Unveiling the Potential of Structure-Preserving for Weakly Supervised Object Localization

Xingjia Pan[1,3*]   Yingguo Gao[1*]   Zhiwen Lin[1*]   Fan Tang[2†]   Weiming Dong[3,4,5]
Haolei Yuan[1]   Feiyue Huang[1]   Changsheng Xu[3,4,5]
[1]Youtu Lab, Tencent   [2]Jilin University   [3]NLPR, Institute of Automation, CAS
[4]School of Artificial Intelligence, UCAS   [5]CASIA-LLVision Joint Lab
{noahpan, yingguogao, xavierzwlin,harryyuan,garyhuang}@tencent.com
tangfan@jlu.edu.cn, {weiming.dong, changsheng.xu}@ia.ac.cn

Abstract

Weakly supervised object localization (WSOL) remains an open problem due to the deficiency of finding object extent information using a classification network. While prior works struggle to localize objects by various spatial regularization strategies, we argue that how to extract object structural information from the trained classification network is neglected. In this paper, we propose a two-stage approach, termed structure-preserving activation (SPA), towards fully leveraging the structure information incorporated in convolutional features for WSOL. In the first stage, a restricted activation module (RAM) is designed to alleviate the structure-missing issue caused by the classification network, based on the observation that the unbounded classification map and global average pooling layer drive the network to focus only on object parts. In the second stage, we propose a post-process approach, termed self-correlation map generating (SCG) module to obtain structure-preserving localization maps on the basis of the activation maps acquired from the first stage. Specifically, we utilize the high-order self-correlation (HSC) to extract the inherent structural information retained in the learned model and then aggregate HSC of multiple points for precise object localization. Extensive experiments on two publicly available benchmarks including CUB-200-2011 and ILSVRC show that the proposed SPA achieves substantial and consistent performance gains compared with baseline approaches.

1. Introduction

Weakly supervised learning refers to methods that utilize training data with incomplete annotations to learn recogni-

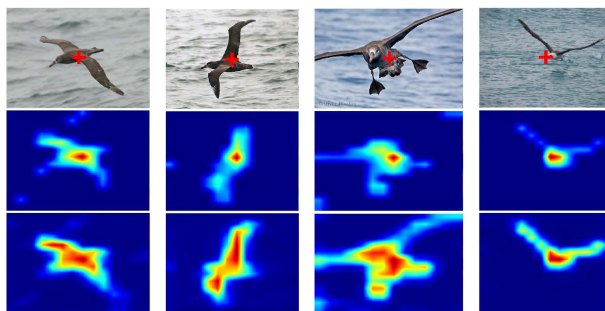*Equal contribution
†Corresponding author

Figure 1. Self-correlation maps corresponding to the image positions masked by red crosses. The 2nd and 3rd rows are the first-order and second-order self-correlation maps, respectively.

tion models. Weakly supervised object localization (WSOL) solely requires the image-level annotations indicating the presence or absence of a class of objects in images to learn localization models [27]. In recent years, WSOL has attracted increasing attention because it can leverage the rich Web images with tags to learn object-level models.

As a work for WSOL, Class Activation Mapping (CAM) [34] uses the intermediate classifier activation to discover discriminative image regions for target object localization [4]. Afterward, divergent activation methods [18, 27, 29] design multiple parallel branches or introduce attention modules to drive networks learning complete object extent. Adversarial erasing methods [5, 13, 25, 30] pursue learning full object extent in a hide-and-seek fashion. Existing methods [13, 25, 30] largely depend on CAM and spatial regularization to localize objects, *e.g.*, expanding activated regions to full object extent; nevertheless, preserving the structure of object is unfortunately neglected.

By experiments, we observed that the head structure of the classification network causes the CAM missing object structure information contained in convolutional features. Specifically, the network driven by sole classification loss

tends to activate a small proportion of features with largest discriminative capability while depressing the majority of object extent. Furthermore, many CAM-based methods employ a *global average pooling* (GAP) [11] layer atop the feature maps to retain the localization information. The GAP layer treats each pixel within the feature map equally, which hinders distinguishing true objects from noisy background.

The inherent spatial correlation in the deep neural network has been widely used in image classification, object detection and segmentation areas [2, 23, 24], but remain unexploited for WSOL. Following the self-attention mechanism, some recent works [24, 32, 33] adopt the first-order pixel-wise correlation to improve object localization. However, pixels within one object usually have dissimilar features for the large appearance variation, which limits the capability of the first-order self-correlation to preserve object structural information, Fig. 1 (2nd row).

In this study, we propose a two-stage WSOL approach, termed the structure-preserving activation (SPA), to pursue accurate object localization using sole image-level category labels as supervision. In the first stage, a restricted activation module (RAM) is designed to alleviate the limitations of the head structure of the classification network. RAM suppresses the response value range of CAM to avoid the misleading by local extremely high response for classification, and differentiates objects from background under the guidance of estimated pseudo-masks. In the second stage, a self-correlation map generating (SCG) module is proposed, with the aim to refine the localization map under the guidance of the structural information extracted from trained features. In SCG, to guide the activation of objects, we propose to use the high-order self-correlation (HSC) which facilities capturing precise spatial layouts of objects by long-range spatial correlations, Fig. 1 (3rd row). We conduct extensive experiments on the CUB-200-2011 [22] and ILSVRC [15] datasets. Our method obtains significant gains compared with baseline methods and achieve comparable results with the SOTAs on both bounding box and mask localization.

The contributions of this study include:

- We unveil the fact that the spatial structure preserving is crucial for discovering the localization information contained in convolutional features for weakly supervised object localization.

- We propose a simple-yet-effective SPA approach to distill the structure-preserving ability of features for accurate object localization.

- With negligible computational complexity and cost overheads, our proposed approach shows consistent and substantial gains across CUB-200-2011 and ILSVRC datasets for both bounding box and mask localization.

2. Related Work

Weakly supervised object localization (WSOL) aims to learn the localization of objects with only image-level labels. A representative work on WSOL is CAM [34], which produces localization maps by aggregating deep feature maps using a class-specific fully connected layer. Hwang and Kim [8] simplified CAM by removing the last fully connected layer. Despite the simplicity and effectiveness of CAM-based methods, they suffer from the weakness of only identifying small discriminative part of objects. To improve the activation of CAMs, HaS [18] and CutMix [18] adopt erasing-based strategy from input images to force the network to focus on more relevant parts of objects. Differently, ACoL [30] and ADL [5] instead erased feature maps corresponding to discriminative regions and used multiple parallel classifiers that were trained adversarially. Apart from the above erasing methods, SPG [31] and I²C [32] increased the quality of localization maps by introducing the constraint of pixel-level correlations into the network. DANet [27] applied a divergent activation to learn complementary and discriminative visual patterns for WSOL. SEM [33] refined the localization maps by using the point-wise similarity within seed regions. GC-Net [12] took geometric shape into account and proposed a multi-task loss function. Since existing methods only focus on expanding activation regions, they are puzzled by the contradiction between precise classification and object localization. The problem about how to leverage a classification network to active and localize full object extent remains unsolved.

Weakly supervised semantic segmentation (WSSS) aims to predict precise pixel-level object masks using weak annotations. The mainstream methods for WSSS with image-level labels train classification networks to estimate object localization maps as pseudo masks which are further used for training the segmentation networks. To generate accurate pseudo masks, [1, 7, 23] resorted to region growing strategy. Meanwhile, some researchers investigated to directly enhance the feature-level activated regions [10, 26]. Others accumulate CAMs by multiple training phases [9], exploring boundary constraint [3], leveraging equivariance for semantic segmentation [24], and mining cross-image semantics [20] to obtain more perfect pseudo masks.

Feature Self-Correlation. Spatial self-correlation is an instantiate of self-attention mechanism for non-sequential data in computer vision. Most WSOL/WSSS methods [24, 32, 33] utilize the similarity of pixels to refine the features or activation maps following self-attention mechanism. Wang *et al.* [23] proposed a non-local block to capture long-range dependency within image pixels. Cao *et al.* [2] found that the global contexts modeled by non-local network are almost the same for query positions and thereby proposed NLNet [23] with SENet [6] for global context modeling. MST [19] proposed the learnable tree filter to leverage the structural
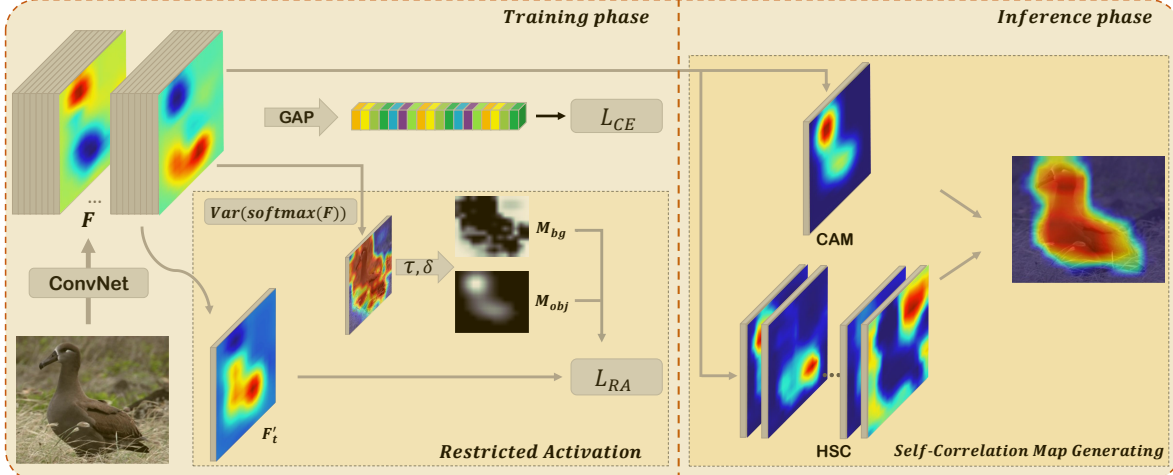
Figure 2. The framework of the proposed SPA approach. During training phase, we design a restricted activation module paralleled with the classification branch on the baseline architecture of simplified CAM [34]. In the inference phase, we propose a self-correlation map generating module, which is a post-process method to refine the localization map by introducing the structural information of objects.

property of minimal spanning tree to model long-range dependencies. DNL [28] disentangled the non-local block into a whitened pairwise term and a unary term to facilitate the learning process. These methods belong to the first-order self-correlation, and acquire the long-range context by stacking numerous modules in different stages. For a single feature layer, they can only retain local structural information.

3. Structure-Preserving Activation

3.1. Overview

Based on the structure-preserving ability of convolutional features, we obtain precise localization maps by proposing the structure-preserving activation (SPA) approach for WSOL. As shown in Fig. 2, we adopt the CAM network [34] as our baseline and remove the last fully-connected layer following ACoL [30]. Overall, the proposed SPA retains the structural information of objects in two stages. In the first stage, as shown in the training phase of Fig .2, we design the restricted activation module (RAM) to alleviate the structure-missing issue of the head structure of CAM. Furthermore, a restricted activation loss (L_{RA}) is proposed to cooperate with cross-entropy loss (L_{CE}) for driving the model to cover object extent during training phase. The total loss of SPA training is defined as:

$$L = L_{CE} + \alpha L_{RA}, \tag{1}$$

where L_{CE} is the multi-class cross entropy loss, and L_{RA} is the restricted activation loss in RAM. α is a regularization factor to balance the two items. In the second stage, as shown in the inference phase of Fig. 2, we propose the self-correlation map generating module (SCG) to obtain accurate

localization maps on the basis of the results of CAM during inference phase. We extract high-order self-correlations (HSCs) for each point of CAM from the convolutional features and aggregate the HSCs to acquire activation maps for object localization.

3.2. Restricted Activation Module

The proposed RAM alleviates the structure-missing issue of CAM from two aspects: suppressing the response value range of CAM to avoid the misleading by the local extremely high response for the classification; discriminating the object from background region with the help of coarse pseudo-masks.

Given a fully convolutional network (FCN), we denote the last convolutional feature maps as $F \in \mathbb{R}^{H \times W \times C}$, where $H \times W$ is the spatial size and C is the number of channels which is equal to the number of target classes. The feature maps are fed into a GAP [11] layer followed by a *softmax* layer for classification, as shown in Fig. 2. In order to ensure that the high activation value after GAP layer is due to broad object activation rather than the local extremely high response, we first suppress the feature value range using the *sigmoid* layer:

$$F_{t}^{'} = sigmoid(F_t), \tag{2}$$

where t is the ground truth label index and F_t is the t_{th} feature map. The *sigmoid* layer can effectively suppress the extremely high response and normalize the values to (0,1). The GAP layer averages the full feature activation and treats each pixel equally, which hinders the model differentiating the object from background. To overcome this issue, we propose a simple method to generate coarse pseudo-masks for guiding the model to focus on the object regions. The mask generation method is based on the observation that the
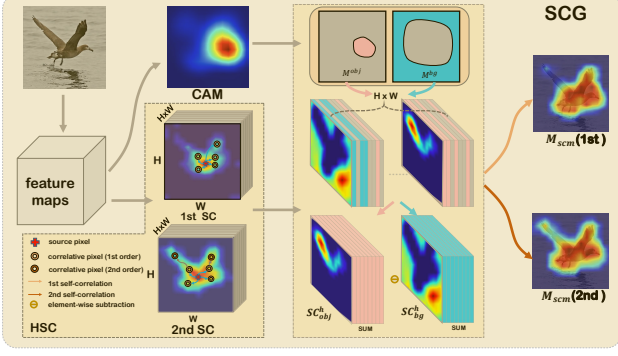
Figure 3. The pipeline of the proposed SCG module. Here we show examples of using first-order and second-order SC to obtain final localization maps respectively.

activation values within the background area are distributed much evenly across all classes, and the activation value of the object region can always be highly responsive in at least one target class. Therefore, we obtain the coarse background mask M_{bg} as:

$$M_{bg} = \mathbb{I}(Var(softmax(F)) < \tau), \quad (3)$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $Var(\cdot)$ denotes the standard deviation of each position on the feature map in the channel dimension. τ is a constant value as the threshold to determinate the background region. We further obtain the coarse object region as:

$$M_{obj} = \mathbb{I}(Var(softmax(F)) > \tau + \sigma), \quad (4)$$

where σ denotes the gap between the background and object regions. On the basis of pseudo-masks, we define the restricted activation loss to guide the model to focus on the object region as:

$$L_{RA} = \frac{1}{HW} \sum_{i,j} (M_{bg} * F_t^{'} + M_{obj} * (1 - F_t^{'}))|_{i,j}, \quad (5)$$

where $*$ indicates element-wise multiplication. The simple L_1 regularization loss can guide the model to suppress the background area while pay much attention to active full object extent. Cooperating with the classification branch, the proposed RAM enables the model to preserve the structural information of the target object.

3.3. Self-Correlation Map Generating

Before introducing SCG, we first analyze the first-order correlation and introduce the concept of spatial HSC which could capture the long-range structural information based on rich context of the object. Then for generating precise localization maps on the basis of HSC, we propose SCG to distill the structure-preserving ability of deep CNN.

First-order Self-Correlation. We refer the relation response directly calculated by pixel-to-pixel similarity as spatial first-order self-correlation. Given a feature map $f \in \mathbb{R}^{HW \times C}$, we use cosine distance to evaluate inter-pixel similarity for feature of index i and j:

$$S(f_i, f_j) = \frac{\theta(f_i)^{\mathrm{T}} \theta(f_j)}{||\theta(f_i)|| \cdot ||\theta(f_j)||}, \quad (6)$$

where $i, j \in \{0, 1, \ldots, HW - 1\}$ indicate the index of features, and $f_i, f_j \in \mathbb{R}^{C \times 1}$ are the feature vectors. In our experiments, we remove the embedding function θ for simplicity. We define the *first-order self-correlation* of f as:

$$SC(f) = [SC(f)_{i,j}], \quad where \ SC(f)_{i,j} = ReLU(S(f_i, f_j)). \quad (7)$$

The similarities $S(\cdot, \cdot)$ are activated by ReLU [14] to suppress negative values and $SC(f) \in \mathbb{R}^{HW \times HW}$. Due to large appearance variation, the pixels within an object are usually dissimilar. The 2$^{\mathrm{nd}}$ row in Fig. 1 shows several examples of self-correlation corresponded to positions masked by the red cross. It shows that first-order self-correlation only can preserve local spatial structure information.

High-order Self-Correlation. To apply the inherent structure-preserving ability of the network for accurate WSOL, we propose to use high-order self-correlation (HSC) to capture the structural information of objects. The h_{th} order similarity between f_i and f_i are formulated as:

$$S^h(f_i, f_j) = \frac{1}{(HW)^{h-1}} \sum_{k^0} \cdots \sum_{k^{h-2}} S(f_i, f_{k^0}) \cdots S(f_{k^{h-2}}, f_j), \quad (8)$$

where $k^0, k^{h-2} \in \Omega$, $h \in \{2, 3, 4 \cdots\}$ and $i \neq k^0 \neq k^{h-2}$. h denotes the order, and Ω denotes the set of all features. The $S^h(f_i, f_j)$ is then normalized to $[0, 1]$ following:

$$\hat{S}^h(f_i, f_j) = \frac{S^h(f_i, f_j) - min_{k \in \Omega} S^h(f_i, f_k)}{max_{k \in \Omega} S^h(f_i, f_k) - min_{k \in \Omega} S^h(f_i, f_k)}, \quad (9)$$

Then, we define HSC as:

$$SC^h(f) = \left[ReLU(\hat{S}^h(f_i, f_j))|_{i,j} \right]. \quad (10)$$

Compared with first-order self-correlation, HSC can preserve the details of the object by considering long-range context. However, the HSC may introduce additional noise. Therefore, we utilize both the first-order and high-order SC, and combine them by element-wise maximum operation in our experiments. For the trade-off between accuracy and efficiency, we employ the SC^2 as the HSC to capture long-range structural information in our experiments. The 3$^{\mathrm{rd}}$ row in Fig. 1 lists numerous examples of SC^2.

CAM [34] can only highlight local region of interest and thus lose the structural information. To acquire accurate

object extent, we propose the *SCG* to refine the localization maps with the help of HSC.

Given the feature map f, we obtain $SC^h(f) \in \mathbb{R}^{HW \times HW}$ following Equ. 10. For clarity of description, we here reshape $SC^h(f)$ to $\mathbb{R}^{H \times W \times H \times W}$. We first employ the CAM to obtain the coarse localization map $M_{cam} \in \mathbb{R}^{H \times W}$ following ACoL [30] by removing the last fully-connected layer. We define a threshold δ_h to discover the coarse object mask $M_{cam}^{obj} = M_{cam} > \delta_h$. Given the indices of object region, we extract the corresponded high-order self-correlation of the object as:

$$SC_{obj}^h = G(SC^h, M_{cam}^{obj}), \tag{11}$$

where $SC_{obj}^h \in \mathbb{R}^{N \times H \times W}$. $G(\cdot)$ denotes the index function, and N is the number of pixels within object region. Then the self-correlation map of the object M_{scg}^{obj} is obtained by aggregating the HSC of each point within object region as :

$$M_{scg}^{obj} = \frac{1}{N} \sum_i SC_{obj}^h[i]. \tag{12}$$

In order to remove the possible background area covered by M_{scg}^{obj}, we define another threshold δ_l and obtain the background self-correlation map M_{scg}^{bg} in a similar way. We acquire the final localization map M_{scg} as:

$$M_{scg} = ReLU(M_{scg}^{obj} - M_{scg}^{bg}) \tag{13}$$

The final self-correlation map M_{scg} is refined by removing the background area and is activated using ReLU to suppress negative values. Algorithm 1 illustrates the procedure of the proposed *SCG* approach.

Algorithm 1 Localization algorithm of SCG.

Input: Coarse localization map $M_{cam} \in \mathbb{R}^{H \times W}$; feature map $f \in \mathbb{R}^{H \times W \times C}$; threshold δ_h and δ_l;

Output: Final localization map M_{scg};

1: Obtain high-order self-correlation $SC^h \in \mathbb{R}^{HW \times HW}$
2: Reshape $SC^h \in \mathbb{R}^{H \times W \times H \times W} \leftarrow reshape(SC^h)$
3: Discover the coarse object region $M_{cam}^{obj} \leftarrow M_{cam} > \delta_h$
4: Extract object HSC $SC_{obj}^h \leftarrow G(SC^h, M_{cam}^{obj})$
5: Obtain the object map $M_{scg}^{obj} \leftarrow sum(SC_{obj}^h)$
6: Discover background region $M_{cam}^{bg} \leftarrow M_{cam} < \delta_l$
7: Extract background HSC $SC_{bg}^h \leftarrow G(SC^h, M_{cam}^{bg})$
8: Obtain the background map $M_{scg}^{bg} \leftarrow sum(SC_{bg}^h)$
9: Obtain localization map $M_{scg} \leftarrow (M_{scg}^{obj} - M_{scg}^{bg})_{(>0)}$
   return M_{scg};

---

4. Experiments

4.1. Experimental Settings.

Datasets. We evaluate the proposed approach on two publicly available benchmarks including CUB-200-2011 [22]
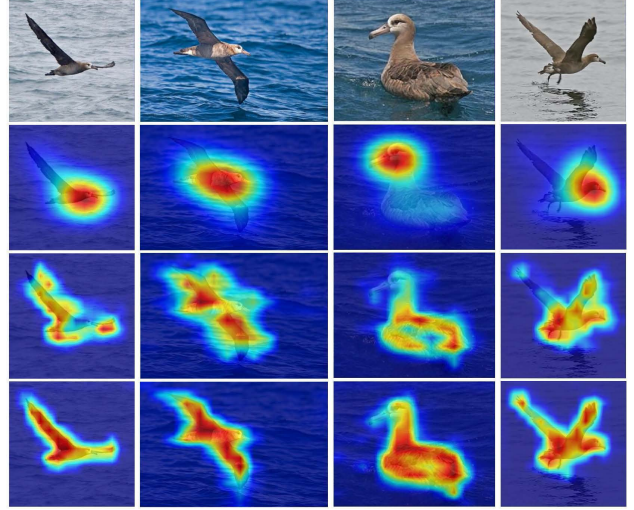

Figure 4. Visualization of the localization maps of CAM [34], SPA with first-order and second-order self-correlation, respectively. The images are from the CUB-200-2011 [22] testing set.

and ILSVRC [15], following the previous SOTAs [4, 13, 32, 33]. CUB-200-2011 is a fine-grained bird dataset of 200 different species, which is splited into the training set of $5,994$ images and the testing set of $5,794$ images. For ILSVRC, there are around 1.2 million images of $1,000$ categories for training and $50,000$ images for validation. Both benchmarks are all only annotated with class labels for training. In addition to class labels, CUB-2000-2011 also provides the tight box and mask labels for images in testing set. For ILSVRC, only tight box labels are provided for validation. Zhang *et al.* [33] annotates the ground-truth masks for the images on the validation set of ILSVRC: $5,729$ images are manually excluded and the rest are splited into validation set ($23,151$ images) and testing set ($21,120$ images).

Metrics. We apply two kinds of metrics to evaluate the localization maps from the bounding box and mask, respectively. For bounding boxes, we follow the baseline methods [15, 31, 34] and report the location error (Loc. Err.). A prediction is positive when it satisfies the following two conditions simultaneously: the predicted classification labels match the ground-truth categories; the predicted bounding boxes have over 50% IoU with at least one of the ground-truth boxes. $Gt\text{-}Known$ indicates it considers localization only regardless of classification. For masks, we mainly utilize Peak-IoU and Peak-T, which are defined in SEM [33], to directly evaluate the localization performance by performing a pixel-wise comparison between the predicted localization map and the ground-truth mask. *Peak-IoU* $\in [0, 1]$ and *Peak-T* $\in [0, 255]$ denote the best IoU score and its corresponding threshold, respectively. A high-quality localization map should meet two requirements: 1) the full object extent

Methods	Backbone	Loc. Err.		
		Top-1	Top-5	Gt-Known
Backprop [16]	VGG16	61.12	51.46	-
CAM [34]	VGG16	57.20	45.14	-
CutMix [29]	VGG16	56.55	-	-
ADL [5]	VGG16	55.08	-	-
ACoL [30]	VGG16	54.17	40.57	37.04
I²C [32]	VGG16	52.59	41.49	36.10
MEIL [13]	VGG16	53.19	-	-
Ours	VGG16	**50.44**	**38.68**	**34.95**
CAM [34]	InceptionV3	53.71	41.81	37.32
SPG [31]	InceptionV3	51.40	40.00	35.31
ADL [5]	InceptionV3	51.29	-	-
ACoL [30]	GoogLeNet	53.28	42.58	-
DANet [27]	GoogLeNet	52.47	41.72	-
MEIL [13]	InceptionV3	50.52	-	-
I²C [32]	InceptionV3	**46.89**	35.87	**31.50**
GC-Net [12]	InceptionV3	50.94	41.91	-
Ours	InceptionV3	47.27	**35.73**	31.67

Table 1. Comparison between our method and the state-of-the-art on the ILSVRC [15] validation set.

Methods	Backbone	Loc. Err.		
		Top-1	Top-5	Gt-Known
CAM [34]	GoogLeNet	58.94	49.34	44.9
SPG [31]	GoogLeNet	53.36	42.8	-
DANet [27]	InceptionV3	50.55	39.54	33.0
ADL [5]	InceptionV3	46.96	-	-
Ours	InceptionV3	**46.41**	**33.50**	**27.86**
CAM [34]	VGG16	55.85	47.84	44.0
ADL [5]	VGG16	47.64	-	-
ACoL [30]	VGG16	54.08	43.49	45.9
DANet [27]	VGG16	47.48	38.04	32.3
SPG [31]	VGG16	51.07	42.15	41.1
I²C [32]	VGG16	44.01	31.6	-
MEIL [13]	VGG16	42.54	-	-
GC-Net [12]	VGG16	**36.76**	**24.46**	**18.9**
Ours	VGG16	39.73	27.5	22.71

Table 2. Comparison between our method and the state-of-the-art on the CUB-200-2011 [22] test set.

can be accurately covered with a specific threshold; 2) brightness values of pixels belonging to object and the background should differ greatly so that the objects can be well visualized [33]. Higher *Peak-IoU* and *Peak-T* values indicates better localization maps.

Implementation Details. We implement the proposed algorithm based on two popular backbone networks, *i.e.* VGG16 [17] and Inception V3 [21]. We make the same modifications on backbones following ACoL [30] and SPG [21], and use the simplified method in ACoL [30] to obtain localization maps. Both networks are fine-tuned on the pre-trained weights of ILSVRC [15]. The input images are randomly cropped to 224×224 pixels after being re-sized to 256×256 pixels. For classification, we average the scores from the softmax layer with 10 crops. We also implement several recent benchmark methods, *i.e.*, CAM [34], HaS [18], ACoL [30], SPG [21], ADL [5], and CutMix [29] according to the codes[1] released by Choe *et al*. [4]. For fair comparisons, we adopt the same training strategy with SEM [33]. The codes for *Peak-IoU* and *Peak-T* are provided on the workshop of *Learning from Imperfect Data (LID)*[2]. To calculate the self-correlation on VGG16, we utilize the features of *stage* 4 and 5, and combine the two SCs by element-wise summation. For Inception V3, we utilize the features of layer *feat4* and *feat5* to calculate SC and sum them element-wisely.

4.2. Experimental Results

Bounding Box Localization. We first compare the proposed approach with the SOTAs on the localization error by using tight bounding boxes. We only show the *Loc.*

---

[1] https://github.com/clovaai/wsolevaluation
[2] https://lidchallenge.github.io/challenge.html

Err., please refer to the supplementary materials for more details. Table 1 reports the results of our method and several baselines on the ILSVRC validation set. Our method on the basis of VGG16 achieves the lowest error rate of 50.44% in Top-1 *Loc. Err.*, significantly surpassing all the baselines. Specifically, we achieve remarkable gains of 3.5% and 4.4% in terms of Top-1 *Loc. Err.* compared with ACoL and ADL. Compared with the state-of-the-art I^2C, we achieve a performance gain of 2.0%, which is a significant margin to the challenging problem. On the InceptionV3, our method obtains comparable results with I^2C and surpasses other methods by a large margin. I^2C leverage pixel-level similarities across different objects to prompt the consistency of object features within the same categories, but it cannot retain the structural information for the objects. Fig. 5 shows several examples of the localization maps by CAM [34] and the proposed SPA. Our results retain the structure of objects well, and cover more extent of the objects.

Table 2 compares the proposed method with various baseline methods on the CUB-200-2011 testing set. All the baselines adopt CAM to obtain localization maps. Our method based on VGG16 surpasses all the baseline methods on Top-1, Top-5, and Gt-Known metrics, yielding the localization error of Top-1 39.73%, and Top-5 27.5%. Compared with the current state-of-the-art I^2C and $MEIL$, we achieve gains of 3.5% and 2.0% in terms of Top-1 *Loc. Err.*, respectively. Fig. 4 shows some examples of the localization map. The 3rd and 4th rows are the results of our method by using first-order and second-order self-correlation, respectively. Compared with CAM [34], the results of our method preserve the structure of objects well. The results of SC^2 obtain more accurate masks than that of first-order, but they get almost the same tight bounding boxes. In order to reveal the superiority of the method, we further evaluate our method by comparing with the ground-truth
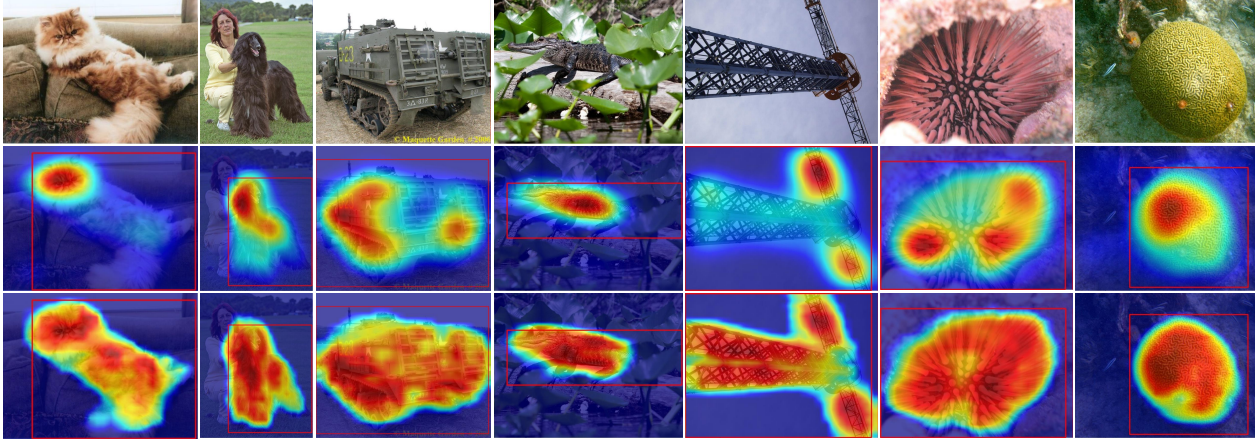
Figure 5. Visualization of the localization maps with CAM [34] (middle row) and the proposed SPA (bottom row). The ground truth boxes are in red. The images are from the ILSVRC [15] validation set.

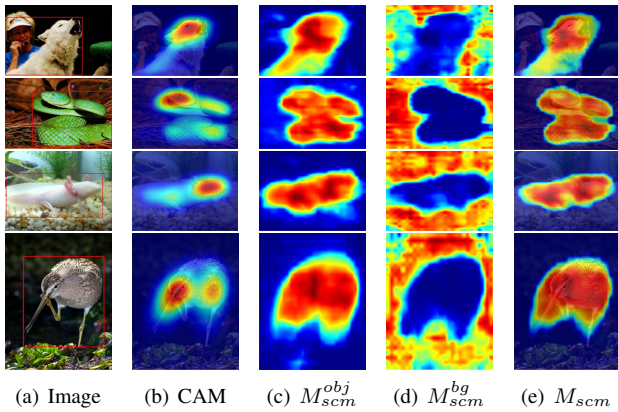| (a) Image | (b) CAM | (c) M_{scm}^{obj} | (d) M_{scm}^{bg} | (e) M_{scm} |

Figure 6. Visualization of process for SCG. Given the input images (a), we employ simplified CAM [34] to obtain the coarse localization maps (b). (c) and (d) are the object and background SC maps by aggregating HSCs of pixels within the corresponding areas on the basis of CAM, respectively. The final SCGs (e) are obtained by subtracting background from object SC maps.

masks below.

Mask Localization. To further verify the effectiveness of our method, we compare the localization map with the ground-truth mask and adopt the *Peak-T* and *Peak-IoU* as metrics following SEM [33]. We also report the *Gt-Known Loc. Acc.* of each method. In this section, we only apply the SCG to the baseline methods without involving RAM for fair comparison with SEM. Given the input images, we employ the simplified CAM following ACoL [30] to obtain the coarse localization maps. Fig. 6 visualizes the detailed process of the proposed SCG. All baseline methods adopt our re-implemented models and surpasses the corresponding results of SEM as show in Table 3. The proposed SCG achieves consistent gains on *Peak-IoU*, *Peak-T* and *Gt-Known*. Specifically, we achieves an improvement by 2.3%

Methods	SEM [33]	SCG	Peak-IoU	Peak-T	GT-Known
CAM [34]			53.59	33	64.09
	✓		51.39	74	62.67
		✓	**56.38**	**79**	**66.79**
HaS [18]			54.99	50	65.32
	✓		51.59	79	63.06
		✓	**57.29**	**92**	**68.31**
ACoL [30]			50.89	52	63.28
	✓		48.92	83	59.92
		✓	**52.45**	**132**	**65.45**
CutMix [29]			54.54	34	64.65
	✓		52.02	79	63.84
		✓	**56.96**	**83**	**68.23**
SPG [31]			53.76	33	64.19
	✓		51.74	84	63.08
		✓	**56.40**	**89**	**66.78**
ADL [5]			52.87	29	63.64
	✓		50.01	72	62.38
		✓	**56.01**	**76**	**66.08**

Table 3. Evaluation results of Peak-T, Peak-IoU and GT-Known *Loc Acc* on ILSVRC validation set. All the methods apply Inception V3 as the backbone network.

Methods	ILSVRC(%)			CUB-2011-200(%)		
	M-Ins	Part	More	M-Ins	Part	More
VGG16	10.65	3.85	9.58	-	21.91	10.53
Ours	9.97	2.83	7.66	-	9.25	6.33
InceptionV3	10.36	3.22	9.49	-	23.09	5.52
Ours	9.48	2.89	7.80	-	12.81	6.83

Table 4. Localization error statistics.

compared with the best baseline HaS [18] in terms of *Peak-IoU*. As for *Peak-T*, our results significantly outperforms all the baselines by improving about 50 points on average. The re-implemented SEM based on the code released by the author performs worse than all the baselines.

Error Analysis. To further reveal the effect of our method, we divide the localization error into five cases: classification error (Cls), multi-instance error (M-Ins), localization part
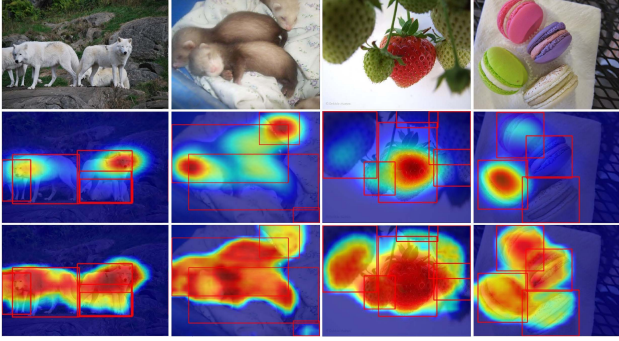
Figure 7. Localization maps of some images from ILSVRC validation set with CAM [34] and the proposed SPA. Ground truth boxes are in red. Both methods cannot separate each instance when there are multiple instances in the scene.

error (Part), localization more error (More), and others (OT). *Part* indicates that the predicted bounding box only cover the parts of object, and IoU is less than a certain threshold. In contrast, *More* indicates that the predicted bounding box is larger than the ground truth bounding box by a large margin. Each metric calculates the percentage of images belonging to corresponding error in the validation/testing set. Table 4 lists localization error statistics of *M-Ins*, *Part*, and *More*. Our method effectively reduce the *M-Ins*, *Part* and *More* errors, which indicates that our localization maps are much accurate. Please refer to supplementary materials for detailed analysis and definitions of each metric.

4.3. Ablation Study

We conduct a series of experiments to verify the effectiveness of the proposed RAM and SCG. Table 5 shows the results on ILSVRC validation set with different configurations. On VGG16, the RAM and SCG improve the baseline by 2% and 1.4%, respectively. It achieves a significant gain of 3.1% when we use both modules simultaneously. On Inception V3, the two modules also achieve remarkable gains, yielding localization error of Top-1 47.29%. In Table 6, we evaluate the performance of the RAM and SCG on CUB-200-2011 testing set. The proposed approach achieves significant improvements. Specifically, the SCG and RAM obtain gains of 11.5% and 8.1% in terms of Top-1 *Loc. Err.* on VGG16 respectively, and it achieves a remarkable improvement of 17.7% when we employ both modules simultaneously. On Inception V3, our method also achieves a significant gain of 9.2% Top-1 *Loc. Err.* The experimental results show that the proposed approach achieves consistent and substantial improvement on different backbones and benchmarks. Please refer to the supplementary materials for more details.

Methods	SCG	RAM	Loc Err.		
			Top-1	Top-5	Gt-Known
VGG16			53.76	42.75	39.21
	✓		51.15	39.57	35.98
		✓	52.33	40.88	37.29
	✓	✓	**50.44**	**38.68**	**34.95**
InceptionV3			49.86	38.86	35.05
	✓		47.38	35.75	31.74
		✓	49.31	38.20	34.29
	✓	✓	**47.27**	**35.73**	**31.67**

Table 5. Localization error on ILSVRC [15] validation set when using different configurations.

Methods	SCG	RAM	Loc Err.		
			Top-1	Top-5	Gt-Known
VGG16			57.49	49.05	46.24
	✓		45.98	35.59	31.33
		✓	49.34	39.14	34.50
	✓	✓	**39.73**	**27.55**	**22.71**
InceptionV3			55.64	44.27	39.14
	✓		48.84	35.80	30.77
		✓	52.04	40.21	35.17
	✓	✓	**46.41**	**33.50**	**27.86**

Table 6. Localization error on CUB-200-2011 [22] test set when using different configurations.

4.4. Limitation

Although the proposed approach achieves much better performance than CAM-based SOTAs, it is challenged when multiple instances come together. Fig. 7 shows localization results with CAM and our approach in the multi-instance scenes. Compared with CAM, our results more precisely cover the object extent. However, due to the lack of instance-level supervision, it is difficult to distinguish different instances. The results in the Table 4 also show that the *M-Ins* error is currently the main source of localization error. The structural information from other images containing only one instance may alleviate this problem. In the future work, it is worth to explore the consistency of structure-preserving across images to achieve *weakly supervised instance localization*.

5. Conclusion

In this study, we unveiled the fact that the spatial structure-preserving is crucial for discovering the localization information contained in convolutional features for WSOL. We accordingly proposed a structure preserving activation (SPA) approach to precisely localize objects. SPA leverages the restricted activation maps to alleviate the structure missing issue of head structure of the classification network. It also utilizes self-correlation generation (SCG) to distill the structure-preserving ability of features for acquiring precise localization maps. In SCG, high-order correlation

is proposed to make up the inability of first-order self-correlation for capturing long-range structural information. Extensive experiments on CUB-200-2011 and ILSVRC benchmarks validated the effectiveness of the proposed approach, in striking contrast with the state-of-the-arts. The structure-preserving activation approach provides a fresh insight to the weakly supervised object localization problem.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 2

[3] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[4] 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. 1, 5, 6

[5] 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. 1, 2, 6, 7

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[7] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[8] Sangheum Hwang and Hyo-Eun Kim. Self-transfer learning for weakly supervised lesion localization. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 239–246, 2016. 2

[9] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[10] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[11] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 2, 3

[12] 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. 2, 6

[13] 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. 1, 5, 6

[14] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 4

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2, 5, 6, 7, 8

[16] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 6

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[18] 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. 1, 2, 6, 7

[19] Lin Song, Yanwei Li, Zeming Li, Gang Yu, Hongbin Sun, Jian Sun, and Nanning Zheng. Learnable tree filter for structure-preserving feature transform. In *Advances in Neural Information Processing Systems*, pages 1711–1721, 2019. 2

[20] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[21] 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. 6

[22] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 5, 6, 8

[23] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[24] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for

weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 2

[25] 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. 1

[26] 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 (CVPR)*, 2018. 2

[27] 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. 1, 2, 6

[28] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. *arXiv preprint arXiv:2006.06668*, 2020. 3

[29] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019. 1, 6, 7

[30] 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. 1, 2, 3, 5, 6, 7

[31] 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 (ECCV)*, pages 597–613, 2018. 2, 5, 6, 7

[32] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 5, 6

[33] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Fei Wu. Rethinking localization map: Towards accurate object perception with self-enhancement maps. *arXiv preprint arXiv:2006.05220*, 2020. 2, 5, 6, 7

[34] 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, 2, 3, 4, 5, 6, 7, 8

Supplementary Material for "Unveiling the Potential of Structure-Preserving for Weakly Supervised Object Localization"

Xingjia Pan[1,3*]   Yingguo Gao[1*]   Zhiwen Lin[1*]   Fan Tang[2†]   Weiming Dong[3,4,5]
Haolei Yuan[1]   Feiyue Huang[1]   Changsheng Xu[3,4,5]
[1]Youtu Lab, Tencent   [2]Jilin University   [3]NLPR, Institute of Automation, CAS
[4]School of Artificial Intelligence, UCAS   [5]CASIA-LLVision Joint Lab
{noahpan, yingguogao, xavierzwlin,harryyuan,garyhuang}@tencent.com
tangfan@jlu.edu.cn, {weiming.dong, changsheng.xu}@ia.ac.cn

1. Quantitative Results

Bounding Box Localization   We here show more results on bounding boxes localization. Table 1 and Table 2 are the localization results on ILSVRC validation and CUB-200-2011 testing sets, respectively. On ILSVRC validation set, we achieve the state-of-the-art with VGG16, and obtain comparable results with current state-of-the-art method I^{2}C [14] with Inception V3. Compared with baseline methods, the proposed SPA is much simpler and almost parameter-free without introducing additional convolutional layers. On CUB-200-2011 testing set, the proposed SPA surpassing all the baseline methods.

Error Analysis.   To further reveal the effect of our method, we define five metrics to perform bad case analysis on the proposed SPA approach. Specifically, we divide the localization error into five cases: classification error (Cls), multi-instance error (M-Ins), localization part error (Part), localization more error (More), and others (OT). For better description, we also define *IoG* and *IoP*. Similar to *IoU*(intersection over Union), IoG means that intersection over ground truth box and IoP means that intersection over predict bounding box.

- *Cls* includes the predictions that are wrongly classified.

- *M-Ins* indicates that the prediction intersects with at least two ground truth boxes, and IoG > 0.3.

- *Part* indicates that the predicted bounding box only cover the parts of object, and IoB > 0.5.

- *More* indicates that the predicted bounding box is larger than the ground truth bounding box by a large margin, and IoG > 0.7.

---

*Equal contribution
†Corresponding author

- *OT* indicates other predictions that do not fall under above mentioned cases.

The five cases defined above are mutually exclusive. We show the details of the definition of five metrics in Algorithm 1 Each metric calculates the percentage of images

---

Algorithm 1 Error Analysis Algorithm.

Input: Predicted bounding box $B_p \in \mathbb{R}^{1\times 4}$; Predicted label C_p; ground truth boxes list $B_g \in \mathbb{R}^{N\times 4}$; ground truth label C_g; Maximum IoU $I\hat{o}U$ between the B_p and B_g;
Output: Cls, M-Ins, Part, More, OT;
1: Set Cls = M-Ins = Part = More = OT = 0
2: **if** $C_p \neq C_b[0]$ **then**
3:      Cls =1
4:      **return** Cls, M-Ins, Part, More, OT
5: **end if**
6: **if** $I\hat{o}U > 0.5$ **then**
7:      **return** Cls, M-Ins, Part, More, OT
8: **end if**
9: Calculate $IoG \in \mathbb{R}^{1\times N}$ between the B_p and B_g
10: **if** $Count(IoG > 0.3) > 1$ **then**
11:      M-Ins = 1
12:      **return** Cls, M-Ins, Part, More, OT
13: **end if**
14: Calculate the maximum $I\hat{o}G$ and $I\hat{o}P$
15: **if** $I\hat{o}P > 0.5$ **then**
16:      Part = 1
17:      **return** Cls, M-Ins, Part, More, OT
18: **end if**
19: **if** $I\hat{o}G > 0.7$ **then**
20:      More = 1
21:      **return** Cls, M-Ins, Part, More, OT
22: **end if**
23: OT = 1
24: **return** Cls, M-Ins, Part, More, OT

Methods	Backbone	Loc Err.			Cls Err.	
		Top-1	Top-5	Gt-Known	Top-1	Top-5
Backprop [5]	VGG16 [6]	61.12	51.46	-	-	-
CAM [15]	VGG16 [6]	57.20	45.14	-	31.2	11.4
CutMix [11]	VGG16 [6]	56.55	-	-	-	-
ADL [1]	VGG16 [6]	55.08	-	-	-	-
ACoL [12]	VGG16 [6]	54.17	40.57	37.04	32.5	12.0
I^{2}C [14]	VGG16 [6]	52.59	41.49	36.10	30.6	10.7
MEIL [3]	VGG16 [6]	53.19	-	-	29.73	-
Ours	VGG16 [6]	**50.44**	**38.68**	**34.95**	**29.49**	**9.95**
CAM [15]	InceptionV3 [8]	53.71	41.81	37.32	26.7	8.2
SPG [13]	InceptionV3 [8]	51.40	40.00	35.31	30.3	9.9
ADL [1]	InceptionV3 [8]	51.29	-	-	27.2	-
ACoL [12]	GoogLeNet [7]	53.28	42.58	-	29.0	11.8
DANet [10]	GoogLeNet [7]	52.47	41.72	-	27.5	8.6
MEIL [3]	InceptionV3 [8]	50.52	-	-	26.69	-
I^{2}C [14]	InceptionV3 [8]	**46.89**	35.87	**31.50**	26.7	8.4
GC-Net [2]	InceptionV3 [8]	50.94	41.91	-	**22.6**	**6.4**
Ours	InceptionV3 [8]	47.29	**35.73**	31.67	26.74	8.19

Table 1. Comparison between our method and the state-of-the-art on the ILSVRC [4] validation set. Our method outperforms all other methods by a large margin for object localization. Here 'ClsErr', 'LocErr' and 'Gt-Known' are short for classification error, location error and Gt-known location error, respectively.

Methods	Backbone	Loc Err.			Cls Err.	
		Top-1	Top-5	Gt-Known	Top-1	Top-5
CAM [15]	GoogLeNet [7]	58.94	49.34	44.9	26.2	8.5
SPG [13]	GoogLeNet [7]	53.36	42.8	-	-	-
DANet [10]	InceptionV3 [8]	50.55	39.54	33.0	28.8	9.4
ADL [1]	InceptionV3 [8]	46.96	-	-	**25.45**	-
Ours	InceptionV3 [8]	**46.41**	**33.50**	**27.86**	26.49	8.61
CAM [15]	VGG16 [6]	55.85	47.84	44.0	**23.4**	**7.5**
ADL [1]	VGG16 [6]	47.64	-	-	34.73	-
ACoL [12]	VGG16 [6]	54.08	43.49	45.9	28.1	-
DANet [10]	VGG16 [6]	47.48	38.04	32.3	24.6	7.7
SPG [13]	VGG16 [6]	51.07	42.15	41.1	24.5	7.9
I^{2}C [14]	VGG16 [6]	44.01	31.6	-	-	-
MEIL [3]	VGG16 [6]	42.54	-	-	25.23	-
Ours	VGG16 [6]	**39.73**	**27.5**	**22.71**	23.89	7.85

Table 2. Comparison between our method and the state-of-the-art on the CUB-200-2011 test set. Here 'ClsErr', 'LocErr' and 'Gt-Known' are short for classification error, location error and Gt-known location error, respectively.

Methods	ILSVRC(%)					CUB-2011-200(%)				
	Cls	M-Ins	Part	More	OT	Cls	M-Ins	Part	More	OT
VGG16	29.47	10.65	3.85	9.58	0.2	23.28	-	21.91	10.53	1.76
Ours	29.49	9.97	2.83	7.66	0.5	23.89	-	9.25	6.33	0.26
InceptionV3	26.73	10.36	3.22	9.49	0.20	26.39	-	23.09	5.52	0.64
Ours	26.74	9.48	2.89	7.80	0.37	26.49	-	12.81	6.83	0.03

Table 3. Localization error statistics.

belonging to corresponding error in the validation/testing set. Table 3 lists localization error statistics. Localization error *Cls* caused by wrong classification dominates the error. But the classification is not the main concern for WSOL. We mainly focus on *M-Ins*, *Part*, and *More* metrics. For ILSVRC, the proposed SPA effectively reduces the three

kinds of localization error. Although there is no explicit solution for multi-instance problem, our method reduces the *M-Ins* error. From Table 3, it shows that the *M-Ins* and *More* are the most important problems to solve. It it worth to explore CAM-based methods for *weakly supervised instance localization* in future work. For CUB-200-2011, there is only one instance in each image. Our method effectively reduces *Part* and *More* errors on VGG16, which indicates that our localization maps are much accurate. On Inception V3, the proposed SPA significantly reduce *Part* by 10.2%.

2. More Examples

Self-correlation Maps. We here show more visualization examples for self-correlation maps in Fig. 1, Fig. 2, Fig. 3, Fig. 4. In each figure, we show five self-correlation maps for corresponding points marked by green cross in original images. The middle row shows the first-order self-correlation maps and the bottom row shows the second-order self-correlation maps.

localization Results. We also show more visualization examples of localization maps in Fig. 5, Fig. 6, Fig. 7. In each figure, the original images with ground truth bounding boxes (in red) are shown in top row. The localization maps with CAM and the proposed SPA are shown in middle and bottom row, respectively.

References

[1] 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] 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.

[3] 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.

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

[5] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

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

[7] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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 (ECCV)*, pages 597–613, 2018.

[14] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[15] 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.
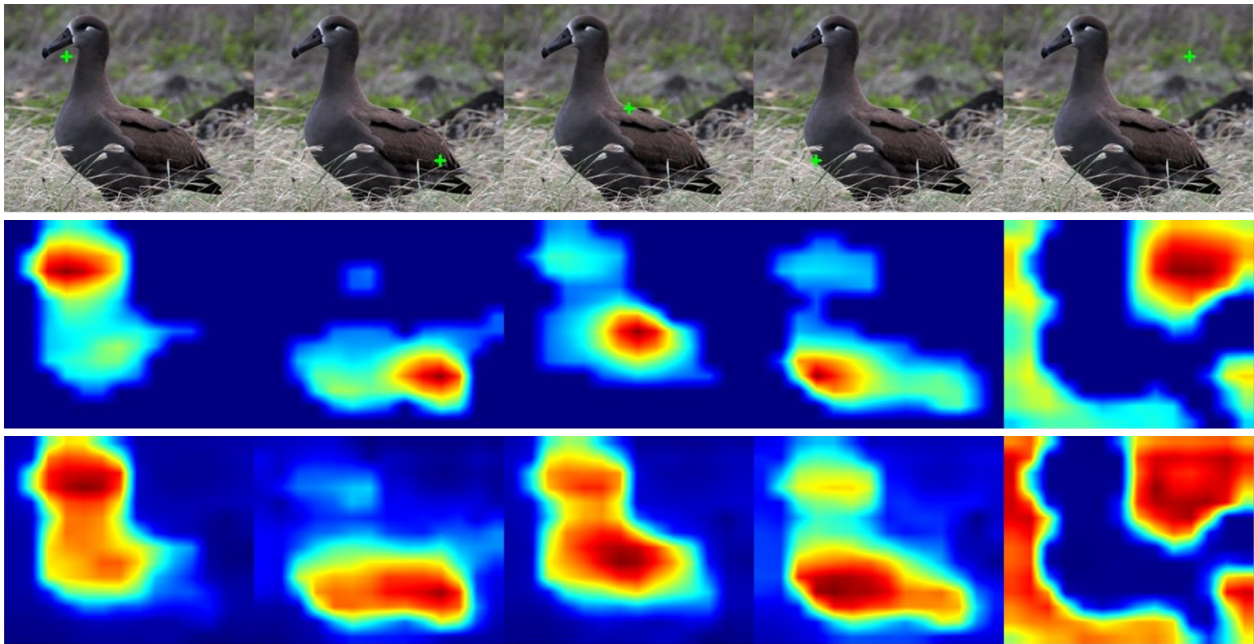
Figure 1. Visualization of the self-correlation maps with first-order similarity (middle row) and second-order similarity(bottom row). The images are from the CUB-200-2011 [9] testing set.
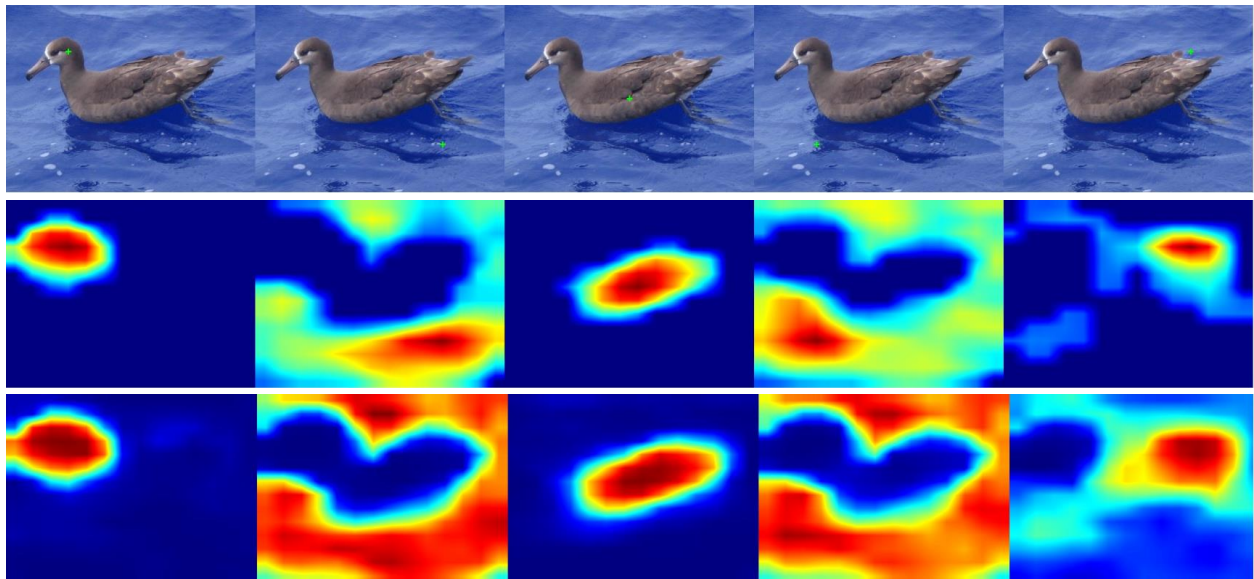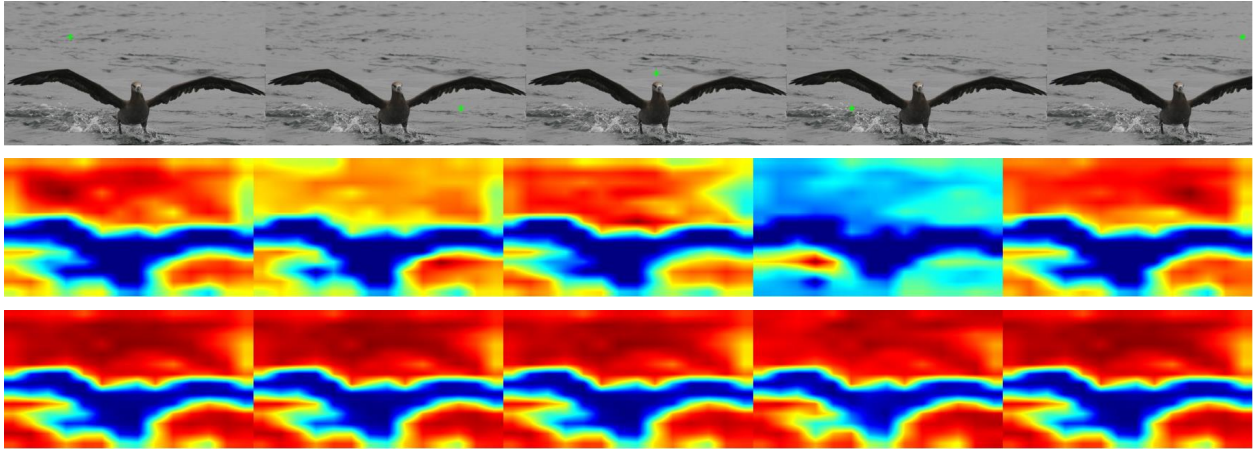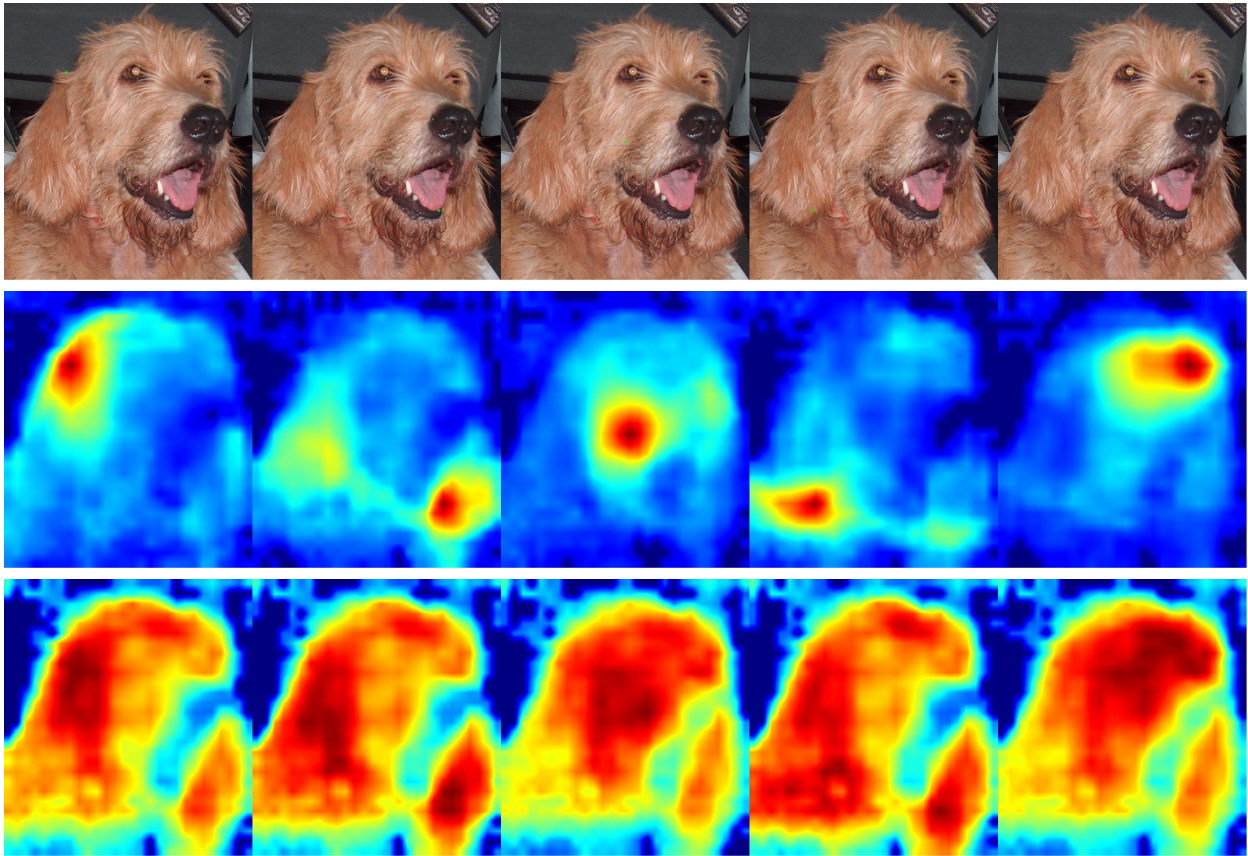

Figure 2. Visualization of the self-correlation maps with first-order similarity (middle row) and second-order similarity(bottom row). The images are from the CUB-200-2011 [9] testing set.

Figure 3. Visualization of the self-correlation maps with first-order similarity (middle row) and second-order similarity(bottom row). The images are from the CUB-200-2011 [9] testing set.

Figure 4. Visualization of the self-correlation maps with first-order similarity (middle row) and second-order similarity(bottom row). The images are from the ILSVRC [4] validation set.

Figure 5. Visualization of the localization maps with CAM [15] (middle row) and the proposed SPA(bottom row). The images are from the ILSVRC [4] validation set.
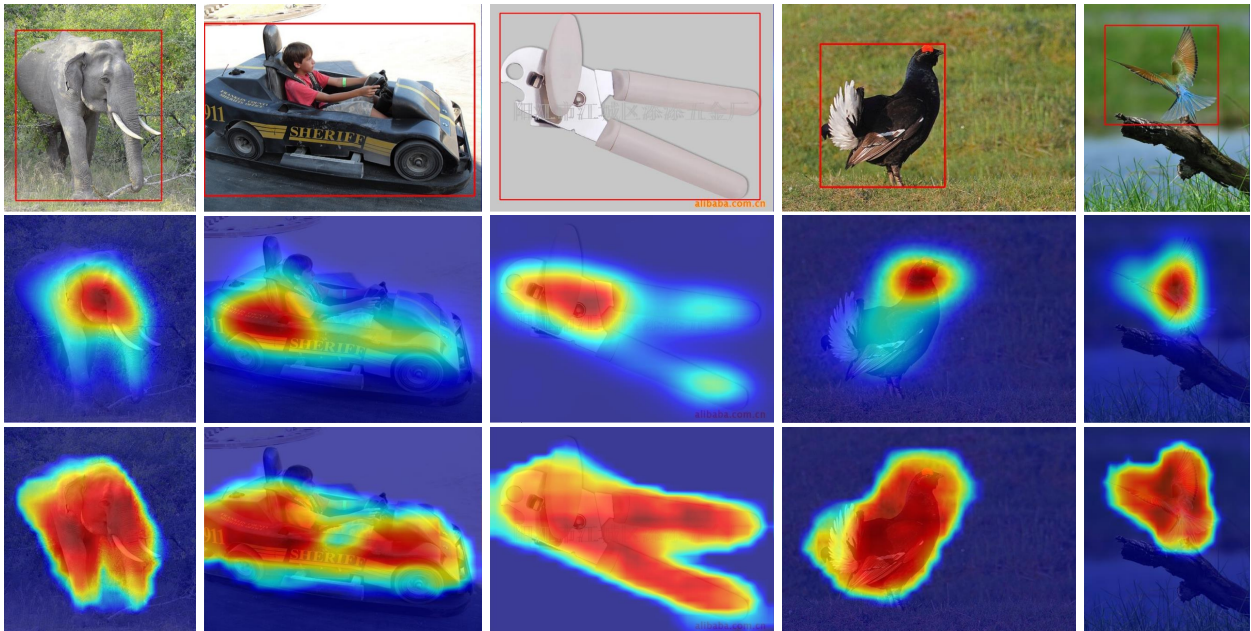

Figure 6. Visualization of the localization maps with CAM [15] (middle row) and the proposed SPA(bottom row). The images are from the ILSVRC [4] validation set.
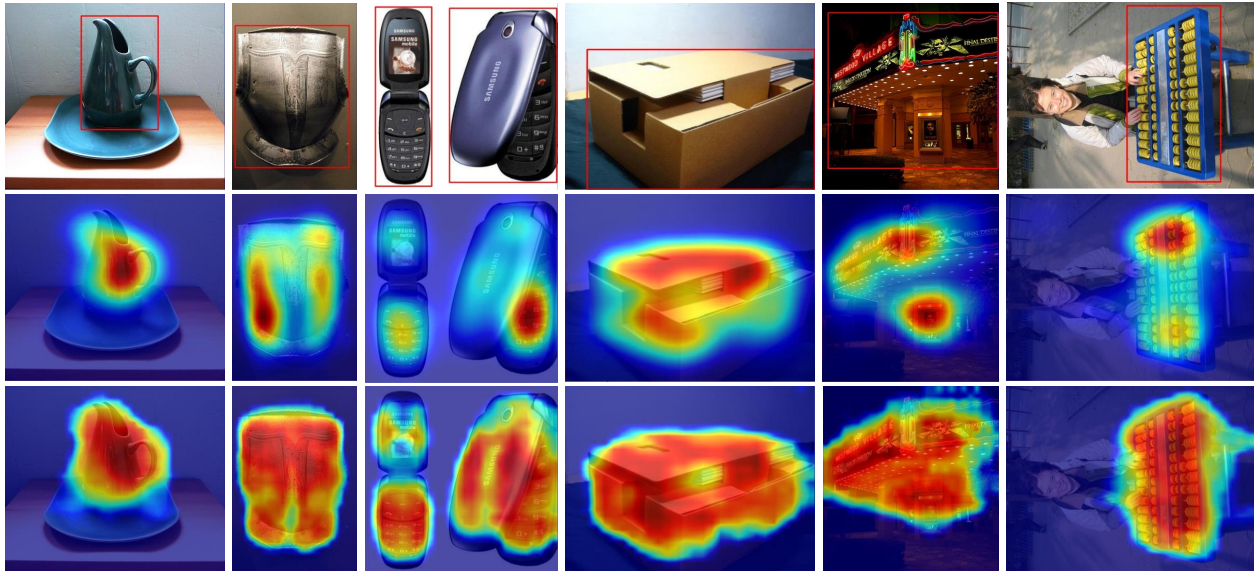
Figure 7. Visualization of the localization maps with CAM [15] (middle row) and the proposed SPA(bottom row). The images are from the ILSVRC [4] validation set.