SPA^{2}Net: Structure-Preserved Attention Activated Network for Weakly Supervised Object Localization

Dong Chen, Xingjia Pan, Fan Tang, Weiming Dong, *Member, IEEE*, and Changsheng Xu, *Fellow, IEEE*

***Abstract*— By exploring the localizable representations in deep CNN, weakly supervised object localization (WSOL) methods could determine the position of the object in each image just trained by the classification task. However, the partial activation problem caused by the discriminant function makes the network unable to locate objects accurately. To alleviate this problem, we propose Structure-Preserved Attention Activated Network (SPA^{2}Net), a simple and effective one-stage WSOL framework to explore the ability of structure preservation of deep features. Different from traditional WSOL approaches, we decouple the object localization task from the classification branch to reduce their mutual influence by involving a localization branch which is online refined by a self-supervised structural-preserved localization mask. Specifically, we employ the high-order self-correlation as structural prior to enhance the perception of spatial interaction within convolutional features. By succinctly combining the structural prior with spatial attention, activations by SPA^{2}Net will spread from part to the whole object during training. To avoid the structure-missing issue caused by the classification network, we furthermore utilize the restricted activation loss (RAL) to distinguish the difference between foreground and background in the channel dimension. In conjunction with the self-supervised localization branch, SPA^{2}Net can directly predict the class-irrelevant localization map while prompting the network to pay more attention to the target region for accurate localization. Extensive experiments on two publicly available benchmarks, including CUB-200-2011 and ILSVRC, show that our SPA^{2}Net achieves substantial and consistent performance gains compared with baseline approaches. The code and models are available at https://github.com/MsterDC/SPA2Net.**

***Index Terms*— High-order self-correlation, class activation map, structure preservation, weakly supervised object localization.**

I. Introduction

DIFFERENT from the task of object localization under full supervision, weakly supervised object localization (WSOL) aims to infer the position information of the target in an image by utilizing image-level annotation. Since category-related annotations are easier to obtain, WSOL can be implemented without intensive annotations, reducing the high cost of labeling and gaining much attention.

As a representative method for WSOL, class activation mapping (CAM) [1] uses the intermediate classifier activation to discover discriminative regions for object localization [2]. Nonetheless, CAM is afflicted with significantly underestimating object regions. This arises because the discriminative regions activated through the classification models are often much smaller than the actual extent of the objects. Several class activation mapping-based solutions [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] have been proposed, mainly emphasizing the expansion of activation regions. However, some of these works [8], [9], [10] may lead to the loss of crucial information, especially when the target and background have similar textures or colors, and they often include background noise and ignore the structural information related to the object [4], [7], [12], [13].

Experiments revealed that the head structure of the classification network causes the CAM to overlook important object structure information found in convolutional features. As shown in the second row of Fig. 1, traditional CAMs tend to pay more attention to the head region of the bird while ignoring the fact that body parts occupy a larger proportion of the image. The network driven by sole classification loss tends to activate a small proportion of features with the largest discriminative capability while depressing the majority of object extent. Moreover, many CAM-based methods [14] use global average pooling atop feature maps, treating each pixel equally and making it more difficult to differentiate true objects from noisy backgrounds. Therefore, the lack of structural information is one of the main reasons why classification networks are difficult to locate.

The inherent spatial correlation reveals the structural information of an image by measuring the similarity of local patches [15]. Early works used self-correlation itself as a robust descriptor for visual correspondence [16], object detection [17] and action recognition [18], [19]. Following the self-attention mechanism, some recent works [20], [21], [22] 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 structural information [23].

In this study, we propose a one-stage WSOL network, termed structure-preserved attention activated net-

Manuscript received 17 May 2022; revised 6 August 2023 and 22 September 2023; accepted 30 September 2023. Date of publication 17 October 2023; date of current version 24 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62102162, Grant U20B2070, Grant 61832016, and Grant 61832002; and in part by the Beijing Natural Science Foundation under Grant L221013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ioannis (Yiannis) Kompatsiaris. *(Corresponding author: Fan Tang.)*

Dong Chen is with the School of Artificial Intelligence, Jilin University, Changchun 130012, China (e-mail: chendong21@mails.jlu.edu.cn).

Xingjia Pan is with Momenta, Beijing 100018, China (e-mail: panxingjia@momenta.ai).

Fan Tang is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: tangfan@ict.ac.cn).

Weiming Dong and Changsheng Xu are with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: weiming.dong@ia.ac.cn; changsheng.xu@ia.ac.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2023.3323793, provided by the authors.

Digital Object Identifier 10.1109/TIP.2023.3323793

1941-0042 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

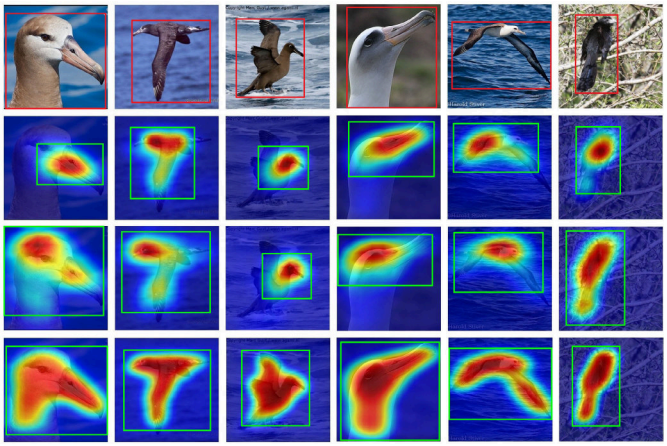


Fig. 1. Visual comparisons of class activation maps (CAM) by different methods. From top to bottom: input images, localization results by original CAM [1], improved CAM [11] and SPA^{2}Net. By fusing high-order correlation information into attention weight, object region can be activated accurately by our approach.

work (SPA^{2}Net), for accurate object localization using image-level category labels as supervision in conjunction with self-supervised pseudo masks. SPA^{2}Net is distinct from two-stage WSOL methods [24], [25] as it directly performs forward inference on input images and outputs object masks once trained, due to its self-supervised trained localization head. We formulate the concept of high-order self-correlation to model the long-range structural information of objects and propose structure-preserved attention for generating complete activation maps. To eliminate the influence of local discriminative activation, a localization branch is added which is supervised by self-generated pseudo masks based on the self-correlation map. Besides, a restricted activation loss is further designed for the classification branch by suppressing the response value range of CAM and differentiating objects from backgrounds. In general, the contributions of this study include:

- We unveil that spatial structure preserving is crucial for discovering the localization information contained in convolutional features for WSOL by formulating the concept of high-order self-correlation (HSC).
- An one-stage WSOL framework named SPA^{2}Net is developed, which directly predicts a localization map supervised by the self-generated pseudo mask.
- A structure-preserved attention is proposed to locate objects more accurately by aggregating HSC as structural prior to the attention mechanism.
- The proposed SPA^{2}Net shows consistent and substantial gains across CUB-200-2011 and ILSVRC datasets for bounding box and mask localization.

Based on the conference version of this work [23], we have made significant extensions in this paper. First, inspired by the improved structural preservation capacity of the higher-order self-correlations, we suggest aggregating HSC with a self-attention mechanism to reinforce the spatial interaction capacity within features and propose a structure-preserved attention (SPA) module. Formally, the attention weight can be aggregated with high-order self-correlations. In our experiments, we discovered that the SPA module effectively diverges the activation region and mitigates the issue of partial activation in the classification network. Secondly, we argue that despite the self-correlation map obtaining performance gains through offline processing in previous work, certain problems still exist: 1) The class activation map and self-correlation map are generated offline, and thus the localization map cannot be flexibly refined online. 2) Target boundary positions are still impacted by the weak response, resulting in threshold sensitivity. To address the above issues, we generate a pseudo object mask based on the HSC and the restricted activation loss and construct a localization branch for self-supervised training. The pixel-to-pixel surveillance from pseudo mask ensures that the network can produce a threshold robust localization map. Based on the above efforts, we propose a one-stage localization framework termed SPA^{2}Net. The key distinctions between SPA^{2}Net and previous work can be summarized as follows: 1) Network Architecture: SPA^{2}Net extends the classification network with new localization components (SPA module, self-generated map, and a localization branch), whereas previous work follows a standard classification network structure. 2) Objective: SPA^{2}Net includes three terms in its total optimization objective (classification loss, restricted activation loss, and localization loss), while previous work includes only the first two terms. 3) End-to-End Framework: SPA^{2}Net is an end-to-end one-stage WSOL framework, directly outputting localization maps through its localization branch. In contrast, previous work accomplishes target localization through the SCG algorithm in an offline post-processing manner. Experimental results show that SPA^{2}Net achieves about 7% and 2% absolute improvement compared to the previous method on CUB-200-2011 [26] and ILSVRC2016 [27] while speeding up the inference speed by approximately two times. The supplementary material provides additional information on the differences between the two methods. Furthermore, more ablation studies are presented to investigate the critical properties of our methods.

In summary, this paper proposes an end-to-end WSOL framework that addresses local activation issues in classification networks by combining high-order self-correlation and self-attention mechanisms. Through self-supervised training, the model directly predicts target masks, achieving threshold-robust object localization. In the remainder of this paper, Section II covers recent WSOL advancements, Section III details the proposed SPA^{2}Net framework, and Section IV presents experimental results demonstrating its superiority.

II. Related Work

Weakly Supervised Object Localization (WSOL) aims to learn the localization of objects with only image-level labels. As one of the most influential works, CAM [1] produces localization maps by aggregating deep feature maps using global average pooling (GAP) and a class-specific fully connected layer. Following CAM, Hwang and Kim [28] proposed to simplify CAM by removing the last fully connected layer.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

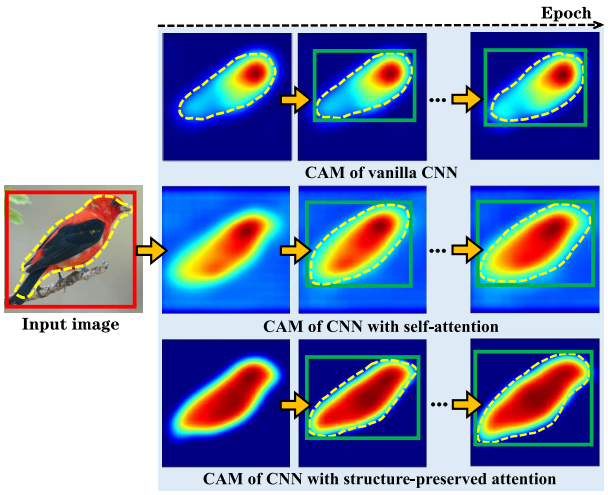


Fig. 2. Snapshots of CAMs obtained by different methods during training and each is optimized only by the cross-entropy loss. In the 1st row, there is a partial activation problem in the classification network. Using a self-attention module can diverge activation, however, the boundary positions are ambiguous (the 2nd row). Aggregating HSC (the 3rd row) with self-attention is able to preserve the structure effectively. Please refer to Sec. III-C for more details.

Although the CAM-based method provides a simple and effective way to solve the WSOL problem, it is also plagued by partial activation problems. To solve this issue, HaS [3] and CutMix [3] adopted an erasing-based strategy to force the network to focus on more relevant parts of objects by removing different parts of images. Furthermore, ACoL [10] and ADL [29] chose to erase the feature map corresponding to the discriminative region and trained multiple parallel classifiers adversarially. Different from the above methods, SPG [6] and I^{2}C [21] introduced the constraint of pixel-level correlations into the network for localization. DANet [4] applied a divergent activation to learn complementary and discriminative visual patterns for WSOL. Zhang et al. [22] proposed to refine the localization maps using the point-wise similarity within seed regions. Lu et al. [30] designed a novel architecture called GC-Net to predict a set of location coefficients. Since only focuses on extended activation, the above method ignores the contradiction between classification and object localization. Leveraging a classification network to activate and localize the complete range of objects is still an open question.

Recently, many works aim to address the issue where classification networks focus only on the more discriminative parts of the target region rather than the entire object region. Meng et al. [31] introduced the Part Discovery Model (PDM) for precise object localization and classification. Their approach involves modeling robust object parts, which aids in learning complementary spatial information and local details for WSOL. Xie et al. [25] proposed a two-stage method C^{2}AM to generate class-agnostic activation maps using cross-image foreground-background contrastive learning. However, the efficiency of C^{2}AM is limited by its two-stage processing method, which heavily depends on high-quality pseudo-bounding boxes and requires robust pre-trained models. Zhu et al.proposed BagCAM [32], a plug-and-play mechanism that derives regional localizers from a well-trained classifier, that can identify object-related factors for each spatial position. Additionally, Zhu et al. [33] formulated WSOL as a domain adaptation task, using a score estimator trained on the image domain to locate objects in the pixel domain. Our method achieves threshold-robust localization results through self-supervised training and online mask generation. SPA^{2}Net enhances deep convolutional features' structural preservation and introduces a novel structure-preserved attention module to capture spatial relations among objects by aggregating high-order correlations.

Spatial Interaction: Self-Correlation and Self-Attention. Several recent works [20], [21], [22] have employed pixel similarity to enhance features or activation maps following the self-attention mechanism [34], [35], [36]. Wang et al. [37] proposed a non-local block to capture long-range dependency within image pixels. Cao et al. [38] found that the global contexts modeled by non-local networks are almost the same for query positions, thereby proposing NLNet [37] with SENet [39] for global context modeling. Song et al. [40] proposed the learnable tree filter to leverage the structural property of a minimal spanning tree to model long-range dependencies. DNL [41] disentangled the non-local block into a whitened pairwise term and a unary term to facilitate the learning process. These methods use first-order self-correlation and stack multiple modules at different stages to capture long-range context. However, they can only retain local structural information for a single feature layer. Furthermore, attention acts as a mechanism to prioritize the allocation of available resources to the most informative parts of an input. The non-local operation [42] was introduced to CNNs in a self-attention manner so that the response at each position is a weighted sum of the features at all (global) positions. Ramachandran et al. [43] verified that self-attention is an effective stand-alone layer for CNNs. Recently, many works [44], [45], [46] introduced a cascaded self-attention mechanism in the transformer model to capture long-range feature dependency. Chen et al. [47] proposes a high-order attention (HOA) module to model and use the complex high-order statistical information in the attention mechanism to solve person re-identification problems. In this work, we use high-order self-correlation of deep features as structural prior to aggregating with the low-level similarity performed in self-attention to enhance the structure preservation ability.

Self-Supervised Learning (SSL) acquires general-purpose image features directly from raw pixel data without manual supervision. These learned networks serve as foundational backbones for various tasks like classification, object detection, and image segmentation [48], [49], [50], [51], [52]. CNN-based SSL has showcased remarkable performance, surpassing fully-supervised pre-training methods [53], [54] on tasks with a restricted number of labeled samples. Recent works have been made by combining SSL with Transformer and self-attention architectures. Caron et al. [55] finds that self-supervised Vision Transformer (ViT) features contain explicit information about the semantic segmentation of an image, which does not emerge as clearly with supervised ViTs, nor with convnets. Aiming to improve the efficiency

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

of Transformer-based SSL, Li et al. [56] presents EsViT by using a multi-stage architecture and a region-based pre-training task for self-supervised representation learning. In the realm of unveiling spatial correlations, the findings of DINO [55] reveal that self-supervised training of ViT features explicitly incorporates scene layout, particularly object boundaries. Interestingly, our approach likewise uncovers that CNNs trained with structural preservation capabilities can also focus on object regions, as illustrated in Fig. 1 and Fig. 2. Differing from learning generic visual representations through self-supervision, our work capitalizes on self-attention and high-order self-correlation to delve into spatial interactions within deep-level convolutional features. These spatial interactions are then harnessed in a supervised training process for the localization branch, extracting target region information from the classification network.

III. Framework

A. Overview

In this section, we start by introducing the problem setting of the weakly supervised object localization (WSOL) task and the pipeline of the proposed SPA^{2}Net, then move on to the definition of high-order self-correlation (Sec. III-B). Finally, we propose the structure-preserved attention (SPA) module (Sec. III-C) and self-supervised localization branch on the basis of high-order self-correlation (HSC) (Sec. III-D). The training details are involved in Sec. IV.

For a set of images with image-level annotations, the goal of WSOL is to locate objects of the specified class. Given the feature map $\mathbf{f} \in \mathbb{R}^{HW \times K}$ of an image, $\mathbf{f}_k$ denotes k-th channel of $\mathbf{f}$ where $k \in (1, \ldots, K)$ and HW represents the spatial size of the feature map. By passing through a GAP layer, each $\mathbf{f}_k$ will be averaged and outputs a pooled feature vector v as follows:

$$v_k = \frac{1}{HW} \sum_{h,w} \mathbf{f}_k^{(h,w)}, \tag{1}$$

where v_k denotes a scalar of v at k-th channel. The feature vector v is then transformed into C-dim logits by a fully connected layer where C is the number of classes. The class activation map $\mathbf{M}_c^{cam}$ corresponding to the class c is calculated in the following form:

$$\mathbf{M}_c^{cam} = \sum_{k=1}^{K} w_{k,c} \cdot \mathbf{f}_k, \tag{2}$$

where $\mathbf{M}_c^{cam} \in \mathbb{R}^{H \times W}$ and $w_{k,c}$ are the weights of fully connected layer. Then, the $\mathbf{M}_c^{cam}$ is usually first resized to the input image size, and a threshold is used to convert the scaled activation map into a binary mask for localization. Typically, classification networks use cross-entropy loss as a discriminative constraint, which makes the network pay more attention to local regions of features, leading to challenging partial activation problems. As shown in the first row of Fig. 2, the activation map of vanilla CNN could not cover the whole object and is only highly activated in the head region.

The overview of the proposed structure-preserved attention activated network (SPA^{2}Net) can be divided into three parts: structure-preserved attention module, classification branch, and localization branch, as shown in the Fig. 3. The localization branch and the classification branch share a similar network structure, with three convolutional layers and two ReLU layers. The distinction lies in the final convolutional layer's output: the classification branch produces feature channels equal to the number of classes, while the localization branch outputs a single-channel localization map. To enlarge the activation region, we try to append a self-attention module (green box in Fig. 3) after the vanilla CNN. We then introduce the HSC (bottom blue box) into the self-attention module, composing the SPA module (gray box) to activate objects more precisely. In Fig. 2, the class activation maps show a considerable enlargement of the activation area with the addition of self-attention. However, this also leads to increased values in the non-object area. The limitation arises from self-attention's reliance on first-order similarity, which struggles to establish accurate pixel-wise semantic connections due to noise interference. In the third row of Fig. 2, replacing self-attention with our SPA module results in clearer object activation areas and effective suppression of background noise (indicated by darker blue regions).

Furthermore, we propose a restricted activation loss (L_{RA} in Fig. 3) in the classification branch, which distinguishes foreground and background driven by channel variance while avoiding being misled by locally abnormally high responses. To further alleviate the local activation problem of CAM, we add a localization branch to directly refine the localization map in a self-supervised manner based on the HSC, while preserving the structure of the object. As shown in Fig. 3, the HSC will be used as a structural prior for generating high-quality pseudo masks (M^{sgm}) to guide the localization branch to output robust localization maps (M^{loc}). In the inference stage (shown in Fig. 4), the localization branch directly outputs the localization map, and our results demonstrate that the method without additional post-processing is concise and effective.

B. High-Order Self-Correlation

We first analyze the first-order correlation and introduce the concept of spatial **H**igh-order **S**elf-**C**orrelation (HSC) which could capture the long-range structural information based on the rich context of the object.

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

$$S(F_i, F_j) = \frac{\theta(F_i)^{\mathrm{T}} \theta(F_j)}{||\theta(F_i)|| \cdot ||\theta(F_j)||}, \tag{3}$$

where $i, j \in \{0, 1, \ldots, HW-1\}$ indicate the index of features, and $F_i, F_j \in \mathbb{R}^K$ are the feature vectors. In experiments, the embedding function θ is removed for simplicity. We define the *first-order self-correlation* of F as:

$$SC^{1st}(F) = ReLU(S(F_i, F_j)). \tag{4}$$

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

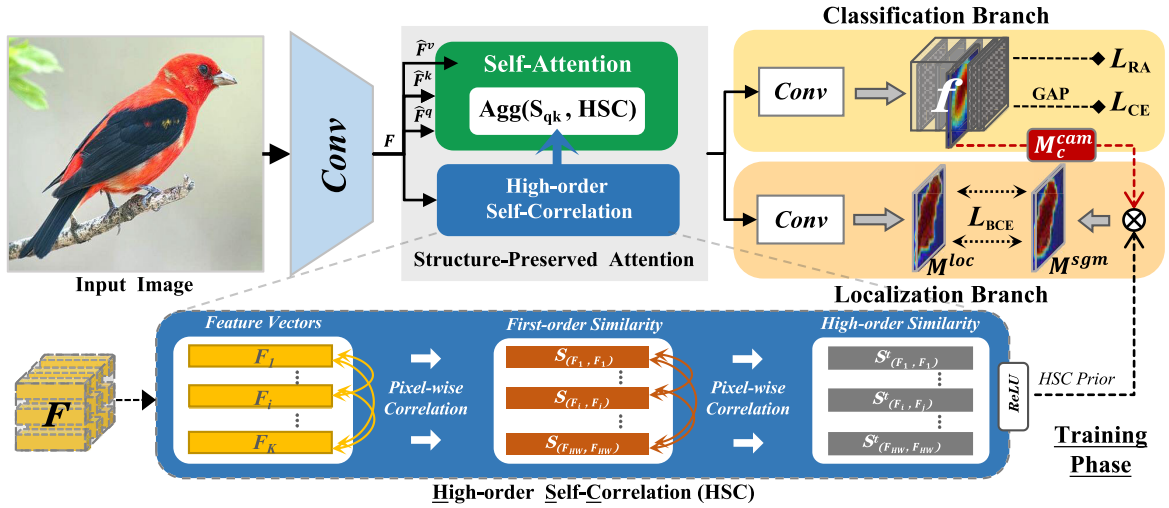


Fig. 3. Overview of the proposed SPA^{2}Net. The SPA module is composed of two parts: the self-attention module and the HSC module. The HSC is used to aggregate attention weights S_{qk} and acquire high-quality pseudo object masks $\hat{M}^{sgm}$ by coupling with the class activation map (M_c^{cam}). The restricted activation loss L_{RA} roughly extracts the target region from the background under the guidance of channel variance and excavates the structure preservation ability in a network. Afterward, the localization branch is guided by the self-generated map (M^{sgm}) for robust object localization.

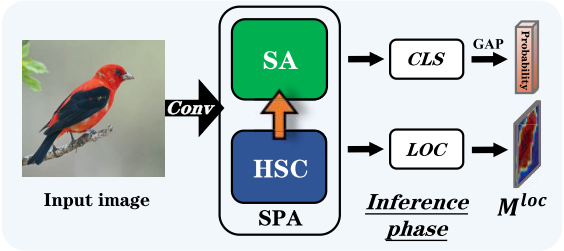


Fig. 4. Inference phase of the proposed SPA^{2}Net. The network directly outputs a category-independent localization map based on the HSC. The *CLS* and *LOC* in the figure represent the classification branch and the localization branch, respectively.

The similarities $S(\cdot,\cdot)$ are activated by ReLU [57] to suppress negative values and $SC^{1st}(F) \in \mathbb{R}^{HW\times HW}$. Due to large appearance variations, the pixels within an object are usually dissimilar, and the low-order self-correlation can capture the structural relationship in a small local area.

b) High-order self-Correlation: Based on the first-order self-correlation, we further explored and defined higher-order self-correlation and used the better inherent structure-preserving capabilities of the network to achieve accurate WSOL.

Specifically, we first define the t_{th}-order similarity between F_i and F_j:

$$S^t(F_i, F_j) = \frac{\sum_{u^0}\cdots\sum_{u^{t-2}} S(F_i, F_{u^0})\cdots S(F_{u^{t-2}}, F_j)}{(HW)^{t-1}}, \tag{5}$$

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

$$\hat{S}^t(F_i, F_j) = \frac{S^t(F_i, F_j) - min_{u\in\Omega} S^t(F_i, F_k)}{max_{u\in\Omega} S^t(F_i, F_k) - min_{u\in\Omega} S^t(F_i, F_k)}, \tag{6}$$

then, we define the t_{th}-order self-correlation as:

$$SC^t(F) = \left[ReLU(\hat{S}^t(F_i, F_j))|_{i,j}\right]. \tag{7}$$

When t is larger than 1, we refer to the corresponding SC^t as HSC. Compared to SC^{1st}, HSC preserves details of the object better when considering long-range contexts. However, we observed that exclusively employing high orders could potentially introduce additional noise. To solve the problem, we fuse SC^{1st} to SC^{2nd} by element-wise maximum operation $Max(SC^{1st}, SC^{2nd})$ in our method. More qualitative analysis for HSC can be found in Sec. IV-C.

C. Structure-Preserved Attention

When continuously approximating the target distribution, the cross-entropy loss seeks to create a hyper-plane to discriminate samples belonging to different categories, and the hyper-plane will expand to include more samples. In this phase, the network will minimize unneeded portions that may contain background noise or are semantically irrelevant. Many works attempt to cover more regions, but they typically ignore the structure of objects. We suggest that the problem can be solved by combining self-attention with the high-order structural priors HSC and propose a structure-preserving attention (SPA) module. Specifically, we feed the deep features $F \in \mathbb{R}^{HW\times K}$ extracted by CNN into a self-attention module, where H and W represent the height and width, K indicates the number of channels. Then, the attention module employs N attention heads and various parameterized linear layers to transform F into various features: $\hat{F}^q$, $\hat{F}^k$ and $\hat{F}^v$, and each from a different head is concatenated, where $\hat{F}^* \in \mathbb{R}^{N\times HW\times K}$. For $\hat{F}^q$ and $\hat{F}^k$, scaled dot product was chosen in our experiments for the compatibility function to compare the two features:

$$s_{qk} = \frac{\hat{F}^q \cdot \left(\hat{F}^k\right)^T}{\sqrt{K}}, \tag{8}$$

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

where $s_{qk} \in \mathbb{R}^{N\times HW\times HW}$. Afterward, softmax and a dropout layer are utilized to calculate the weight coefficient α_{qk}:

$$\alpha_{qk} = Dropout(softmax(s_{qk})). \tag{9}$$

The output feature $\hat{F}^o$ is computed as weighted sum by another transformed feature $\hat{F}^v$:

$$\hat{F}^o = \alpha_{qk} \cdot \hat{F}^v, \tag{10}$$

where $\hat{F}^o \in \mathbb{R}^{N\times HW\times K}$. The output is linearly transformed for classification and localization in both branches. However, low-level similarity (Eq. 8) has limited capacity to quantify the relations between features, notably when they are not close enough in semantic space. Since higher-order correlations preserve structure better than low-order similarities, we aggregate the high-order structure prior (*i.e.* HSC) to extract the object effectively:

$$s_{qk} \leftarrow Agg(s_{qk}, HSC), \tag{11}$$

where $Agg(\cdot)$ represents the element-wise summation operation between the attention weight s_{qk} and the HSC. We believe that directly inputting deep-level features into the self-attention module can result in the loss of object structure information and accumulating errors as training progresses. This loss is most noticeable at object boundaries, leading to heightened responses in certain background regions (as shown in the second row of Fig. 2). Deep-level features, on the other hand, effectively preserve the target's structure. Thus, we calculate HSC from these deep-level features and include them as supplementary information in the self-attention module. This approach proves beneficial for preserving target structure information during training (as shown in the last row of Fig. 2).

Fig. 5 shows the aggregation procedure of structure priors and the attention module. As previously stated, the potential of spatial interaction is improved in deep features to extract the full object region, which assists in localization. The efficiency of the aggregation strategy is demonstrated in Sec. IV.

D. Towards Self-Supervised Localization

In the previous section, we elaborated that the SPA module can preserve object structure and expand the activation of CAM simultaneously. In fact, the inter-pixel interactions within the feature are not sufficient to achieve high-quality localization due to the obvious appearance changes occurring inside the object and the widely differing backgrounds. To address this issue, we propose to output the localization map in a self-supervised manner instead of directly utilizing the improved CAM to perform localization.

We mainly achieve high-quality self-supervised localization from the following two aspects. First, we propose a restricted activation loss (*RAL*) by suppressing the range of CAM response values and distinguishing objects from background regions with the help of coarse pseudo masks. After that, we couple HSC as a structural prior with CAM to obtain a structure-preserved pseudo localization mask. In order to distinguish from the pseudo mask in *RAL*, it is called a self-generated map, which is employed as auxiliary supervision

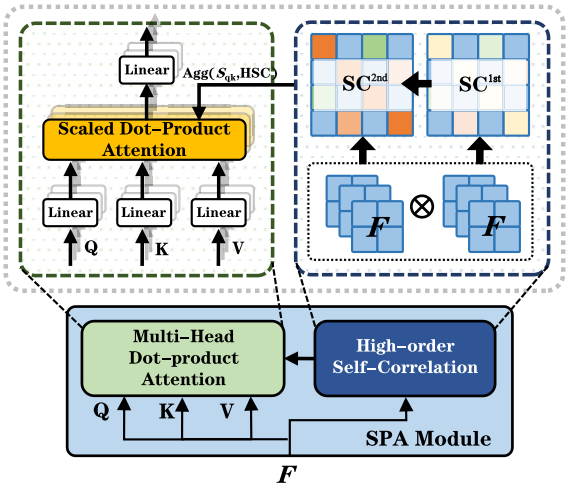


Fig. 5. Detail construction of the proposed SPA module. The self-attention mechanism has the same input as the HSC module. The attention mechanism aims to discover important positions, and we further incorporate HSC into the attention weights to better capture spatial long-range dependencies, ensuring the accuracy and completeness of the activated object regions.

information for the training of the localization branch after reassignment.

Assume the dimension of the convolutional feature $F \in \mathbb{R}^{HW\times C}$, the HSC can be viewed as (HW, HW), and the class activation map M^{cam} is reshaped to $(HW, 1)$, a dot-product operation between CAM and HSC can be formulated as:

$$M_j^{sgm} = \sum_{i=1}^{HW} M_i^{cam} \cdot HSC\left(F_i, F_j\right), \tag{12}$$

where M_i^{cam} represents the activation response at position i in CAM, and M_j^{sgm} represents the confidence score in self-generated map at position j. Each pixel in M^{sgm} merges the correlation information between the current pixel with all other pixels according to M^{cam}, and the coupling operation enables the activation to be quantified and continuous. Fig. 6 shows an example of this process.

Even if the self-generated map covers most of the target area, the confidence of object boundaries cannot be determined well, which makes the self-generated map sensitive to thresholds. We develop a localization branch that directly predicts a single-channel localization map to decrease sensitivity while enabling the localization map to be refined online. It is constructed in a similar way as the classification branch. Specifically, the localization branch is composed of several convolutional and $ReLU$ layers. For each image, a single-channel localization map is finally output, and the value of each position represents the confidence score that belongs to the object. We simply use two thresholds μ_l and μ_h to divide M^{sgm} into three parts:

$$\hat{M}_{(i,j)}^{sgm} = \begin{cases} 0, & \text{if } \mu_l \geq M_{(i,j)}^{sgm}, \quad 0 < \mu_l < 1 \\ 1, & \text{if } \mu_h \leq M_{(i,j)}^{sgm}, \quad 0 < \mu_h < 1 \\ M_{(i,j)}^{sgm}, & \text{otherwise}, \end{cases} \tag{13}$$

where $\hat{M}_{(i,j)}^{sgm}$ is the pseudo localization mask and (i, j) represents spatial location. We separate M^{sgm} into foreground,

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

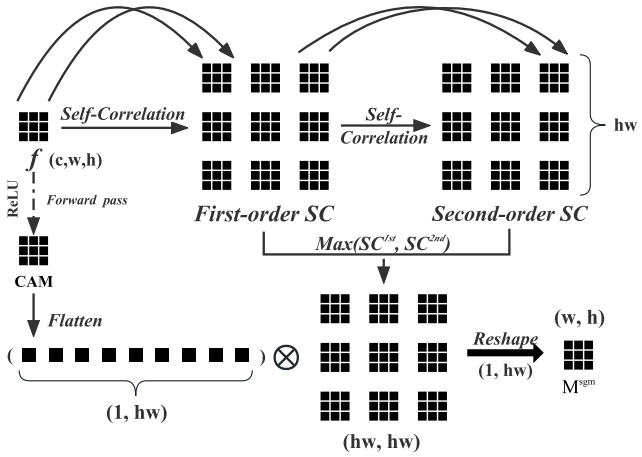


Fig. 6. Example for generating the self-generated map M^{sgm} by coupling CAM and the HSC. We exploit the element-wise maximum of first- and second-order self-correlations to extract the object's structure precisely.

background, and uncertain regions, respectively. Without modifying the original values, the uncertainty region can be treated as an adaptive decision space to allow the network to fit the score at the boundary positions. The binary cross-entropy loss L_{BCE} is employed to measure pixel-wise divergence between the predicted localization map M^{loc} and $\hat{M}^{sgm}$ at position (i, j):

$$L_{BCE} = -\frac{1}{NHW}\sum_{i,j,n}\left(\hat{M}^{sgm}_{(i,j,n)}\log M^{loc}_{(i,j,n)} + \left(1-\hat{M}^{sgm}_{(i,j,n)}\right)\log\left(1-M^{loc}_{(i,j,n)}\right)\right). \quad (14)$$

In general, the localization branch predicts the probability that each pixel belongs to the foreground for localizing, *i.e.*, confidence scores are close to 1 for foreground pixels and close to zero for backgrounds.

E. Restricted Activation Loss

The proposed restricted activation loss is based on such observations: for a certain class, regions with large variances in the channel dimension are considered more likely to be objects. In contrast, areas with relatively small variance are considered suppressed background regions. It 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; and discriminating the object from the background region with the help of coarse pseudo-masks.

We denote the last convolutional feature maps in the classification branch 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 object classes. The feature maps are fed into a Global Average Pool (GAP) [14] layer followed by a *softmax* function for classification, as shown in Fig. 3. To ensure that the high activation value after the 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*:

$$f'_t = sigmoid(f_t), \quad (15)$$

where t is the ground truth label index and f_t is the t_{th} feature map, and $f'_t \in \mathbb{R}^{H\times W\times 1}$. 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, hindering the model from differentiating the object from the 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 is based on the observation that the 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:

$$M^{bg} = \mathbb{I}(Var(softmax(f)) < \tau), \quad (16)$$

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 determine the background region. We further obtain the coarse object region as:

$$M^{obj} = \mathbb{I}(Var(softmax(f)) > \tau + \sigma), \quad (17)$$

where σ denotes the gap between the background and object regions. Combined with Eqs.15-17, we define the restricted activation loss to guide the model to focus on the object region as:

$$L_{RA} = \frac{1}{HW}\sum_{i,j}(M^{bg} * f'_t + M^{obj} * (1 - f'_t))|_{i,j}, \quad (18)$$

where $*$ indicates element-wise multiplication. and M^{bg} $(M^{obj}) \in \mathbb{R}^{H\times W\times 1}$ The simple L_1 regularization loss can guide the model to suppress the background area while paying attention to the active entire object extent. Cooperating with the classification branch, the proposed *RAL* enables the model to better preserve the structural information of the object.

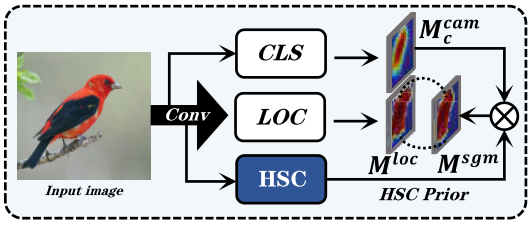
F. Joint Optimization

After the proposed SPA module, our network is divided into two branches for classification and localization, respectively. For classification, cross-entropy (CE) loss is employed:

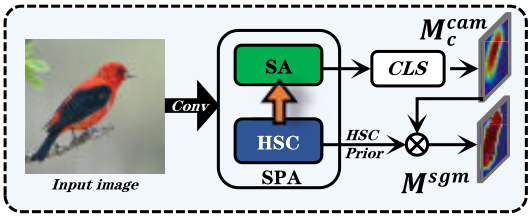
$$L_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=0}^{C-1} y_i^{(n)}\log\left(\hat{y}_i^{(n)}\right), \quad (19)$$

where C represents the total number of categories, N denotes the batch size, y and $\hat{y}$ are one-hot representations of ground truth and the predicted probability vectors, respectively. Following ACoL [10], $ReLU$ is appended for normalizing the output map to generate a class activation map that has a similar performance to CAM [1]. Due to the existence of SPA, the structural information of deep features is well preserved, and the activation regions of the CAM are divergent, which alleviates the partial activation problem of the classification network.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.



(a) CNN with the localization branch



(b) CNN with the SPA module

Fig. 7. Two different variants of the SPA^{2}Net. (a) We use the self-generated map M^{sgm} to guide the training of the localization branch without the SPA module. (b) Without the localization branch, we only employ the pseudo mask M^{sgm} to locate the object.

Algorithm 1 Training and Inference Procedure for SPA^{2}Net

Input: Feature map $F \in \mathbb{R}^{H\times W\times K}$ and $f \in \mathbb{R}^{H\times W\times C}$ of classification branch; thresholds τ, σ, μ_l and μ_h;
Output: Localization map M^{loc};
Compute SC^{2nd} by Eqs. 3-7;
Compute $\hat{F}^q$, $\hat{F}^k$, $\hat{F}^v$ of self-attention module;
Compute s_{qk} according Eq. 8;
Update s_{qk} according to Eq. 11;
Compute α_{qk} and $\hat{F}^o$ by Eq. 9 and Eq. 10, respectively;
while is $training$ **do**
    Compute M^{sgm} according to Eq. 12;
    Output $M^{loc} \leftarrow LOC(\hat{F}^o)$;
    Compute L_{CE} by Eq. 19;
    Compute L_{RA} with τ, σ by Eqs. 15-18
    Compute L_{BCE} with μ_l and μ_h by Eqs. 13, 14;
    Update full model θ_{full} according to Eq. 20;
end while
if is $inference$ **then**
    Output $M^{loc} \leftarrow LOC(\hat{F}^o)$;
end if

In general, our optimization objective consists of three parts, and the entire training procedure is carried out under the joint optimization of Eqs. 14, 18 and 19:

$$L = L_{CE} + \lambda_1 L_{RA} + \lambda_2 L_{BCE}, \tag{20}$$

where λ_1 and λ_2 are the weight hyperparameters to balance the training process. In our framework, the classification loss L_{CE} is given primary importance during training. Determining the approximate target location through this loss enables better foreground-background separation and more accurate pseudo masks. Therefore, we ensure the dominance of the classification loss while keeping the other loss terms consistent (both λ_1 and λ_2 are set to 0.5 across all experiments). To perform object localization, bilinear interpolation is used for upsampling the localization map to the input image size. As in previous work, we identify discriminative regions by hard thresholding. The detected bounding box is the coverage of the largest connected region obtained using threshold truncation on M^{loc}.

Our model training involves an adversarial process: the cross-entropy loss encourages learning local discriminative features for class distinction, while the SPA module enhances interactions between pixels, diverging activations from local to the entire object. This adversarial training preserves the structure of the object, crucial for generating pseudo object masks for self-supervised training. By utilizing *RAL*, the foreground and background can be further roughly separated, and the self-generated map is refined for self-supervised training to predict the high-quality localization map $M^{loc} \leftarrow LOC(\hat{F}^o)$ by L_{BCE}, where *LOC* represents the forward process of the localization branch. Algorithm 1 describes details of the training and inference procedure of our SPA^{2}Net.

IV. Experiments

A. Experimental Settings

1) Implementation Details: Following [23], we implement the SPA^{2}Net based on two popular backbone networks, *i.e.* VGG-16 [58] and Inception-v3 [59]. We adopt the same modifications on backbones following ACoL [10] and SPG [59] and use the simplified method in ACoL to obtain the class activation maps. The scaled dot-product attention (SA) [60] aggregates attention weight with the proposed high-order self-correlation information. Following [60], to obtain complete attention information, the self-attention mechanism in the SPA module adopts a multi-head self-attention formulation. In all experiments, the number of self-attention heads is set to more to balance performance and efficiency simultaneously. The classification branch and the localization branch have the same network structure. The difference is the output dimension of classification is equal to the number of categories, and the localization branch directly predicts a single-channel localization map. The input images are randomly cropped to 224 × 224 pixels after re-sized to 256 × 256 pixels. Both networks are fine-tuned on the pre-trained weights of ILSVRC [27]. For classification, we average the scores from the softmax layer with 10 crops for testing. For fair comparisons, following [23], we utilize the features of *stage* 4 and 5 in VGG-16, and combine the two HSCs by element-wise summation. As for Inception-v3, the features of the last two layers are used to calculate HSC respectively and sum them element-wise. We utilize four Titan X (Pascal) GPUs for training. In the testing phase, when only one GPU is used, the inference speed of our method is about 262.11 Frames-Per-Second (FPS). Compared with 132.49 FPS of the previous work [23], SPA^{2}Net shows obvious advantages of the one-stage localization method.

2) Datasets: We evaluate the proposed SPA^{2}Net on two publicly available benchmarks, including CUB-200-2011 [26] and ILSVRC [27], following the previous state-of-the-arts (SOTA) [2], [8], [21], [22]. CUB-200-2011 is a fine-grained bird dataset of 200 different species, which is split 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

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

TABLE I

THE COMPARISON BETWEEN OUR METHOD AND THE PREVIOUS STATE-OF-THE-ART APPROACHES ON BOTH THE ILSVRC [27] AND THE CUB-200-2011 [26] DATASETS

Method	ILSVRC				CUB-200-2011			
	Backbone	*Loc. Err.*			Backbone	*Loc. Err.*		
		Top-1	Top-5	Gt-Known		Top-1	Top-5	Gt-Known
CAM(*2016*) [1]	Inception-v3	53.71	41.81	37.32	GoogLeNet	58.94	49.34	44.9
SPG(*2018*) [6]	Inception-v3	51.40	40.00	35.31	GoogLeNet	53.36	42.8	-
ADL(*2019*) [29]	Inception-v3	51.29	-	-	Inception-v3	46.96	-	-
ACoL(*2019*) [10]	GoogLeNet	53.28	42.58	-	-	-	-	-
DANet(*2019*) [4]	GoogLeNet	52.47	41.72	-	Inception-v3	50.55	39.54	33.0
RCAM(*2020*) [11]	GoogLeNet	52.30	-	37.24	GoogLeNet	48.95	-	34.9
MEIL(*2020*) [8]	Inception-v3	50.52	-	-	-	-	-	-
GC-Net(*2020*) [30]	Inception-v3	50.94	41.91	-	-	-	-	-
I^{2}C(*2020*) [21]	Inception-v3	46.89	35.87	31.50	-	-	-	-
SPA(*2021*) [23]	Inception-v3	47.27	35.73	31.67	Inception-v3	46.41	33.50	27.86
BagCAM(*2022*) [32]	Inception-v3	47.78	-	30.80	Inception-v3	42.07	-	20.56
DAOL(*2022*) [33]	Inception-v3	47.30	-	30.89	Inception-v3	43.71	-	19.97
SPA^{2}Net (*ours*)	Inception-v3	**46.31**	**34.50**	**30.23**	Inception-v3	**39.14**	**23.63**	**16.88**
CAM(*2016*) [1]	VGG-16	57.20	45.14	-	VGG-16	55.85	47.84	44.0
ACoL(*2018*) [10]	VGG-16	54.17	40.57	37.04	VGG-16	54.08	43.49	45.9
CutMix(*2019*) [5]	VGG-16	56.55	-	-	-	-	-	-
ADL(*2019*) [29]	VGG-16	55.08	-	-	VGG-16	47.64	-	-
RCAM(*2020*) [11]	VGG-16	54.60	-	37.32	VGG-16	38.70	-	19.28
MEIL(*2020*) [8]	VGG-16	53.19	-	-	VGG-16	42.54	-	-
I^{2}C(*2020*) [21]	VGG-16	52.59	41.49	36.10	VGG-16	44.01	31.6	-
CI-CAM(*2021*) [12]	VGG-16	51.29	41.24	37.64	VGG-16	41.61	29.46	24.32
SPA(*2021*) [23]	VGG-16	50.44	38.68	34.95	VGG-16	39.73	27.5	22.71
PDM(*2022*) [31]	VGG-16	48.90	36.20	30.70	VGG-16	32.70	17.80	17.80
C^{2}AM(*2022*) [25]	VGG-16	55.97	43.82	38.46	VGG-16	36.41	24.70	19.01
SPA^{2}Net (*ours*)	VGG-16	**48.73**	**36.52**	**31.64**	VGG-16	**32.71**	**20.90**	**15.74**

1,000 categories for training and 50,000 images for validation. Both benchmarks are only annotated with class labels for training. CUB-2000-2011 also provides tight box and mask labels for images in the testing set. For ILSVRC, only tight box labels are provided for validation. Zhang et al. [22] annotate the ground-truth masks for the images on the validation set of ILSVRC: 5,729 images are manually excluded, and the rest are split into validation set (23,151 images) and testing set (21,120 images).

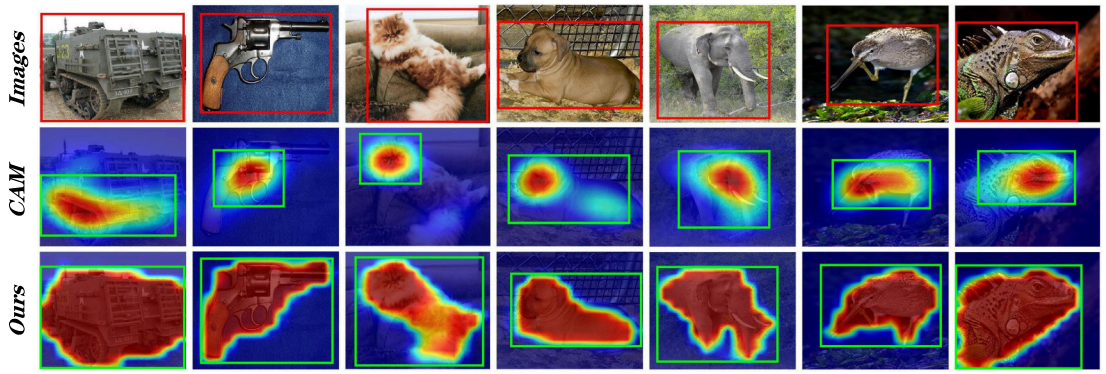
3) Metrics: Following the previous methods [23], we apply both bounding box and mask metrics to evaluate the performance of our SPA^{2}Net. For bounding boxes, we follow the baseline methods [1], [6], [23], [27] and report the localization error (*Loc. Err.*). The *Loc. Err.* is calculated based on the Top1 accuracy and location accuracy. A prediction is correct only if both localization($i.e.$ IoU $\geqslant 0.5$) and classification are correct. For *Gt-Known* metric, it considers localization only regardless of classification. For masks, we mainly report the *Peak-IoU* and *Peak-T* [22] to evaluate the 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 entire object extent can be accurately covered with a specific threshold; 2) brightness values of pixels belonging to the object and the background should differ greatly so that the objects can be well visualized [22]. As a result, higher *Peak-IoU* and *Peak-T* values indicate better localization maps with the complete and compact activated region. The codes for *Peak-IoU* and *Peak-T* are provided in the workshop of *Learning from Imperfect Data (LID)*.[1]

B. Experimental Results

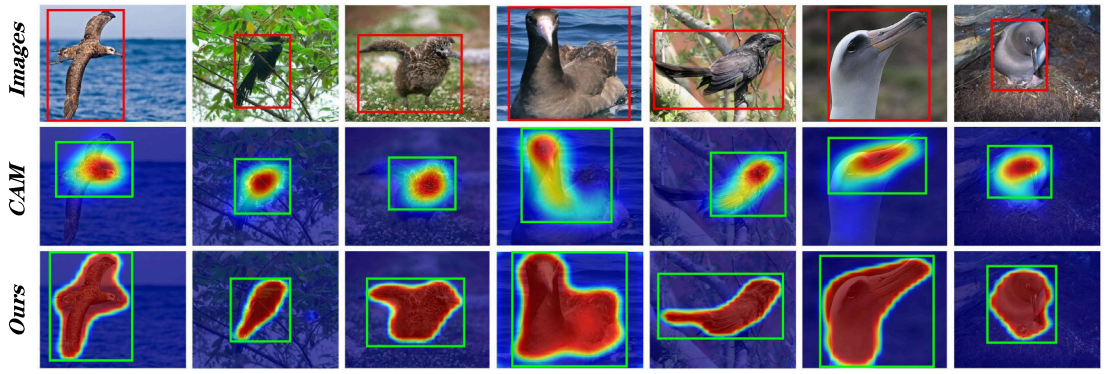
1) Bounding Box Localization: We first compare the proposed approach with other baseline methods on the localization error by using tight bounding boxes. For a fair comparison, many of the baselines are CAM-based weakly supervised object localization methods, such as the erasure-based methods [8], [10] and some other well-known works. Table I reports the results of our method and several baselines on the ILSVRC validation set and the CUB-200-2011 testing set. From Table I, we can observe the following findings: 1) SPA^{2}Net outperforms all baseline methods in terms of Top-1 *Loc. Err.*, indicating its advantage in simultaneously considering classification accuracy and localization precision. 2) SPA^{2}Net also surpasses most methods in terms of Gt-Known *Loc. Err.*, demonstrating that the localization branch trained through self-supervised learning achieves excellent localization accuracy. 3) Compared with recently published baseline methods, SPA^{2}Net maintains a competitive edge, highlighting the potential of leveraging spatial similarity in deep features to expand the scope of local activation regions. For instance,

[1] https://lidchallenge.github.io/challenge.html

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.



(a) Visualization results on ILSVRC2016 [27] validation set



(b) Visualization results on CUB-200-2011 [26] test set

Fig. 8. Visualization of the localization maps with CAM [1] and our proposed *SPA^{2}Net* on different datasets. (a) Visualization results on ILSVRC2016 validation set. (b) Visualization results on CUB-200-2011 test set.

when comparing our method to the two-stage approach C^{2}AM, we excel in both localization accuracy (7.24% higher on ILSVRC and 3.7% higher on CUB-200-2011) and training efficiency (25 epochs for C^{2}AM on ILSVRC compared to 20 for SPA^{2}Net, and 120 epochs for C^{2}AM on CUB-200-2011 compared to 100 for SPA^{2}Net). 4) Compared to our previous work SPA, SPA^{2}Net shows a significant performance boost, mainly attributed to the further extension of high-order self-correlation. This improvement is achieved through the aggregation of self-attention mechanisms, generating high-quality pseudo masks, and utilizing them in self-supervised training of the localization branch. More specifically, our method based on VGG-16 achieves the lowest error rate of 48.73% in Top-1 *Loc. Err.* on ILSVRC, surpassing all the baselines. SPA^{2}Net achieves remarkable gains of 3.86% and 2.56% in terms of Top-1 *Loc. Err.* compared with I^{2}C [21] and CI-CAM [12] on ILSVRC. Our method based on VGG-16 surpasses all the baseline methods on Top-1, Top-5, and Gt-Known metrics when testing on CUB-200-2011, yielding the localization error of Top-1 32.71%, Top-5 20.90% and Gt-known 15.74%. Compared with I^{2}C [21] and CI-CAM [12], we achieve gains of 11.3% and 8.9% in terms of Top-1 *Loc. Err.*, respectively. As for Inception-v3, our method can also obtain comparable results and surpass other methods by a large margin.

2) 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 [22]. In addition, we also report the *Gt-Known Loc. Acc.* of each method in Table II. Our approach achieves consistent gains on *Peak-IoU*, *Peak-T*, and *Gt-Known*. It can be seen that compared to the baseline method, our approach has a better positioning effect on the mask and can obtain the best result at a higher threshold ratio, which indicates that the object mask is more accurate and the highlight region is more compact. In Fig. 8, we show the object mask (including the inferred bounding box) on CUB-200-2011 and ILSVRC datasets. Intuitively, it can be found that the mask quality obtained by our method is better, and the inferred bounding boxes (green) are closer to the labeled boxes (red).

C. Ablation Study

1) Effectiveness of the Structure-Preserved Attention: The SPA module is designed to capture remote dependencies while maintaining spatial structure for localization. In Table III and Table IV, we show the localization results evaluated by CAM and the self-generated map (SGM) with different networks on the CUB-200-2011 and ILSVRC dataset, respectively. Three different methods were evaluated: the baseline method using VGG-16 or Inception-v3, the baseline network augmented with the scaled dot-product attention (SA), and the network incorporating the proposed SPA (as in Fig. 7(b)). The difference between SA and SPA is that the attention weights of SPA contain high-order self-correlation priors from convolutional features F. Experiments were performed under the same configuration conditions with restricted activation loss. As shown

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

TABLE II

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	*Peak-IoU*	*Peak-T*	*GT-Known*
CAM [1]	53.59	33	64.09
HaS [3]	54.99	50	65.32
ACoL [10]	50.89	52	63.28
CutMix [5]	54.54	34	64.65
SPG [6]	53.76	33	64.19
ADL [29]	52.87	29	63.64
SEM [22]	56.05	93	-
SPA^{2}Net(*ours*)	**61.45**	**217**	**68.34**

TABLE III

ABLATION STUDY OF THE SPA MODULE. *SGM* REPRESENTS THE SELF-GENERATED MAP. WE USE TWO BACKBONE NETWORK TESTS ON THE CUB-200-2011 TEST SET, RESPECTIVELY

Method	*CAM*	*SGM*	*Loc.Err.* Top-1	Top-5	Gt-Known
VGG16	✓		49.34	39.14	34.50
		✓	39.28	25.70	20.31
VGG16 with SA	✓		40.78	27.65	23.58
		✓	37.54	22.90	17.50
VGG16 with SPA	✓		38.33	24.15	18.78
		✓	**36.59**	**21.45**	**15.12**
Inception v3	✓		57.90	46.19	42.06
		✓	53.68	40.78	35.97
Incept.v3 with SA	✓		46.48	33.36	28.55
		✓	44.37	29.63	23.66
Incept.v3 with SPA	✓		42.96	28.67	23.11
		✓	**41.94**	**27.05**	**21.11**

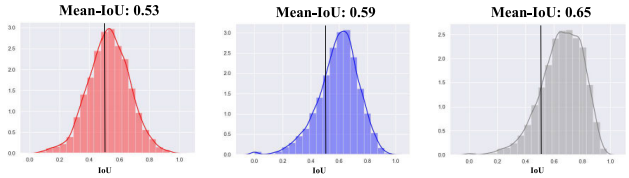
in Table III, with Inception-v3 as backbone, the Top-1 *Loc. Err.* evaluated by the CAM and the SGM are dropped from 46.48% to 42.96%, and from 44.37% to 41.94%, respectively. Additionally, Table IV shows the results on the ILSVRC dataset, using VGG-16 as the evaluation backbone. From this table, we can observe that simply adding the self-attention mechanism did not result in the expected improvement on ILSVRC as observed in CUB-200-2011. Instead, it led to a slight degradation in performance (CAM's Top-1 localization error rate increased from 57.20% to 58.61%, and SGM's Top-1 localization error rate increased from 52.72% to 53.93%). We analyzed that this could be attributed to the diverse range of appearances and textures present in the large-scale dataset, which makes it challenging for the classification network to precisely locate the target regions. Consequently, incorporating attention mechanisms might exacerbate localization errors. Interestingly, replacing SA with the proposed SPA module effectively enhanced the performance (with a decrease of 3.52% and 2.25% in Top-1 *Loc. Err.* for CAM and SGM, respectively). Moreover, in comparison to CAM, SGM consistently demonstrated superior improvement across all settings. These results indicate that whether utilizing CAM or SGM, the localization accuracy can be improved after aggregating the HSC priors with SA, which is helpful in locating objects more precisely.

To reveal the improvements more intuitively, we also investigate the intersection over union (IoU) distribution between the predicted bounding boxes and the ground-truth boxes.

TABLE IV

ABLATION STUDY OF THE SPA MODULE ON THE ILSVRC DATASET. *SGM* REPRESENTS THE SELF-GENERATED MAP

Method	*CAM*	*SGM*	*Loc.Err.* Top-1	Top-5	Gt-Known
VGG16	✓		57.20	45.14	42.36
		✓	52.72	41.61	37.19
VGG16 with SA	✓		58.61	48.88	45.07
		✓	53.93	42.77	38.45
VGG16 with SPA	✓		55.09	44.77	40.78
		✓	**51.68**	**40.41**	**35.99**



(a) Vanilla CNN (b) CNN with SPA (c) SPA^{2}Net

Fig. 9. IoU statistics of bounding boxes predicted by different methods. (a) Statistical results of Vanilla CNN. (b) Statistical results by CNN with the SPA module. (c) Statistical result of the predicted localization map by SPA^{2}Net is reported.

TABLE V

ABLATION STUDY FOR THE RESTRICTED ACTIVATION LOSS (*RAL*) ON THE CUB-200-2011 TEST SET.*SPA* REPRESENTS THE SELF-ATTENTION MODULE WHICH AGGREGATES WITH HIGH-ORDER STRUCTURAL PRIORS AND *LOC* REFERS TO THE LOCALIZATION BRANCH

Method	*RAL*	*Loc.Err.* Top-1	Top-5	Gt-Known
Vanilla CNN	-	55.85	47.84	44.00
	✓	49.33	38.26	33.78
CNN with SPA	-	38.26	29.50	27.29
	✓	36.59	21.45	15.12
CNN with LOC	-	37.75	24.77	19.71
	✓	36.66	23.44	18.07
SPA^{2}Net	-	37.80	27.96	25.32
	✓	32.71	20.90	15.74

As shown in Fig. 9(a), the mean IoU of CAM [1] is boosted from 0.53 to 0.59 (maximum value is 1) with 6% gains when applying the SPA module (Fig. 9(b)). Moreover, the mean IoU is further boosted to 0.65 with 12% gains by SPA^{2}Net as shown in Fig. 9(c). The above analysis indicates that the object can be better activated when combined with the proposed structure-preserved attention module. There is a continuous improvement when applying the localization branch for predicting.

2) Effectiveness of Various Modules in SPA^{2}Net: We conducted a series of experiments to verify the effectiveness of different modules. In the experiments, we found that the restricted activation loss has a positive effect on suppressing background noise and preserving structure to focus on the object. Since the purity of the foreground is guaranteed, the potential of high-order attention can be better utilized. To show the effectiveness of the proposed restricted activation loss (*RAL*), we try to add or delete it in different architectures and report their respective localization errors on CUB-200-2011 and ILSVRC in Table V and Table VI. As shown in Table V, for the baseline method, after adding the proposed

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

TABLE VI

ABLATION STUDY FOR THE RESTRICTED ACTIVATION LOSS (*RAL*) ON THE ILSVRC DATASET. *SPA* REPRESENTS THE SELF-ATTENTION MODULE WHICH AGGREGATES WITH HIGH-ORDER STRUCTURAL PRIORS AND *LOC* REFERS TO THE LOCALIZATION BRANCH

Method	*RAL*	*Loc.Err.*		
		Top-1	Top-5	Gt-Known
Vanilla CNN	-	55.33	42.26	39.78
	✓	52.72	41.61	37.19
CNN with SPA	-	52.87	42.11	38.08
	✓	51.68	40.41	35.99
CNN with LOC	-	52.02	40.79	36.47
	✓	51.63	40.28	35.65
SPA^{2}Net	-	49.88	38.54	32.90
	✓	48.73	36.52	31.64

TABLE VII

ABLATION STUDY FOR DIFFERENT MODULES ON THE CUB-200-2011 TEST SET.ONCE THE *LOC* IS NOT SPECIFIED, RESULTS ARE OBTAINED BY THE SELF-GENERATED MAP (*SGM*), OTHERWISE EVALUATED BY THE PREDICTED LOCALIZATION MAP

Backbone	*SGM*	*SPA*	*LOC*	*Loc.Err.*		
				Top-1	Top-5	Gt-Known
VGG16				49.34	39.14	34.50
	✓			39.28	25.70	20.31
	✓	✓		36.59	21.45	15.12
	✓		✓	36.66	23.44	18.07
	✓	✓	✓	**32.71**	**20.90**	**15.74**
InceptionV3				57.90	46.19	42.06
	✓			53.68	40.78	35.97
	✓	✓		41.94	27.05	21.11
	✓		✓	51.83	40.27	34.52
	✓	✓	✓	**39.14**	**23.63**	**16.88**

TABLE VIII

ABLATION STUDY FOR DIFFERENT MODULES ON ILSVRC DATASET.ONCE THE *LOC* IS NOT SPECIFIED, RESULTS ARE OBTAINED BY THE SELF-GENERATED MAP (*SGM*), OTHERWISE EVALUATED BY THE PREDICTED LOCALIZATION MAP

SGM	*SPA*	*LOC*	*Loc.Err.*		
			Top-1	Top-5	Gt-Known
VGG16 (CAM)			57.20	45.14	42.36
✓			52.72	41.61	37.19
✓	✓		51.68	40.41	35.99
✓		✓	51.63	40.28	35.65
✓	✓	✓	**48.73**	**36.52**	**31.64**

RAL, the Top-1 *Loc. Err.* decreased by 3.86% and Top-5 decreased by 6.27%, *Gt-Known* decreased by 7.16%. When using a self-attention mechanism that aggregates high-order self-correlation priors (*i.e*. *SPA* module), the proposed *RAL* brings about a 2% improvement. At the same time, the full model also gets about 5% performance improvement from the *RAL*. On the ILSVRC dataset (Table VI), the addition of *RAL* to different network structures demonstrated consistent performance improvements. Each of the four network variants achieved a reduction in localization error rates of 2.61%, 1.19%, 0.39%, and 1.15%, respectively, resulting in an average decrease in localization error rate of 1.34%. It can be seen that when combined with various variant structures (as shown in Fig. 7), the *RAL* shows continuous improvement in localization.

Table VII and Table VIII show the localization performance of different modules in SPA^{2}Net on the CUB-200-2011 and ILSVRC dataset, respectively. The baseline methods are trained using the cross-entropy loss. Once *LOC* is not specified, we report the location error evaluated by the self-generated map (SGM). As demonstrated in Table VII, while our initial intention was to dynamically generate localization maps during training by combining high-order self-correlation and CAM, the results reveal that the SGM can still significantly enhance localization performance when utilized in the inference phase (approximately 10% improvement). The ablation study results on the ILSVRC dataset (Table VIII) also indicate that SGM effectively reduces localization errors compared to the original CAM (by 4.48%). Upon integrating the SPA module, which expands activations while retaining target structural information, the localization error is further reduced (by 1.04%). The complete SPA^{2}Net framework, with the inclusion of the localization branch and self-supervised training for predicting threshold-robust localization maps, achieves the best localization performance. Other results in Table VII show that the proposed method achieves consistent and substantial improvements under different modules.

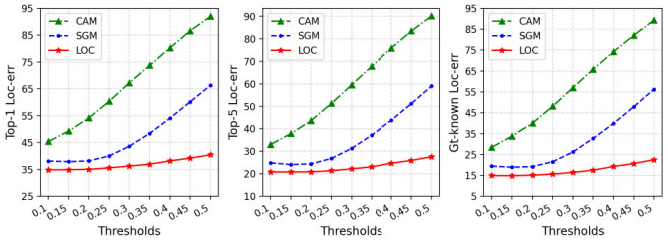


Fig. 10. Threshold sensitive statistics for different activation maps. CAM is generated by the classification branch and *LOC* is the output of the localization branch. Compared with CAM and SGM, the predicted localization map is more stable and robust to different thresholds.

3) Improvement of the Class Activation Map: With the help of the proposed SPA module, the network will focus on the object from a global perspective, which alleviates the partial activation problem in previous work [1] and the CAM obtained by SPA^{2}Net can highlight the object more accurately. Fig. 1 shows CAMs generated by baseline methods [1], [11] and our approach. Compared with the previous method, it can be seen that the activation region of our localization map can extract objects from the background more accurately. In Table III, we also report the location results evaluated by our CAM. Our method, with attention and higher-order structure priors in class activation maps, greatly improves localization performance and demonstrates promising structure preservation capabilities.

4) Sensitivity Analysis: To perform object localization, we identify the object bounding boxes by a hard threshold as in previous works [1], [6]. The threshold value is crucial for accurate localization. A robust localization map should be stable against threshold variations within a limited range. Threshold robustness also ensures the map's resistance to interference at boundary positions, where activation values are typically lower than in the central region. Based on the above reasons, we conduct sensitivity analyses for the localization map on the CUB-200-2011 dataset, and the results

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

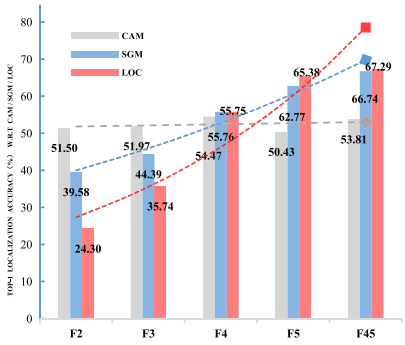


Fig. 11. Comparison of localization performance when using different stage features as input to HSC. The horizontal axis represents different input features of HSC. For example, 'F2' means to calculate the affinity matrix using the features of the second stage of the backbone network.

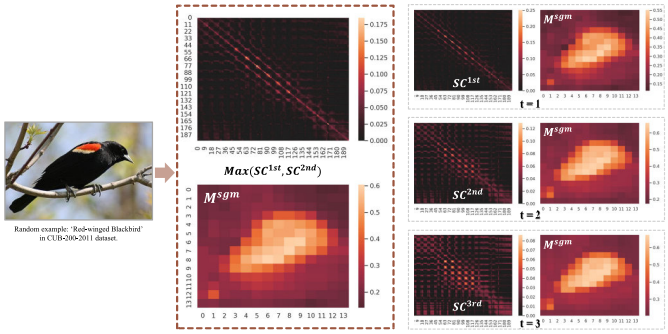


Fig. 12. A visual example of self-correlation matrices and the SGM (M^{sgm}) at various orders t (see Sec. III-B). Left: Random sample from CUB-200-2011. Middle: Fused self-correlation by $Max(SC^{1st}, SC^{2nd})$, and the corresponding SGM used in our method. Right: First-order self-correlation (SC^{1st}), second self-correlation (SC^{2nd}), and third-order self-correlation (SC^{3rd}) matrices and their SGMs. Higher orders lead to more noise and computational complexity. *Zoom in* to observe the details.

are presented in Fig. 10. The performances of our method are stable when the threshold varies from 0.1 to 0.5, which proves that the localization map (*i.e. LOC* in Fig. 10) can substantially reduce the number of uncertain pixels (0.5) and improve the robustness of our method to threshold values.

5) The Selection of Features for HSC: In this work, the HSC plays a crucial role, serving both as a higher-order structural prior within the self-attention mechanism and facilitating pseudo-mask generation for self-supervised training. Our approach employs deep-level features from CNNs for both HSC and self-attention mechanisms. In classification tasks, the discriminative loss function encourages the network to focus on local regions of the target, which benefits our method in preserving target structures through spatial interactions. In this section, we extensively analyze the effect of using features from different stages on the HSC matrix and localization performance. Our aim is to confirm the importance of deep-level features in capturing essential target structural information. We expect that utilizing deeper features, closer to the classifier, will lead to a purer HSC and improved object localization. To verify this hypothesis, we aim to derive the affinity matrix by separately using convolutional features from different stages as inputs for the HSC module.

Fig. 11 shows the Top-1 localization accuracy of CAM, SGM (M^{sgm}), and localization map (M^{loc}) using different features. Consistently keeping other parameters, we observe that as the depth of input features increases, both M^{sgm} and M^{loc} achieve higher localization accuracy. Deeper features enable more accurate pseudo-masks for self-supervised training. Using feature fusion from the fourth and fifth stages as input further reduces background noise and improves localization performance. This motivates the use of deep features in the HSC module to enhance the effectiveness of the final weighted pseudo-mask's supervision.

In the supplementary materials, we present visualizations of self-correlation maps using features from various stages. It is clear that deeper features are less affected by background noise and better preserve structural information compared to shallower features.

6) Qualitative Analysis of HSC: In the experiments, SPA^{2}Net utilizes the element-wise maximum fusion of first- and second-order self-correlations, serving as the higher-order self-correlation for generating the self-generated map (SGM). This approach is driven by two primary considerations: 1) Addressing additional noise: As shown in Fig. 12, we visualized the first-order (SC^{1st}), second-order (SC^{2nd}), and third-order (SC^{3rd}) self-correlations (right), along with the fused self-correlation (middle) and their corresponding SGMs. Higher-order self-correlation matrices exhibit higher similarity in non-diagonal regions, indicating a higher correlation in noisy positions of feature maps. By fusing the first- and second-order self-correlations, we effectively reduce noise while extending the activated region of the target. 2) Managing computational complexity: Higher-order self-correlations have stronger spatial correlations, but come with higher computational costs, especially for larger feature maps. (please refer to Eq. 5).

In our experiments, we explored high-order correlation and achieved better localization results. High-order correlation preserves structure more effectively due to its ability to mine deep features' rich semantic information and enhance interactions between distant pixels. We successfully integrated HSC into a popular self-attention mechanism, leading to natural divergence around target objects. Additionally, the localization branch reduced threshold sensitivity and enhanced the network's focus on pseudo object areas, further improving localization.

V. Conclusion

In this paper, we propose a novel one-stage WSOL framework named **Structure-Preserved Attention Activated Network** (SPA^{2}Net), which aims to explore spatial high-order interactions to capture long-range dependencies and preserve object structure in an end-to-end manner. Different from previous erasing-based methods, which drive the model to explore more areas related to the target by erasing the discriminative parts, SPA^{2}Net unveils the essential of the high-order correlation within the image for localization. By equipping the self-attention mechanism with structural prior, the spatial interaction of deep features is enhanced, and the superiority of structural preservation is shown. In addition, the foreground and background regions are roughly separated by the restricted

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

activation loss, and it can precisely predict the localization map under the guidance of the self-generated map. Experiments on both CUB-200-2011 and ILSVRC demonstrate the effectiveness of the proposed approach, in striking contrast with the state-of-the-art.

REFERENCES

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[2] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim, "Evaluating weakly supervised object localization methods right," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3130–3139.

[3] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3544–3553.

[4] H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji, and Q. Ye, "DANet: Divergent activation for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6588–6597.

[5] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[6] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang, "Self-produced guidance for weakly-supervised object localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 597–613.

[7] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7268–7277.

[8] J. Mai, M. Yang, and W. Luo, "Erasing integrated learning: A simple yet effective approach for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8763–8772.

[9] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6488–6496.

[10] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1325–1334.

[11] W. Bae, J. Noh, and G. Kim, "Rethinking class activation mapping for weakly supervised object localization," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2020, pp. 618–634.

[12] F. Shao et al., "Improving weakly-supervised object localization via causal intervention," 2021, *arXiv:2104.10351*.

[13] S. Yang, Y. Kim, Y. Kim, and C. Kim, "Combinational class activation maps for weakly supervised object localization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2930–2938.

[14] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.

[15] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[16] A. Torabi and G.-A. Bilodeau, "Local self-similarity-based registration of human Rois in pairs of stereo thermal-visible videos," *Pattern Recognit.*, vol. 46, no. 2, pp. 578–589, Feb. 2013.

[17] T. Deselaers and V. Ferrari, "Global and efficient self-similarity for object classification and detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1633–1640.

[18] I. N. Junejo, E. Dexter, I. Laptev, and P. Púrez, "Cross-view action recognition from temporal self-similarities," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2008, pp. 293–306.

[19] I. N. Junejo, E. Dexter, I. Laptev, and P. Pérez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 172–185, Jan. 2011.

[20] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12272–12281.

[21] X. Zhang, Y. Wei, and Y. Yang, "Inter-image communication for weakly supervised localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 271–287.

[22] X. Zhang, Y. Wei, Y. Yang, and F. Wu, "Rethinking localization map: Towards accurate object perception with self-enhancement maps," 2020, *arXiv:2006.05220*.

[23] X. Pan et al., "Unveiling the potential of structure preserving for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11637–11646.

[24] C.-L. Zhang, Y.-H. Cao, and J. Wu, "Rethinking the route towards weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13457–13466.

[25] J. Xie, J. Xiang, J. Chen, X. Hou, X. Zhao, and L. Shen, "C2 AM: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 979–988.

[26] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[27] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[28] S. Hwang and H.-E. Kim, "Self-transfer learning for weakly supervised lesion localization," in *Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer, 2016, pp. 239–246.

[29] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2214–2223.

[30] W. Lu, X. Jia, W. Xie, L. Shen, Y. Zhou, and J. Duan, "Geometry constrained weakly supervised object localization," 2020, *arXiv:2007.09727*.

[31] M. Meng, T. Zhang, W. Yang, J. Zhao, Y. Zhang, and F. Wu, "Diverse complementary part mining for weakly supervised object localization," *IEEE Trans. Image Process.*, vol. 31, pp. 1774–1788, 2022.

[32] L. Zhu, Q. Chen, L. Jin, Y. You, and Y. Lu, "Bagging regional classification activation maps for weakly supervised object localization," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, Tel Aviv, Israel, vol. 13670. Berlin, Germany: Springer, Oct. 2022, pp. 176–192.

[33] L. Zhu, Q. She, Q. Chen, Y. You, B. Wang, and Y. Lu, "Weakly supervised object localization as domain adaption," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 14617–14626.

[34] Y. Xu et al., "Transformers in computational visual media: A survey," *Comput. Vis. Media*, vol. 8, no. 1, pp. 33–62, Mar. 2022.

[35] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.

[36] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Comput. Vis. Media*, vol. 9, no. 4, pp. 733–752, Dec. 2023.

[37] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1354–1362.

[38] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[40] L. Song et al., "Learnable tree filter for structure-preserving feature transform," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1711–1721.

[41] M. Yin et al., "Disentangled non-local neural networks," 2020, *arXiv:2006.06668*.

[42] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

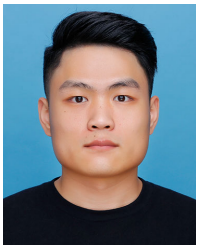
[43] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," 2019, *arXiv:1906.05909*.

[44] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[45] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.

[46] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," 2021, *arXiv:2101.11986*.
[47] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 371–381.
[48] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1734–1747, Sep. 2016.
[49] X. Zhan, Z. Liu, P. Luo, X. Tang, and C. Loy, "Mix-and-match tuning for self-supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 7534–7541.
[50] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[51] Y. Li, D. Huang, D. Qin, L. Wang, and B. Gong, "Improving object detection with selective self-supervised self-training," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 589–607.
[52] C. Yang, Z. Wu, B. Zhou, and S. Lin, "Instance localization for self-supervised detection pretraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3986–3995.
[53] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.
[54] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9912–9924.
[55] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.
[56] C. Li et al., "Efficient self-supervised vision transformers for representation learning," 2021, *arXiv:2106.09785*.
[57] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.
[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[59] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[60] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.



Dong Chen received the M.Eng. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2021. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Jilin University. His research interests include computer vision and machine learning.

Xingjia Pan received the B.S. degree in automation and finance from Nankai University in 2015 and the Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2020. He is currently a Senior Research and Development Engineer with Momenta. His research interests include computer vision and machine learning.

Fan Tang received the B.Sc. degree in computer science from North China Electric Power University in 2013 and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2019. He is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics, computer vision, and machine learning.

Weiming Dong (Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from Tsinghua University, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from the University of Lorraine, France, in 2007. He is currently a Professor with the Sino-European Laboratory in Computer Science, Automation and Applied Mathematics (LIAMA) and the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences.

Changsheng Xu (Fellow, IEEE) is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He is an IAPR Fellow. He received the Best Associate Editor Award from *ACM Transactions on Multimedia Computing, Communications and Applications* in 2012 and the Best Editorial Member Award from *ACM/Springer Multimedia Systems Journal* in 2008. He is an ACM Distinguished Scientist.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:36 UTC from IEEE Xplore. Restrictions apply.