Spatial-Aware Token for Weakly Supervised Object Localization

Pingyu Wu¹, Wei Zhai¹‡, Yang Cao¹,³, Jiebo Luo², Zheng-Jun Zha¹
¹ University of Science and Technology of China  ² University of Rochester
³ Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

Abstract

Weakly supervised object localization (WSOL) is a challenging task aiming to localize objects with only image-level supervision. Recent works apply visual transformer to WSOL and achieve significant success by exploiting the long-range feature dependency in self-attention mechanism. However, existing transformer-based methods synthesize the classification feature maps as the localization map, which leads to optimization conflicts between classification and localization tasks. To address this problem, we propose to learn a task-specific spatial-aware token (SAT) to condition localization in a weakly supervised manner. Specifically, a spatial token is first introduced in the input space to aggregate representations for localization task. Then a spatial aware attention module is constructed, which allows spatial token to generate foreground probabilities of different patches by querying and to extract localization knowledge from the classification task. Besides, for the problem of sparse and unbalanced pixel-level supervision obtained from the image-level label, two spatial constraints, including batch area loss and normalization loss, are designed to compensate and enhance this supervision. Experiments show that the proposed SAT achieves state-of-the-art performance on both CUB-200 and ImageNet, with 98.45% and 73.13% GT-known Loc, respectively. Even under the extreme setting of using only 1 image per class from ImageNet for training, SAT already exceeds the SOTA method by 2.1% GT-known Loc. Code and models are available at https://github.com/wpy1999/SAT.

1. Introduction

Weakly supervised object localization (WSOL) aims to localize objects with only image-level labels available. Since no expensive bounding box or pixel-level annotation is required, WSOL significantly reduces the cost of manual annotations [21, 44, 12, 13, 4, 43, 22, 45] and has attracted increasing attention in the research community [40, 42, 41, 39, 10, 27, 29, 28, 14, 38, 23].

As a representative work, CAM [48] extracts class activation maps from the classifier as localization maps. However, CAM is usually coarse and focuses on the most discriminative regions, leading to imprecise and incomplete localization results. To solve these problems, many CNN-based methods have been proposed, such as adversarial erasure [22, 45, 6, 16], divergent activation [36, 40, 22], seed region growing [30, 46], regularization [18, 13, 47], feature refining [1, 37, 30], regression-based [35, 43, 9].

Recently, transformer [7, 24] has been introduced to the field of computer vision with great success. Benefiting from the long-range feature dependency, the attention map upon the class token usually can capture the object region relevant to classification in the whole image, thus it is widely used for localization in the transformer-based WSOL methods. The pioneering work TS-CAM [8] accumulates attention maps of each layer and mixes them with semantic-aware maps to achieve localization task. Further, to address the problem that transformer lacks inherent spatial coherence
of the object, LCTR [3] and SCM [2] propose to consider cross-patch information and use activation diffusion to increase the local continuity of localization map, respectively.

Existing transformer-based methods synthesize the feature maps learned by the classification task, such as attention maps, as the localization map, and try to increase its connectivity and completeness. However, this convenient approach results in optimization conflicts between classification and localization tasks. 1) For classification, making classification feature maps learn more object regions with low discrimination will reduce the classification ability. 2) For localization, the learning of the localization map is limited by the properties of the feature maps. For example, the attention map is generated by the softmax function, which is difficult to produce a balanced and comprehensive response over the object. Therefore, to achieve a more promising localization performance and avoid optimization conflicts, it is necessary to construct task-specific parameters for the generation and learning of localization map.

Based on transformer architecture, a straightforward idea is to introduce a spatial token in the input space to aggregate global representation for the localization task. And the localization map can be further obtained by interacting the spatial token with each patch in the forward propagation. To this end, we propose a Spatial-Query Attention (SQA) module that takes the spatial token as a query to calculate the similarity with different patches and produces the localization map efficiently. Meanwhile, to make the spatial-aware token obtain localization supervision from image-level labels, the localization map is treated as a visual cue to participate in the calculation of cross attention, thus the generation and learning of localization map can be well adapted to the transformer-based classification model.

Building upon this efficient SQA module, we propose a simple but effective Spatial-Aware Token (SAT) approach, as shown in Fig. 2. Specifically, a task-specific spatial token is introduced in the input space. To achieve the localization task, several SQA modules are applied in the transformer to produce localization maps learned from different layers and aggregate them together. In this way, the final localization map can maximally capture the localization knowledge from the whole classification model. However, the pixel-level supervision generated by image-level labels is sparse and unbalanced. To compensate and strengthen this supervision, we propose two spatial constraints, including batch area loss and normalization loss. Batch area loss aims to provide a sparse area supervision with prior knowledge to compensate for the insufficient supervision. Normalization loss is employed to enhance pixel-level supervision by encouraging the localization map to be more discriminatory.

Benefiting from the task-specific spatial token and avoiding optimization conflicts, SAT releases the potential of the model and achieves excellent performance in both classification and localization tasks. Besides, unlike synthetic classification feature maps as localization results, the generation and learning of localization map in SAT rely mainly on an extra lightweight token, which brings advantages of data-efficiency and tuning-efficiency. As illustrated in Fig. 1, SAT outperforms SOTA method SCM by 2.1% GT-known Loc with less than 0.1% training data on ImageNet.

In summary, the contributions of this paper include:

1) We propose to learn a task-specific spatial token to implement the localization task, instead of synthesizing classification feature maps, thus avoiding optimization conflicts between classification and localization tasks.
2) We propose a simple but effective spatial-aware token (SAT) pipeline for WSOL, which utilizes a spatial-query attention (SQA) module to generate localization map through a spatial token and supervise it with two spatial constraints, including batch area loss and normalization loss.
3) Extensive experiments show that SAT outperforms SOTA methods by a large margin on multiple benchmarks and performs excellently even under extreme settings.

2. Related work

CNN-based methods. WSOL aims to learn the localization of objects with only image-level labels. CAM [48] proposes to synthesize depth feature maps with fully connected weights to obtain class activation maps (CAM). To increase the localization ability of CAM, HaS [22] adopts a random erasure strategy forcing the network to attend to different object regions. ACoL [45] and EIL [16] use adversarial erasing to achieve learning of complementary regions. ADL [4] drops the most discriminative or highlighted regions in the different layers during forward propagation. Besides, SPA [18] designs a post-process approach to extract structure-preserving localization map from the feature maps. SPG [46] and SPOL [30] facilitate learning of localization by selecting and spreading confidence regions as seeds. PSOL [43] and C2AM [34] divide WSOL into independent classification task and class-agnostic object localization. I2C [47] and ISIC [31] consider feature similarities across different objects to achieve more complete and robust localization. ORNet [33], FAM [17] and BAS [32] propose to generate a foreground prediction map (FPM) to implement localization. Unlike FPM-based methods that require complex and heavy structures, and the learning of generator relies on a specific feature layer of the classification network. Our method simply and efficiently obtains localization maps from different layers by a task-specific token.

Transformer-based methods. Different from CNNs, transformer [7, 25] can capture global cues with the advantage of long-range dependency in the self-attention mechanism, effectively alleviating the partial activation problem. TS-CAM [8] mixes the attention maps of different layers and combines them with semantic-aware tokens to produce
semantic-aware localization results. Based on TS-CAM, to address the lack of spatial consistency in transformer, LCTR [3] proposes to enhance the local perception capability among long-range feature dependencies by considering cross-patch information. SCM [2] increases the spatial coherence of attention maps by diffusing the semantic and spatial connections between patch tokens. Notably, they all generate localization map by synthesizing feature maps. Instead, we produce it by a task-specific spatial token.

3. Methodology

3.1. Overview

The overall network architecture of SAT is shown in Fig. 2. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, a plain vision transformer splits it into a sequence of non-overlapping patches. Then the divided patches are flattened and transformed into patch tokens $\{\mathbf{x}_n \in \mathbb{R}^{1 \times D}, n = 1, 2, ..., H \times W\}$ by linear projection, where $H = h/P$, $W = w/P$, $P$ is the patch size, and $D$ is the number of channels. For simplicity, we omit the description of the batch size $B$. After grouping the class token and the extra spatial-aware token $\mathbf{x}_{spa}$ with patch tokens, this token sequence $\mathbf{X}^0 \in \mathbb{R}^{(N+2) \times D}$ is sent into stacked transformer blocks and spatial aware transformer blocks for subsequent representation learning.

The proposed spatial aware transformer block inherits the typical transformer block structure. Thus the entire network is characterized by alternating attention layers and MLP layers. For a given input token subsequent $\mathbf{X}^{l-1}$, the overall equations of the $l$-th block are defined as follows:

$$\mathbf{X}^l = \text{Layer-Norm} \left( \mathbf{X}^{l-1} \right),$$

$$\mathbf{Z}^l = \mathbf{X}^{l-1} + \text{Attention} \left( \mathbf{X}^l \right),$$

$$\mathbf{F}^l = \mathbf{Z}^l + \text{MLP} \left( \text{Layer-Norm} \left( \mathbf{Z}^l \right) \right).$$

3.2. Spatial Aware Transformer Block

The proposed spatial aware transformer block mainly contains a spatial-query attention (SQA) module to generate localization map and extract localization knowledge. Besides, two spatial constraints, including batch area loss $\mathcal{L}_{ba}$, and normalization loss $\mathcal{L}_{norm}$ are designed as the complement to the insufficient localization supervision.

Spatial-query Attention. The proposed SQA module is based on self-attention module and aims at converting the spatial-aware token $\mathbf{x}_{spa}$ into a localization map $\mathbf{M}^l_{spa}$ by querying different patches. Meanwhile, the generated $\mathbf{M}^l_{spa}$ is applied to the cross-attention calculation in the form of dot product, thus obtaining localization supervision from the image-level labels. Specifically, the SQA module first linearly projects the token sequence $\mathbf{X}^{l-1}$ to the query matrix $\mathbf{Q}$, key matrix $\mathbf{K}$, value matrix $\mathbf{V}$, where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{(N+2) \times D}$. Then the query vector corresponding to the spatial token $\mathbf{Q}_{spa} \in \mathbb{R}^{1 \times D}$ is selected from $\mathbf{Q}$ and used to query $\mathbf{K}$ to obtain the query results for each token. The similarity map between the spatial query vector and key matrix in the $l$-th layer is calculated as follows:

$$\mathbf{S}^l(\mathbf{Q}_{spa}, \mathbf{K}) = \frac{\mathbf{Q}_{spa}\mathbf{K}^T}{\sqrt{D}} \in \mathbb{R}^{1 \times (N+2)},$$

where $\sqrt{D}$ is employed as a scaling factor. After applying sigmoid activation, the similarity map is transformed into foreground probability $\mathbf{M}^l_{pro}$ with a 0 to 1 interval distribution. This process can be expressed as follows:

$$\mathbf{M}^l_{spa} = \text{Sigmoid} \left( \mathbf{S}^l(\mathbf{Q}_{spa}, \mathbf{K}) \right) \in \mathbb{R}^{1 \times (N+2)}.$$  

The generated $\mathbf{M}^l_{spa}$ is served as a visual cue to point out the object region and combined in the calculation of attention with the form of dot product. In this way, the whole SQA module can be formulated in the following form:

$$\text{SQA}(\mathbf{X}^l) = \text{Softmax} \left( \frac{\mathbf{QK}^T}{\sqrt{D}} \right) \ast \mathbf{M}^l_{spa} \mathbf{V},$$

![Diagram of SAT block](image-url)
Figure 3. Visualization comparison. The ground-truth bounding boxes are in red, and the predicted bounding boxes are in green.

where $*$ denotes element-wise multiplication.

By cropping and reshaping, the part of the $M_p^l$ to the patch tokens can be transformed into a localization map $M^l \in \mathbb{R}^{H \times W}$. After obtaining $M^l$ learned from different layers, we take their averages as the final localization map $M$ to increase the completeness and robustness. Although this approach can maximally capture the localization information in the classification network, the pixel-level supervision obtained from image-level labels is sparse and unbalanced. To compensate and strengthen this supervision, we design batch area loss and normalization loss.

**Batch Area.** To alleviate the insufficient localization supervision caused by the absence of dense labels in WSOL, we propose to use batch area loss as a complement to localization supervision. It aims to constrain the average area of localization maps $\{M_b \in \mathbb{R}^{H \times W}, b = 1, 2, \ldots, B\}$ within a batch $B$ to a hyperparameter $\lambda$, as follows:

$$L_{ba} = \frac{1}{H \times W} \sum_{b} \sum_{i} \sum_{j} (\lambda - \frac{M_b(i, j)}{B \times H \times W}),$$

where $\lambda$ is a sparse area supervision with prior knowledge. The $\lambda$ is set to 0.25 and 0.35 on CUB-200 and ImageNet.

**Normalization.** For unbalanced pixel-level supervision, a normalization loss is applied to enhance the pixel-level supervision based on the pixel response intensity. It encourages pixels to be more distinguishable about whether they are foreground or background, as presented by Eq. 8. Before penalizing the uncertainty value of the localization map $M$, to further enhance its local consistency, we adopt a gaussian filter with a kernel size of 3x3 to establish spatial connections between adjacent patches. The standard deviation of gaussian kernel is set to 6 on all datasets. After applying gaussian filtering, the value of each patch is redistributed according to the adjacent patch values, and $M$ is converted to $M^*$. By computing the normalization loss with $M^*$ can further encourage $M$ to have local consistency and avoid noise. It can be expressed as follows:

$$L_{norm} = \frac{1}{H \times W} \sum_{i} \sum_{j} M^*(i, j)(1 - M^*(i, j)).$$

### 3.3. Total loss

We follow the design of classification head in the baseline method TS-CAM [8] to obtain the predicted probability $y$ and the cross-entropy classification loss $L_{cls}$. By jointly optimizing classification loss, batch area loss, and normalization loss, the total loss function is defined as follows:

$$L = L_{cls} + L_{ba} + L_{norm},$$

where no extra hyperparameter is used to balance each loss.

### 4. Experiment

#### 4.1. Experimental Setup

**Datasets.** We evaluate the proposed SAT on three popular datasets, including CUB-200 [26], ImageNet [20], OpenImages [5]. CUB-200 is a fine-grained dataset with 200 bird species, which consists of 5,994 images for training and 5,794 images for testing. ImageNet contains about 1.2 million training images and 50,000 validation images from 1,000 different classes. OpenImages contains 100 categories, it has 29,819, 2,500, and 5,000 samples in training, val, and test sets, respectively. Apart from the image-level
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
</tr>
<tr>
<td>CAM [48]</td>
<td>CVPR16</td>
<td>VGG16</td>
<td>41.06</td>
<td>50.66</td>
</tr>
<tr>
<td>ORNet [33]</td>
<td>ICCV21</td>
<td>VGG16</td>
<td>67.73</td>
<td>80.77</td>
</tr>
<tr>
<td>BAS [32]</td>
<td>CVPR22</td>
<td>VGG16</td>
<td>71.33</td>
<td>85.33</td>
</tr>
<tr>
<td>Kim et al. [11]</td>
<td>CVPR22</td>
<td>VGG16</td>
<td>70.83</td>
<td>88.07</td>
</tr>
<tr>
<td>CREAM [35]</td>
<td>CVPR22</td>
<td>VGG16</td>
<td>70.44</td>
<td>85.67</td>
</tr>
<tr>
<td>GCNet [15]</td>
<td>ECCV20</td>
<td>InceptionV3</td>
<td>58.58</td>
<td>71.00</td>
</tr>
<tr>
<td>SPA [18]</td>
<td>CVPR21</td>
<td>InceptionV3</td>
<td>53.59</td>
<td>66.50</td>
</tr>
<tr>
<td>FAM [17]</td>
<td>ICCV21</td>
<td>InceptionV3</td>
<td>70.67</td>
<td>–</td>
</tr>
<tr>
<td>CREAM [35]</td>
<td>CVPR22</td>
<td>InceptionV3</td>
<td>71.76</td>
<td>86.37</td>
</tr>
<tr>
<td>BAS [32]</td>
<td>CVPR22</td>
<td>InceptionV3</td>
<td>73.29</td>
<td>86.31</td>
</tr>
<tr>
<td>BagCAMS [49]</td>
<td>ECCV22</td>
<td>InceptionV3</td>
<td>60.07</td>
<td>–</td>
</tr>
<tr>
<td>SPOL [30]</td>
<td>CVPR21</td>
<td>ResNet50</td>
<td>80.12</td>
<td>93.44</td>
</tr>
<tr>
<td>BAS-WSOL [50]</td>
<td>CVPR22</td>
<td>ResNet50</td>
<td>66.65</td>
<td>–</td>
</tr>
<tr>
<td>BAS [32]</td>
<td>CVPR21</td>
<td>ResNet50</td>
<td>77.25</td>
<td>90.08</td>
</tr>
<tr>
<td>Kim et al. [11]</td>
<td>CVPR21</td>
<td>ResNet50</td>
<td>73.16</td>
<td>86.68</td>
</tr>
<tr>
<td>CREAM [35]</td>
<td>CVPR22</td>
<td>ResNet50</td>
<td>76.03</td>
<td>–</td>
</tr>
<tr>
<td>BagCAMS [49]</td>
<td>ECCV22</td>
<td>ResNet50</td>
<td>69.67</td>
<td>–</td>
</tr>
<tr>
<td>ISIC [31]</td>
<td>ECCV22</td>
<td>ResNet50</td>
<td>80.68</td>
<td>94.08</td>
</tr>
<tr>
<td>TS-CAM [8]</td>
<td>ICCV21</td>
<td>Deit-S</td>
<td>71.30</td>
<td>83.80</td>
</tr>
<tr>
<td>LCTR [3]</td>
<td>AAI122</td>
<td>Deit-S</td>
<td>79.20</td>
<td>89.90</td>
</tr>
<tr>
<td>SCM [2]</td>
<td>ECCV22</td>
<td>Deit-S</td>
<td>76.40</td>
<td>91.60</td>
</tr>
<tr>
<td>SAT (ours)</td>
<td>This Work</td>
<td>Deit-S</td>
<td>80.96</td>
<td>94.13</td>
</tr>
</tbody>
</table>

Table 1. Comparison with state-of-the-art methods. The best results are highlighted in bold, second are underlined.

<table>
<thead>
<tr>
<th>Setting</th>
<th>$\mathcal{L}<em>{cls}$ $\mathcal{L}</em>{ba}$ $\mathcal{L}_{norm}$ $F.$</th>
<th>Loc Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>✓</td>
<td>56.86</td>
</tr>
<tr>
<td>(b)</td>
<td>✓ ✓</td>
<td>58.21</td>
</tr>
<tr>
<td>(c)</td>
<td>✓ ✓ ✓</td>
<td>59.86</td>
</tr>
<tr>
<td>(d)</td>
<td>✓ ✓ ✓ ✓</td>
<td><strong>60.15</strong></td>
</tr>
</tbody>
</table>

Table 2. Ablation study of SAT on ImageNet. $F.$: Gaussian filter.

...labels, both CUB-200 and OpenImages provide pixel-level mask labels, which are only used in the testing phase.

**Metrics.** Following [32, 18, 50], for localization, we utilize GT-known localization accuracy (GT-known Loc), Top-1/Top5 localization accuracy (Top-1/Top-5 Loc), and maximal box accuracy (MaxBoxAccV2) [5] as evaluation metrics. GT-known Loc is correct indicating that the intersection over union (IoU) of the predicted bounding box and the ground-truth bounding box is 50% or more. Top-1/Top-5 Loc is correct when the ground-truth class belongs to Top-1/Top-5 prediction categories and GT-known Loc is correct. For mask, the peak intersection over union (pIoU) [43] and pixel average precision (PxAP) [5] are adopted as metrics.

**Implementation Details.** We evaluate our method on the Deit-S [24] pre-trained on ImageNet [20]. On CUB-200 [26], the training process lasts 30 epochs with a batch size of 256. For ImageNet [20], we train 5 epochs and set batch size to 512. In the training phase, the input images are resized to 256×256 and then randomly cropped to 224×224. In the inference phase, following [32, 9, 43], we adopt ten crop augmentations to obtain classification results and replace random crop with center crop for localization.

**4.2. Ablation Study**

In this subsection, we implement a series of ablation experiments with Deit-S [24] as the backbone. All experiments are conducted on ImageNet [20], a universal dataset, to increase the generalizability of the experimental results.

**Ablation studies of SAT components.** Table 2 shows the localization accuracy of SAT with different compositions, where we retain the structure of SAT as the baseline. It can be noted that the proposed baseline can effectively activate the object region and already exceeds SOTA method SCM with only $\mathcal{L}_{cls}$. Based on the baseline, the addition of $\mathcal{L}_{ba}$ significantly improves the GT-known Loc with a 1.78% gain, by limiting the expansion of the localization map and forcing it to focus on the regions relevant to the classifi-
cution. Then the use of $L_{\text{norm}}$ can effectively increase the distinction between the foreground and background regions, resulting in a remarkable increase of 1.71% GT-known Loc. Finally, applying gaussian filter can further eliminate noise and increase local connectivity, by considering the consistency of adjacent patches in the calculation of $L_{\text{norm}}$.

**Hyperparameters.** Fig. 4 presents the sensitivity of localization quality to the hyperparameters $\lambda$ and batch size in $L_{\text{ba}}$ (Eq. 7). $\lambda$ denotes the constraint for the average area of the localization maps. A larger $\lambda$ encourages localization map to learn more regions relevant to the classification. Conversely, when $\lambda$ is small, the region learned by the localization map is small but reliable. Fig. 4 (a) shows that the model behaves insensitive to $\lambda$, since GT-known Loc remains stable over a large interval. Batch size affects the calculation of the average area of the localization map within a batch. A sufficiently large batch size can ensure tolerance of area variation between instances. As shown in Fig. 4 (b), when the batch size is larger than 64, the effect of batch on GT-known Loc is limited, with a change of less than 0.5%.  

### 4.3. Performance

**Main Results.** Table 1 compares SAT with state-of-the-art methods on CUB-200 [26] and ImageNet [20]. Experiments show that SAT achieves stable and consistent improvements on both benchmark datasets and significantly outperforms all methods using various backbones. Compared to the CNN counterpart ResNet50-ISIC [31], which uses three separate networks, while we use only one network and surpass ResNet50-ISIC by 1.13% and 3.12% GT-known Loc on the CUB-200 and ImageNet, respectively. These quantitative results demonstrate the superiority of the proposed SAT in terms of simplicity and effectiveness. In addition, on CUB-200 [26], SAT achieves a remarkable performance of 80.96%/98.45% on Top-1/GT-known Loc, exceeding the baseline method TS-CAM [8] by 9.66%/10.75%. On ImageNet [20], we surpass the SOTA method SCM [2] by a large margin, achieving a 4.33% boost in GT-known Loc under the same Deit-S [24] backbone. To further demonstrate the effectiveness of SAT, we visualize the localization maps of our method and TS-CAM in Fig. 3. In addition to generating sharper and more complete localization maps than TS-CAM, SAT demonstrates robust localization ability in various complex and challenging scenarios. More accuracy results and visualizations on six datasets are provided in the supplementary materials.

**Localization Quality.** In Table 3, we compare the localization performance with SOTA methods under the MaxBoxAccV2 [5] criterion. Despite the limitation of localization map resolution ($14 \times 14$), we achieve excellent results under different IoU thresholds. Notably, at $\delta$ of 0.7, a strict criterion for the localization quality, the proposed SAT exceeds the previous best method by a significant margin, with 7.61% and 7.58% improvement on CUB-200 [26] and ImageNet [20], respectively. We further plot the IoU distribution of correct bounding boxes in Fig. 5, following DANet [36]. Compared to the latest SCM [2], with the same Deit-S [24] backbone, we obtain a median IoU gain of 2.9% and 6.1% on CUB-200 and ImageNet. The IoU distribution map demonstrates that the localization results generated by SAT have higher localization quality, especially on ImageNet [20], a universal and challenging dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Venue</th>
<th>Backbone</th>
<th>CUB-200 [26]</th>
<th>ImageNet [20]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\delta = 0.3$</td>
<td>$\delta = 0.5$</td>
</tr>
<tr>
<td>SAT (ours)</td>
<td>This Work</td>
<td>Deit-S</td>
<td>98.88 (0.24)</td>
<td>98.45 (1.85)</td>
</tr>
</tbody>
</table>

Table 3. **Localization quality.** Comparison of localization accuracy under different IoU thresholds ($\delta$). Best results are highlighted in **bold**, second are **underlined**. The last line indicates the improvement (decrease) of our method over the previous best method.
Segmentation Quality. We evaluate the segmentation performance of SAT on CUB-200 [26] and OpenImages [5] since pixel-level labels are available during evaluation. As shown in Table 4, the proposed SAT substantially exceeds all methods in pIoU and PxAP metrics. Especially compared to the baseline method TS-CAM [8], we achieve a remarkable increase of 10.66% and 15.89% pIoU on CUB-200 and OpenImages. To further analyze the segmentation quality, we draw the density distribution map about IoU and object size on CUB-200 in Fig. 6. It can be observed the proposed SAT achieves an overall improvement and more stable performance on arbitrary sized objects compared to TS-CAM. Nevertheless, there still exists an imbalance segmentation ability of different sized objects. This is mainly due to the limitation of the localization map resolution, which hinders the acquisition of fine segmentation results.

Threshold Sensitivity. Fig. 7 illustrates the curves of GT-known Loc under different thresholds on CUB-200 and ImageNet. Obviously, the curves corresponding to the proposed SAT entirely cover the curves of TS-CAM [8] and SCM [2], indicating that SAT performs stable and achieves better results than other methods at arbitrary threshold. Besides, SAT is not sensitive to thresholds and maintains high localization accuracy over a large threshold interval. This phenomenon shows that the localization map generated by SAT has high confidence and excellent visualization.

Generalizability. To validate the generalizability of the proposed SAT, we implement experiments on different Transformers and various scales. As reported in Table 6, the proposed approach can be well applied to Deit [24] and Conformer [19] on CUB-200 [26]. On Deit-T/Deit-B, we achieve 3.2%/4.3% GT-known Loc improvement compared to SCM [2], and 3.3%/2.1% Top-1 classification accuracy gain over TS-CAM [8]. Besides, SAT also achieves state-of-the-art results on Conformer-S. Quantitative experiments show that the proposed SAT achieves excellent performance in both tasks on various transformer backbones.

4.4. Discussions

Why spatial token-based structure is critical? To clarify this question, we follow TS-CAM [8] to construct a attention map-based baseline and apply the proposed losses on it for comparison. As shown in Table 5, the attention map-based method substantially reduces the classification and localization accuracy, with 2.52% and 3.39% decrease in Top-1 Cls and GT-known Loc, respectively. To explore how this structural difference affects the accuracy, we plot the curves of spatial losses during training in Table 5. It can be noticed the potential optimization conflicts of attention map-based method will lead to poor convergence of losses and insufficient learning, thus limiting the model to benefit from the losses. Experimental results show that the proposed spatial token-based structure can release the potential of the model by more efficient localization learn-
ing and avoiding optimization conflicts between two tasks, which are essential for the accuracy improvement.

To further verify the effectiveness of the proposed structure in generating localization map, we reproduce two FPM-based methods ORNet [33] and BAS [32] on the Deit-S. Both of them produce localization maps by a generator. As shown in Table 7, first, we note that the original BAS is not suitable for the transformer due to the skip-connection in transformer causing inaccurate evaluation of the masking effect. Second, both FPM-based methods suffer from incomplete localization problem. This may be because they can only mask the input of a specific layer, resulting in over-fitting the features that the specific feature layer focuses on. While the proposed structure can extract localization knowledge from different layers in an efficient form, thus improving localization ability and robustness. Quantitative and qualitative results demonstrate the effectiveness of SAT in terms of localization map generation and learning.

Data-efficiency and tuning-efficiency. To validate the data-efficiency of the proposed method, we evaluate the localization performance of SAT and TS-CAM [8] under few-shot setting in Fig. 8. Compared to TS-CAM, SAT achieves promising results with only a few training data. Even in the extreme setting of only 1 training image per class, we exceed TS-CAM trained with full data by 3.6% and 3.3% GT-known Loc on CUB-200 and ImageNet, respectively. The box plot verifies the effectiveness and efficiency of the proposed SAT that implements localization task by an extra lightweight spatial token and enables high degree of network parameter sharing, thus requiring only a few data to achieve excellent localization results. In contrast, TS-CAM requires large amounts of data to fine-tune the classification feature maps to serve the localization task.

To explore the tuning-efficiency of SAT, we freeze different parts of the network and evaluate the GT-known Loc of the proposed method under few-shot setting. All experiments are performed with ImageNet pre-trained weights. As shown in Table 8, SAT can still obtain competitive results when freezing most parameters of the pre-trained classification model. It can be analyzed that the task-specific spatial token can be well adapted to the classification features and plays a major role in the localization learning, which allows efficient implementation of the localization task with fewer parameters to be fine-tuned. Experiments demonstrate both the data-efficiency and the tuning-efficiency of SAT. When freezing 81% of the parameters and using only 1 image per class, SAT exceeds the SOTA method SCM [2] by 2.1% GT-known Loc on ImageNet.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>91.3</td>
<td>70.9</td>
<td>70.9</td>
</tr>
<tr>
<td>3</td>
<td>95.1</td>
<td>71.9</td>
<td>71.9</td>
</tr>
<tr>
<td>5</td>
<td>97.3</td>
<td>72.6</td>
<td>72.6</td>
</tr>
<tr>
<td>10</td>
<td>98.5</td>
<td>72.9</td>
<td>72.9</td>
</tr>
<tr>
<td>Full</td>
<td>3.6%</td>
<td>73.1</td>
<td>73.1</td>
</tr>
<tr>
<td>(a)</td>
<td>1%</td>
<td>70.9</td>
<td>70.9</td>
</tr>
<tr>
<td>(b)</td>
<td>24%</td>
<td>71.6</td>
<td>71.6</td>
</tr>
<tr>
<td>(c)</td>
<td>48%</td>
<td>71.8</td>
<td>71.8</td>
</tr>
<tr>
<td>(d)</td>
<td>71%</td>
<td>71.3</td>
<td>71.3</td>
</tr>
<tr>
<td>(e)</td>
<td>91%</td>
<td>70.9</td>
<td>70.9</td>
</tr>
</tbody>
</table>

Table 8. Frozen rate w.r.t. few-shot learning. The frozen parts are as follows: (a) None. (b) Attention layer of transformer blocks. (c) MLP layer of transformer blocks. (d) Transformer blocks. (e) Position embedding, projection, transformer blocks, and MLP layer of spatial aware transformer blocks. Fro.: Frozen.

![Localization error analysis](image)

Figure 9. Localization error analysis on CUB-200.

4.5. Limitations

As shown in Fig. 9, we count all localization errors (90 images) on CUB-200 test set (5,794 images) and classify the error causes into six categories, including object occlusion, localization more, water reflection, localization part, multiple instances, and label error. Among them, object occlusion causes the object to be split into two or more parts, resulting in incomplete localization results. Localization more is usually due to the positive effect of co-occurrence context on classification, leading to localizing confounding background regions. Water reflection is an inherent problem of WSOL and difficult to be solved with only image-level labels. Therefore, future work needs to consider more the interaction between object and context to overcome the problems of object occlusion and localization more.

5. Conclusion

This paper proposes to introduce a spatial-aware token (SAT) specific to the localization task, instead of synthesiz-
ing the classification feature map, thus avoiding optimization conflicts between classification and localization tasks. To this end, we construct a spatial-query attention (SQA) module to generate localization map and extract localization knowledge from the classification task. Meanwhile, two spatial constraints, including batch area loss and regularization loss, are designed to compensate for the insufficient supervision. Extensive experiments on multiple benchmarks verify the effectiveness and efficiency of the proposed SAT that surpasses previous methods by a large margin.

Acknowledgments. This work is supported by National Key R&D Program of China under Grant 2020AAA0105700, National Natural Science Foundation of China (NSFC) under Grants 62225207, U19B2038 and 62121002.

References


[40] 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/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 1


