PsyNet: Self-Supervised Approach to Object Localization Using Point Symmetric Transformation

Kyungjune Baek,* Minhyun Lee,* Hyunjung Shim
School of Integrated Technology, Yonsei University, Incheon, South Korea
{bkjbkj12, lmh315, kateshim}@yonsei.ac.kr

Abstract

Existing co-localization techniques significantly lose performance over weakly or fully supervised methods in accuracy and inference time. In this paper, we overcome common drawbacks of co-localization techniques by utilizing self-supervised learning approach. The major technical contributions of the proposed method are two-fold. 1) We devise a new geometric transformation, namely point symmetric transformation and utilize its parameters as an artificial label for self-supervised learning. This new transformation can also play the role of region-drop based regularization. 2) We suggest a heat map extraction method for computing the heat map from the network trained by self-supervision, namely class-agnostic activation mapping. It is done by computing the spatial attention map. Based on extensive evaluations, we observe that the proposed method records new state-of-the-art performance in three fine-grained datasets for unsupervised object localization. Moreover, we show that the idea of the proposed method can be adopted in a modified manner to solve the weakly supervised object localization task. As a result, we outperform the current state-of-the-art technique in weakly supervised object localization by a significant gap.

1 Introduction

Object localization and detection aim to identify and locate the object of interest within an image. Recently, deep convolutional neural networks (CNNs) trained with the large scale annotated datasets achieve remarkable performance in object localization and detection task (Redmon et al. 2016; Liu et al. 2016). These models adopt fully supervised learning, thus require pixel-level annotations such as bounding box or segmentation mask. Unfortunately, pixel-level annotations are quite expensive and time-consuming to acquire, thereby often unavailable in practice. To relax the need for pixel-level annotations, several recent works (Zhou et al. 2016; Zhang et al. 2018; Choe and Shim 2019) have developed weakly supervised object localization (WSOL) techniques, which only utilize image-level labels.

Moving toward much weaker supervision, several recent studies suggest solving a co-localization problem, which only has an image collection of a single class object without any assumptions. Existing co-localization techniques extract hand-crafted features or learned features, obtain many object proposals from the features, and then refine them for localization (Tang et al. 2014; Joulin, Tang, and Fei-Fei 2014; Wei et al. 2019). However, these techniques commonly have significant drawbacks in accuracy compared to fully or weakly supervised techniques. Also, they are computationally very expensive because of generating region proposals.

Figure 1: Demonstrating our point symmetric transformation (PST). Left: original input image, middle: result of applying large translation (*i.e.* the half of image resolution) to the direction A as depicted in Figure 3 with zero padding, right: result of PST (*i.e.* applying large translation with reflection padding).

In this paper, we focus on tackling unsupervised object localization through co-localization. The goal of our method is 1) to improve the co-localization accuracy and 2) to achieve the computational efficiency for real-time performance. To this end, we utilize the CNN classifier trained with self-supervised learning approach. The proposed technique is inspired by the fact that existing WSOL techniques significantly improve the localization accuracy and inference speed by analyzing the discriminative features extracted from CNN classifiers. As a natural extension, we follow the design principle of Class Activation Mapping (CAM) (Zhou et al. 2016), one of the pioneer WSOL techniques, which mines and tracks the interesting object from the feature of input images. Whereas WSOL techniques utilize image-level labels to train the CNN classifier, our method uses no prior knowledge (*e.g.* labels) to train the CNN. Instead, we employ the self-supervised learning paradigm, one of the most

*indicates equal contribution.

Figure 2: Effects of point symmetric transformation under various settings. The leftmost is the original input image. Each image from the second-left to the rightmost is the results of applying PST as shown in Figure 1, where translation parameter $t_x(= t_y)$ varies from 1.0 to 3.0, respectively.

powerful training schemes in unsupervised learning.

The key idea of self-supervised learning is to generate artificial labels of data for providing surrogate supervision. Therefore, the network model learns to extract meaningful representations for predicting those labels. Among various self-supervised techniques, our method is particularly motivated by (Gidaris, Singh, and Komodakis 2018; Zhang et al. 2019a) because of its impressive performance in a classification task. Similar to RotNet (Gidaris, Singh, and Komodakis 2018) but specialized for localization problem, we devise a new geometric transformation, namely point symmetric transformation (PST), and train the network model for predicting artificial labels (*i.e.* parameters of the geometric transformation). We highlight that PST is designed to implicitly perform the regional dropout by erasing the large portions of objects during transformation. Because the regional dropout is a well-known technique for improving the localization, our PST is a particularly effective transformation for self-supervising the localization task. Afterward, similar to CAM, we track the features that contribute the most for the prediction and localize their regions in the image. However, because co-localization does not involve to predict class labels, unlike CAM, we cannot track the class activations. Instead, we extract class-agnostic activations by aggregating the contributions from all parameters for generating the heat map. We call this method class-agnostic activation mapping (CAAM), and it is simply done by computing the spatial attention of feature map. By combining self-supervised learning approach and CAAM, we construct PsyNet, a novel and effective solution for unsupervised object localization. Moreover, the proposed model with a simple modification can be adopted to WSOL problem and also substantially improves the accuracy. The major contributions of this study are summarized as follows.

- On three benchmarks, we achieve substantial improvements over existing state-of-the-arts in unsupervised object localization. More surprisingly, PsyNet outperforms the state-of-the-art weakly supervised techniques, in terms of the GT-known localization accuracy.
- To the best of our knowledge, we are the first to introduce self-supervised learning approach in the problem of unsupervised object localization. Beyond simple adaptation, we introduce PST, a transformation for addressing the common issue in object localization.
- We show that our CAAM is effective to generate heat map from the CNN trained with self-supervised learning.
- PsyNet with a simple modification is also an effective solution for weakly supervised object localization. As a result, we achieve the new state-of-the-art performance in weakly supervised object localization.

The source code for this work is available at https://github.com/FriedRonaldo/PsyNet.

2 Related Work

2.1 Self-supervised Learning

Self-supervised representation learning is one of the most powerful training schemes in unsupervised learning. These self-supervised learning methods usually utilize artificial labels, which can be easily generated from the data. Then, meaningful feature representations are learned while the network is trained with artificial labels. For example, in (Dosovitskiy et al. 2014), various transformations are applied to each image, and convolutional neural networks (CNNs) are trained to classify the label of transformed images. (Doersch, Gupta, and Efros 2015) predict the relative position of two randomly given patches in an image. Other than that, several studies train CNNs to solve Jigsaw puzzles (Noroozi and Favaro 2016) or to colorize gray-scale image (Larsson, Maire, and Shakhnarovich 2016).

Recently, (Gidaris, Singh, and Komodakis 2018) propose to train CNNs by identifying the rotational transformation applied to the input images. Interestingly, with this simple transformation, they reported an impressive performance on an image classification task. We are motivated by the success of (Gidaris, Singh, and Komodakis 2018) and devise a geometric transformation for extracting meaningful representations without the annotations.

2.2 Object Localization with Weak Supervision

Fully supervised object localization techniques need expensive annotations (*i.e.* pixel-level annotations) to train the CNN classifier. WSOL addresses this issue with weak supervision (*i.e.* an image label). The most representative technique is class activation mapping (CAM), which utilizes the CNN classifier to mine and track the most discriminative features of the target object. CAM is used to extract a heat map from most of WSOL techniques (Zhou et al. 2016; Zhang et al. 2018; Singh and Lee 2017; Choe and Shim 2019).

Because WSOL techniques rely on the classifier, they commonly face the issue that the network only focuses on the most discriminative part of the object. To address this problem, (Singh and Lee 2017) propose Hide-and-Seek,

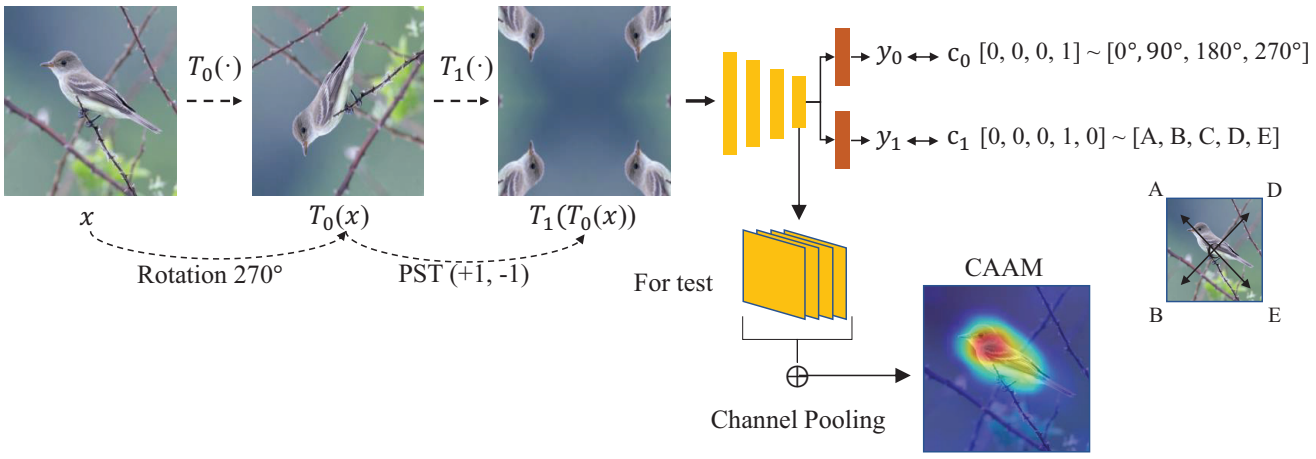

Figure 3: Overview of training procedure of the proposed method. $T_0(\cdot)$ and $T_1(\cdot)$: rotation by 270° and point symmetric transformation by $(+1, -1)$, D in this example, respectively, y_i:predicted label of each transformation $T_i(\cdot)$, c_i: ground truth label of each transformation.

which randomly drops patches of the input image to enforce the network for covering the integral object. Attention-based Dropout Layer (ADL) (Choe and Shim 2019) utilizes an attention map to learn the integral extent of the object.

Co-localization tackles the task by utilizing much weaker supervision, a set of images containing a single class object. For example, (Tang et al. 2014) perform co-localization in real-world settings by utilizing SIFT (Lowe 2004). (Li et al. 2016) devise the pre-trained fully connected network as a feature extractor to conduct co-localization. (Wei et al. 2019) localize objects by evaluating the correlations between descriptors of the samples extracted by the pre-trained model.

3 Self-supervised Object Localization

To enjoy the strong performance gain reported in recent unsupervised representation learning algorithms (Noroozi and Favaro 2016; Gidaris, Singh, and Komodakis 2018), we employ self-supervised learning approach for unsupervised object localization. As with (Gidaris, Singh, and Komodakis 2018; Zhang et al. 2019a; Feng, Xu, and Tao 2019), we design a network model that predicts geometric transformations applied to a single image. Once the transformation parameters are determined, they serve as labels, and the image-label pairs can be arbitrarily generated. Utilizing those of artificial labels as surrogate supervision, it is possible to lead the network model to learn meaningful representation.

Compared to existing self-supervised techniques (Gidaris, Singh, and Komodakis 2018; Zhang et al. 2019a), the proposed method is novel in two aspects. 1) We propose a novel geometric transformation, namely point symmetric transformation (PST), for self-supervised learning. This new transformation is specialized to address the common issue in a fine-grained object localization task. 2) We introduce class-agnostic activation mapping (CAAM), a new and effective method of generating the heat map from self-supervised network model. State-of-the-art techniques in weakly-supervised object localization utilize class activation mapping (CAM) (Zhou et al. 2016) to obtain a heat map from the network model. CAAM is inspired by CAM but modified to handle the model trained by self-supervised learning. The details of the proposed method will be presented in the following sections.

3.1 Point Symmetric Transformation

To alleviate the overfitting problem of neural networks, various studies (Singh and Lee 2017; DeVries and Taylor 2017; Choe and Shim 2019) have proposed model regularization techniques through regional dropout. These techniques drop several regions of the input image or feature map at random or intentionally during training. Because they tend to hide the important image regions (*i.e.* the most discriminative parts) during training, regional dropout techniques prevent the network from focusing only on the most discriminative parts. Instead, the model is guided to learn the less discriminative part as all. By applying the regional dropout techniques, (DeVries and Taylor 2017) achieve the meaningful gain in the classification accuracy while (Choe and Shim 2019) significantly improve the localization accuracy.

The problem of focusing only on the most discriminative parts is particularly more serious in the fine-grained object localization. For example, to distinguish *red-winged blackbird* from the *rusty blackbird*, the network should focus on the color of wings to improve the classification accuracy. This clearly degrades the accuracy of object localization that should cover the integral extent of objects. Therefore, we should recognize such common challenges of the fine-grained object localization and design the self-supervised technique to address this problem effectively. To this end, we propose a point symmetric transformation (PST) for self-supervised learning, which induces the effect of the regional dropout.

PST is defined as follows. First of all, a large translation (*i.e.* greater than the half of image resolution) is applied to the input image. Afterward, an undefined region after shifting is filled with reflection padding. We visualize PST in

Table 1: Effects of different heat map extraction methods and different geometric transformations. Transformations include rotation (R) by [0°, 90°, 180°, 270°], point symmetric transformation (S) and scaling (C) by [0.7, 1.0, 1.3]. In the following experiments, we use this notation.

	Average (Multi-CAM)	Max (Multi-CAM)	Spatial Attention
R(·)	60.82	60.82	**80.51**
S(·)	36.35	36.35	**66.98**
S(R(·))	26.34	63.84	**83.78**
S(R(C(·)))	21.87	42.27	**58.66**

Figure 1. As seen from this figure, PST effectively erases the large portions of the target object, but only leaves the small square regions of the target object. Consequently, applying PST can be interpreted as a transformation of conducting translation and the regional dropout simultaneously.

Moreover, we claim that our PST smartly handles the issue of how to set the value of undefined regions. As also discussed by (Singh and Lee 2017), regional dropout techniques are applied only during training, not testing stage. Then, the statistic of the activations of convolutional layers for training data does not match that of test data because the value of undefined regions can introduce the bias in training data. Existing techniques simply fill the undefined region with the average pixel value by assuming that the activation distributions follow Gaussians. However, this assumption is not realistic, thus still unresolved in most of the dropout techniques. Our PST easily resolves it by utilizing reflection padding, which is simple but very effective to reflect the real data statistics during training. Therefore, PST performs regional dropout without distorting the original activation distributions.

Finally, the proposed PST can be utilized for a self-supervised learning method because we can generate a pair of image-label by applying PST. The method of defining an artificial label is based on (Gidaris, Singh, and Komodakis 2018; Zhang et al. 2019a). Specifically, we assign five labels according to the direction of vertical or horizontal translation, and train CNN to predict these labels. In addition to PST, we utilize the rotation transformation for additional supervisions like (Gidaris, Singh, and Komodakis 2018). We present how to obtain heat map from trained CNN in Section 3.2 and our training scheme in Section 3.3.

3.2 Class-agnostic Activation Mapping

Existing WSOL techniques utilize CAM to obtain heat map. They perform localization by tracking the spatial distribution of activations, which are feature responses corresponding to the target label. On the other hand, unsupervised object localization techniques should generate the heat map without image-level labels. Therefore, we extract the class-agnostic activations by aggregating all feature responses for predicting the geometric transformation. We name this method as class-agnostic activation mapping (CAAM). Assuming that the target object appears more frequently in the fine-grained datasets, we expect that our network model learns the features of target objects.

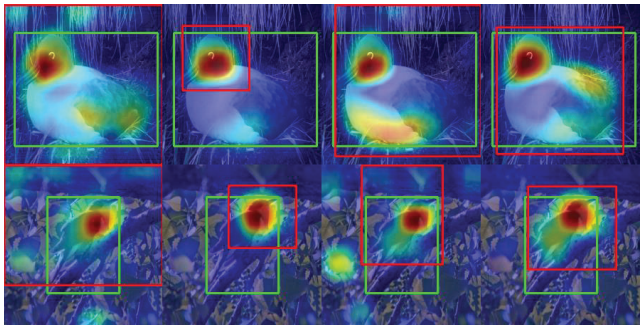

Figure 4: Effects of combination of various transformations. Each image from the leftmost to the rightmost shows the localization result of pre-trained only, rotation transformation, PST and PST + rotation transformation, respectively.

To extract activation for unsupervised object localization, we explore two simple approaches. 1) Spatial attention map with channel pooling (Zagoruyko and Komodakis 2016) is used as a heat map (See Figure 3). This attention map can be computed by applying an average pooling along the channel axis on the last feature map of the model. In this way, we aggregate feature responses to estimate the entire labels, regardless of 1 (*i.e.* target label) or 0 (*i.e.* non-target label). 2) We name the second approach by Multi-CAM because it focuses on the activations from the target labels like CAM. That is, the activation map is obtained for each target label similarly to CAM; tracking the spatial distribution of activations corresponding to a single target label. Afterward, multiple activation maps for various target labels (*e.g.* two target labels are used for representing the combination of translation and rotations) are then aggregated by either taking an average or choosing the maximum value. Unlike 1), this approach only considers the information of the target labels.

Table 1 demonstrates the localization performance among three different heat map extraction methods mentioned above. We observe that utilizing the spatial attention map achieves better localization accuracy than two variants of Multi-CAM on CUB-200-2011 dataset. We conjecture that it is because the activation for the entire labels encodes both negative (such as 0 labels) and positive (1 labels) information whereas Multi-CAM only uses positive information corresponding to the target label. Hence, in the remaining experiments, we decide to use the spatial attention for computing heat map. We note that other strategy for aggregation such as log-sum-exp does not improve the performance.

3.3 Unsupervised Object Localization via Self-supervised Approach

We slightly modify existing network models to handle multi-labels (*i.e.* allowing the output of network model to have multiple logits). After this modification, the network is trained to predict the artificial labels that represent geo-

metric transformations applied to the input image (see Figure 3). Formally, we denote x, F_θ, and $T_i(0 \leq i \leq n)$ as the input image, network model parameterized by θ and selected transformations, respectively. The parameters of randomly selected transformations T_i are determined by randomly sampling them from a candidate list of the parameter of each transformation, and then applied to the input image, sequentially. Given the transformed image $T_n(...T_0(x))$, the network F_θ predicts the parameters of transformations. In a meantime, we extract the spatial attention map $\mathbf{A}$ from the feature map of the last convolutional layer. Formally,

$$[y_0, ..., y_n, \mathbf{A}] = F_\theta(T_n(...T_0(x)). \quad (1)$$

where y_i is the estimated parameter of T_i. The network can be trained by minimizing a loss function L such that

$$L = \sum_{i=0}^{n} \mathcal{H}(y_i, c_i). \quad (2)$$

where c_i is the artificial label of each transformation T_i and $\mathcal{H}$ indicates the cross entropy function, *i.e.* $\mathcal{H}(p, q) = -\int_x p(x) \log q(x)$. We use SGD as the optimizer to train this self-supervised model. Once the network is trained, we can process the input image in the testing phase as follows. First, we pass the input to the network and obtain obtain the spatial attention map $\mathbf{A}$. Then, we extract the bounding box from $\mathbf{A}$ using the same method as presented in (Zhou et al. 2016). This bounding box is our final result, indicating the object of interest.

4 Experiments

In this section, we analyze the effects of various transformations that consist of our point symmetric transformation, and its hyperparameter via ablation study. Then, we evaluate the performance of the proposed method compared to existing techniques, including the state-of-the-art.

Backbone network and hyperparameters. For the find-grained object localization task, we initialize the network using an ImageNet pre-trained model before training because it is a common practice in existing studies (Wei et al. 2017; Zhang et al. 2019b; Wei et al. 2019). As the backbone networks, VGG16 (Simonyan and Zisserman 2015) and SE-ResNet50 (He et al. 2016; Hu, Shen, and Sun 2018) with the batch normalization (Ioffe and Szegedy 2015) are chosen. The use of batch normalization helps stabilize training but has no gain in the performance of networks. Then, the network is trained to predict multi-class artificial labels (*i.e.* transformation parameters).

Dataset and evaluation protocol. We conduct experiments on the fine-grained datasets: CUB200-2011 (Wah et al. 2011), Stanford Cars (Krause et al. 2013) and FGVC-Aircraft (Maji et al. 2013). We do not use the object discovery benchmark dataset (Rubinstein et al. 2013) because the size of the dataset is relatively small, thus not suitable to train the deep neural network. Therefore, we choose the dataset that is commonly used to train deep neural networks for weakly supervised object localization. For evaluating the unsupervised localization accuracy, we use GT-Known localization (*GT-Known Loc*), which is a standard metric for unsupervised localization problem. *GT-Known Loc* measures the intersection between the known ground truth bounding box and its estimated bounding box, normalized by their union. If this ratio is greater than or equals to 0.5, then the estimated bounding box is considered as correct.

Table 2: *GT-Known Loc* performance according to the parameter of point symmetric transformation.

Param. of S	GT-Known	Param. of S	GT-Known
1.0	**72.06**	1.2	67.82
1.5	52.66	1.8	32.51
2.1	47.38	2.4	53.47
2.7	67.73	3.0	69.07
[1, 3)	**75.34**	-	-

Toward weakly supervised learning. The key idea of the proposed method is 1) spatial attention based activation mapping and 2) geometric transformation for self-supervised learning. We employ two ideas and slightly tune them to solve weakly supervised object localization. First of all, we replace CAM with our spatial attention based activation mapping as-is. In addition, we use our geometric transformation (*i.e.* rotation with PST) as a data augmentation rule, generating the augmented data. To evaluate the weakly supervised localization accuracy, we use three metrics as reported in (Zhou et al. 2016; Zhang et al. 2018; Choe and Shim 2019) : GT-Known localization, Top-1 classification accuracy (*Top-1 Cls*) and Top-1 localization accuracy (*Top-1 Loc*). *Top-1 Cls* measures whether the predicted label exactly matches the ground truth label. *Top-1 Loc* determines that the answer is correct if both *GT-Known Loc* and *Top-1 Cls* are correct.

4.1 Ablation Study

Effects of point symmetric transformation parameters. PST is implemented by performing the large translation followed by filling an undefined region with the reflected remaining part of the image. We use five possible directions for translation, which is depicted as [A, B, C, D, E] in Figure 3 (they are not normalized for the sake of simplicity). Per training example, these five translations with reflection padding are applied and their artificial labels are assigned accordingly. The control parameters of PST are x- and y-translation and they are denoted as t_x and t_y. In this paper, we set t_x and t_y to be the same absolute value of t.

To investigate the effect of the parameters of PST (t_x and t_y), we change the absolute values from 1.0 to 3.0 while the size of an input image is normalized in-between -1 to 1. Figure 2 illustrates the transformed images when $[t_x, t_y] = [+t, +t]$ $(1 \leq t \leq 3)$ is applied, and the corresponding *GT-Known Loc* performance is shown in Table 2. We note that the transformation with 1.0 indicates the translation with the half of image resolution and it is equivalent to the transformation with 3.0 (because 2 is the full image resolution). We observe that the network performs well when the parameter approaches 1.0 (or 3.0). We believe

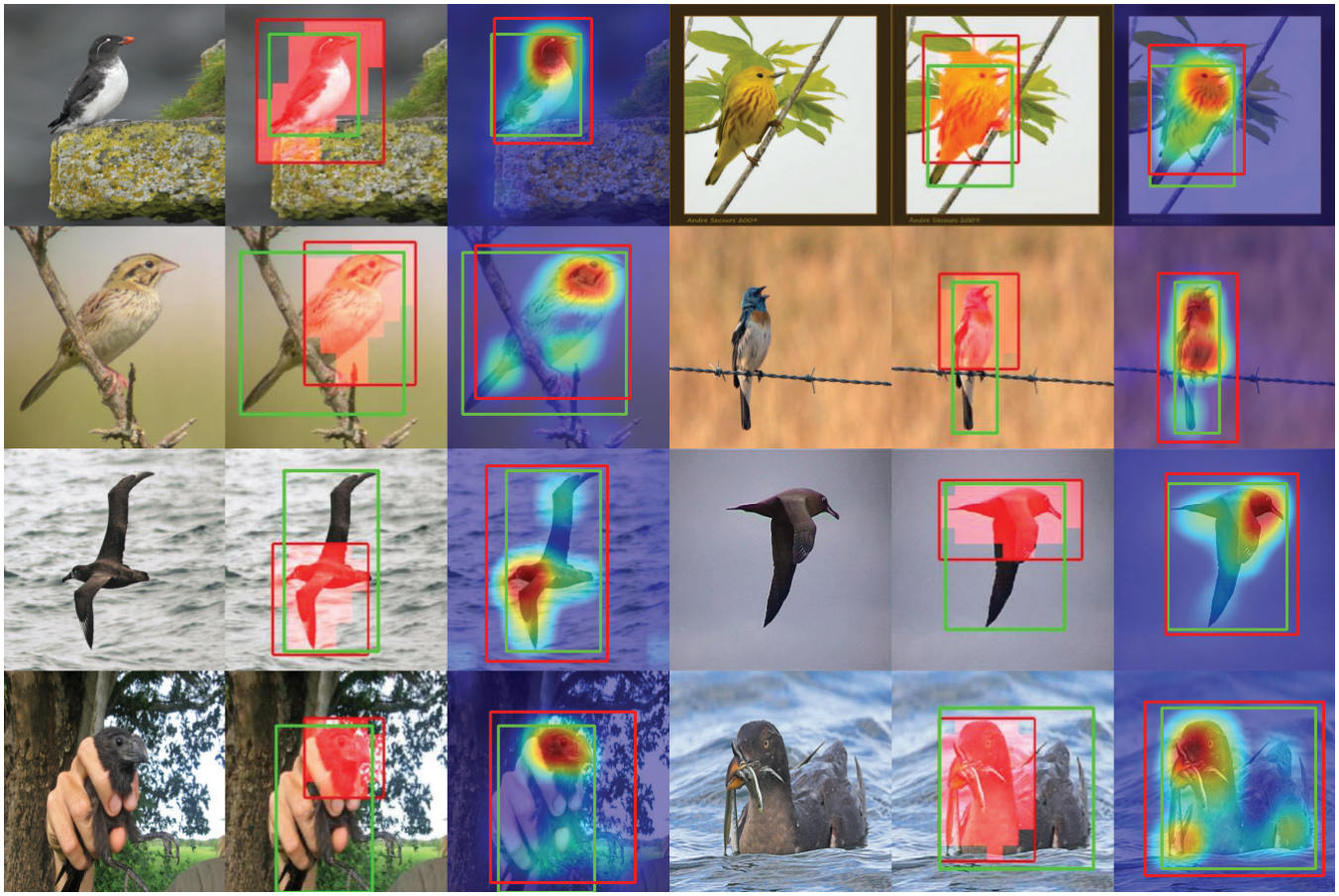

Figure 5: Qualitative comparison with DDT (Wei et al. 2019). The green and red box are the ground truth and predicted bounding box, respectively. For each triplet, result from the original, DDT and the proposed method are listed from left to right.

that the higher performance with $t = 1.0$ comes from the higher drop rate. That is, the original image is erased maximally when $t = 1.0$. Instead of the fixed parameter, we also examine the effect of the randomly sampled parameter and find that the random selection performs better than others. However, the performance of the random selection is degraded when it is combined with other transformations. Therefore, we set the absolute value of $t_x = t_y$ as 1.0 in the following experiments.

Effects of combining geometric transformations. We analyze the effects of combining various geometric transformations. For that, five different transformations are chosen, which are shear, flip, scale, rotation and proposed PST. The label for rotation transformation represents four possible rotation angles, $[0, 90, 180, 270]$ degrees, as same as (Gidaris, Singh, and Komodakis 2018). Similarly, we set three possible labels for shear and scale, and two possible labels for flip. Due to the page limit, we report the partial results of the experiment in Table 3.

As described in Table 3, the combination of PST and rotation outperforms any other combinations or any single transformations across all the datasets. In Figure 4, we visualize three localization results: 1) trained with PST only, 2) trained with rotation only, and 3) trained with the combination of rotation and PST. Interestingly, the heat map from rotation tends to have narrower activations. Thus, the resulting bounding box is shrunk. On the other hand, the map from PST tends to have wider activations, which expands the bounding box. From this study, we confirm that rotation or PST itself is also effective, but their combination increases

Table 3: *GT-Known Loc* performance according to the combination of the transformations. During the entire paper, Bold text refers the best localization accuracy in each table. H and E denotes horizontal flip and shear, respectively.

Combination	CUB200	Cars	Aircraft
R(·)	80.51	89.21	87.67
S(·)	66.98	92.89	92.41
C(·)	41.68	83.34	70.48
H(·)	52.59	77.44	87.43
E(·)	51.17	85.86	90.19
S(R(·))	**83.78**	**95.59**	**96.61**
S(R(C(·)))	58.66	89.96	71.80

Table 4: Comparison with previous works in terms of *GT-Known Loc* performance. Note that "*" indicates *GT-Known Loc* performance of weakly supervised method.

Dataset	Method	*GT-Known Loc*
CUB-200-2011	CAM*	51.09
	ACoL*	64.86
	ADL*	75.41
	Cho et al.	69.37
	SCDA	76.79
	DDT	82.26
	MO	80.45
	PsyNet	**83.78**
	PsyNet (SE-Res50)	**85.10**
FGVC-Aircraft	Cho et al.	36.23
	SCDA	94.91
	DDT	92.53
	MO	94.94
	PsyNet	**95.59**
	PsyNet (SE-Res50)	**97.81**
Stanford Cars	Cho et al.	93.05
	SCDA	90.96
	DDT	71.33
	MO	92.51
	PsyNet	**96.61**
	PsyNet (SE-Res50)	**98.81**

the benefits further. We conjecture that focusing on the common patterns among classes is advantageous to predict the correct rotation label. On the other hand, PST encourages the network to learn the entire extent of an object by dropping a part of the input object. Therefore, we believe that the proper combination of two transformations encourages the network to consider the integral object from the input image well.

4.2 Comparison with the State-of-the-arts

We evaluate the object localization performance of the proposed method qualitatively and quantitatively. For the quantitative evaluation, we compare our method to unsupervised and weakly supervised object localization methods using *GT-Known Loc*. Specifically, we compare the proposed method with 1) CAM (Zhou et al. 2016), 2) ACoL (Zhang et al. 2018) and 3) ADL (Choe and Shim 2019), 4) DDT (Wei et al. 2019), 5) Cho et al. (Cho et al. 2015), 6) SCDA (Wei et al. 2017) and 7) mining object (MO) (Zhang et al. 2019b). Note that 1), 2) and 3) are the most important techniques for weakly supervised learning, 4) is the competitor for tackling the co-localization problem, and 5), 6) and 7) are the unsupervised object localization techniques. (MO is not yet published elsewhere but on arxiv.) The weakly supervised methods utilize the image-level label during training, and they predict the bounding box and label simultaneously as output. Among CAM, ACoL, and ADL, ADL is the current state-of-the-art technique. On the other hand, the proposed algorithm considers the co-localization problem, where a set of images containing the target objects from the same class is given as a training set. Thus, we compare our results with DDT, which is the state-of-the-art technique in co-localization techniques. Unsupervised object localization techniques utilize a single image without additional data, and SCDA and MO achieve the best accuracy among unsupervised techniques except for PsyNet.

Table 4 summarizes the comparison results on three fine-grained datasets. We observe that our method outperforms all the existing techniques, including the state-of-the-art in weakly, co-localization, and unsupervised localization. DDT, MO, and Cho et al. achieve the second-best localization accuracy in CUB-200-2011, FGVD-Aircraft, and Stanford Cars, respectively. In these datasets, the proposed method achieves the localization accuracy to 83.78%, 95.59% and 96.61% with VGG16 network, respectively. It means that our method earns the performance gain by 3.3%, 0.6%, and 3.5% over the current state-of-the-art performance at each dataset, respectively. By adopting more powerful backbone network, such as SE-ResNet50 (He et al. 2016; Hu, Shen, and Sun 2018), we can enjoy the additional performance benefit of 2 %.

Figure 5 visualizes the localization results and qualitatively compare the proposed method with DDT on CUB-200-2011 dataset. We observe that the proposed method generates more accurate localization maps and bounding boxes than DDT. For example, DDT tends to capture the part of the object (only body part is captured in the second row at the second column) or out of the boundary (in the first row at the second column) as shown in Figure 5. However, the proposed method can localize the bird more accurately than DDT in that the background objects are excluded from the heat map. Through quantitative and qualitative comparison, we demonstrate that the proposed method is effective for unsupervised object localization.

Comparing the inference time. Additionally, we try to compare the inference time of our method with MO and DDT. MO stated that their execution time is totally about 0.24 second/image on GPU, given CUB-200-2011 dataset. DDT stated that the average deep descriptor transforming time is 0.0333 second/image on GPU. On the other hand, the inference time of our method is about 0.0067 second/image on GPU, which at least five times faster than DDT and 30 times faster than MO. This shows that the efficiency of our method in real-time applications.

4.3 Toward WSOL Techniques

We slightly modify the two key ideas of the proposed method and use them to improve WSOL method. They are 1) CAAM instead of CAM, and 2) a novel training strategy to perform self-supervised learning using the newly proposed PST. First, CAAM can be directly applied to the WSOL method. Next, the idea of novel PST needs a slight modification. Concretely, PST can be used as a data augmentation policy. In order to further improve performance, PST and other geometric transformation are applied as data augmentation policies. By introducing two simple training tactics, we achieve the new state-of-the-art performance in CUB-200-2011 dataset, with nearly 10 % of performance gain over the current state-of-the-art. The performance compar-

Table 5: Modified PsyNet for WSOL is compared with existing WSOL methods. "M" indicates that the proposed heat map extraction is used. "A" denotes that PST is used as a data augmentation policy.

Method	*GT-Known Loc*	*Top-1 Cls*	*Top-1 Loc*
CAM	51.09	67.55	34.41
ACoL	64.86	71.90	45.92
HaS	70.77	69.59	50.80
ADL	75.41	65.27	52.36
PsyNet(M)	**80.32**	69.67	57.97
PsyNet(A)	61.56	**77.25**	48.10
PsyNet(M+A)	77.39	75.04	**59.37**

isons with the existing WSOL techniques are demonstrated in Table 5. For WSOL task, we use ImageNet pre-trained network as the previous WSOL works.

To investigate the effects from two different ideas, we first apply the heat map extraction method (denoted as "M" in the table), and then add the data augmentation using PST over the heat map extraction method (denoted as "A" in the table). We observe that our heat map extraction method of computing a spatial attention map significantly improves the performance in terms of both *GT-Known Loc* and *Top-1 Loc*. Previous literatures stated (Singh and Lee 2017; Choe and Shim 2019) that there exists a trade-off relationship between localization and classification accuracy. Surprisingly, our heat map extraction method does not sacrifice the *Top-1 Cls* for improving *GT-Known Loc*. As a result, it is possible to fully enjoy the performance gain in *Top-1 Loc*. Then, PST and other geometric transformations are used as data augmentation policies. Although the transformation based data augmentation moderately degrades *GT-Known Loc*, it achieves best *Top-1 Loc* by significantly improving the classification accuracy. Based on this experiment, we confirm that the proposed method can be successfully applied to WSOL on the fine-grained dataset.

5 Conclusion

We proposed a new fine-grained co-localization technique using self-supervised learning approach. To this end, this paper introduced two important ideas. First, for self-supervised learning, we devised a novel point symmetric transformation implicitly possessing the property of regional dropout. This new transformation is then applied to the input, and its label is used as an artificial label to train the model. Secondly, we suggested utilizing the spatial attention map for computing a heat map. This scheme allows us to extract the heat map without object label, thus being suitable for the network trained with self-supervision. Based on extensive evaluations, we confirmed that the proposed method outperforms existing unsupervised object localization techniques, including the current state-of-the-art. Additionally, we applied the proposed method with a small modification to weakly supervised object localization setting. Consequently, we could achieve the new state-of-the-art performance among the same kinds. For the future work, we will enable the PsyNet to conduct the localization without the ImageNet pre-trained network and on the ImageNet dataset.

Acknowledgement

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the MSIP (NRF-2019R1A2C2006123), MSIT(Ministry of Science and ICT), Korea, under the "ICT Consilience Creative Program" (IITP-2019-2017-0-01015) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation), and ICT R&D program of MSIP/IITP. [R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding]

References

Cho, M.; Kwak, S.; Schmid, C.; and Ponce, J. 2015. 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*, 1201–1210.

Choe, J., and Shim, H. 2019. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2219–2228.

DeVries, T., and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 1422–1430.

Dosovitskiy, A.; Springenberg, J. T.; Riedmiller, M.; and Brox, T. 2014. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, 766–774.

Feng, Z.; Xu, C.; and Tao, D. 2019. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10364–10374.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456.

Joulin, A.; Tang, K.; and Fei-Fei, L. 2014. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision*, 253–268. Springer.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*.

Larsson, G.; Maire, M.; and Shakhnarovich, G. 2016. Learning representations for automatic colorization. In *European Conference on Computer Vision*, 577–593. Springer.

Li, Y.; Liu, L.; Shen, C.; and van den Hengel, A. 2016. Image co-localization by mimicking a good detector's confidence score distribution. In *European Conference on Computer Vision*, 19–34. Springer.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2):91–110.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Noroozi, M., and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, 69–84. Springer.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Rubinstein, M.; Joulin, A.; Kopf, J.; and Liu, C. 2013. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1939–1946.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Singh, K. K., and Lee, Y. J. 2017. 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)*, 3544–3553. IEEE.

Tang, K.; Joulin, A.; Li, L.-J.; and Fei-Fei, L. 2014. Co-localization in real-world images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1464–1471.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wei, X.-S.; Luo, J.-H.; Wu, J.; and Zhou, Z.-H. 2017. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing* 26(6):2868–2881.

Wei, X.-S.; Zhang, C.-L.; Wu, J.; Shen, C.; and Zhou, Z.-H. 2019. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition* 88:113–126.

Zagoruyko, S., and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Zhang, X.; Wei, Y.; Feng, J.; Yang, Y.; and Huang, T. S. 2018. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1325–1334.

Zhang, L.; Qi, G.-J.; Wang, L.; and Luo, J. 2019a. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2547–2555.

Zhang, R.; Huang, Y.; Pu, M.; Guan, Q.; Zhang, J.; and Zou, Q. 2019b. Mining objects: Fully unsupervised object discovery and localization from a single image. *arXiv preprint arXiv:1902.09968*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.