Category-aware Allocation Transformer for Weakly Supervised Object Localization

Zhiwei Chen[1] Jinren Ding[1] Liujuan Cao[1*] Yunhang Shen[2]
Shengchuan Zhang[1] Guannan Jiang[3] Rongrong Ji[1]
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China. [2]Tencent Youtu Lab, China. [3]CATL, China.
zhiweichen.cn@gmail.com, dingjinren@stu.xmu.edu.cn, odysseyshen@tencent.com,
{caoliujuan, zsc_2016, rrji}@xmu.edu.cn, jianggn@catl.com

Abstract

Weakly supervised object localization (WSOL) aims to localize objects based on only image-level labels as supervision. Recently, transformers have been introduced into WSOL, yielding impressive results. The self-attention mechanism and multilayer perceptron structure in transformers preserve long-range feature dependency, facilitating complete localization of the full object extent. However, current transformer-based methods predict bounding boxes using category-agnostic attention maps, which may lead to confused and noisy object localization. To address this issue, we propose a novel ***C****ategory-aware* ***A****llocation* ***TR****ansformer (CATR) that learns category-aware representations for specific objects and produces corresponding category-aware attention maps for object localization. First, we introduce a Category-aware Stimulation Module (CSM) to induce learnable category biases for self-attention maps, providing auxiliary supervision to guide the learning of more effective transformer representations. Second, we design an Object Constraint Module (OCM) to refine the object regions for the category-aware attention maps in a self-supervised manner. Extensive experiments on the CUB-200-2011 and ILSVRC datasets demonstrate that the proposed CATR achieves significant and consistent performance improvements over competing approaches.*

1. Introduction

Weakly supervised learning utilizes minimal supervision or coarse annotations for training. In particular, weakly supervised object localization (WSOL) aims to locate objects using only image-level annotations, making it an attractive research area in various applications due to the elimination of the need for costly bounding box annotations [4, 35, 8, 14, 1].

To tackle the challenge of WSOL with only image-level labels, most approaches [36, 32, 19, 3, 21, 9] rely on Class Activation Maps (CAM) [38] to discover discriminative im-

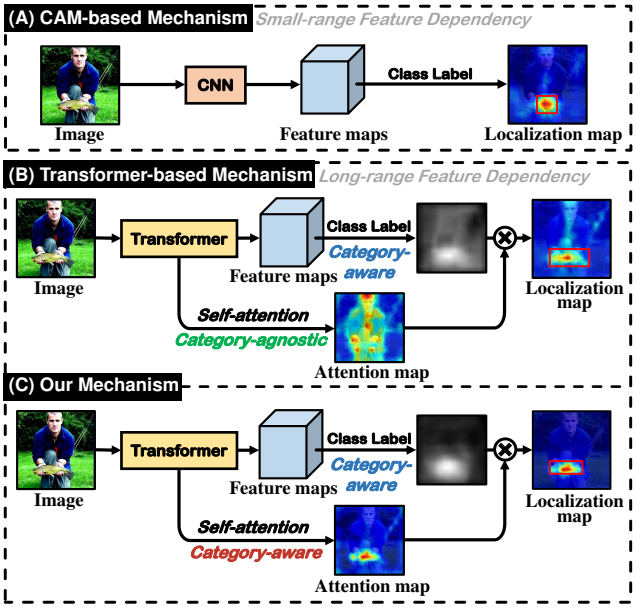

Figure 1. The comparison of different mechanisms for generating localization maps: (A) The CAM-based mechanism [19, 3] uses feature maps of the class label, which tends to capture the most discriminative regions in the localization map. (B) The transformer-based mechanism [8, 5] combines the category-agnostic attention map with the category-aware feature map, which brings *person* noise to the localization map. (C) Our mechanism investigates category awareness in self-attention maps to generate a category-aware attention map, which is then coupled with the category-aware feature map to produce an accurate localization map. The predicted bounding boxes are in red. Best viewed in color.

*Corresponding author.

age regions for localizing potential target objects. However, these methods tend to grossly underestimate object regions and produce much smaller bounding boxes than the actual range, as illustrated in Fig. 1 (A). To capture more object parts, different techniques have been proposed to enhance CAM, such as graph propagation [40], data augmentation [33, 15], adversarial erasing [3, 6, 19] and spatial relation activation [32, 37, 9]. Despite their promising success, existing approaches still exhibit limited performance to completely localize objects, owing to the inherent characteristic of convolutional neural networks (CNNs) [22, 5], *i.e*., failing to explore the global feature relations properly.

Recently, visual transformers [26] have made significant breakthroughs in computer vision, demonstrating that pure transformers can be as effective as CNNs for feature extraction. Gao *et al*. [8] first combined the semantic-aware tokens with the semantic-agnostic attention map to generate a localization map. Chen *et al*. [5] and Bai *et al*. [1] further explored the local-continuous visual patterns and semantic similarities of transformers for object localization, respectively. The current transformer-based methods couple the attention map upon the class tokens with the feature map of the specific category to generate a localization map, as illustrated in Fig. 1 (B). The feature map of the specific category is obtained according to the ground-truth image-level class label, while the attention map upon the class tokens that captures the long-range feature dependency is category-agnostic (not distinguishable to object classes) and is not competent to category-aware localization [8]. As a result, it brings category-agnostic noise to the localization maps, which affects the generation of bounding boxes.

Based on the above analysis, we propose a simple but effective framework called **C**ategory-aware **A**llocation **TR**ansformer (CATR), which employs category information to exploit category-aware transformer attention. As illustrated in Fig. 1 (C), our approach focuses on generating category-aware attention maps, which can effectively learn the discriminative representation of specific object classes. Two category-aware maps are combined to generate the localization maps, which significantly reduces the impact of background clutter. Specifically, we introduce a Category-aware Stimulation Module (CSM) into the transformer attention mechanism, which induces a learning bias to associate self-attention maps with a specific category. CSM can be viewed as auxiliary supervision to guide the learning of more effective transformer representations and establishes a strong one-to-one connection between the self-attention maps and the corresponding classes. Moreover, we design an Object Constraint Module (OCM) to refine object regions for category-aware attention maps in a self-supervised manner. OCM restricts background clutter and generates pixel-level pseudo labels guided by the self-attention maps, which helps to activate precise object regions. Furthermore, we apply an automatic weighted loss mechanism [16] to adjust the loss weights during training. To validate the effectiveness of the proposed CATR, we conduct comprehensive experiments on challenging WSOL benchmarks. The contributions of this work are as follows:

- We propose the Category-aware Allocation Transformer (CATR) for weakly supervised object localization, which significantly enhances the category awareness of self-attention maps among long-range feature dependency.
- We introduce the Category-aware Stimulation Module (CSM) to learn a specific category for the self-attention maps among the different transformer blocks.
- We propose an Object Constraint Module (OCM) to refine the object regions for category-aware attention maps in a self-supervised manner.
- CATR achieves new state-of-the-art performance on the CUB-200-2011 and ILSVRC datasets with 79.62% and 56.90% Top-1 localization accuracy, respectively.

2. Related Work

CNN-based Methods for WSOL. Most methods rely on the CAM [38] pipeline, which generates object bounding boxes using the class activation map from a classification network. However, due to the lack of localization supervision, it easily becomes trapped in the most discriminative parts rather than the whole object. To address this issue, some methods use augmentation strategies on images or features to highlight non-discriminative parts of objects. For example, Singh *et al*. [15] hid the image patches randomly in the training phase to discover different object parts. Yun *et al*. [33] proposed to cut and paste the patches among training images to attend to non-discriminative parts of objects. Additionally, Zhang *et al*. [36] and Mai *et al*. [19] mined different discriminative regions by two adversary classifiers. Junsk *et al*. [6] erased discriminative spatial positions on the feature map to capture the integral extent of the object. Chen *et al*. [3] combined erasing and maxout learning strategies to highlight foreground objects without losing information. Besides the augmentation strategies, there are some other methods that highlight the spatial relationships among object parts to obtain integral regions. Xue *et al*. [32] utilized a discrepant activation method to learn complementary and discriminative visual patterns. Zhang *et al*. [37] adopted constraints to prompt the consistency of object features within the same categories. Pan *et al*. [21] leveraged structural information incorporated into convolutional features to distill the structure-preserving ability of features. Apart from the above works, Xie *et al*. [30] proposed a new paradigm that learns a foreground

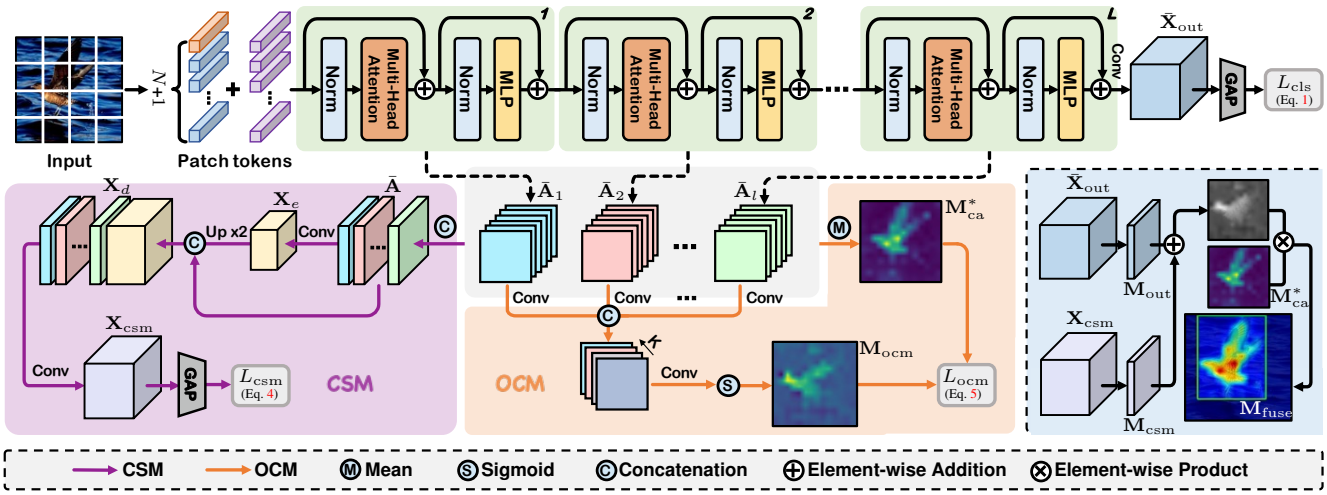

Figure 2. The architecture of the proposed CATR. It consists of a vision transformer backbone, a category-aware stimulation module (CSM), and an object constraint module (OCM). In the inference phase, we add two specific-category supervised maps together, namely $\mathbf{M}_{\text{out}}$ and $\mathbf{M}_{\text{csm}}$, and multiply them by the category-aware map $\mathbf{M}^*_{\text{ca}}$ to generate a localization map $\mathbf{M}_{\text{fuse}}$, as shown in the bottom right.

prediction map to achieve localization. Wu *et al*. [29] proposed a background activation suppression strategy to learn foreground prediction maps for object localization. Zhu *et al*. [39] modeled WSOL as a domain adaption task, where the score estimator trained on the source/image domain is tested on the target/pixel domain to locate objects. Some other methods, such as Zhang *et al*. [34], Guo *et al*. [9] and Wei *et al*. [28], divided WSOL into two independent subtasks, including classification and the class-agnostic localization. Such methods are not end-to-end and have separated training phases, which may be inefficient for WSOL.

Most of the aforementioned methods rely on convolutional neural networks, which have the inherent limitation of failing to capture global information and are thus susceptible to focusing on local discriminative object regions [8, 22]. To address this issue, we propose a transformer-based method for WSOL.

Transformer-based Methods for WSOL. Recent advancements in computer vision have shown the great potential of transformers for various tasks [13, 7, 31, 11, 2, 12, 24], as they excel in capturing long-range dependencies and perform better than convolutional neural networks.

Dosovitskiy *et al*. [7] demonstrated that a pure transformer performs exceptionally well on image classification tasks when applied directly to sequences of image patches. In the context of weakly supervised object localization, transformer-based models have also shown promising results. For instance, Gao *et al*. [8] combined semantic-aware tokens with the semantic-agnostic attention map, which could use both semantic and positioning information from a visual transformer to find objects. Chen *et al*. [5] highlighted local details of global representations using learnable kernels and cross-patch information guided by the class-token attention map. Gupta *et al*. [10] improved localization maps by incorporating a patch-based attention dropout layer into the transformer attention blocks. Bai *et al*. [1] considered the semantic similarities of patch tokens and their spatial relationships for WSOL.

However, the above methods directly use the semantic-agnostic (*i.e*., category-agnostic) attention maps, which tend to bring background noise into object localization. In this paper, we propose to inject category-aware information into these semantic-agnostic attention maps, which allows for pinpointing specific objects.

3. Methodology

In this section, we first provide an overview of the proposed Category-aware Allocation TRansformer (CATR), followed by detailed descriptions of the Category-aware Stimulation Module (CSM) and Object Constraint Module (OCM). We then incorporate the modules with the transformer structure into a joint optimization framework, as illustrated in Fig. 2.

3.1. Framework Overview

Consider an input image $I \in \mathbb{R}^{H\times W\times M}$, where H, W, M denote its height, width, and the number of channels, respectively. We first split I into $w \times h$ patches, which are then flattened and linearly projected into a sequence of patch tokens $\mathbf{T}_{\text{p}} \in \mathbb{R}^{N\times D}$, where D is the dimension of each patch and $N = w \times h$. An extra learnable class token $\mathbf{T}_{\text{cls}} \in \mathbb{R}^{1\times D}$ is then prepended to the tokens, together with a position embedding $\mathbf{T}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$, to form the input sequence $\mathbf{X} \in \mathbb{R}^{(N+1)\times D}$, which is fed into the transformer encoder with L consecutive transformer blocks.

To allocate the semantic information, the proposed CSM

is applied to all the self-attention maps, resulting in a feature map $\mathbf{X}_{\text{csm}} \in \mathbb{R}^{C\times w\times h}$, where C is the number of classes. An auxiliary classification loss $\mathcal{L}_{\text{csm}}$ is further added following $\mathbf{X}_{\text{csm}}$ to promote category awareness. Subsequently, we apply OCM to all the self-attention maps to refine the object regions and obtain the object constraint loss $\mathcal{L}_{\text{ocm}}$.

Denote $\mathbf{X}_{\text{out}} \in \mathbb{R}^{(N+1)\times D}$ as the output feature map, we discard the class token and apply a convolutional layer to it, as in [8], resulting in the feature map $\bar{\mathbf{X}}_{\text{out}} \in \mathbb{R}^{C\times w\times h}$. Finally, we generate the class probability distribution $\hat{y}$ for classification prediction by applying a global average pooling layer. With the corresponding image-level one-hot encoding label y, the classification loss function is as follows:

$$\mathcal{L}_{\text{cls}} = -\sum_{i}^{C} y_i \log\left(\frac{e^{\hat{y}_i}}{\sum_j^C e^{\hat{y}_j}}\right). \tag{1}$$

During training, we apply an automatic weighted mechanism [16] to dynamically balance the importance of different modules. The overall training loss can be defined as:

$$\begin{aligned}\mathcal{L} = &\frac{1}{2\lambda_1^2}\mathcal{L}_{\text{cls}} + \log(1+\lambda_1^2) + \frac{1}{2\lambda_2^2}\mathcal{L}_{\text{csm}} + \log(1+\lambda_2^2) \\ &+ \frac{1}{2\lambda_3^2}\mathcal{L}_{\text{ocm}}^{\text{s}} + \log(1+\lambda_3^2) + \frac{1}{2\lambda_4^2}\mathcal{L}_{\text{ocm}}^{\text{a}} + \log(1+\lambda_4^2),\end{aligned} \tag{2}$$

where λ_1-λ_4 are learnable parameters initialized to 1.

During testing, we first obtain an object map $\mathbf{M}_{\text{out}} \in \mathbb{R}^{w\times h}$ from $\bar{\mathbf{X}}_{\text{out}}$ based on the predicted class. Similarly, we obtain another object map $\mathbf{M}_{\text{csm}} \in \mathbb{R}^{w\times h}$ from $\mathbf{X}_{\text{csm}}$. The final localization map is then computed by:

$$\mathbf{M}_{\text{fuse}} = \mathbf{M}_{\text{ca}}^{*} \otimes (\mathbf{M}_{\text{out}} + \mathbf{M}_{\text{csm}}), \tag{3}$$

where $\otimes$ denotes element-wise multiplication. $\mathbf{M}_{\text{ca}}^{*}$ is a category-aware attention map generated from self-attention maps, as described in Sec. 3.3. $\mathbf{M}_{\text{fuse}}$ is then resized to the same size as the original image using linear interpolation. Finally, the predicted box is obtained by finding the tight bounding box covering the largest connected area in the foreground pixels, as done in previous works [38, 8].

3.2. Category-aware Stimulation Module

To mitigate the impact of category-agnostic attention maps in the localization map generation stage, we propose to inject category-aware information into the transformer. Assuming that the attention matrix of the multi-head attention module in a transformer block as $\mathbf{A}_{\text{am}} \in \mathbb{R}^{L\times S\times(N+1)\times(N+1)}$. Note that L represents the number of transformer blocks and S denotes the number of heads in the multi-head attention mechanism. We extract and reshape the class-token attention vector from $\mathbf{A}_{\text{am}}$ to obtain

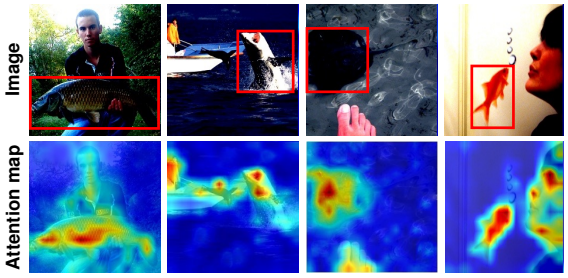

Figure 3. Visualization of attention maps in the transformer. The ground-truth bounding boxes are in red.

an attention map $\bar{\mathbf{A}} \in \mathbb{R}^{L\times S\times w\times h}$, which aggregates the feature representation of all transformer blocks and heads. However, $\bar{\mathbf{A}}$ acquires information from different representation subspaces at various positions in the input image, encompassing all regions of interest in the image. This makes it difficult to locate a specific object, as shown in Fig. 3. For instance, the first-column sample reveals that both the fish and people are activated, despite the ground-truth label being solely assigned to the fish. This is because $\bar{\mathbf{A}}$ lacks category supervision during training. To cope with this issue, we propose CSM to optimize $\bar{\mathbf{A}}$.

Specifically, we reshape $\bar{\mathbf{A}}$ into $\mathbf{X}_A \in \mathbb{R}^{(LS)\times w\times h}$ and apply a CNN-based encoder-decoder to extract more precise semantic information for distinguishing different interest regions. The encoder part comprises a max-pooling operation with stride 2 and two 3×3 convolutional layers, which generate a downsampled feature map $\mathbf{X}_e \in \mathbb{R}^{G\times\frac{w}{2}\times\frac{h}{2}}$. Note that we set the number of feature channels to G. The decoder then upsamples $\mathbf{X}_e$ and concatenates it with the input feature map $\mathbf{X}_A$. The concatenated features pass through two 3×3 convolutional layers to obtain the output feature map $\mathbf{X}_d$. Next, we add an auxiliary classification head following $\mathbf{X}_d$ to promote category awareness. The classification head comprises a 1×1 convolutional layer and a global average pooling layer, and produces the output feature map $\mathbf{X}_{\text{csm}} \in \mathbb{R}^{C\times w\times h}$ and the class probability distribution $\hat{y}^c$. The loss of the CSM is formulated as:

$$\mathcal{L}_{\text{csm}} = -\sum_{i}^{C} y_i \log\left(\frac{e^{\hat{y}_i^c}}{\sum_j^C e^{\hat{y}_j^c}}\right). \tag{4}$$

Here, y_i denotes the ground-truth label of class i, and C denotes the number of classes.

3.3. Object Constraint Module

CSM incorporates category information into the category-agnostic attention maps, thereby reducing the background interference of the localization maps. Then, for generating precise attention maps on the basis of CSM, we propose the Object Constraint Module (OCM) to refine the object regions, which leverages the discriminative power of

attention maps and generates pseudo labels to supervise the attention maps in a self-supervised manner.

OCM consists of noise suppression and object awakening mechanisms that restrict the background clutter of the attention maps and enhance object regions, respectively. The joint objective function for OCM training is optimized with two loss terms: the noise suppression loss $\mathcal{L}^{\mathrm{s}}_{\mathrm{ocm}}$ and the object awakening loss $\mathcal{L}^{\mathrm{a}}_{\mathrm{ocm}}$. The overall loss for OCM is formulated as:

$$\begin{aligned}\mathcal{L}_{\mathrm{ocm}} = &\frac{1}{2\lambda_3^2}\mathcal{L}^{\mathrm{s}}_{\mathrm{ocm}} + \log(1+\lambda_3^2) \\ &+ \frac{1}{2\lambda_4^2}\mathcal{L}^{\mathrm{a}}_{\mathrm{ocm}} + \log(1+\lambda_4^2).\end{aligned} \tag{5}$$

Noise suppression mechanism. We employ the average operator to the attention map $\bar{\mathbf{A}} \in \mathbb{R}^{L\times S\times w\times h}$ along both the block and head dimensions. We acquire the category-aware attention map $\mathbf{M}^*_{\mathrm{ca}} \in \mathbb{R}^{w\times h}$, which is formulated as:

$$\mathbf{M}^*_{\mathrm{ca}} = \frac{1}{LS}\sum_{l}^{L}\sum_{s}^{S}\bar{\mathbf{A}}_{(l,s)}. \tag{6}$$

Since long-range dependency is preserved in $\mathbf{M}^*_{\mathrm{ca}}$, we treat values below the p-percentile as background clutter, and suppress them to zero. The loss function for noise suppression is defined as:

$$\mathcal{L}^{\mathrm{s}}_{\mathrm{ocm}} = \frac{1}{hw}\sum_{i}^{h}\sum_{j}^{w}\bar{\mathbf{M}}^*_{ca(i,j)}, \tag{7}$$

$$\bar{\mathbf{M}}^*_{\mathrm{ca}(i,j)} = \begin{cases} 0, & \text{if } \mathbf{M}^*_{\mathrm{ca}(i,j)} > \mathbf{\Phi}_p\left(\mathbf{M}^*_{\mathrm{ca}}\right), \\ \mathbf{M}^*_{\mathrm{ca}(i,j)}, & \text{otherwise,} \end{cases} \tag{8}$$

where $\mathbf{\Phi}_p(\cdot)$ denotes a function that finds p-th percentile from the given values. Specifically, $\mathbf{M}^*_{\mathrm{ca}}$ is flattened and ranked in ascending order, and the p-th percentile value is selected as the threshold for noise suppression.

Object awakening mechanism. To learn the favorable object regions, a key step is to derive reliable attention pseudo labels as supervision. We first split the attention map $\bar{\mathbf{A}}$ along the block dimension to obtain $\bar{\mathbf{A}}_l \in \mathbb{R}^{S\times w\times h}$ in the l-th block. Then, we apply a 1×1 convolutional layer to derive a new feature map $\mathbf{X}_l \in \mathbb{R}^{1\times w\times h}$, which approximates the spatial distribution of the object in each block. Given that different blocks learn diverse representations, we concatenate the new feature maps from the K transformer blocks and pass them through a 3×3 convolutional layer with a sigmoid activation function, resulting in a pixel-level pseudo map $\mathbf{M}_{\mathrm{ocm}} \in \mathbb{R}^{w\times h}$. This pseudo map further supervises the category-aware attention map during training. The activation loss is formulated as:

$$\mathcal{L}^{\mathrm{a}}_{\mathrm{ocm}} = \frac{1}{hw}\sum_{i}^{h}\sum_{j}^{w}\left(\mathbf{M}_{\mathrm{ocm}(i,j)} - \mathbf{M}^*_{\mathrm{ca}(i,j)}\right)^2. \tag{9}$$

Methods	Backbone	Loc. Acc		
		Top-1	Top-5	GT-known
CAM [38]	GoogLeNet	41.06	50.66	55.10
DANet [32]	GoogLeNet	49.45	60.46	67.03
ADL [6]	InceptionV3	53.04	–	–
SPA [21]	InceptionV3	53.59	66.50	72.14
FAM [20]	InceptionV3	70.67	–	87.25
CAM [38]	VGG16	44.15	52.16	56.00
ADL [6]	VGG16	52.36	–	75.41
ACoL [36]	VGG16	45.92	56.51	62.96
DANet [32]	VGG16	52.52	61.96	67.70
MEIL [19]	VGG16	57.46	–	73.84
SPA [21]	VGG16	60.27	72.50	77.29
FAM [20]	VGG16	69.26	–	89.26
ORNet [30]	VGG16	67.73	80.77	86.20
BAS [29]	VGG16	71.33	85.33	91.00
BGC [14]	VGG16	70.83	88.07	93.17
TS-CAM [8]	Deit-S	71.30	83.80	87.70
LCTR [5]	Deit-S	79.20	89.90	92.40
SCM [1]	Deit-S	76.40	91.60	**96.60**
CATR (Ours)	Deit-S	**79.62**	**92.08**	94.94

Table 1. Comparison of CATR with the state-of-the-art methods on the CUB-200-2011 [27] test set.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate the effectiveness of our proposed method on two widely-used WSOL benchmarks: CUB-200-2011 [27] and ILSVRC [23]. The CUB-200-2011 dataset comprises 200 bird categories with $5,994$ training images and $5,794$ testing images. In contrast, the ILSVRC dataset is a more extensive dataset that comprises $1,000$ classes with $1,281,197$ training images and $50,000$ validation images. We train our model using only the training set and evaluate it on the validation set, where the bounding box annotations are solely used for evaluation purposes.

Evaluation Metrics. Following previous methods [38, 5, 29], five evaluation metrics are adopted for evaluation, including Top-1/Top-5 localization accuracy (Top-1/Top-5 Loc), GT-known localization accuracy (GT-known Loc) and Top-1/Top-5 classification accuracy (Top1/Top-5 Cls). Concretely, Top-1 Loc is the fraction of images with the correct predictions of classification and more than 50% intersection over union (IoU) with the ground-truth bounding boxes. Top-5 Loc is the fraction of images with class labels belonging to Top-5 predictions and more than 50% IoU with the ground-truth bounding boxes. GT-known Loc is the fraction of images for which the predicted boxes have more than 50% IoU with the ground-truth bounding boxes.

Implementation Details. We construct our proposed CATR using the Deit-S backbone [25] pre-trained on ILSVRC [23] and adopt TS-CAM [8] as our baseline method. Specifically, we replace the MLP head with a con-

Methods	Backbone	Loc. Acc		
		Top-1	Top-5	GT-known
CAM [38]	VGG16	42.80	54.86	59.00
ADL [6]	VGG16	44.92	–	–
ACoL [36]	VGG16	45.83	59.43	62.96
I^{2}C [37]	VGG16	47.41	58.51	63.90
MEIL [19]	VGG16	46.81	–	–
FAM [20]	VGG16	51.96	–	**71.73**
ORNet [30]	VGG16	52.05	63.94	68.27
BAS [29]	VGG16	52.96	65.41	69.64
BGC [14]	VGG16	49.94	63.25	68.92
CAM [38]	InceptionV3	46.29	58.19	62.68
ADL [6]	InceptionV3	48.71	–	–
DANet [32]	GoogLeNet	47.53	58.28	–
I^{2}C [37]	InceptionV3	53.11	64.13	68.50
GC-Net [18]	InceptionV3	49.06	58.09	–
SPA [21]	InceptionV3	52.73	64.27	68.33
FAM [20]	InceptionV3	55.24	–	68.62
TS-CAM [8]	Deit-S	53.40	64.30	67.60
LCTR [5]	Deit-S	56.10	65.80	68.70
SCM [1]	Deit-S	56.10	66.40	68.80
CATR (Ours)	Deit-S	**56.90**	**66.64**	69.25

Table 2. Comparison of CATR with state-of-the-art methods on the ILSVRC [23] validation set.

volutional layer and add a global average pooling layer on top of it. The input images are resized to 256×256 and then randomly cropped to 224×224. We use an AdamW optimizer [17] with ϵ=1e-8, β_1=0.9, β_2=0.99 and weight decay of 5e-4, to train our network. For the experiments on CUB-200-2011, we train the network for 80 epochs with a batch size of 128 and a learning rate of 5e-5. For the experiments on ILSVRC, we train the network for 14 epochs with a batch size of 128 and a learning rate of 5e-4.

4.2. Comparison with State-Of-The-Arts

Quantitative Comparison. Tab. 1 presents a comparison between CATR and other methods on the CUB-200-2011 dataset [27]. The experimental results indicate that CATR outperforms the baseline TS-CAM [8] in terms of Top-1/Top-5/GT-known Loc metrics, achieving Top-1 Loc accuracy of 79.62% and GT-known Loc accuracy of 94.94%. Furthermore, compared with the CNN-based state-of-the-art methods (BGC [14] and BAS [29]), CATR respectively achieves improvements of 8.79% and 8.29% in terms of Top-1 Loc. Additionally, CATR exhibits 2.54% improvement in GT-known Loc compared to the transformer-based state-of-the-art method LCTR [5], which further highlights the potential of transformers for WSOL.

Tab. 2 reports the localization accuracy on the ILSVRC dataset [23]. The proposed CATR surpasses the baseline TS-CAM [8] by 3.50% and 1.65% in terms of Top-1 Loc and GT-knwon Loc, respectively. Remarkably, CATR achieves a Top-1 Loc accuracy of 56.90%, outperforming all transformer-based methods. Compared to the CNN-based methods, the proposed CATR achieves state-of-the-art performance, with only a slightly lower GT-Known Loc than FAM [20].

	Baseline	CSM	OCM	Loc. Acc		
				Top-1	Top-5	GT-known
(a)	√			74.16	86.42	89.28
(b)	√	√		77.59	90.35	93.22
(c)	√	√	√	**79.62**	**92.08**	**94.94**

Table 3. Comparison of the object localization performance of CATR using different modules.

G	N/A	2	4	8	16	32
Top-1	74.16	72.20	72.80	**77.59**	75.71	74.18
Top-5	86.42	85.15	85.43	**90.35**	88.39	88.16
GT-k.	89.28	87.81	88.29	**93.22**	90.84	90.10

Table 4. Performance analyses of hyperparameter G in CSM. Note that 'N/A' indicates the baseline.

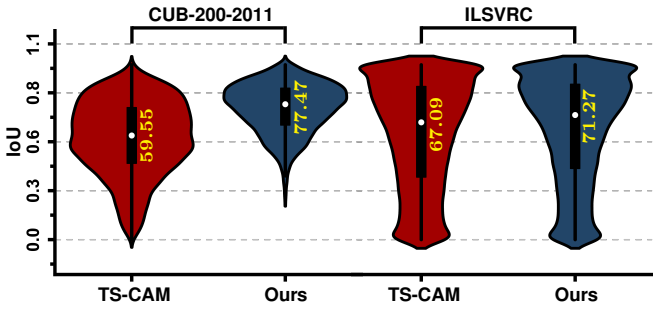

Figure 4. Statistical analysis of the IoU between predicted bounding boxes and ground-truth bounding boxes. The median (%) is displayed in yellow text.

Visual Comparison. We compare the localization results of the proposed CATR and TS-CAM [8] on CUB-200-2011 [27] and ILSVRC [23] in Fig. 5. Our proposed CATR consistently generates more accurate and refined localization maps that encompass category-aware object regions, exhibiting sharper and more compact boundaries compared to TS-CAM. For example, as seen from the second-column sample in the upper part of Fig. 5, TS-CAM only highlights the tail part of the blocked *bird*, while our method accurately captures the entire body except for the occlusions. Moreover, as evidenced by the sixth-column sample in the lower part of Fig. 5, TS-CAM focuses exclusively on the top of the *pole*, while our method successfully localizes the entire object regions.

Localization Quality. We present a statistical analysis of the IoU between predicted bounding boxes and ground-truth bounding boxes, as depicted in Fig. 4. On the CUB-200-2011 dataset, we achieve 77.47% median IoU, exceeding the baseline TS-CAM [8] by 17.92%, and correspondingly by 4.18% on ILSVRC. From these results, we can find that the proposed CATR improves localization quality on both datasets.

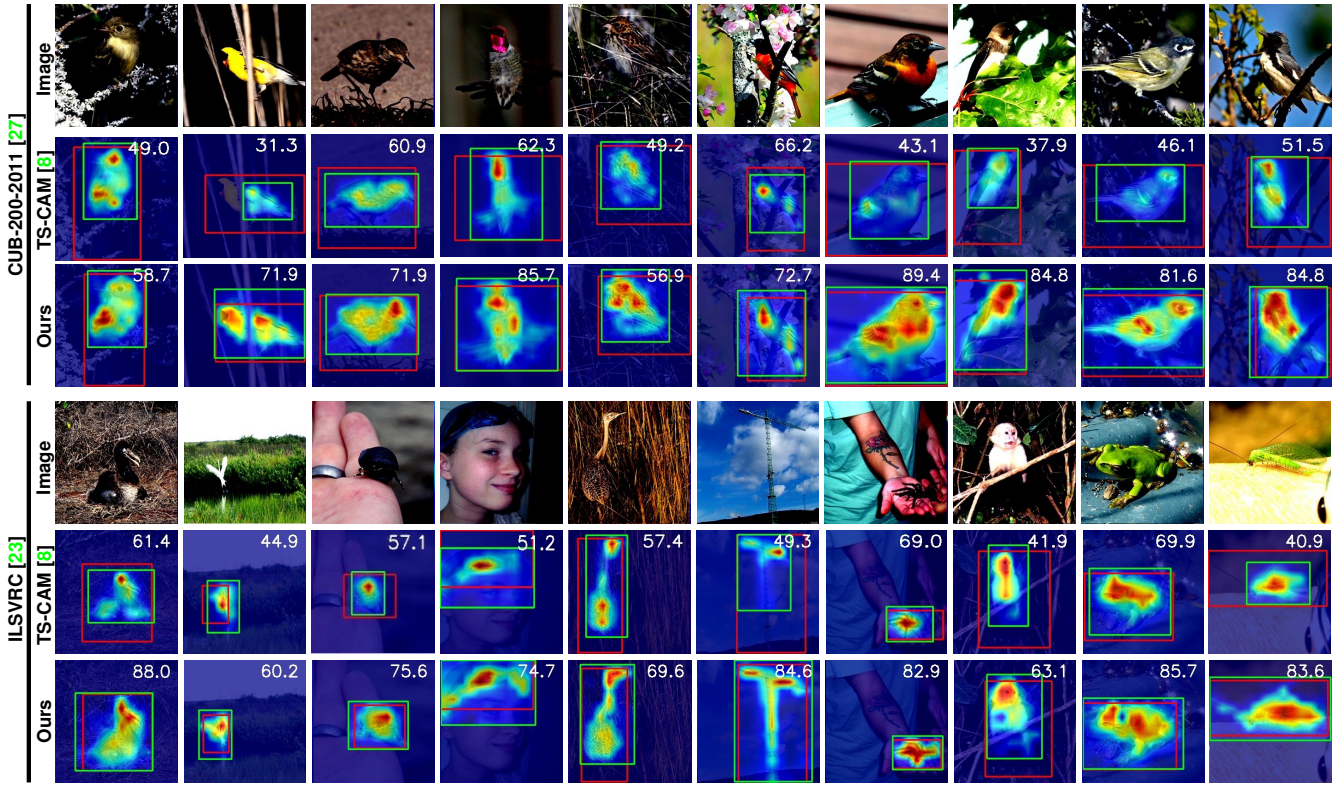

Figure 5. Visualization comparison with the baseline TS-CAM [8] method on CUB-200-2011 [27] and ILSVRC [23]. The ground-truth bounding boxes are highlighted in red, the predicted bounding boxes are highlighted in green, and the corresponding IoU values (%) are displayed in white text.

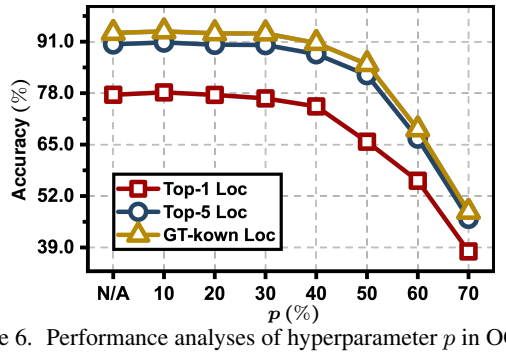

Figure 6. Performance analyses of hyperparameter p in OCM.

K	N/A	2	3	4	5	6
Top-1	78.20	76.80	77.36	78.12	78.43	78.56
Top-5	90.82	88.68	89.21	90.76	90.25	90.99
GT-k.	93.61	91.46	91.75	93.54	93.13	93.87
K	7	8	9	10	11	12
Top-1	78.55	78.87	**79.62**	76.13	74.97	65.55
Top-5	91.27	91.30	**92.08**	88.47	87.26	76.82
GT-k.	93.78	94.17	**94.94**	91.06	89.88	79.18

Table 5. Performance analyses of hyperparameter K in OCM. Note that 'N/A' indicates the baseline.

Dataset	ALM [16]	Loc. Acc		
		Top-1	Top-5	GT-known
CUB-200-2011 [27]	×	79.41	91.80	94.70
	√	**79.62**	**92.08**	**94.94**
ILSVRC [23]	×	56.13	65.99	68.60
	√	**56.90**	**66.64**	**69.25**

Table 6. Performance analyses of ALM [16].

4.3. Ablation Study

In this subsection, we perform a series of experiments to validate the effectiveness of various combinations of the proposed components on the two datasets.

First, we compare the performance of the proposed approach using different modules. Note that we re-implement TS-CAM [8] and take it as our baseline. Tab. 3 shows that the CSM, which builds a connection between the attention map and the category information, obtains gains of $3.43\%, 3.94\%$ in terms of Top-1 Loc and GT-known Loc, respectively. The model achieves a remarkable Top-1 Loc accuracy of 79.62% when we employ both CSM and OCM.

Next, we perform the sensitivity analysis on all hyperparameters of CATR through extensive experiments.

Hyperparameter G in CSM. We first investigate the effect of G in terms of Top-1/Top-5/GT-known Loc met-

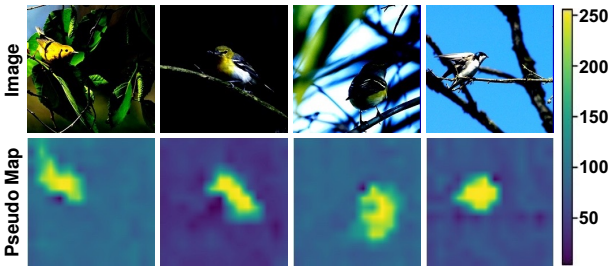

Figure 7. Visualization of the pixel-level pseudo map $\mathbf{M}_{\text{ocm}}$.

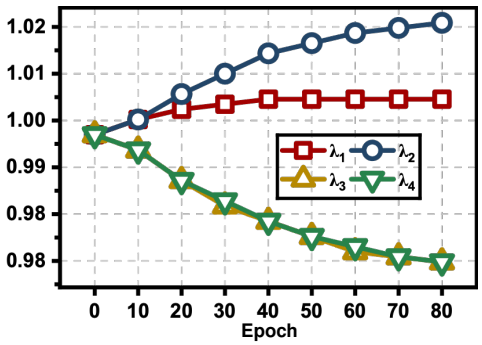

Figure 8. Analyses of loss weights in the training phase.

rics. Here, G represents the number of channels in the encoder part of CSM. As shown in Tab. 4, we vary G values to study the localization performance changes. From these results, we observe that the best localization accuracy can be achieved when $G = 8$. However, setting a larger G may lead to performance degradation, which we believe is caused by parameter redundancy.

Hyperparameter p in OCM. We then explore the effect of the hyperparameter p. Fig. 6 summarizes the experimental results on CUB-200-2011. We observe that the localization accuracy reaches the highest performance at $p = 10\%$ and then degrades. Note that p determines how much background noise needs to be suppressed. As p gets larger, the model suppresses more regions, and even suppresses the foreground. The results in Fig. 6 also support it, the performance drops sharply when $p = 70\%$.

Hyperparameter K in OCM. We further show the effect of the hyperparameter K, which indicates the number of attention maps aggregated from top-K transformer blocks. The experimental results are summarized in Tab. 5. The best localization accuracy can be achieved when K is set to 9. A larger K means more self-attention maps, which give more insights to generate the pixel-level pseudo map. However, the performance decreased when K exceeded 9, possibly due to overfitting.

Third, we visualize the pixel-level pseudo map (*i.e.*, $\mathbf{M}_{\text{ocm}}$) of OCM in Fig. 7. We observe that $\mathbf{M}_{\text{ocm}}$ contains the class-specific features, which highlight the robust object regions. Note that these pseudo-maps are generated based on the self-attention maps in the training phase without any pixel-level supervision. In this case, OCM effectively refines the object regions for the category-aware attention map in a self-supervised manner, resulting in the precise activation of object regions.

Lastly, we investigate the effects of the automatic weighted loss mechanism (ALM) [16] on our losses from two aspects. On the one hand, we present the performance when the ALM is not used, and λ_1-λ_4 are fixed to 1 for training. The results in Tab. 6 demonstrate that using the ALM can slightly improve localization performance. On the other hand, we study the changes in four learnable parameters (*i.e.*, λ_1-λ_4) in Eq. 2 during the training process. Please note that these learnable parameters are all initialized to 1. As shown in Fig. 8, we observe that λ_2 is consistently the largest during training, even though the losses weighted by λ_1 and λ_2 are the same classification loss (*i.e.*, the cross-entropy loss). We attribute this to the category-aware information brought by CSM, which is beneficial to classification. This finding supports the argument that CSM establishes the connection between the attention maps and the specific classes. Moreover, the results suggest that OCM plays a supporting role in refining the object regions, as the values of λ_3 and λ_4 decrease in the training phase.

5. Conclusion

In this paper, we introduce the **C**ategory-aware **A**llocation **TR**sanformer (CATR), a novel weakly supervised object localization method that leverages the self-attention mechanism of the transformer to deliver category information. We first propose a category-aware stimulation module (CSM) that learns the category information for the self-attention maps, providing auxiliary supervision to guide the learning of more effective transformer representations. Besides, we propose an object constraint module (OCM) to refine the object regions for category-aware attention maps in a self-supervised manner. Extensive experiments conducted on the CUB-200-2011 and ILSVRC benchmarks demonstrate the effectiveness of the proposed CATR, outperforming the state-of-the-art methods.

Acknowledgements

This work was supported by National Key R&D Program of China (No.2022ZD0118201) , the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

References

[1] Haotian Bai, Ruimao Zhang, Jiong Wang, and Xiang Wan. Weakly supervised object localization via transformer with implicit spatial calibration. *ECCV*, pages 612–628, 2022. 1, 2, 3, 5, 6

[2] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *IEEE ICCV*, pages 357–366, 2021. 3

[3] Zhiwei Chen, Liujuan Cao, Yunhang Shen, Feihong Lian, Yongjian Wu, and Rongrong Ji. E2net: Excitative-expansile learning for weakly supervised object localization. In *ACM MM*, pages 573–581, 2021. 1, 2

[4] Zhiwei Chen, Rongrong Ji, Jipeng Wu, and Yunhang Shen. Multi-scale features for weakly supervised lesion detection of cerebral hemorrhage with collaborative learning. In *ACM MM Asia*, pages 1–7. 2019. 1

[5] Zhiwei Chen, Changan Wang, Yabiao Wang, Guannan Jiang, Yunhang Shen, Ying Tai, Chengjie Wang, Wei Zhang, and Liujuan Cao. Lctr: On awakening the local continuity of transformer for weakly supervised object localization. In *AAAI*, volume 36, pages 410–418, 2022. 1, 2, 3, 5, 6

[6] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *IEEE CVPR*, pages 2219–2228, 2019. 2, 5, 6

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020. 3

[8] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *IEEE ICCV*, pages 2886–2895, 2021. 1, 2, 3, 4, 5, 6, 7

[9] Guangyu Guo, Junwei Han, Fang Wan, and Dingwen Zhang. Strengthen learning tolerance for weakly supervised object localization. In *IEEE CVPR*, 2021. 1, 2, 3

[10] Saurav Gupta, Sourav Lakhotia, Abhay Rawat, and Rahul Tallamraju. Vitol: Vision transformer for weakly supervised object localization. In *IEEE CVPR*, pages 4101–4110, 2022. 3

[11] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *IEEE ICCV*, pages 15013–15022, 2021. 3

[12] Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Yan Wang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. Istr: End-to-end instance segmentation with transformers. *arXiv preprint arXiv:2105.00637*, 2021. 3

[13] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s):1–41, 2022. 3

[14] Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, and Sungroh Yoon. Bridging the gap between classification and localization for weakly supervised object localization. In *IEEE CVPR*, pages 14258–14267, 2022. 1, 5, 6

[15] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *IEEE ICCV*, pages 3524–3533, 2017. 2

[16] Lukas Liebel and Marco Körner. Auxiliary tasks in multi-task learning. *arXiv preprint arXiv:1805.06334*, 2018. 2, 4, 7, 8

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2017. 6

[18] Weizeng Lu, Xi Jia, Weicheng Xie, Linlin Shen, Yicong Zhou, and Jinming Duan. Geometry constrained weakly supervised object localization. In *ECCV*, pages 481–496, 2020. 6

[19] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *IEEE CVPR*, pages 8766–8775, 2020. 1, 2, 5, 6

[20] Meng Meng, Tianzhu Zhang, Qi Tian, Yongdong Zhang, and Feng Wu. Foreground activation maps for weakly supervised object localization. In *IEEE ICCV*, pages 3385–3395, 2021. 5, 6

[21] Xingjia Pan, Yingguo Gao, Zhiwen Lin, Fan Tang, Weiming Dong, Haolei Yuan, Feiyue Huang, and Changsheng Xu. Unveiling the potential of structure preserving for weakly supervised object localization. In *IEEE CVPR*, pages 11642–11651, 2021. 1, 2, 5, 6

[22] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *IEEE ICCV*, pages 367–376, 2021. 2, 3

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. In *IJCV*, volume 115, pages 211–252. Springer, 2015. 5, 6, 7

[24] Lei Tan, Pingyang Dai, Rongrong Ji, and Yongjian Wu. Dynamic prototype mask for occluded person re-identification. In *ACM MM*, pages 531–540, 2022. 3

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357. PMLR, 2021. 5

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017. 2

[27] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. (CNS-TR-2011-001), 2011. 5, 6, 7

[28] Jun Wei, Sheng Wang, S Kevin Zhou, Shuguang Cui, and Zhen Li. Weakly supervised object localization through inter-class feature similarity and intra-class appearance consistency. In *ECCV*, pages 195–210, 2022. 3

[29] Pingyu Wu, Wei Zhai, and Yang Cao. Background activation suppression for weakly supervised object localization. In *IEEE CVPR*, pages 14228–14237, 2022. 3, 5, 6

[30] Jinheng Xie, Cheng Luo, Xiangping Zhu, Ziqi Jin, Weizeng Lu, and Linlin Shen. Online refinement of low-level feature based activation map for weakly supervised object localization. In *IEEE CVPR*, pages 132–141, 2021. 2, 5, 6

[31] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *IEEE CVPR*, pages 4310–4319, 2022. 3

[32] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *IEEE ICCV*, pages 6589–6598, 2019. 1, 2, 5, 6

[33] 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 *IEEE ICCV*, pages 6023–6032, 2019. 2

[34] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. In *IEEE CVPR*, pages 13460–13469, 2020. 3

[35] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: a survey. *IEEE TPAMI*, 44(9):5866–5885, 2021. 1

[36] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *IEEE CVPR*, pages 1325–1334, 2018. 1, 2, 5, 6

[37] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *ECCV*, pages 271–287, 2020. 2, 6

[38] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE CVPR*, pages 2921–2929, 2016. 1, 2, 4, 5, 6

[39] Lei Zhu, Qi She, Qian Chen, Yunfei You, Boyu Wang, and Yanye Lu. Weakly supervised object localization as domain adaption. In *IEEE CVPR*, pages 14637–14646, 2022. 3

[40] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *IEEE ICCV*, pages 1841–1850, 2017. 2