Cross-token Guidance Transformer for Weakly Supervised Object Localization

Zhiwei Chen[1]*, Yiran Nie[2], Ruize Han[3], and Qinqin Zhou[4,5]

[1] School of Artificial Intelligence, Nanchang University, China
[2] School of Artificial Intelligence, Jiangxi Normal University, China
[3] Faculty of Computer Science and Artificial Intelligence, Shenzhen University of Advanced Technology, China
[4] School of Computer and Data Science, Fuzhou University, China
[5] Key Laboratory of Education Blockchain and Intelligent Technology (Guangxi Normal University), Ministry of Education, China
zhiweichen@ncu.edu.cn

Abstract. Weakly supervised object localization (WSOL) aims to train an object locator using only image-level annotations. Recent progress in WSOL has been predominantly driven by visual transformer architectures, which effectively model long-range feature dependencies through self-attention mechanisms and multilayer perceptrons. However, existing transformer-based approaches typically depend solely on location tokens for localization, neglecting valuable semantic insights provided by class tokens, resulting in diffuse activations and imprecise localization. In this paper, we propose a novel framework named Cross-token Guidance TRansformer (CGTR), which enhances semantic coherence and localization precision by facilitating effective interaction between class and location tokens. Specifically, we introduce an Attention Regulation Module (ARM) to globally refine spatial activations using semantic guidance from class tokens, mitigating contextual biases. Additionally, we develop a Filter Regulation Module (FRM) that applies local structural refinement guided by semantic information, enriching the semantic representation of localization maps. Extensive experimental evaluations demonstrate the effectiveness of the proposed CGTR, achieving substantial and consistent performance gains compared with previous approaches on both the CUB-200-2011 and ILSVRC datasets.

Keywords: Object localization · Weakly supervised learning · Visual transformer · Self-attention maps

1 Introduction

Weakly supervised object localization (WSOL) is a fundamental yet challenging task in computer vision, aiming to identify object locations in images using only image-level annotations during training. Compared to fully supervised object localization methods, which require extensive and costly annotations such

* Corresponding author.

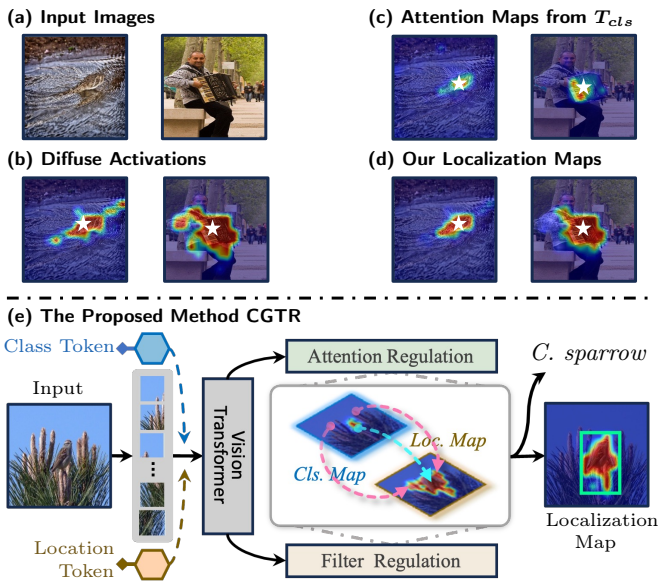


Fig. 1: (a) Input images. (b) Diffuse activation problem. Attention maps from the location token exhibit uncontrolled activations on target objects (*e.g.*, *Diamondback* and *Accordion*), as indicated by stars. (c) Attention maps obtained from the class token ($\mathbf{T}_{cls}$) effectively capture long-range feature dependencies but provide only coarse delineations of object regions. (d) The proposed CGTR precisely localizes and activates object regions. (e) To overcome the diffuse activation issue, we introduce an end-to-end CGTR framework that reduces diffuse activations by attention and filter regulation between location and class tokens. Predicted bounding boxes are depicted in green. Best viewed in color.

as bounding boxes or pixel-wise masks, WSOL significantly reduces the annotation requirements, thereby offering substantial advantages for large-scale visual recognition applications [6, 8, 39, 47, 52, 56].

The seminal work of Zhou *et al.* [56] introduced the Class Activation Map (CAM), leveraging features learned during classification to provide localization cues. However, CAM primarily activates the most discriminative object parts, resulting in incomplete localization. To overcome this limitation, numerous CAM-based methods [19, 27, 41, 42] have employed various strategies, including data augmentation, adversarial erasure, and spatial refinement techniques. Despite these advances, convolutional neural network (CNN)-based methods inherently suffer from limited receptive fields and insufficient global context modeling capabilities, leading to partial object activation issues [7, 12, 18, 28].

Vision transformers [33] recently emerged as powerful alternatives, excelling at modeling long-range dependencies within visual data [5, 12, 13, 30, 48]. More recently, to circumvent optimization conflicts between classification and localization within transformer architectures, Wu *et al.* [38] proposed utilizing a dedicated spatial token separate from the class token for object localization. However, such separation may neglect the beneficial semantic insights inherent

to classification, particularly without class token information during localization. Consequently, relying solely on the location token often leads to diffuse activations, as contextual noise from irrelevant elements negatively impacts localization precision, as shown in Fig. 1 (b). From these localization maps, it is evident that activated regions are prone to being perturbed by co-occurring contextual elements, such as {*Diamondback*, *Riverbed*}, {*Accordion*, *Person*}, {*Chipping Sparrow*, *Bushes*}. Fundamentally, the diffuse activation problem arises from the entangled contextual elements that co-occur within the image during feature representation. Consequently, this entanglement induces contextual biases, leading to diffuse activations in the localization maps. In contrast, we observe that attention maps derived via the class token can effectively disentangle these co-occurring contexts, as depicted in Fig. 1 (c). To attain more precise localization and alleviate the diffuse activation problem, it is crucial to establish effective interactions between location and class tokens.

Based on the analysis above, we propose a simple yet effective framework, termed **C**ross-token **G**uidance **TR**ansformer (CGTR), which explicitly exploits effective semantic guidance from classification for localization. As illustrated in Fig. 1 (e), our method aims to mitigate diffuse activations through robust interaction between location and class tokens. Specifically, we introduce an Attention Regulation Module (ARM) that utilizes global semantic information from the class token to refine spatial attention distributions within location tokens. By leveraging the knowledge from the class token, ARM effectively aligns the location token with category-specific regions, suppressing irrelevant contextual activations. Conversely, Filter Regulation Module (FRM) is further proposed to apply local structural refinement via semantic class token-guided filtering. FRM smoothes object structures identified in location tokens in accordance with class tokens, thereby generating more robust and precise localization maps. ARM and FRM thus synergistically operate at complementary scales—global semantic alignment and local spatial refinement—to comprehensively address diffuse activations. Importantly, both modules introduce no additional training parameters, significantly enhancing computational efficiency. To evaluate the effectiveness of the proposed CGTR, we perform extensive experiments on widely-used datasets. The primary contributions of this study are as follows:

- We introduce the Cross-token Guidance Transformer (CGTR) framework for weakly supervised object localization, explicitly bridging classification and localization tasks through enhanced cross-token interactions, effectively addressing diffuse activations.
- We propose an Attention Regulation Module (ARM) that leverages global semantic guidance from classification to localization, refining spatial activations and reducing contextual biases.
- We design a Filter Regulation Module (FRM) that performs detailed local refinement guided by semantic class tokens, ensuring precise and robust localization maps.
- The proposed CGTR shows consistent and substantial gains across the CUB-200-2011 and ILSVRC datasets.

2 Related Work

2.1 CNN-based methods for WSOL

Most weakly supervised object localization (WSOL) methods rely on the class activation maps (CAMs) [56], which generate object bounding boxes based on a classification network. However, these approaches typically focus only on the most discriminative object regions, often failing to capture its entire extent. To mitigate this limitation, various augmentation strategies have been proposed at both the image and feature levels, aiming to emphasize non-discriminative regions. For instance, Yun *et al*. [49] proposed a cut-and-paste strategy using patches from different images, pushing the model to focus on less discriminative regions. Similarly, Singh *et al*. [20] introduced random hiding of image patches during training, encouraging the model to explore different object parts. Junsk *et al*. [10] employed an erasure strategy on discriminative spatial locations to promote more comprehensive object coverage, while Mai *et al*. [23] utilized adversarial classifiers to mine diverse discriminative regions. Chen *et al*. [4] combined erasure with maxout learning to highlight the foreground without losing critical information. Other methods have focused on enhancing spatial relationships between object parts to achieve holistic localization. Pan *et al*. [26] leveraged structural information to distill structure-preserving features from convolutional outputs. Xue *et al*. [46] utilized discrepant activation learning to capture complementary and discriminative visual patterns, while Zhang *et al*. [53] enforced consistency constraints to improve the coherence of features within the same category. Additionally, Wu *et al*. [37] developed a background activation suppression strategy to enhance foreground prediction accuracy, and Xie *et al*. [40] employed foreground prediction maps for localization guidance. Zhu *et al*. [57] took a different approach by training a score estimator in the image domain and applying it to the pixel domain for object localization. Another line of WSOL research [14, 36, 44, 51, 55] has framed the task as two independent sub-tasks: classification and class-agnostic localization, aiming to decouple their respective challenges. Despite these efforts, CNN-based methods are still prone to capturing only partial object features. The inherent limitation in global feature perception often restricts these methods from activating the entire object, hindering complete object localization.

2.2 Transformer-based methods for WSOL

Recent advancements in vision transformers have demonstrated superior performance over convolutional neural networks (CNNs) in various computer vision tasks, primarily due to their exceptional ability to capture long-range dependencies [3, 11, 17, 18, 43, 54]. Dosovitskiy *et al*. [11] revealed that transformers can achieve outstanding image classification performance by directly processing sequences of image patches. Extending this foundation, Gao *et al*. [12] were the first to employ this architecture for WSOL by combining semantic-aware tokens with semantic-agnostic attention maps. Chen *et al*. [7] introduced learnable

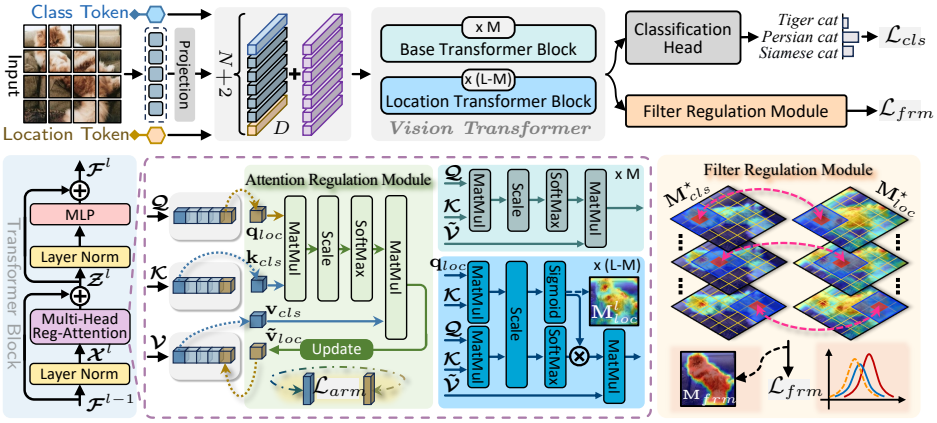


Fig. 2: Overview of the proposed CGTR. It consists of a vision transformer backbone, an attention regulation module (ARM), and a filter regulation module (FRM). The input sequence is processed by M base transformer blocks, followed by $(L-M)$ location transformer blocks. The final localization map is obtained by fusing all location attention maps $\mathbf{M}_{loc}^{l}$ with the regulatory localization map $\mathbf{M}_{frm}$.

kernels guided by class-token attention to emphasize local details within global representations. Bai *et al*. [1] modeled semantic similarities and spatial relationships among patch tokens. Gupta *et al*. [15] introduced a patch-based attention dropout mechanism to refine localization maps. Cao *et al*. [2] leveraged low-level image cues to explore both global and local representations within transformers. Su *et al*. [30] exploited the spatial-semantic consistency between shallow and deep transformer features. To address optimization conflicts, Wu *et al*. [38] proposed a task-specific spatial-aware token for object localization. Nevertheless, many existing transformer-based approaches neglect the valuable interplay between classification and localization. Our proposed CGTR effectively integrates cross-cue information among tokens, closely coupling classification with localization to enhance WSOL performance significantly.

3 Methodology

3.1 Overview

The overall network architecture of CGTR is depicted in Fig. 2. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, where H and W denote its height and width, respectively. Initially, the input image $\mathbf{I}$ is partitioned into $w \times h$ patches. These patches are subsequently flattened and projected through a linear transformation to generate a sequence of patch tokens $\mathbf{T}_p \in \mathbb{R}^{N \times D}$, where D denotes the dimension of each patch and $N = w \times h$ is the total number of patches. A learnable class token $\mathbf{T}_{cls} \in \mathbb{R}^{1 \times D}$ and a location token $\mathbf{T}_{loc} \in \mathbb{R}^{1 \times D}$ are then appended to the token sequence, along with a position embedding $\mathbf{T}_{pos} \in \mathbb{R}^{(N+2) \times D}$. Thus, the resulting input sequence $\boldsymbol{\mathcal{F}}^{0} \in \mathbb{R}^{(N+2) \times D}$ is fed into the transformer encoder, comprising

M base transformer blocks, followed by $(L - M)$ location transformer blocks. Here, L denotes the total number of blocks in the transformer encoder.

As illustrated in Fig. 2, given the token sequence $\boldsymbol{\mathcal{F}}^{l-1}$ from the preceding transformer block, we first apply a Layer Normalization [21] to obtain $\boldsymbol{\mathcal{X}}^{l} \in \mathbb{R}^{(N+2)\times D}$. The resulting sequence $\boldsymbol{\mathcal{X}}^{l}$ is subsequently processed by the Multi-Head Reg-Attention (MHRA). Specifically, MHRA consists of the proposed Attention Regulation Module (Sec. 3.2) along with self-attention layers. In the base transformer blocks, multi-head attention layers [33] are employed to capture long-range feature dependencies. In contrast, the location transformer blocks incorporate spatial-query attention layers [38], where the query vector of the location token dictates the attention output for each token. Next, a residual connection is established between the output of MHRA and $\boldsymbol{\mathcal{F}}^{l-1}$, resulting in $\boldsymbol{\mathcal{Z}}^{l}$. Subsequently, $\boldsymbol{\mathcal{Z}}^{l}$ passes through a Layer Normalization [21] and a Multi-Layer Perceptron (MLP), followed by a residual connection. This process yields the output sequence $\boldsymbol{\mathcal{F}}^{l}$ for the l-th transformer block. Following the transformer encoder, the Filter Regulation Module (Sec. 3.3) is introduced to refine the location token using a local filtering operation. Concurrently, the transformer encoder forwards the processed sequence to the classification head [12], producing the final image classification results.

For the object localization process, we derive the localization map by calculating the element-wise average between the location attention map $\mathbf{M}^{\star}_{loc}$ (Sec. 3.2) and the regulatory localization map $\mathbf{M}_{frm}$ (Sec. 3.3). Subsequently, the localization map is resized to the original image dimensions using linear interpolation. To generate the predicted bounding box, we identify the tightest bounding box that encloses the largest connected component of foreground pixels within the localization map, following previous approaches [6, 12, 56].

3.2 Attention Regulation Module

Although the self-attention mechanisms [33, 38] have demonstrated efficacy in expanding the localization map to activate non-discriminative regions, they often result in the unintended activation of co-occurring contextual elements. To address this, we propose an Attention Regulation Module (ARM) to suppress contextual activations by modulating the spatial relationships between location and class tokens.

To refine the attention distribution of the location token, as depicted in Fig. 2, the sequence $\boldsymbol{\mathcal{X}}^{l}$ first undergoes a linear projection to generate the query matrix $\boldsymbol{\mathcal{Q}}$, key matrix $\boldsymbol{\mathcal{K}}$, and value matrix $\boldsymbol{\mathcal{V}}$, where $\boldsymbol{\mathcal{Q}}, \boldsymbol{\mathcal{K}}, \boldsymbol{\mathcal{V}} \in \mathbb{R}^{(N+2)\times D}$. For simplicity, the batch dimension is omitted from these notations. Afterward, these matrices are reshaped to $\mathbb{R}^{S\times(N+2)\times P}$. Here, S represents the number of heads in the self-attention layers, and $P = D/S$. The transformed matrices are then passed through the proposed ARM for contextual suppression. Specifically, we extract the location query $\mathbf{q}_{loc}$ from $\boldsymbol{\mathcal{Q}}$, along with the class key $\mathbf{k}_{cls}$ and class value $\mathbf{v}_{cls}$ from $\boldsymbol{\mathcal{K}}$ and $\boldsymbol{\mathcal{V}}$, respectively. Note that $\mathbf{q}_{loc}, \mathbf{k}_{cls}, \mathbf{v}_{cls} \in \mathbb{R}^{S\times P}$. Next, we compute the attention scores between $\mathbf{q}_{loc}$ and $\mathbf{k}_{cls}$ using scaled dot-product

attention, formulated as:

$$\mathbf{s}_{lc} = \texttt{Softmax}\left(\frac{\mathbf{q}_{loc}\mathbf{k}_{cls}^{\top}}{\sqrt{D}}\right), \tag{1}$$

where $\sqrt{D}$ is a scaling factor and $\top$ denotes the transpose operator. After that, the spatial distribution of the location value $\mathbf{v}_{loc}$ is refined by incorporating category-specific knowledge. This process can be formulated as:

$$\tilde{\mathbf{v}}_{loc} = \mathbf{v}_{loc} + \mathbf{s}_{lc}\mathbf{v}_{cls}. \tag{2}$$

In this light, we update the value matrix $\boldsymbol{\mathcal{V}}$ by substituting $\mathbf{v}_{loc}$ with $\tilde{\mathbf{v}}_{loc}$, resulting in a category-aware value matrix $\tilde{\boldsymbol{\mathcal{V}}} \in \mathbb{R}^{S\times(N+2)\times P}$. By leveraging the prior knowledge that the class token in single-category images corresponds primarily to the associated category semantics, the spatial distribution of the location token is trained with enhanced semantic awareness.

Afterward, we apply the category regulation to the location token. In particular, $\boldsymbol{\mathcal{Q}}$, $\boldsymbol{\mathcal{K}}$, and $\tilde{\boldsymbol{\mathcal{V}}}$ are fed into the multi-head attention layers [33] or spatial-query attention layers [38]. To generate the location attention map, we follow the setup of [38], querying different patches with the location token $\mathbf{T}_{loc}$ in the location transformer blocks. Mathematically, given the location query matrix $\mathbf{q}_{loc} \in \mathbb{R}^{1\times D}$ and key matrix $\boldsymbol{\mathcal{K}} \in \mathbb{R}^{(N+2)\times D}$ in the l-th location transformer block, the location attention map $\mathbf{M}_{loc}^{l}$ can be computed as follows:

$$\mathbf{M}_{loc}^{l} = \mathit{\Gamma}^{w\times h}\left(\texttt{Sigmoid}\left(\frac{\mathbf{q}_{loc}\boldsymbol{\mathcal{K}}^{\top}}{\sqrt{D}}\right)\right), \tag{3}$$

where $\mathit{\Gamma}^{w\times h}(\cdot)$ represents the reshape operator, which extracts the attention vector ($\mathbb{R}^{1\times N}$) and converts it into a spatial map ($\mathbb{R}^{w\times h}$). Subsequently, an averaging operation is applied to the location attention maps from $(L-M)$ location transformer blocks, yielding the final location attention map $\mathbf{M}_{loc}^{\star} \in \mathbb{R}^{w\times h}$. Similarly, querying different patches with the class token $\mathbf{T}_{cls}$ produces the class attention map:

$$\mathbf{M}_{cls}^{l} = \mathit{\Gamma}^{w\times h}\left(\texttt{Softmax}\left(\frac{\mathbf{q}_{cls}\boldsymbol{\mathcal{K}}^{\top}}{\sqrt{D}}\right)\right), \tag{4}$$

where $\mathbf{q}_{cls} \in \mathbb{R}^{1\times D}$ denotes the class query derived from $\mathcal{Q}$. The final class attention map $\mathbf{M}_{cls}^{\star} \in \mathbb{R}^{w\times h}$ is computed by averaging $\mathbf{M}_{cls}^{l}$ across $(L-M)$ location transformer blocks. Finally, an area loss [24,38,40] is employed to enforce alignment between attention maps and the corresponding object regions. This can be formulated as follows:

$$\mathcal{L}_{arm} = \left|\left(\frac{1}{hw}\sum_{i}^{h}\sum_{j}^{w}\left|\mathbf{M}_{loc(i,j)}^{\star} - \mathbf{M}_{cls(i,j)}^{\star}\right|\right) - \xi\right|, \tag{5}$$

where $|\cdot|$ represents the absolute value operation, and ξ is a hyper-parameter.

3.3 Filter Regulation Module

To mitigate the diffuse activation problem, we further propose a Filter Regulation Module (FRM), which transfers structural properties from the class token to the location token, thereby enhancing the object semantic representation.

In order to obtain the long-range structural cues, we derive the location attention map $\mathbf{M}_{loc}^{\star}$ and class attention map $\mathbf{M}_{cls}^{\star}$ from the location token $\mathbf{T}_{loc}$ and the class token $\mathbf{T}_{cls}$, respectively, as detailed in Sec. 3.2. Then, we perform the local filtering to learn the relationship between $\mathbf{M}_{loc}^{\star}$ and $\mathbf{M}_{cls}^{\star}$ in local patches. Inspired by the guided filtering [16], we first compute the local means $\mathbf{m}_{loc}^{avg}$ and $\mathbf{m}_{cls}^{avg}$ of $\mathbf{M}_{loc}^{\star}$ and $\mathbf{M}_{cls}^{\star}$, respectively, through an average pooling operation ($\texttt{AVG}$). The pooling process is formulated as follows:

$$\mathbf{m}_{loc}^{avg} = \texttt{AVG}_k\left(\mathbf{M}_{loc}^{\star}\right),\ \mathbf{m}_{cls}^{avg} = \texttt{AVG}_k\left(\mathbf{M}_{cls}^{\star}\right), \tag{6}$$

where $\texttt{AVG}_k\left(\cdot\right)$ represents an average pooling operation with a $k \times k$ kernel. Next, we calculate the variance $\mathbf{m}_{cc}^{var}$ of $\mathbf{M}_{cls}^{\star}$, which captures the intensity variation within the kernel window, as follows:

$$\mathbf{m}_{cc}^{var} = \texttt{AVG}_k\left(\mathbf{M}_{cls}^{\star} \cdot \mathbf{M}_{cls}^{\star}\right) - \mathbf{m}_{cls}^{avg} \cdot \mathbf{m}_{cls}^{avg}, \tag{7}$$

where $\cdot$ denotes element-wise multiplication. The covariance $\mathbf{m}_{lc}^{cov}$, which represents the relationship between $\mathbf{M}_{loc}^{\star}$ and $\mathbf{M}_{cls}^{\star}$, is computed as follows:

$$\mathbf{m}_{lc}^{cov} = \texttt{AVG}_k\left(\mathbf{M}_{loc}^{\star} \cdot \mathbf{M}_{cls}^{\star}\right) - \mathbf{m}_{loc}^{avg} \cdot \mathbf{m}_{cls}^{avg}. \tag{8}$$

Subsequently, the linear coefficients of the local filter are determined as follows:

$$\mathbf{w}_{loc} = \frac{\mathbf{m}_{lc}^{cov}}{\mathbf{m}_{cc}^{var} + \varepsilon}, \tag{9}$$

$$\mathbf{b}_{loc} = \mathbf{m}_{loc}^{avg} - \mathbf{w}_{loc} \cdot \mathbf{m}_{cls}^{avg}, \tag{10}$$

where ε is a small constant (e^{-6}) used to prevent overfitting to specific local regions. The computed coefficients $\mathbf{w}_{loc}$ and $\mathbf{b}_{loc}$ are further processed through $\texttt{AVG}_k\left(\cdot\right)$ to obtain $\mathbf{w}_{loc}^{avg}$ and $\mathbf{b}_{loc}^{avg}$. The regulatory localization map $\mathbf{M}_{frm} \in \mathbb{R}^{w \times h}$ is then obtained by applying these coefficients to the class attention map $\mathbf{M}_{cls}^{\star}$, as defined by:

$$\mathbf{M}_{frm} = \Phi\left(\mathbf{w}_{loc}^{avg} \cdot \mathbf{M}_{cls}^{\star} + \mathbf{b}_{loc}^{avg}\right), \tag{11}$$

where $\Phi\left(\cdot\right)$ denotes a clamping operation to ensure the values are normalized within the range $[0, 1]$. This spatially-varying linear filtering process refines $\mathbf{M}_{loc}^{\star}$ using structural information from $\mathbf{M}_{cls}^{\star}$, resulting in a more complete and precise activation of $\mathbf{M}_{frm}$.

Finally, we introduce an entropy regularization loss to penalize misalignment with the expected activation pattern in the regulatory localization map, defined as:

$$\begin{aligned}\mathcal{L}_{frm} = -\Big(&\mathbf{M}_{frm} \times \log\left(\mathbf{M}_{frm} + \epsilon\right) \\ &+ \left(1 - \mathbf{M}_{frm}\right) \times \log\left(1 - \mathbf{M}_{frm} + \epsilon\right)\Big).\end{aligned} \tag{12}$$

Here, ϵ is a small constant (e^{-6}) to ensure numerical stability.

3.4 Training Objectives

As shown in Fig. 2, CGTR training objectives comprise three components: $\mathcal{L}_{cls}$, $\mathcal{L}_{arm}$ and $\mathcal{L}_{frm}$. In particular, $\mathcal{L}_{cls}$ is the cross-entropy loss for classification. The total loss function for our CGTR is defined as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{arm} + \lambda \mathcal{L}_{frm}, \tag{13}$$

where λ is a hyper-parameter for balancing the total loss.

4 Experiment

4.1 Settings

Datasets. The proposed method is evaluated on two widely-used benchmarks: CUB-200-2011 [34] and ILSVRC [29]. CUB-200-2011 comprises 200 bird species, with $5,994$ training images and $5,794$ testing images. ILSVRC contains approximately 1.2 million training images and $50,000$ validation images across $1,000$ categories. Only image-level labels are employed in the training phase, both image-level labels and bounding box annotations are leveraged for evaluation. Moreover, CUB-200-2011 provides pixel-level mask annotations, which are solely used for evaluation.

Evaluation Metrics. Following previous studies [9, 12, 46, 56], we utilize GT-known localization accuracy (GT-known Loc), Top-1/Top-5 localization accuracy (Top-1/Top-5 Loc), Top-1/Top-5 classification accuracy (Top-1/Top-5 Cls), and maximal box accuracy (MaxBoxAccV2) as the main evaluation metrics. GT-known Loc measures the proportion of images where the predicted bounding box has an intersection over union (IoU) greater than 50% with the ground-truth box. Top-1 Loc measures the proportion of images where the predicted class (*i.e.*, Top-1 Cls) is correct, and the predicted bounding box overlaps with the ground-truth bounding box by more than 50% IoU. Top-5 Loc measures the proportion of images where the true class label is among the Top-5 predictions, and the predicted bounding box has an IoU above 50% with the ground-truth bounding box. Besides, we use the pixel average precision (PxAP) [9] to evaluate the mask quality of the localization map on CUB-200-2011.

Implementation Details. The proposed CGTR is constructed using the Deit-S backbone [32], which is pre-trained on ILSVRC [29]. In particular, we replace the MLP head with a convolutional layer, followed by a global average pooling layer, similar to the baseline approach [12]. The input images are resized to 256×256 and then randomly cropped to 224×224. The network is trained using AdamW optimizer [22], with hyper-parameters ϵ=$1e^{-8}$, β_1=0.9, β_2=0.99, and a weight decay of $5e^{-4}$. On CUB-200-2011, the network is trained for 30 epochs with a batch size of 256 and an initial learning rate of $1e^{-4}$. On ILSVRC, training is performed for only one epoch with a batch size of 512 and an initial learning rate of $1.5e^{-5}$. All experiments are conducted on A800 GPUs.

Table 1: Comparison of localization accuracy and classification accuracy with the state-of-the-art methods.

Methods	Backbone	CUB-200-2011					ILSVRC				
		Loc. Acc			Cls. Acc		Loc. Acc			Cls. Acc	
		Top-1	Top-5	GT-k.	Top-1	Top-5	Top-1	Top-5	GT-k.	Top-1	Top-5
CAM [CVPR16] [56]	VGG16	44.15	52.16	56.00	76.60	92.50	42.80	54.86	59.00	66.60	88.60
SPA [CVPR21] [26]	VGG16	60.27	72.50	77.29	76.11	92.15	–	–	–	–	–
ORNet [CVPR21] [40]	VGG16	67.73	80.77	86.20	77.00	93.00	52.05	63.94	68.27	71.60	90.40
FAM [ICCV21] [24]	VGG16	69.26	–	89.26	77.26	–	51.96	–	71.73	70.90	–
BGC [CVPR22] [19]	VGG16	70.83	88.07	93.17	–	–	49.94	63.25	68.92	–	–
DGL [TMM22] [31]	VGG16	65.05	78.58	84.29	–	–	49.08	60.62	66.23	–	–
TAFormer [TPAMI22] [25]	VGG16	72.02	85.94	90.84	79.13	–	53.42	67.73	74.02	70.67	–
BAS [IJCV23] [50]	VGG16	70.90	85.36	91.04	–	–	52.94	65.38	69.66	–	–
IDC [TIP24] [35]	VGG16	72.36	86.06	93.36	–	–	53.88	67.05	72.45	–	–
AZL [TNNLS25] [8]	VGG16	72.04	86.16	91.47	77.77	93.56	53.00	65.50	69.88	70.97	90.54
CAM [CVPR16] [56]	InceptionV3	41.06	50.66	55.10	73.80	91.50	46.29	58.19	62.68	73.30	91.80
I^2C [ECCV20] [53]	InceptionV3	–	–	–	–	–	53.11	64.13	68.50	73.30	91.60
SPA [CVPR21] [26]	InceptionV3	53.59	66.50	72.14	73.51	91.39	52.73	64.27	68.33	73.26	91.81
DGL [TMM22] [31]	InceptionV3	54.92	68.10	74.11	–	–	53.91	64.64	68.91	–	–
TAFormer [TPAMI22] [25]	InceptionV3	73.32	84.08	88.66	82.22	–	56.00	66.49	69.79	77.76	–
BAS [IJCV23] [50]	InceptionV3	72.09	88.11	94.63	–	–	58.50	69.03	72.07	–	–
IDC [TIP24] [35]	InceptionV3	75.10	87.75	93.83	–	–	59.58	70.17	73.12	–	–
AZL [TNNLS25] [8]	InceptionV3	73.74	88.23	93.66	77.99	93.76	58.62	69.30	72.14	78.00	94.12
CAM [CVPR16] [56]	ResNet50	46.71	54.44	57.35	80.26	–	38.99	49.47	51.86	–	–
FAM [ICCV21] [24]	ResNet50	73.74	–	85.73	82.72	–	54.46	–	64.56	76.48	–
DAL [CVPR22] [57]	ResNet50	66.65	–	81.83	–	–	55.84	–	70.27	–	–
BGC [CVPR22] [19]	ResNet50	73.16	86.68	91.60	–	–	53.76	65.75	69.89	–	–
CREAM [CVPR22] [42]	ResNet50	76.03	–	89.88	–	–	55.66	–	69.31	–	–
DGL [TMM22] [31]	ResNet50	62.24	75.15	80.82	–	–	55.03	64.64	68.64	–	–
TAFormer [TPAMI22] [25]	ResNet50	74.96	87.78	91.96	82.91	–	57.45	69.86	75.54	76.84	–
BAS [IJCV23] [50]	ResNet50	76.75	90.04	95.41	–	–	58.50	69.03	72.07	–	–
IDC [TIP24] [35]	ResNet50	77.64	91.37	95.62	–	–	57.74	69.32	72.90	–	–
AZL [TNNLS25] [8]	ResNet50	77.56	89.43	94.37	81.31	94.27	57.20	68.60	71.60	76.10	93.23
TS-CAM [ICCV21] [12]	Deit-S	71.30	83.80	87.70	80.30	94.80	53.40	64.30	67.60	74.30	92.10
LCTR [AAAI22] [7]	Deit-S	79.20	89.90	92.40	**85.00**	**97.10**	56.10	65.80	68.70	77.10	93.40
SCM [ECCV22] [1]	Deit-S	76.40	91.60	96.60	78.50	94.50	56.10	66.40	68.80	76.70	93.00
SAT [ICCV23] [38]	Deit-S	80.96	94.13	98.45	–	–	60.15	70.52	73.13	78.41	–
CATR [ICCV23] [5]	Deit-S	79.62	92.08	94.94	83.72	96.82	56.90	66.64	69.25	77.25	93.64
CIAT [IJCAI24] [30]	Deit-S	77.90	92.20	97.10	–	–	59.80	69.90	72.10	**78.60**	94.20
TMT [TMM24] [45]	Deit-S	80.26	93.84	98.00	–	–	58.22	68.54	71.30	–	–
CDTR [TPAMI25] [6]	Deit-S	81.33	94.06	96.89	83.87	97.03	58.20	68.05	70.72	77.66	93.67
CGTR (ours)	**Deit-S**	**81.80**	**94.40**	**98.68**	82.29	95.65	**60.35**	**70.68**	**73.33**	78.39	**94.34**

4.2 Comparison with State-Of-The-Arts

Main Results. Tab. 1 presents the localization performance of the proposed CGTR compared to recent state-of-the-art methods on CUB-200-2011 [34] and ILSVRC [29]. The results demonstrate that CGTR achieves stable and consistent improvements across both datasets, outperforming all compared methods with various backbones. In particular, CGTR demonstrates superior localization performance over CNN-based methods, such as CAM [56] and BGC [19], validating the effectiveness of capturing long-range dependencies to improve localization accuracy. Moreover, on CUB-200-2011 [34], CGTR achieves a Top-1 Loc of 81.80% and a GT-known Loc of 98.68%, surpassing the baseline TS-CAM [12] by 10.50% and 10.98%, respectively. While CIAT [30] and SAT [38] exhibit strong GT-known Loc performance, achieving 97.10% and 98.45%, respectively, CGTR remains highly competitive with a GT-known Loc of 98.68%. Tab. 1 also provides a comparison of localization performance on ILSVRC [29]. CGTR achieves Top-1 Loc and GT-known Loc of 60.35% and 73.33%, surpassing all CNN-based and transformer-based competitors. Notably, compared to the baseline TS-CAM [12], CGTR yields a notable increase of 6.95% in Top-1 Loc, demonstrating the effectiveness of our approach.

Fig. 3: Visualization of localization maps. Ground-truth boxes are in red, predictions are in green, with IoU values (%) in white text. Please refer to the supplementary material for more visualization.

Table 2: Localization quality on ILSVRC. Comparison of localization accuracy under different IoU thresholds ($\boldsymbol{\delta}$).

Method	Backbone	MaxBoxAccV2 [9]			
		$\boldsymbol{\delta}$=0.3	$\boldsymbol{\delta}$=0.5	$\boldsymbol{\delta}$=0.7	Mean
TS-CAM [ICCV21] [12]	Deit-S	82.11	67.60	44.76	64.82
SCM [ECCV22] [1]	Deit-S	83.64	68.80	45.30	65.91
SAT [ICCV23] [38]	Deit-S	84.42	73.13	**56.83**	71.46
CGTR (ours)	**Deit-S**	**84.66**	**73.33**	56.52	**71.50**

Visual Comparison. Fig. 3 provides a visual comparison of the localization maps between our proposed CGTR and the baseline TS-CAM [12] on CUB-200-2011 [34] and ILSVRC [29]. The qualitative results indicate that our method produces more precise localization maps, successfully capturing the entire spatial extent of target objects. For instance, in the final column of Fig. 3, TS-CAM predominantly focuses on the head of the cat, whereas our approach accurately highlights the complete cat, thus providing a more comprehensive feature representation.

Localization Quality. We conduct a comprehensive statistical analysis of the Intersection over Union (IoU) between predicted and ground-truth bounding boxes, as illustrated in Fig. 4, following DANet [46]. On CUB-200-2011 [34], CGTR achieves a median IoU of 82.40%, outperforming the baseline TS-CAM [12] by 7.36%. Similarly, on ILSVRC [29], our method shows an improvement of 5.70%. Compared to the recent SAT [38], our method yields a median IoU improvement of 1.21% and 1.24% on CUB-200-2011 and ILSVRC, respectively. These results demonstrate the effectiveness of our method in enhancing localization precision across both datasets. Furthermore, we use the MaxBoxAccV2 [9] criterion to evaluate localization quality on ILSVRC [29], as summarized in

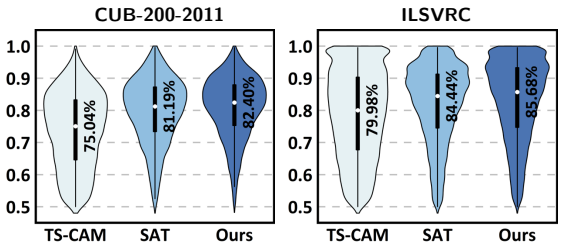


Fig. 4: Statistical analysis of the IoU between predicted and ground-truth boxes.

Table 3: Mask quality of localization maps on CUB-200-2011.

Method	Backbone	PxAP [9]
TS-CAM [ICCV21] [12]	Deit-S	81.49
SCM [ECCV22] [1]	Deit-S	88.15
SAT [ICCV23] [38]	Deit-S	89.87
CGTR (ours)	**Deit-S**	**90.22**

Table 4: Abalation study of CGTR. Top: ILSVRC, Bottom: CUB-200-2011.

	Vanilla	ARM	FRM	Loc. Acc		
				Top-1	Top-5	GT-known
(a)	✓			56.86	66.56	69.10
(b)	✓	✓		58.57	68.39	70.91
(c)	✓		✓	58.22	68.11	70.58
CGTR	✓	✓	✓	**60.35**	**70.68**	**73.33**
(d)	✓			76.25	87.87	91.77
(e)	✓	✓		79.43	91.89	95.96
(f)	✓		✓	78.36	90.46	94.39
CGTR	✓	✓	✓	**81.80**	**94.40**	**98.68**

Tab. 2. For the localization performance across varying IoU thresholds (*i.e.*, $\delta \in \{0.3, 0.5, 0.7\}$), the proposed CGTR achieves highly competitive results. Specifically, our method attains the best localization accuracy at δ=0.3 and δ=0.5, with slightly lower accuracy than SAT [38] at δ=0.7. Moreover, it is noteworthy that CGTR achieves the best average localization accuracy of 71.50%, which is a more robust metric for assessing overall localization quality.

Mask Quality. The predicted bounding box is obtained by generating a tight enclosure around the foreground pixels within the localization map; consequently, a high-quality localization map more closely approximates the ground-truth mask of the object. Following Choe *et al.* [9], we evaluate the mask quality of the localization map on CUB-200-2011 [34], which provides the ground-truth masks for comparison. As presented in Tab. 3, the proposed CGTR achieves a PxAp of 90.22%, outperforming all competing methods. Notably, compared to the baseline TS-CAM [12], our method yields a substantial improvement of 8.73% in PxAp. Furthermore, we visualize the localization maps and the ground-truth mask in the right panel of Tab. 3. It is observed that TS-CAM [12] has difficulty capturing the long-range feature dependencies on the wings of *Frigatebird*, whereas SAT [38] activates the object but lacks precision, particularly around the tail (yellow dashed rectangle). In contrast, CGTR effectively reduces background noise and enhances the precision around challenging regions.

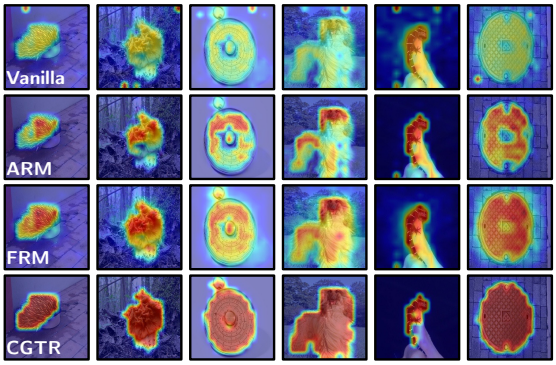


Fig. 5: Abalation study of CGTR w.r.t. localization maps.

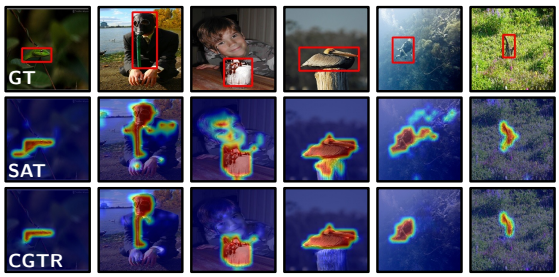
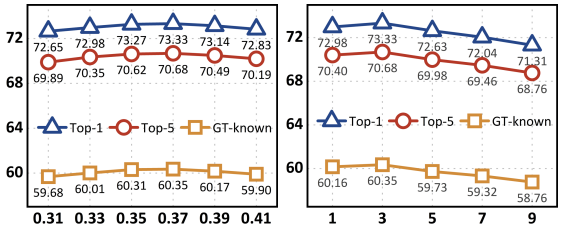


Fig. 6: Localization maps for co-contexts between the proposed CGTR and SAT [38]. Ground-truth bounding boxes are in red.

4.3 Ablation Study

Efficacy of Key Components. Ablation experiments to assess the key components of CGTR are summarized in Tab. 4. We only retain the multi-head attention mechanism [33] and the spatial-query attention mechanism [38] as the vanilla. It is observed that ARM yields an improvement of 1.81% in GT-known Loc compared to the vanilla. Furthermore, applying FRM results in a Top-1 Loc of 58.22%, surpassing the vanilla by 1.36%. The best localization accuracy is attained when both ARM and FRM are utilized. Additionally, we visualize the localization maps under different settings in Fig. 5. Compared with the vanilla, ARM effectively mitigates contextual bias by exploiting the object-centric properties of the class token. Moreover, FRM refines semantic object representations through content-aware local token filtering, enhancing the delineation of object structures. Consequently, the combination of ARM and FRM enables CGTR to generate superior localization maps with clearly defined object regions.

Effectiveness in Tackling Diffuse Activation. We present visualizations of localization maps generated by the proposed CGTR and the recent SAT [38], as shown in Fig. 6. For the target objects highlighted by the red boxes (*Vine snake*, *Gasmask*, *Ice cream*, *Pelican*, *Scuba diver*, *Ptarmigan*), SAT exhibits limitations in addressing the issue of diffuse activation. It frequently misidentifies co-occurring contextual elements as part of the target, resulting in spurious activations of associated background regions such as *Leave*, *Person*, *Wood*, *Algae*, and



(a) Hyper-parameter ξ (b) Hyper-parameter k

Fig. 7: Sensitivity analysis w.r.t. $\boldsymbol{\xi}$ and $\boldsymbol{k}$ on ILSVRC.

Table 5: Sensitivity analysis w.r.t. $\boldsymbol{\lambda}$.

λ	ILSVRC Loc. Acc			CUB-200-2011 Loc. Acc		
	Top-1	**Top-5**	**GT-known**	**Top-1**	**Top-5**	**GT-known**
0.40	60.28	70.61	73.29	80.81	93.58	98.12
0.45	**60.35**	**70.68**	**73.33**	**81.80**	**94.40**	**98.68**
0.50	60.34	70.66	73.30	80.76	93.55	98.21
0.55	60.31	70.61	73.28	80.48	93.01	97.98

Grass. This phenomenon stems from an insufficient suppression of co-occurrence noise during feature representation. In contrast, our CGTR, benefiting from the interactions between location and class tokens, covers more object regions.

Hyper-parameter Sensitive Analysis. We first investigate the sensitivity of localization quality with ξ in $\mathcal{L}_{arm}$ (Eq. 5) and k in `AVG` (Eq. 6), as illustrated in Fig. 7. The hyper-parameter ξ controls the regularization strength on the discrepancy between the localization and class attention maps. By encouraging consistency between the two attention maps, ξ indirectly influences the spatial extent (*i.e.*, the average activation area) of the localization responses. As shown in Fig. 7 (a), optimal localization accuracy is achieved when ξ=0.37. The hyper-parameter k represents the kernel size of the pooling operation applied to the location and class attention maps. Fig. 7 (b) shows that larger kernel sizes negatively affect localization performance. We attribute this to the fact that an overly large local window may introduce redundant, uncertain attention from the location and class tokens. It is observed that CGTR achieves the best performance when k is set to 3. Next, we explore the sensitivity of localization quality with λ, which is the weighting factor of the total loss $\mathcal{L}$ (Eq. 13). As shown in Tab. 5, the model achieves optimal localization accuracy when λ is set to 0.45. Furthermore, the results indicate that CGTR behaves insensitive to λ, as the performance remains stable over a large interval.

5 Conclusion

In this paper, we present the **C**ross-token **G**uidance **TR**ansformer (CGTR), a novel framework designed to enhance collaboration between location and class tokens, thereby effectively addressing the diffuse activation problem in weakly supervised object localization. Specifically, we introduce an Attention Regulation Module (ARM) that refines spatial relationships among tokens, significantly mitigating the excessive contextual activations. In addition, we propose a Filter

Regulation Module (FRM) to harness the structural knowledge encoded in the class token, guiding the selective filtering process of the location token. Extensive experiments validate the effectiveness of the proposed CGTR, consistently demonstrating superior performance over previous methods.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 62506146), Jiangxi Provincial Natural Science Foundation (No. 20252BAC200196), Early-Career Young Scientists and Technologists Project of Jiangxi Province (No. 20252BEJ730022), Natural Science Foundation of Fujian Province of China (No. 2026J001252), and the Fund of Key Laboratory of Education Blockchain and Intelligent Technology (Guangxi Normal University), Ministry of Education (No. EBME2026B08).

References

1. Bai, H., Zhang, R., Wang, J., Wan, X.: Weakly supervised object localization via transformer with implicit spatial calibration. ECCV pp. 612–628 (2022)
2. Cao, X., Zheng, X., Shen, Y., Li, K., Chen, J., Lu, Y., Tian, Y.: Locloc: Low-level cues and local-area guides for weakly supervised object localization. In: ACM MM. pp. 5655–5664 (2023)
3. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: ICCV. pp. 357–366 (2021)
4. Chen, Z., Cao, L., Shen, Y., Lian, F., Wu, Y., Ji, R.: E2net: Excitative-expansile learning for weakly supervised object localization. In: ACM MM. pp. 573–581 (2021)
5. Chen, Z., Ding, J., Cao, L., Shen, Y., Zhang, S., Jiang, G., Ji, R.: Category-aware allocation transformer for weakly supervised object localization. In: ICCV. pp. 6643–6652 (2023)
6. Chen, Z., Shen, Y., Cao, L., Zhang, S., Ji, R.: Clip-driven transformer for weakly supervised object localization. TPAMI (2025)
7. Chen, Z., Wang, C., Wang, Y., Jiang, G., Shen, Y., Tai, Y., Wang, C., Zhang, W., Cao, L.: Lctr: On awakening the local continuity of transformer for weakly supervised object localization. In: AAAI. vol. 36, pp. 410–418 (2022)
8. Chen, Z., Wang, S., Cao, L., Shen, Y., Ji, R.: Adaptive zone learning for weakly supervised object localization. TNNLS (2024)
9. Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. In: CVPR. pp. 3133–3142 (2020)
10. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: CVPR. pp. 2219–2228 (2019)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2020)
12. Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., Ye, Q.: Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In: ICCV. pp. 2886–2895 (2021)

13. Guichemerre, A., Belharbi, S., Mayet, T., Murtaza, S., Shamsolmoali, P., McCaffrey, L., Granger, E.: Source-free domain adaptation of weakly-supervised object localization models for histology. In: CVPR. pp. 33–43 (2024)
14. Guo, G., Han, J., Wan, F., Zhang, D.: Strengthen learning tolerance for weakly supervised object localization. In: CVPR (2021)
15. Gupta, S., Lakhotia, S., Rawat, A., Tallamraju, R.: Vitol: Vision transformer for weakly supervised object localization. In: CVPR. pp. 4101–4110 (2022)
16. He, K., Sun, J., Tang, X.: Guided image filtering. TPAMI **35**(6), 1397–1409 (2012)
17. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. In: ICCV. pp. 15013–15022 (2021)
18. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. CSUR **54**(10s), 1–41 (2022)
19. Kim, E., Kim, S., Lee, J., Kim, H., Yoon, S.: Bridging the gap between classification and localization for weakly supervised object localization. In: CVPR. pp. 14258–14267 (2022)
20. Kumar Singh, K., Jae Lee, Y.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: ICCV. pp. 3524–3533 (2017)
21. Lei Ba, J., Kiros, J.R., Hinton, G.E.: Layer normalization. ArXiv e-prints pp. arXiv–1607 (2016)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. ICLR (2017)
23. Mai, J., Yang, M., Luo, W.: Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In: CVPR. pp. 8766–8775 (2020)
24. Meng, M., Zhang, T., Tian, Q., Zhang, Y., Wu, F.: Foreground activation maps for weakly supervised object localization. In: ICCV. pp. 3385–3395 (2021)
25. Meng, M., Zhang, T., Zhang, Z., Zhang, Y., Wu, F.: Task-aware weakly supervised object localization with transformer. TPAMI **45**(7), 9109–9121 (2022)
26. Pan, X., Gao, Y., Lin, Z., Tang, F., Dong, W., Yuan, H., Huang, F., Xu, C.: Unveiling the potential of structure preserving for weakly supervised object localization. In: CVPR. pp. 11642–11651 (2021)
27. Pan, Y., Yao, Y., Cao, Y., Chen, C., Lu, X.: Coarse2fine: local consistency aware re-prediction for weakly supervised object localization. In: AAAI. vol. 37, pp. 2002–2010 (2023)
28. Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q.: Conformer: Local features coupling global representations for visual recognition. In: ICCV. pp. 367–376 (2021)
29. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. vol. 115, pp. 211–252. Springer (2015)
30. Su, H., Yang, M.: A consistency and integration model with adaptive thresholds for weakly supervised object localization. In: IJCAI. pp. 1281–1289 (2024)
31. Tan, C., Gu, G., Ruan, T., Wei, S., Zhao, Y.: Dual-gradients localization framework with skip-layer connections for weakly supervised object localization. TMM **25**, 4933–4942 (2022)
32. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML. pp. 10347–10357. PMLR (2021)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. vol. 30 (2017)
34. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (CNS-TR-2011-001) (2011)

35. Wang, C., Xu, R., Xu, S., Meng, W., Wang, R., Zhang, X.: Exploring intrinsic discrimination and consistency for weakly supervised object localization. TIP (2024)
36. Wei, J., Wang, S., Zhou, S.K., Cui, S., Li, Z.: Weakly supervised object localization through inter-class feature similarity and intra-class appearance consistency. In: ECCV. pp. 195–210 (2022)
37. Wu, P., Zhai, W., Cao, Y.: Background activation suppression for weakly supervised object localization. In: CVPR. pp. 14228–14237 (2022)
38. Wu, P., Zhai, W., Cao, Y., Luo, J., Zha, Z.J.: Spatial-aware token for weakly supervised object localization. In: ICCV. pp. 1844–1854 (2023)
39. Wu, Z., Xu, Y., Yang, J., Li, X.: Misclassification in weakly supervised object detection. TIP (2024)
40. Xie, J., Luo, C., Zhu, X., Jin, Z., Lu, W., Shen, L.: Online refinement of low-level feature based activation map for weakly supervised object localization. In: CVPR. pp. 132–141 (2021)
41. Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., Shen, L.: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In: CVPR. pp. 989–998 (2022)
42. Xu, J., Hou, J., Zhang, Y., Feng, R., Zhao, R.W., Zhang, T., Lu, X., Gao, S.: Cream: Weakly supervised object localization via class re-activation mapping. In: CVPR. pp. 9437–9446 (2022)
43. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Multi-class token transformer for weakly supervised semantic segmentation. In: CVPR. pp. 4310–4319 (2022)
44. Xu, R., Luo, Y., Hu, H., Du, B., Shen, J., Wen, Y.: Rethinking the localization in weakly supervised object localization. In: ACM MM. pp. 5484–5494 (2023)
45. Xu, W., Wang, C., Xu, R., Xu, S., Meng, W., Zhang, M., Zhang, X.: Token masking transformer for weakly supervised object localization. TMM (2024)
46. Xue, H., Liu, C., Wan, F., Jiao, J., Ji, X., Ye, Q.: Danet: Divergent activation for weakly supervised object localization. In: ICCV. pp. 6589–6598 (2019)
47. Yasuki, S., Taki, M.: Cam back again: Large kernel cnns from a weakly supervised object localization perspective. In: CVPR. pp. 341–351 (2024)
48. Yoon, S.H., Kwon, H., Kim, H., Yoon, K.J.: Class tokens infusion for weakly supervised semantic segmentation. In: CVPR. pp. 3595–3605 (2024)
49. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV. pp. 6023–6032 (2019)
50. Zhai, W., Wu, P., Zhu, K., Cao, Y., Wu, F., Zha, Z.J.: Background activation suppression for weakly supervised object localization and semantic segmentation. IJCV pp. 1–26 (2023)
51. Zhang, C.L., Cao, Y.H., Wu, J.: Rethinking the route towards weakly supervised object localization. In: CVPR. pp. 13460–13469 (2020)
52. Zhang, D., Han, J., Cheng, G., Yang, M.H.: Weakly supervised object localization and detection: a survey. TPAMI **44**(9), 5866–5885 (2021)
53. Zhang, X., Wei, Y., Yang, Y.: Inter-image communication for weakly supervised localization. In: ECCV. pp. 271–287 (2020)
54. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: ICCV. pp. 16259–16268 (2021)
55. Zhao, Y., Ye, Q., Wu, W., Shen, C., Wan, F.: Generative prompt model for weakly supervised object localization. In: ICCV. pp. 6351–6361 (2023)
56. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)

57. Zhu, L., She, Q., Chen, Q., You, Y., Wang, B., Lu, Y.: Weakly supervised object localization as domain adaption. In: CVPR. pp. 14637–14646 (2022)