours	**85.88**	**97.26**	**96.60**	**82.82**
SE	PsyNet [2]	85.10	98.81	97.81	77.84
	Ours	**85.93**	**98.95**	**98.75**	**80.32**

Table 3: The ablation study of the main configurations of our method in terms of $\texttt{CorLoc}^{\texttt{IoUs}}$ for rotation task. $\mathcal{L}_{\text{cls}}$: classification loss. $\mathcal{L}_{\text{caml}}$: contrastive attention map loss (CAML). $\mathbf{A}_{\text{PTAP}}$: pixel-wise top-k attention pooling (PTAP).

Method	Dataset: CUB			$\texttt{CorLoc}^{\texttt{IoUs}}$			
	$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{caml}}$	$\mathbf{A}_{\text{PTAP}}$	0.3	0.5	0.7	Mean
Baseline	✓	✗	✗	96.75	77.21	28.66	67.54
Ours	✓	✓	✓	97.30	83.65	38.64	**73.20**
$-\mathcal{L}_{\text{cls}}$	✗	✓	✓	94.73	77.25	40.24	70.74
$-\mathcal{L}_{\text{caml}}$	✓	✗	✓	96.65	78.97	32.70	69.44
$-\mathbf{A}_{\text{PTAP}}$	✓	✓	✗	96.73	79.49	32.27	69.50

Table 4: $\texttt{CorLoc}^{\texttt{IoUs}}$ comparison according to the feature maps or attention maps used to calculate our contrastive attention map loss $\mathcal{L}_{\text{caml}}$.

Method: rotation	$\texttt{CorLoc}^{\texttt{IoUs}}$			
	0.3	0.5	0.7	Mean
$\mathcal{L}_{\text{caml}}$ w/ feature maps	95.82	76.97	30.96	67.92
$\mathcal{L}_{\text{caml}}$ w/ attention maps	**97.30**	**83.65**	**38.64**	**73.20**

feature representations or attention maps to calculate our $\mathcal{L}_{\text{caml}}$ (Table 4). Both cases train our model using Equation (4), but generating positive and negative pairs is different. $\mathcal{L}_{\text{caml}}$ with feature representations multiplies the attention map generated using our PTAP with the weighted feature map $\mathbf{F}_w$ (element-wise multiplication over the channel dimension). Specifically, $\mathcal{L}_{\text{caml}}$ with attended features achieves 5.28% lower performance than $\mathcal{L}_{\text{caml}}$ with attention maps. Consequently, the proposed method achieves better performance when operating on attention maps rather than feature representations.

Necessity of both positive and negative pairs. Recent contrastive learning work [6] proposes a simple siamese network that maximizes the similarity only between positive pairs without negative pairs. We note that [6] is not proposed for image co-localization. However, we believe that the insight of them needs to be considered as a possible design choice for co-localization. Therefore, we validate the idea of only using positive pairs or negative pairs in terms of co-localization.

Table 5: $\texttt{CorLoc}^{\texttt{IoUs}}$ on the proposed method according to the negative sampling or not. We train our network by calculating the L2 loss between positive pairs in the case of w/o neg pairs (first row). Ours w/o pos pairs (second row) train the network so that the negative pairs are separated from each other. The margins used in the second and third rows are applied equally.

Method: rotation	$\texttt{CorLoc}^{\texttt{IoUs}}$			
	0.3	0.5	0.7	Mean
Ours w/o neg pairs	96.94	80.60	32.94	70.16
Ours w/o pos pairs	96.65	78.11	28.99	67.92
Ours	**97.30**	**83.65**	**38.64**	**73.20**

In the Table 5, we compare the performance according to the composition of pairs to calculate our loss using PTAP. We observe that using negative and positive sampling together plays an important role in improving performance. The performance particularly boosts in terms of $\texttt{CorLoc}^{\texttt{IoU=0.7}}$; thus, the model covers the region of the target object more accurately.

Effect of pixel-wise top-k sampling. In Table 6, we also measure $\texttt{CorLoc}^{\texttt{IoUs}}$ according to how we pool the attention map. Ours with the max pooling uses only the maximum value from one location; thus, one exceptionally high activation may overwhelm the entire attention, leading to an exaggerated focus on the most discriminative part. Using the top-70% pooling improves the mean accuracy by 20% compared to the bottom-30% pooling. Figure 4 is the visualization of generating final activation maps from our model with averaging all channels, top-70% channels, and bottom-30% channels. Bottom-30% pooling with relatively low weights activates mainly around objects or background

Table 6: $\texttt{CorLoc}^{\texttt{IoUs}}$ comparisons of PTAP for rotation task. The bottom-30% attention values of the pixel unit are the main factor in the performance drop due to the backgrounds. Excluding them to generate attention maps shows better performance.

Method: rotation	$\texttt{CorLoc}^{\texttt{IoUs}}$			
	0.3	0.5	0.7	Mean
Ours w/ bottom-30%	89.04	57.90	17.74	54.89
Ours w/ max	91.09	62.35	20.60	58.01
Ours w/ top-70%	**97.30**	**83.65**	**38.64**	**73.20**
Ours w/ average	96.30	81.89	34.65	70.94

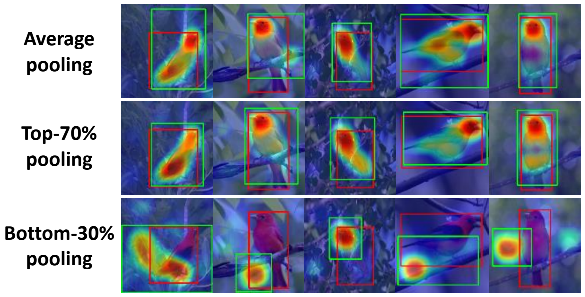

Figure 4: Qualitative comparisons of activation map and localization results according to the pooling methods of our PTAP (average, top-70%, and bottom-30%). The red boxes are the ground-truth and the green boxes are the predictions.

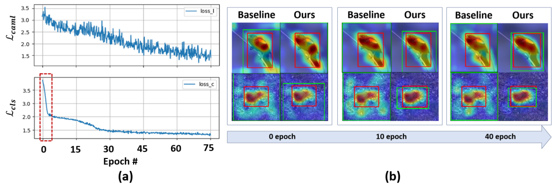

Figure 5: (a) **Loss curve** ($\mathcal{L}_{\text{caml}}$ and $\mathcal{L}_{\text{cls}}$), (b) **Activation maps** from the baseline and ours as training proceeds.

areas. Average pooling tends to suppress the target object area where strong activation appears due to a channel with low reliability, or rather, the background contributes to the generation of an attention map. On the other hand, the top-70% pooling method can cover the extent of the object more accurately. In summary, we confirm that our design choice of PTAP is effective to improve localization performance effectively by both qualitative and quantitative evaluations.

Choice of image transformations. We measure the $\texttt{CorLoc}^{\texttt{IoU=0.5}}$ on various image transformations. Specifically, For that, six different transformations are chosen, which are rotation [11], scale, translation, horizontal flip (Hflip), vertical flip (Vflip), and recently proposed PST [2].

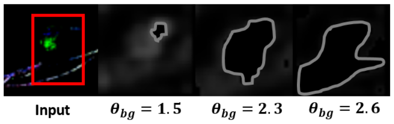

Figure 6: $\mathbf{M}_{\mathbf{bg}}$ regarding θ_{bg}. red box: ground truth.

Table 7 shows the experimental results. Note that the baseline method trains the network using only the classification loss.

On all six pretext tasks, we show that our method improves localization performance compared to baseline across all the datasets. Especially, there are large performance gains in scale and translation tasks where undefined regions occur after transformation. We fill the undefined regions with reflection (reflected remaining part of the image) and zeros for scale and translation tasks, respectively.

We observe that our method consistently improves localization performance overall image transformations. Among them, our method achieves the highest performance in the rotation prediction task [11], except for the Stanford Dogs dataset. To successfully predict the rotation of an image, the RotNet [11] encourages to accomplish the rotation prediction task it learns to focus on high-level object parts (*e.g.*, eyes, tails, heads, *etc*.). Learning the feature representation with these properties helps to localize the target object.

Early behavior analysis. Figure 5a plots the two losses. The rotation classification loss $\mathcal{L}_{\text{cls}}$ quickly drops in the early phase (red box) and our contrastive attention map loss $\mathcal{L}_{\text{caml}}$ slowly follows. The early training dynamics neither heavily fluctuates nor saturates. Qualitatively, the attention maps start from rough coverage and gradually fit to the object extent while the baseline barely changes (Figure 5b).

Performance variation regarding θ_{bg}. Lower θ_{bg} leaves more foreground attention on the negative counterpart, shrinking the attention map to be dissimilar from it (Figure 6). According to our experimental results on the rotation task (same setting with Table 6), we observe that our hyperparameter is quite robust in that our method still surpasses the baseline in a large margin with $1.5 \leq \theta_{\text{bg}} \leq 2.6$.

4.3. Qualitative results

Figure 7 illustrates attention maps and estimated bounding boxes from PsyNet [2] and ours. Since PsyNet [2] learns using only classification loss and has no background constraints, it is challenging to cover less discriminative regions (*e.g.*, leg, beak, tail) accurately. In contrast, our method constrains the background region during training, so it can cover the full extent of the object effectively. More examples are available in the supplemental material. Unfortunately, some examples in Figure 8 degenerate when the objects are under the shadow, overlapping, unclear visual clue, or mirror image. Part of the background is activated due to

Table 7: $\texttt{CorLoc}^{\texttt{IoU=0.5}}$ comparisons with the base task (VGG16) on each dataset. Baseline: learns using only classification loss (*e.g.*, rotation [11]: 0°, 90°, 180°, 270°). The results of PST [2] are reproduced using official code. **Best entries**: boldface + underline. Second-best entries: underline.

Task	CUB-200-2011		Stanford Cars		FGVC-Aircraft		Stanford Dogs	
	Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours
Rotation [11]	77.21	**83.65** (+6.4)	88.69	**97.41** (+8.7)	91.86	**97.14** (+5.3)	77.79	82.03 (+4.2)
Scale	44.66	83.62 (+39.0)	77.41	91.69 (+14.3)	65.70	96.69 (+31.0)	67.30	82.63 (+15.3)
Translation	22.98	83.05 (+60.1)	61.37	97.40 (+36.0)	47.13	96.36 (+49.2)	58.05	66.72 (+8.7)
Hflip	73.76	75.12 (+1.4)	91.45	96.81 (+5.4)	87.15	94.83 (+7.7)	78.18	80.88 (+2.7)
Vflip	75.94	80.51 (+4.6)	91.96	96.08 (+4.1)	92.85	94.77 (+1.9)	77.07	80.44 (+3.4)
PST [2]	48.92	76.80 (+27.9)	92.89	95.68 (+2.8)	92.41	96.51 (+4.1)	67.91	**86.64** (+18.7)

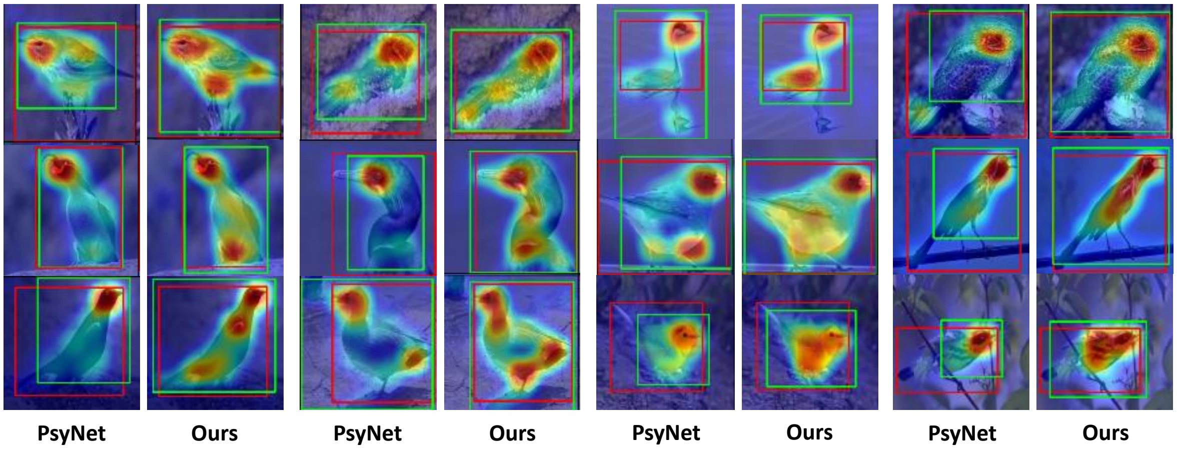

Figure 7: Activation maps and localization outputs of PsyNet [2] and ours on the CUB dataset. The red boxes are the ground-truth and the green boxes are the predictions.

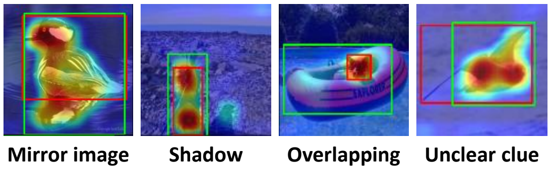

Figure 8: Some degenerated localization outputs on the CUB dataset. A mirror image means that an object is reflected in the water.

the factors mentioned above, and as a result, the model outputs a bounding box larger than the ground truth box.

5. Conclusion

In this paper, we propose a novel method for improving image co-localization performance based on a contrastive learning framework. To encode the input image for our contrastive framework, we propose a pixel-wise top-k attention pooling (PTAP) method that utilizes only the important channels in the feature map. Then, we build positive and negative pairs with the encoded images, where we call them *contrastive attention maps*. Finally, the proposed contrastive attention map loss (CAML) encourages consistency between the contrastive attention maps and also penalizes the background regions during the training phase. In this way, our method can induce the model to learn the full extent of the object accurately. Based on the extensive evaluation, we confirm that the proposed method is effective for improving co-localization performance. Specifically, our method achieves state-of-the-art performance on CUB-200-2011, Stanford Cars, FGVC-Aircraft, and Stanford Dogs datasets.

Acknowledgements. This work was supported by the National Research Foundation of Korea grant funded by Korean government (No. NRF-2019R1A2C2003760) and Artificial Intelligence Graduate School Program (YONSEI UNIVERSITY) under Grant 2020-0-01361.

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 (ECCV)*, 2020. 3

[2] Kyungjune Baek, Minhyun Lee, and Hyunjung Shim. Psynet: Self-supervised approach to object localization using point symmetric transformation. In *AAAI*, pages 10451–10459, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020. 2, 3, 4, 5

[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 3

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3, 5

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 5, 6

[7] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1201–1210, 2015. 6

[8] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3133–3142, 2020. 4, 5

[9] 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. 2, 3, 5

[10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 3

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 3, 7, 8

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[13] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2, 3

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 3, 4, 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 6

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5, 6

[17] Armand Joulin, Kevin Tang, and Li Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision*, pages 253–268. Springer, 2014. 1

[18] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, 2011. 2, 4, 5

[19] Minsong Ki, Youngjung Uh, Wonyoung Lee, and Hyeran Byun. In-sample contrastive learning and consistent attention for weakly supervised object localization. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3, 5

[20] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 2, 4

[21] Bo Li, Zhengxing Sun, Qian Li, Yunjie Wu, and Anqi Hu. Group-wise deep object co-segmentation with co-attention recurrent neural network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8519–8528, 2019. 2

[22] Yao Li, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Image co-localization by mimicking a good detector's confidence score distribution. In *European Conference on Computer Vision*, pages 19–34. Springer, 2016. 2

[23] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 4

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg. SSD: Single shot multibox detector. In *ECCV*, pages 21–37, 2016. 1

[25] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8766–8775, 2020. 3, 5

[26] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2, 4

[27] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 3

[28] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5898–5906, 2017. 3

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-Time object detection. In *CVPR*, pages 779–788, 2016. 1

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-Time object detection with region proposal net-

works. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 1

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[32] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pages 3544–3553. IEEE, 2017. 2

[33] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1464–1471, 2014. 1

[34] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 4, 5

[35] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11534–11542, 2020. 4

[36] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 26(6):2868–2881, 2017. 2, 5, 6

[37] Xiu-Shen Wei, Chen-Lin Zhang, Yao Li, Chen-Wei Xie, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Deep descriptor transforming for image co-localization. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017. 1, 2

[38] Xiu-Shen Wei, Chen-Lin Zhang, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition*, 88:113–126, 2019. 2, 5, 6

[39] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 3

[40] Kaihua Zhang, Jin Chen, Bo Liu, and Qingshan Liu. Deep object co-segmentation via spatial-semantic network modulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12813–12820, 2020. 3

[41] Runsheng Zhang, Yaping Huang, Mengyang Pu, Jian Zhang, Qingji Guan, Qi Zou, and Haibin Ling. Object discovery from a single unlabeled image by mining frequent itemsets with multi-scale features. *IEEE Transactions on Image Processing*, 29:8606–8621, 2020. 5, 6

[42] 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. 5

[43] 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. 1, 3, 4, 5