TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization

Wei Gao[1] Fang Wan[1]* Xingjia Pan[2,3] Zhiliang Peng[1] Qi Tian[4]
Zhenjun Han[1] Bolei Zhou[5] Qixiang Ye[1]*
[1]University of Chinese Academy of Sciences [2]Youtu Lab, Tencent
[3]NLPR, Institute of Automation, CAS [4]Huawei Cloud AI
[5]The Chinese University of Hong Kong
{vasgaowei, xjia.pan}@gmail.com, {wanfang, hanzhj, qxye}@ucas.ac.cn
tian.qi1@huawei.com, bzhou@ie.cuhk.edu.hk

Abstract

Weakly supervised object localization (WSOL) is a challenging problem when given image category labels but requires to learn object localization models. Optimizing a convolutional neural network (CNN) for classification tends to activate local discriminative regions while ignoring complete object extent, causing the partial activation issue. In this paper, we argue that partial activation is caused by the intrinsic characteristics of CNN, where the convolution operations produce local receptive fields and experience difficulty to capture long-range feature dependency among pixels. We introduce the token semantic coupled attention map (TS-CAM) to take full advantage of the self-attention mechanism in visual transformer for long-range dependency extraction. TS-CAM first splits an image into a sequence of patch tokens for spatial embedding, which produce attention maps of long-range visual dependency to avoid partial activation. TS-CAM then re-allocates category-related semantics for patch tokens, enabling each of them to be aware of object categories. TS-CAM finally couples the patch tokens with the semantic-agnostic attention map to achieve semantic-aware localization. Experiments on the ILSVRC/CUB-200-2011 datasets show that TS-CAM outperforms its CNN-CAM counterparts by 7.1%/27.1% for WSOL, achieving state-of-the-art performance. Code is available at https://github.com/vasgaowei/TS-CAM

1. Introduction

Weakly supervised learning refers to methods that utilize training data with incomplete annotations to learn recognition models. Weakly supervised object localization (WSOL) solely requires the image-level annotations indicating the presence or absence of a class of objects in images to learn localization models [23, 24, 28, 46]. WSOL has attracted increasing attention as it can leverage the rich Web images with tags to learn object-level models [46].

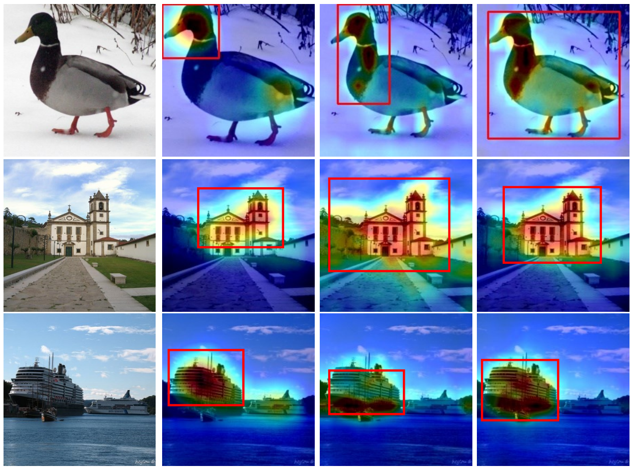

Figure 1. Comparison of weakly supervised object localization results. (a) Input Image. (b) Class Activation Map (CAM). (c) TransAttention: Transformer-based Attention. (d) TS-CAM. Object localization boxes are in red. (Best viewed in color)

As the cornerstone of WSOL [7], the Class Activation Mapping (CAM) [55] utilizes the activation map from the last convolution layer to generate semantic-aware localization maps for object bounding-box estimation. However, CAM suffers from severe underestimation of object regions because the discriminative regions activated through the classification models are often much smaller than objects' actual extent [2]. Local discriminative regions are capable of minimizing image classification loss, but experience

*Corresponding authors

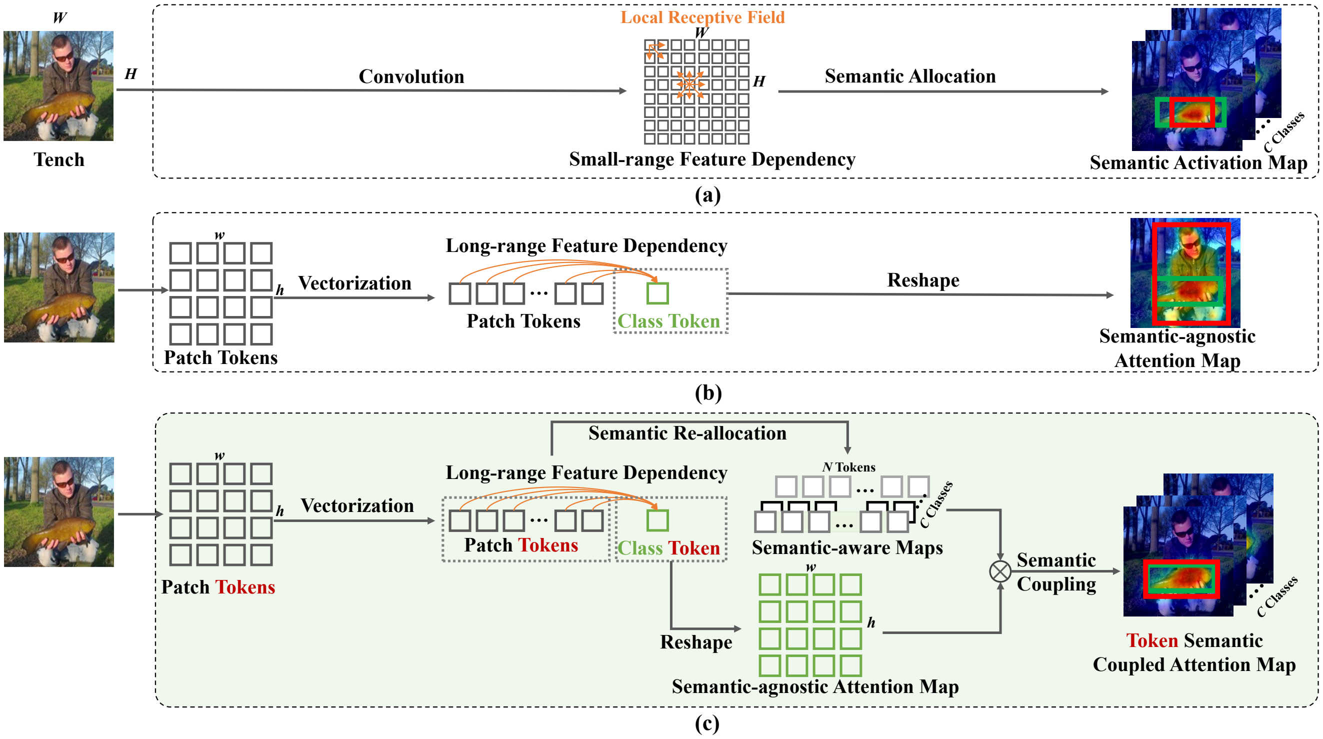

Figure 2. Comparison of the mechanisms of (a) CNN-based CAM, (b) Transformer-based Attention and (c) the proposed TS-CAM. The CNN-based CAM method is limited by the small-range feature dependency and the transformer-based attention is limited by the semantic-anostic issue. TS-CAM is able to produce semantic coupled attention maps for complete object localization. (Best viewed in color)

difficulty for accurate object localization [46], Fig. 1(b). Much effort has been made to solve this problem by proposing various regularizations [50, 51, 46, 20, 21, 6], divergent activation [33, 46, 48] or adversarial training [8, 21, 43, 50, 48, 33]. However, there is very little work to pay attention to fundamentally solving the inherent defects of CNN's local representation, Fig. 2(a). Capturing the long-range feature dependency, which can be interpreted as the semantic correlation between features in different spatial locations, is critical for WSOL.

Recently, visual transformer has been introduced to the computer vision area. Visual transformer [10] constructs a sequence of tokens by splitting an input image into patches with positional embedding and applying cascaded transformer blocks to extract visual representation. Thanks to the self-attention mechanism and Multilayer Perceptron (MLP) structure, visual transformers can learn complex spatial transforms and reflect long-range semantic correlations adaptively, which is crucial for localizing full object extent, Fig. 1(d). Nevertheless, visual transformer cannot be directly mitigated to WSOL for the following two reasons: (1) When using patch embeddings, the spatial topology of the input image is destroyed, which hinders the generation of activation maps for object localization. (2) The attention maps of visual transformers are semantic-agnostic (not distinguishable to object classes) and are not competent to semantic-aware localization, Fig. 2(b).

In this study, we propose the token semantic coupled attention map (TS-CAM), making the first attempt for weakly supervised object localization with visual transformer. TS-CAM introduces a semantic coupling structure with two network branches, Fig. 2(c), one performs semantic re-allocation using the patch tokens and the other generates semantic-agnostic attention map upon the class tokens. Semantic re-allocation, with class-patch semantic activation, enables the patch tokens to be aware of object categories. The semantic-agnostic attention map aims to capture long-distance feature dependency between patch tokens by taking the advantages of the cascaded self-attention modules in transformer. TS-CAM finally couples the semantic-aware maps with the semantic-agnostic attention map for object localization, Fig. 2(c).

The contributions of this work are as follows:

- We propose the token semantic coupled attention map (TS-CAM), as the first solid baseline for WSOL using visual transformer by leveraging the long-range feature dependency.
- We propose the semantic coupling module to combine the semantic-aware tokens with the semantic-agnostic attention map, providing a feasible way to leverage

both semantics and positioning information extracted by visual transformer for object localization.

- TS-CAM achieves a substantial improvement over previous methods on two challenging WSOL benchmarks, fully exploiting the long-range feature dependency in the visual transformer.

2. Related Work

Weakly Supervised Object Localization (WSOL) aims to learn object localizations given solely image-level category labels. A representative study of WSOL is CAM [55], which produces localization maps by aggregating deep feature maps using a class-specific fully connected layer. By removing the last fully connected layer, CAM can also be implemented by fully convolutional networks [15].

Despite the simplicity and effectiveness of CAM-based methods, they suffer identifying small discriminative parts of objects. To improve the activation of CAMs, HaS [33] and CutMix [33] adopted adversarial erasing on input images to drive localization models focusing on extended object parts. ACoL [50] and ADL [8] instead erased feature maps corresponding to discriminative regions and used adversarially trained classifiers to reconvert missed parts. SPG [51] and I^{2}C [52] increased the quality of localization maps by introducing the constraint of pixel-level correlations into the network. DANet [46] applied a divergent activation to learn complementary visual cues for WSOL. SEM [53] and SPA [22] refined the localization maps by using the point-wise similarity within seed regions. GC-Net [20] took geometric shapes into account and proposed a multi-task loss function for WSOL.

Most of the above methods struggled to expand activation regions by introducing sophisticated spatial regularization techniques to CAM. However, they remain puzzled by the contradiction between image classification and object localization. As observed by the visualization approaches [3, 49], CNNs tend to decompose an object into local semantic elements corresponding local receptive fields. Activating a couple of the semantic elements could bring good classification results. The problem about how to collect global cues from local receptive fields remains.

Weakly Supervised Detection and Segmentation are vision tasks closely related to WSOL. Weakly supervised detection train networks to simultaneously perform image classification and instance localization [41, 39, 27]. Given thousands of region proposals, the learning procedure selects high-scored instances from bags while training detectors. In a similar way, weakly supervised segmentation trains classification networks to estimate pseudo masks which are further used for training the segmentation networks. To generate accurate pseudo masks,[17, 1, 14, 41, 56] resorted to a region growing strategy. Meanwhile, some researchers investigated to directly enhance the feature-level activated regions [18, 44]. Others accumulate CAMs by training with multiple phases [16], exploring boundary constraint [5], leveraging equivalence for semantic segmentation [42], and mining cross-image semantics [35] to refine pseudo masks.

Similar to WSOL, many weakly supervised detection and segmentation approaches are prone to localize object parts instead of full object extent. There is a requirement to explore new classification models to solve the partial activation problem in a systematic way.

Long-Range Feature Dependency. CNNs produce a hierarchical ensemble of local features with different reception fields. Unfortunately, most CNNs [31, 12] are good at extracting local features but experience difficulty to capture global cues.

To alleviate such a limitation, one solution is to utilize pixel similarity and global cues to refine activation maps [41, 42, 52, 53]. Cao *et al*. [4] found that the global contexts modeled by non-local networks are almost the same for query positions and thereby proposed NLNet [40] with SENet [13] for global context modeling. MST [34] proposed the learnable tree filter to capture the structural property of minimal spanning tree to model long-range dependencies. The other solution is the attention mechanism [40, 26, 54]. The non-local operation [40] was introduced to CNNs in a self-attention manner so that the response at each position is a weighted sum of the features at all (global) positions. SASA [26] verified that self-attention is an effective stand-alone layer for CNNs. Relation Networks [9] proposed to process a set of objects simultaneously through interaction between their features and geometry, allowing modeling the spatial relations between objects. Recent studies introduced a cascaded self-attention mechanism in the transformer model to capture long-range feature dependency [45, 37, 47, 25].

3. Methodology

In this section, we first give the preliminaries for visual transformer. We then introduce the TS-CAM method.

3.1. Preliminaries

For visual transformer [10], an input image x of $W \times H$ resolution is divided to $w \times h$ patches, where $w = W/P, h = W/P$ and P denotes the width/height of a patch. The divided patches are flattened and linearly projected to construct $N = w \times h$ patch tokens $\{t_n^0 \in \mathbb{R}^{1\times D}, n = 1, 2, ..., N\}$ and a class token $t_*^0 \in \mathbb{R}^{1\times D}$, Fig. 3. D stands for the dimension of each token embedding. The class token t_*^0 is learnable with random initialization. Each token is added with a learnable position embedding in an element-wised manner. These tokens are fed into L cascaded transformer blocks, each of which consists of a multi-head self-attention layer and a Multilayer Perceptron (MLP) block.

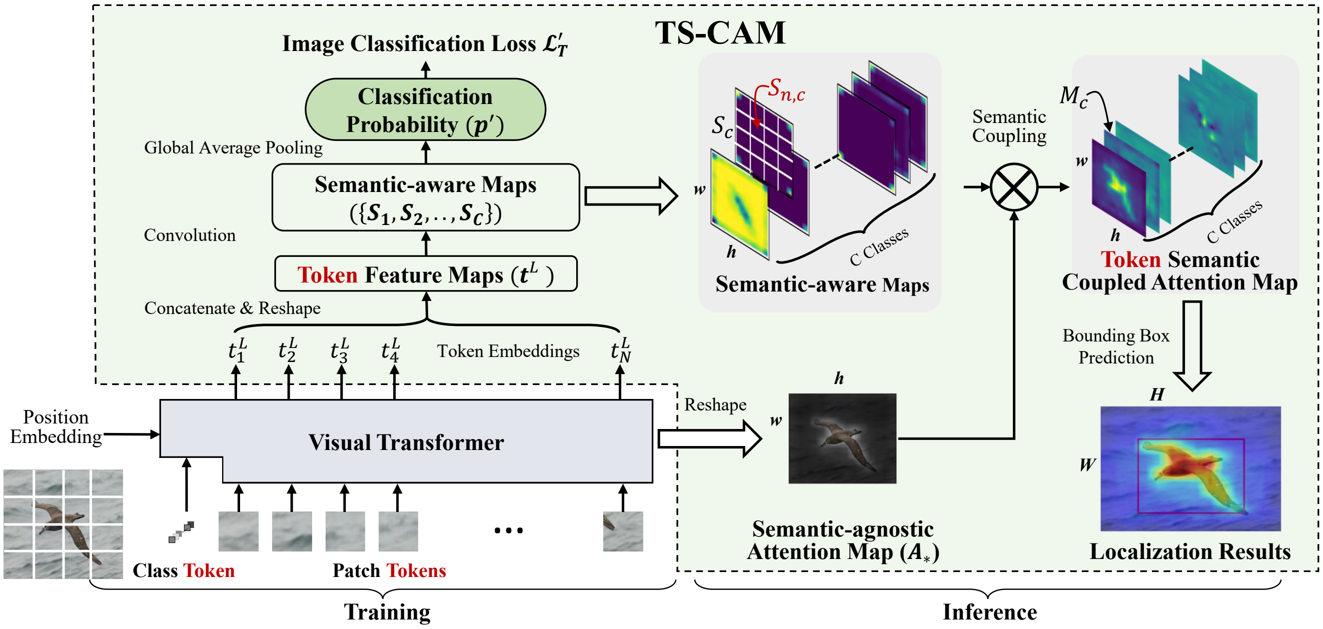
TS-CAM
Image Classification Loss $\mathcal{L}'_T$
Classification Probability (p')
Global Average Pooling
Semantic-aware Maps ($\{S_1, S_2, .., S_C\}$)
Convolution
Token Feature Maps (t^L)
Concatenate & Reshape
t_1^L t_2^L t_3^L t_4^L Token Embeddings t_N^L
$S_{n,c}$
S_c
w
h
C Classes
Semantic-aware Maps
Semantic Coupling
M_c
w
h
C Classes
Token Semantic Coupled Attention Map
Bounding Box Prediction
Position Embedding
Visual Transformer
Reshape
h
w
H
W
Class Token
Patch Tokens
...
Semantic-agnostic Attention Map (A_*)
Localization Results
Training
Inference

Figure 3. TS-CAM framework, which consists of a visual transformer for feature extraction, a semantic re-allocation branch and a semantic coupling module. Note that there is no gradient back-propagation along the semantic re-allocation branch.

Denote t_n^l and t_*^l as the n-th patch token and the class token of the l-th transformer block, respectively. The last embedded class token t_*^L is fed to an MLP block to predict the classification probability, as

$$p = \text{Softmax}\big(\text{MLP}(t_*^L)\big), \tag{1}$$

where $p \in \mathbb{R}^{1\times C}$ and C denotes the number of classes. p_c denotes the predicted probability to class c. MLP$(\cdot)$ denotes the classification function implemented by the MLP block. Denote the ground-truth label for image x as $y \in \{1, 2, ..., C\}$, the classification loss function is defined as

$$\mathcal{L}_T = -\log p_y, \tag{2}$$

which is used to train the visual transformer.

3.2. TS-CAM

We propose the TS-CAM method to generate semantic-aware localization maps upon the trained visual transformer, Fig. 3. In the visual transformer, however, only the class token is semantic-aware while the patch tokens are semantic-agnostic. To fulfill semantic-aware localization, we introduce the semantic re-allocation branch to transfer semantics from the class token to patch tokens and generate semantic-aware maps. Such semantic-aware maps are coupled with the semantic-agnostic attention maps to generate the semantic-aware localization maps.

Semantic Re-allocation. Visual transformer uses the class token to predict image categories (semantics), while using semantic-agnostic patch tokens to embed object spatial locations and reflects feature spatial dependency. To generate semantic-aware patch tokens, we propose to re-allocate the semantics from the class token t_*^L to the patch tokens $\{t_1^L, t_2^L, ..., t_N^L\}$.

As shown in Fig. 3, patch token embeddings of the L-th visual transformer block are concatenated and transposed as $\mathbf{t}^L \in \mathbb{R}^{D\times N}$. They are then reshaped to token feature maps $\mathbf{t}^L \in \mathbb{R}^{D\times w\times h}$, where $\mathbf{t}_d^L, d \in \{1, 2, ..., D\}$ denotes the d-th feature map. The semantic aware map S_c of class c is calculated by convolution as

$$S_c = \sum_d \mathbf{t}_d^L * k_{c,d}, \tag{3}$$

where $k \in \mathbb{R}^{C\times D\times 3\times 3}$ denotes the convolution kernel and $k_{c,d}$ is a 3×3 kernel map indexed by c and d. $*$ is the convolution operator. To produce semantic-aware maps, the loss function defined in Eq. 2 is updated to

$$\begin{aligned}\mathcal{L'}_T &= -\log p'_y \\ &= -\log \frac{\exp\left(\sum_n S_{n,y}/N\right)}{\sum_c \exp\left(\sum_n S_{n,c}/N\right)},\end{aligned} \tag{4}$$

where $S_{n,c}$ is the semantic of the n-th patch token for class c. While optimizing Eq. 2 allocates the semantics to class token t_*^L, minimizing Eq. 4 re-allocates the semantics to patch tokens $\{t_1^L, t_2^L, ..., t_N^L\}$, generating semantic-aware maps for WSOL.

Semantic-agnostic Attention Map. To fully exploit the long-range feature dependency of visual transformer, we propose to aggregate the attention vectors of the class token to generate the semantic-agnostic attention map. Denoting $\mathbf{t}^l \in \mathbb{R}^{(N+1)\times D}$ as the input of transformer block l, which is calculated by concatenating the embeddings of all tokens (including class token and all patch tokens). In the self-attention operation in the (l)-th transformer block, the embedded tokens $\tilde{\mathbf{t}}^l$ is computed as

$$\begin{aligned}\tilde{\mathbf{t}}^l &= \mathrm{Softmax}\Big((\mathbf{t}^l\theta_q^l)(\mathbf{t}^l\theta_k^l)^\top/\sqrt{D}\Big)(\mathbf{t}^l\theta_v^l) \\ &= A^l(\mathbf{t}^l\theta_v^l),\end{aligned} \tag{5}$$

where θ_q^l, θ_k^l and θ_v^l respectively denote parameters of the linear transformation layers of self-attention operation in (l)-th transformer block. $\top$ is a transpose operator. $A^l \in \mathbb{R}^{(N+1)\times(N+1)}$ is the attention matrix while $A_*^l \in \mathbb{R}^{1\times(N+1)}$ is the attention vector of the class token. In the multi-head attention layer where K heads are considered, D in Eq. 5 is updated as D', where $D' = D/K$. A_*^l is then updated as the average of attention vectors from K heads.

Eq. 5 implies that A_*^l records the dependency of the class token to all tokens by the matrix multiplication operation. Eq. 5 implies that the embedding $\tilde{t}_*^l$ of class token of the self-attention operation is calculated by multiplying its attention vector A_*^l with the embedding $\mathbf{t}^l$ in the (l)-th transformer block. $\tilde{t}_*^l$ is therefore able to "see" all patch tokens, where A_*^l implies how much attention is paid on each token. When Eq. 4 is optimized, the attention vector A_*^l is driven to focus on object regions ($e.g.$, long-range features of semantic correction) for image classification. The final attention vector A_* is defined as

$$A_* = \frac{1}{L}\sum_l A_*^l, \tag{6}$$

which aggregates attention vectors (A_*^l) and collects feature dependency from cascaded transformer blocks to indicate full object extent.

Semantic-Attention Coupling. As the attention vector A_* is semantic-agnostic, we use an element-wise multiplication to couple it with the semantic-aware maps to obtain the semantic-coupled attention map M_c for each class c, Fig. 3. The coupling procedure is formulated as

$$M_c = \Gamma^{w\times h}(A_*) \otimes S_c, \tag{7}$$

where $\otimes$ denotes an element-wise multiplication and addition operations. $\Gamma^{w\times h}(\cdot)$ denotes the reshape function which converts the attention vector ($\mathbb{R}^{1\times N}$) to the attention map ($\mathbb{R}^{w\times h}$). M_c is up-sampled to a semantic-aware localization map, which is used for object bounding box prediction with a thresholding approach [51].

4. Experiments

4.1. Experimental Settings

Datasets. TS-CAM is evaluated on two commonly used benchmarks, $i.e.$, CUB-200-2011 [38] and ILSVRC [29]. CUB-200-2011 is a fine-grained bird dataset with 200 different species, which is split into the training set of $5,994$ images and the test set of $5,794$ images. In ILSVRC, there are around 1.2 million images about $1,000$ categories for training and $50,000$ images for validation. The model is trained on the training set and evaluated on the validation set where the bounding box annotations are solely used for evaluation.

Evaluation Metrics. Top-1/Top-5 classification accuracy (Top-1/Top-5 *Cls. Acc*), Top-1/Top-5 localization accuracy (Top-1/Top-5 *Loc. Acc*), and Localization accuracy with ground-truth class (Gt-Known *Loc. Acc*) are adopted as evaluation metrics following baseline methods [29, 51, 55]. For localization, a prediction is positive when it satisfies: the predicted classification is correct; the predicted bounding boxes have over 50% IoU with at least one of the ground-truth boxes. $Gt\text{-}Known$ indicates that it considers localization regardless of classification.

Implementation Details. TS-CAM is implemented based on the Deit backbone [37], which is pre-trained on ILSVRC [29]. Each input image is re-scaled to 256×256 pixels, and randomly cropped by 224×224 pixels. We remove the MLP head, and add one convolution layer with kernel size 3×3, stride 1, pad 1 with 200 output units (1000 units for ILSVRC). The newly added layer is initialized following He's approach [11]. When training WSOL models, we use AdamW [19] with ϵ=1e-8, β_1=0.9 and β_2=0.99 and weight decay of 5e-4. On CUB-200-2011, the training procedure lasts 60 epochs with learning rate 5e-5 and batch-size 128. On ILSVRC dataset, training carries out 12 epochs with learning rate 5e-4 and batch-size 256.

4.2. Performance

Main Results. Table 1 compares TS-CAM with other methods on the CUB-200-2011. TS-CAM with a Deit-S backbone [37] outperforms the baseline methods on Top-1, Top-5, and Gt-Known metrics by a surprisingly large margin, yielding the localization accuracy of Top-1 71.3%, and Top-5 83.8%. Compared with the state-of-the-art methods (RCAM [53] and MEIL [21]), it respectively achieves gains of 12.3% and 13.8% in terms of Top-1 *Loc. Acc*. The left part of Fig. 4 compares localization examples by CAM [55], Transformer-based Attention, and TS-CAM on the CUB-200-2011. TS-CAM preserves global structures and covers more extent of objects. By solely utilizing the attention map from transformer structure, *TransAttention* highlights most object parts, but fails to precisely localize full objects due to the lack of category semantics.

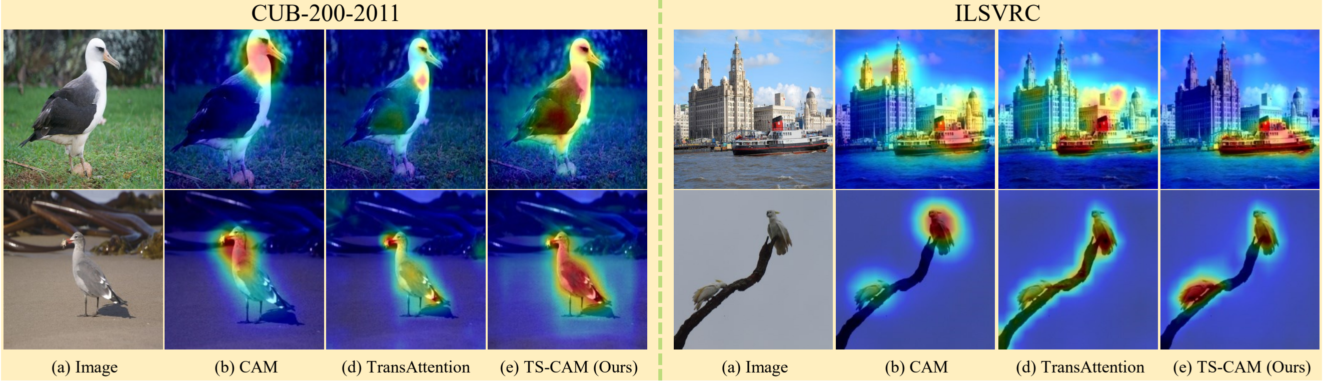

Figure 4. Visualization of localization maps on CUB-200-2011 and ILSVRC datasets. (a) Input Image. (b) Class Activation Map (CAM). (c) TransAttention: Transformer-based Attention. (d) TS-CAM (ours). (Best Viewed in Color)

Methods	Backbone	Loc. Acc		
		Top-1	Top-5	Gt-Known
CAM [55]	GoogLeNet	41.1	50.7	55.1
SPG [51]	GoogLeNet	46.7	57.2	-
RCAM [53]	GoogleNet	44.8	-	61.7
DANet [46]	InceptionV3	49.5	60.5	67.0
ADL [8]	InceptionV3	53.0	-	-
CAM [55]	VGG16	44.2	52.2	56.0
ADL [8]	VGG16	52.4	-	75.4
ACoL [50]	VGG16	45.9	56.5	59.3
DANet [46]	VGG16	52.5	62.0	67.7
SPG [51]	VGG16	48.9	57.2	58.9
I^2C [52]	VGG16	56.0	68.4	-
MEIL [21]	VGG16	57.5	-	73.8
RCAM [53]	VGG-16	59.0	-	76.3
TS-CAM (Ours)	Deit-S	**71.3**	**83.8**	**87.7**

Table 1. Comparison of TS-CAM with the state-of-the-art on the CUB-200-2011 [38] test set.

Methods	Backbone	Loc. Acc		
		Top-1	Top-5	Gt-Known
Backprop [30]	VGG16	38.9	48.5	-
CAM [55]	VGG16	42.8	54.9	59.0
CutMix [48]	VGG16	43.5	-	-
ADL [8]	VGG16	44.9	-	-
ACoL [50]	VGG16	45.8	59.4	63.0
I^2C [52]	VGG16	47.4	58.5	63.9
MEIL [21]	VGG16	46.8	-	-
RCAM [53]	VGG-16	44.6	-	60.7
CAM [55]	InceptionV3	46.3	58.2	62.7
SPG [51]	InceptionV3	48.6	60.0	64.7
ADL [8]	InceptionV3	48.7	-	-
ACoL [50]	GoogLeNet	46.7	57.4	-
DANet [46]	GoogLeNet	47.5	58.3	-
RCAM [53]	GoogleNet	44.8	-	61.7
MEIL [21]	InceptionV3	49.5	-	-
I^2C [52]	InceptionV3	53.1	64.1	**68.5**
GC-Net [20]	InceptionV3	49.1	58.1	-
TS-CAM (Ours)	Deit-S	**53.4**	**64.3**	67.6

Table 2. Comparison of TS-CAM with state-of-the-art methods on the ILSVRC [29] validation set.

In Table 2, we compare TS-CAM with its CNN counterparts (CAM) and the SOTAs on the localization accuracy by using tight bounding boxes on the ILSVRC. TS-CAM respectively outperforms CAM on the VGG16 [32] by 10.6% and 9.4% in terms of Top-1 *Loc. Acc* and Top-5 *Loc. Acc*. Compared with SOTAs with the VGG16 backbone [32], TS-CAM outperforms by $\sim$ 6% and $\sim$ 4% in terms of Top-1 *Loc. Acc* and Top-5 *Loc. Acc*. Compared with I^2C, TS-CAM achieves performance gains of 6.0% Top-1 *Loc. Acc* and 5.8% Top-5 *Loc. Acc*, which are significant margins to the challenging problem. Compared with SOTAs on the well-designed Inception V3 [36], TS-CAM also achieves the best performance. Specifically, TS-CAM achieves performance gains of 7.1% and 6.1% in terms of Top-1 and Top-5 *Loc. Acc* compared with CAM. Compared with I^2C which leverages pixel-level similarities across different objects to prompt the consistency of object features within the same categories, TS-CAM achieves comparable results with a cleaner and simpler pipeline. The right half of Fig. 4 illustrates examples of localization maps on ILSVRC. CAM [55] tends to activate local discriminative regions and cannot retain the object structure well. Due to the lack of category semantics, *TransAttention* activates almost the salient objects within images (*e.g.*, the building in the first image and the branch in the second image).TS-CAM takes the advantage of self-attention mechanism in visual transformer and thus activates the full extent of objects.

In Fig 5, we compare localization accuracy among variant CAM methods under different IoUs on CUB-200-2011 [38]. TS-CAM outperforms the CAM [55] and RCAM [53] under each IoU by large margins. In addition, TS-CAM achieves larger gains as IoU threshold increases, which indicates that the localization maps of our method cover the object extent accurately.

Parameter Complexity. Under similar parameter complexity and computational cost overhead, TS-CAM (with 25.1M

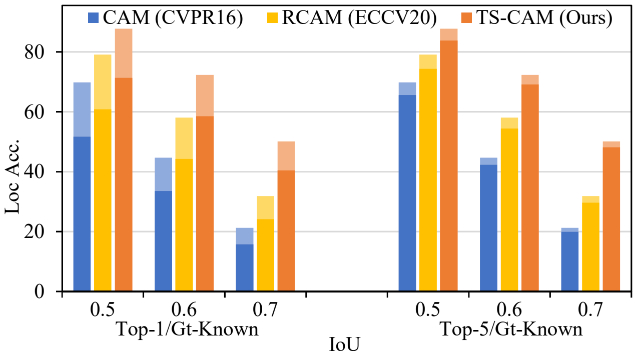

Figure 5. Comparison of localization accuracy under IoUs on CUB-200-2011 [38]. Lighter color means Gt-Known *Loc Acc*.

Methods	Image size	#Params (M)	MACs (G)	Top-1 Loc Acc. %
VGG16-CAM	224^2	19.6	16.3	44.2
GoogleNet-CAM	224^2	16.7	13.5	41.1
TS-CAM (Ours)	224^2	25.1	5.29	71.3

Table 3. Comparison of parameters and MACs. TS-CAM is implemented based on Deit-S [37]. And Top-1 *Loc Acc.* is evaluated on the CUB-200-2011 test set [38].

Methods	ILSVRC(%)			CUB-2011-200(%)		
	M-Ins	Part	More	M-Ins	Part	More
VGG16	10.65	3.85	9.58	-	21.91	10.53
InceptionV3	10.36	3.22	9.49	-	23.09	5.52
TS-CAM (Ours)	9.13	3.78	7.65	-	6.30	2.85

Table 4. Localization error statistics.

parameters and 5.29G MACs) respectively outperforms VGG16-CAM (with 19.6M parameters and 16.3G MACs) by 27.1% (71.3% vs. 44.2%) and GoogleNet-CAM (with 16.7M parameters and 13.5G MACs) by 27.2% (71.3% vs. 41.1%) in Table 3.

Error Analysis. To further reveal the effect of TS-CAM, we categorize the localization errors into five as in [22]: classification error (Cls), multi-instance error (M-Ins), localization part error (Part), localization more error (More), and others (OT). *Part* indicates that the predicted bounding box only cover the parts of object, and IoU is less than a certain threshold. *More* indicates that the predicted bounding box is larger than the ground truth bounding box by a large margin. Each metric calculates the percentage of images belonging to corresponding error in the validation/test set. Table 4 lists localization error statistics of *M-Ins*, *Part*, and *More*. TS-CAM effectively reduces the *M-Ins*, *Part* and *More* errors on both benchmarks, which indicates more accurate localization maps. For CUB-200-2011, TS-CAM significantly reduces both *Part* and *More*-type errors by $\sim$ 17% and $\sim$ 3% compared with CAM [55] on the basis of well-designed Inception V3.

Methods	Backbone	Loc Acc.		
		Top-1	Top-5	Gt-Known
CAM [55]	VGG-16	44.2	52.2	58.0
	GoogleNet	41.2	51.7	55.1
TransAttention	Deit-S	58.9	69.7	73.0
TransCam	Deit-S	17.7	18.3	18.3
TS-CAM (Ours)	Deit-S	71.3	83.8	87.8

Table 5. Ablation studies of TS-CAM components on the CUB-200-2011 test set [38].

Methods	Backbone	Loc Acc.		
		Top-1	Top-5	Gt-Known
CAM	VGG-16	42.8	54.9	59.0
	InceptionV3	46.3	58.2	62.7
TransAttention	Deit-S	43.0	51.9	54.7
TransCam	Deit-S	34.9	42.9	46.0
TS-CAM (Ours)	Deit-S	53.4	64.3	67.6

Table 6. Ablation study of TS-CAM components on the ILSVRC validation set [29].

L	Top-1	Top-5	Gt-Known
8	65.2	75.8	79.0
9	68.5	80.0	83.6
10	70.2	81.2	85.5
11	71.2	83.7	87.7
12	**71.3**	**83.8**	**87.7**

Table 7. Ablation results of TS-CAM when attention maps(A_*^l) from different layers are summed on CUB-200-2011 [38] test set.

4.3. Ablation Study

Attention and activation. Using Deit-S as the backbone, we conduct ablation studies to verify the components in TS-CAM. Specifically *TransAttention* solely uses semantic-agnostic attention map (A_*) for object localization, while *TransCAM* solely uses semantic-aware maps (S_c). A_* and S_c are respectively generated by Eq.6 and Eq.3, and are illustrated in Fig. 3. In Table 5, we evaluate the performance of the TS-CAM on CUB-200-2011 and observed significant improvements over TS-CAM components. Specifically, TS-CAM obtains gains of 12.4%, 14.1%, and 14.8% in terms of Top-1, Top-5 and Gt-Known *Loc. Acc* compared with *TransAttention*. Using the semantic-aware map, *TransCAM* struggles from distinguishing an object from the background due to the destruction of topology. Taking advantage of both modules, TS-CAM generates semantic-aware localization maps by coupling semantic-agnostic attention from transformer and token semantics from the classifier.

Table 6 shows the results on ILSVRC validation set with different configurations. Following CAM [55], *TransCAM* only utilizes the token semantic-aware map from the classifier to capture the object localization. Since the topology of input image is destroyed, *TransCAM* cannot

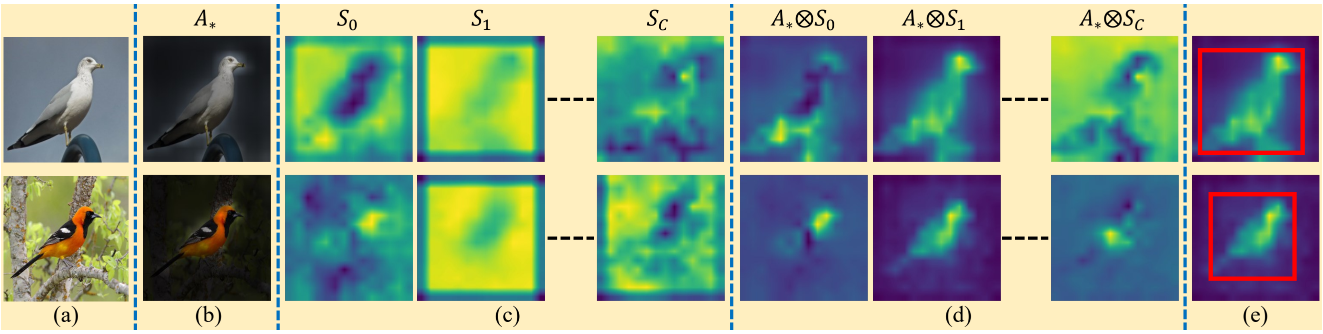

Figure 6. Visualization of semantic-attention coupling. (a) Input image. (b) Semantic-agnostic attention Map. (c) Token semantic-aware maps. (d) Token semantic coupled attention maps. (e) Localization results. (Best viewed in color)

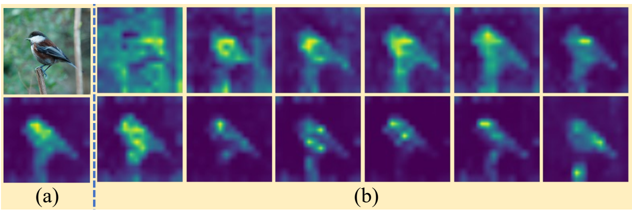

Figure 7. Attention maps from different transformer layers. (a) Input Image and semantic-agnostic attention map (A_*). (b) Attention maps (A_*^l) from different transformer layers. (Best viewed in color)

generate structure-preserving activation maps and thus cannot differentiate the objects from background. It obtains a significant performance degradation compared with TS-CAM and CAM. *TransAttention* achieves 10.4% and 12.4% performance degradation compared TS-CAM in terms of Top-1 and Top-5 *Loc. Acc*. The class-agnostic features puzzled *TransAttention* toward false localization, Fig. 2(b).

Why summarize all attention maps? In Fig. 7, we visualize attention maps A_*^l from all layers and semantic-agnostic attention map A_*. As $\{A_*^1, ..., A_*^L\}$ are complementary, we summarize them for full object extent localization. Ablation study in Table 7 on CUB-200-2011 dataset demonstrates that summing all A_*^l achieves the highest localization accuracy.

Why attention instead of activation? The reasons are two folds: (1) Visual transformer leverages embeddings of a low-resolution class token for image classification and can not produce high-resolution CAM. (2) The semantic-aware maps in TS-CAM, which are calculated by re-allocating the semantics from the class token to patch tokens, fail to discriminatively activate the object regions. By visualizing the similarities among patch token embeddings $\{t_1^L, ..., t_N^L\}$ in Fig. 8 right, we observed that the patch token embeddings are similar with each other, which implies that the activation results (semantic-aware activation maps) generated by these embeddings experience difficulty to discriminate objects from their backgrounds, Fig. 6(c).

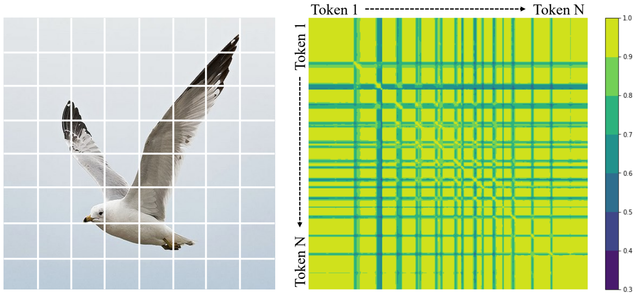

Figure 8. **Left**:Input patch tokens. **Right**: Visualization of the similarity matrix for patch token embeddings. Each row/column represents the cosine similarities between a patch token embedding with all patch token embeddings.

5. Conclusion

We proposed the token semantic coupled attention map (TS-CAM) for weakly supervised object localization. TS-CAM takes full advantage of the cascaded self-attention mechanism in the visual transformer for long-range feature dependency extraction and object extent localization. To solve the semantic agnostic issue of the patch tokens, we proposed to re-allocate category-related semantics for patch tokens, enabling each of them to be aware of object categories. We proposed the semantic coupling strategy to fuse the patch tokens with the semantic-agnostic attention map to achieve semantic-aware localization results. Experiments on the ILSVRC/CUB-200-2011 datasets show that TS-CAM significantly improved the WSOL performance, in striking contrast with its CNN counterpart (CAM). As the first and solid baseline with transformer, TS-CAM provides a fresh insight to the challenging WSOL problem.

Acknowledgment. This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61836012, 61771447 and 62006216, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA27000000, and the China National Postdoctoral Program for Innovative Talents under Grant BX20190323.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE CVPR*, 2018. 3

[2] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, pages 618–634, 2020. 1

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE CVPR*, pages 3319–3327, 2017. 3

[4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *IEEE ICCV Workshops*, pages 1971–1980, 2019. 3

[5] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *ECCV*, volume 12371, pages 347–362, 2020. 3

[6] Nenglun Chen, Xingjia Pan, Runnan Chen, Lei Yang, Zhiwen Lin, Yuqiang Ren, Haolei Yuan, Xiaowei Guo, Feiyue Huang, and Wenping Wang. Distributed attention for grounded image captioning. *arXiv preprint arXiv:2108.01056*, 2021. 2

[7] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *IEEE CVPR*, pages 3133–3142, 2020. 1

[8] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *IEEE CVPR*, pages 2219–2228, 2019. 2, 3, 6

[9] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *IEEE ICCV*, pages 7022–7031, 2019. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2, 3

[11] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 5

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 3

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE CVPR*, pages 7132–7141, 2018. 3

[14] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *IEEE CVPR*, pages 7014–7023, 2018. 3

[15] Sangheum Hwang and Hyo-Eun Kim. Self-transfer learning for weakly supervised lesion localization. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 239–246, 2016. 3

[16] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *IEEE ICCV*, pages 2070–2079, 2019. 3

[17] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, pages 695–711. Springer, 2016. 3

[18] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *IEEE CVPR*, pages 5267–5276, 2019. 3

[19] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[20] Weizeng Lu, Xi Jia, Weicheng Xie, Linlin Shen, Yicong Zhou, and Jinming Duan. Geometry constrained weakly supervised object localization. *arXiv preprint arXiv:2007.09727*, 2020. 2, 3, 6

[21] 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. 2, 5, 6

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11642–11651, June 2021. 3, 7

[23] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *IEEE ICCV*, pages 1742–1750, 2015. 1

[24] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *IEEE ICCV*, pages 1796–1804, 2015. 1

[25] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. *ArXiv*, abs/2105.03889, 2021. 3

[26] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019. 3

[27] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G. Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *IEEE CVPR*, pages 10595–10604, 2020. 3

[28] Anirban Roy and Sinisa Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *IEEE CVPR*, pages 3529–3538, 2017. 1

[29] 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. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 5, 6, 7

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

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

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

[33] 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 3544–3553, 2017. 2, 3

[34] Lin Song, Yanwei Li, Zeming Li, Gang Yu, Hongbin Sun, Jian Sun, and Nanning Zheng. Learnable tree filter for structure-preserving feature transform. In *Advances in Neural Information Processing Systems*, pages 1711–1721, 2019. 3

[35] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, volume 12347, pages 347–365, 2020. 3

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE CVPR*, pages 2818–2826, 2016. 6

[37] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 3, 5, 7

[38] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5, 6, 7

[39] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-MIL: continuation multiple instance learning for weakly supervised object detection. In *IEEE CVPR*, pages 2199–2208, 2019. 3

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE CVPR*, pages 7794–7803, 2018. 3

[41] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *IEEE CVPR*, pages 1354–1362, 2018. 3

[42] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *IEEE CVPR*, pages 12275–12284, 2020. 3

[43] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE CVPR*, pages 1568–1576, 2017. 2

[44] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *IEEE CVPR*, pages 7268–7277, 2018. 3

[45] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020. 3

[46] 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, 3, 6

[47] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 3

[48] 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, 6

[49] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, volume 8689, pages 818–833, 2014. 3

[50] 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. 2, 3, 6

[51] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *ECCV*, pages 597–613, 2018. 2, 3, 5, 6

[52] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *ECCV*, volume 12364, pages 271–287, 2020. 3, 6

[53] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Fei Wu. Rethinking localization map: Towards accurate object perception with self-enhancement maps. *arXiv preprint arXiv:2006.05220*, 2020. 3, 5, 6

[54] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[55] 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, 3, 5, 6, 7

[56] Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, and Jianbin Jiao. Learning instance activation maps for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3