A Consistency and Integration Model with Adaptive Thresholds for Weakly Supervised Object Localization

Hao Su[1,2] , **Meng Yang**[1,2,3*]
[1]School of Computer Science and Engineering, Sun Yat-sen University
[2]State Key Laboratory of Integrated Services Networks (Xidian University)
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, SYSU
suhao8@mail2.sysu.edu.cn, yangm6@mail.sysu.edu.cn

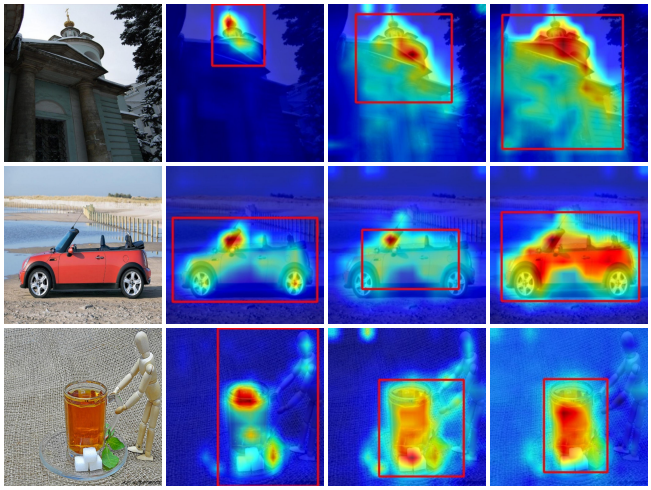
Abstract

Weakly Supervised Object Localization (WSOL) is a challenging task, which aims to learn object localization with less costly image-level labels. Existing convolution neural network (CNN) based methods tend to focus on discriminative regions of objects, while transformer-based methods overemphasize deep global features powerful for classification and lack the capability to perceive object details, leading to prediction results far from the object boundary. In this paper, we propose a novel Consistency and Integration Model with Adaptive Thresholds (CIAT) that exploits the spatial-semantic consistency between shallow and deep features to activate more object regions and detects the object regions adaptively in different images. First, we introduce a simple plug-and-play consistency and integration module of shallow-deep features (CISD), which utilizes shallow features efficiently to enhance the entire object perception. Then, we design an online adaptive threshold (OAT) based on Bayesian decision theory, which computes a reasonable segmentation threshold adaptive for the localization map of each image, making the predicted bounding box closer to the ground truth. Extensive experiments on two widely used CUB-200-2011 and ILSVRC datasets verify the effectiveness of our methods.

1 Introduction

Object detection and localization based on the fully supervised network [Liu *et al.*, 2016; Bochkovskiy *et al.*, 2020; Wang *et al.*, 2023] has achieved great success in recent years, but requires costly bounding box annotations. Weakly supervised object localization (WSOL) determines the position and size of objects through only image-level labels, which attracts wide attention due to its low cost and labor savings.

As a pioneering work, [Zhou *et al.*, 2016] aggregated features from the last convolutional layer in classification networks to learn the class activation map (CAM) for localizing objects. The classification network tends to focus on the most discriminative regions, while the localization task requires the extraction of the entire area of objects. The range of activation regions in CAM is significantly smaller than the actual range of objects, leading to poor localization accuracy. Thus, how to expand the activation areas of objects has become an urgent problem to be solved.



(a) Image (b) TS-CAM (c) SCM (d) CIAT (Ours)

Figure 1: Comparison of visual localization results on different methods: (a) Input images. (b) TS-CAM captures long-range dependencies while missing some object details. (c) SCM enhances the semantic and spatial correlation between the patches but images have various activation values. (d) The proposed CIAT further explores both the high response regions and weak response regions. These predicted bounding boxes are in red. Best viewed in color.

Many subsequent CAM-based works used adversarial erasing [Kumar Singh and Jae Lee, 2017; Zhang *et al.*, 2018a; Choe and Shim, 2019; Mai *et al.*, 2020], spatial relationship activation [Xue *et al.*, 2019; Zhang *et al.*, 2020; Guo *et al.*, 2021], and foreground prediction map [Xie *et al.*, 2021; Wu *et al.*, 2022; Xie *et al.*, 2022] to alleviate the problem of insufficient activation. However, the essence of the above methods is to obtain the local features first and then conduct the activation diffusion, which cannot solve the problem of the restricted activation regions due to local representations captured by the limited receptive field of the convolutional neural network (CNN).

In recent years, Vision Transformer (ViT) [Dosovitskiy *et*

*Corresponding author.

al., 2020] has achieved amazing results in computer vision due to the excellent extraction capability of long-distance feature dependencies. ViT used the multiple transformer blocks with a multi-head self-attention mechanism to successfully extract global features for classification. TS-CAM [Gao *et al.*, 2021] first applied the transformer structure [Touvron *et al.*, 2021] in WSOL, which learned the long-distance dependencies between pixels and alleviated the partial activation problem in CNN networks. However, most of the current researches consider using the deep features in the network for object localization, while ignoring the shallow features that contain rich image details. Deep features preserve the most discriminative information for classification but lose object details (e.g., boundaries and object parts insignificant for classification), as shown in Figure 1(b). Moreover, most existing methods [Bai *et al.*, 2022] segment the localization maps through a fixed threshold for all input images, which may lead to an inappropriate localization result (smaller or bigger one), as shown in Figure 1(c).

Based on the above analysis, we proposed a novel method, namely the Consistency and Integration model with Adaptive Thresholds (CIAT), which promotes and integrates the consistency of shallow and deep features and derives an adaptive threshold for the detection of object regions. CIAT contains a new consistency and integration module of shallow-deep features (CISD) and an online adaptive threshold (OAT) for the final localization. CISD designs an efficient shallow feature extraction module and uses shallow-deep feature consistency to determine the attention value for each class channel of the semantic feature map, forcing the model to explore both the weak response regions and high response areas, as shown in Figure 1(d). OAT is adopted to estimate the segmentation threshold of each localization map, which aims to predict closer bounding boxes to the ground truths, as shown in Figure 3. OAT segments the foreground and background from the localization map and determines an adaptive threshold (i.e., the decision boundary) via Bayesian decision theory. The theoretical threshold is approximately realized by the combination of the mean of foreground and background centers and an adjustment item for compensation, segmenting localization maps with different activation distributions.

Our main contributions can be summarized as:

- We propose the new Consistency and Integration Model with Adaptive Thresholds (CIAT) for WSOL task, which learns accurate bounding boxes to greatly improve the localization performance.
- We introduce a simple consistency and integration module of shallow-deep features (CISD) that integrates the advantages of different features to enhance the entire object perception capability of the model.
- We design an online adaptive threshold (OAT) to obtain an approximation of the theoretical threshold, adaptively and effectively detecting more complete object regions in different images.
- To validate the effectiveness of the proposed methods, we perform a series of experiments on two challenging datasets and have achieved better results than some representative methods.

2 Related Work

2.1 CNN-based Methods for WSOL

As a representative method of WSOL, [Zhou *et al.*, 2016] observed that the global average pooling (GAP) layer [Lin *et al.*, 2014] can be applied for localizing discriminative regions in images. They inserted a GAP layer behind the last convolution layer and aggregated deep features to generate the class activation map (CAM) for object localization. To expand the activation regions in CAM, HaS [Kumar Singh and Jae Lee, 2017] randomly erased a certain amount of patches with high activation in the input images, forcing the network to capture non-salient features. ACoL [Zhang *et al.*, 2018a] and ADL [Choe and Shim, 2019] erased some discriminative patches in feature maps to further expand the activation regions. MEIL [Mai *et al.*, 2020] utilized two parallel CNN networks with shared weights while simultaneously exploring both class-specific regions and non-salient areas.

Besides the erasing methods described above, DANet [Xue *et al.*, 2019] proposed a divergent activation approach to get better localization maps. I^2C [Zhang *et al.*, 2020] introduced the intra-class similarity of different images to obtain the complete localization maps. SLT-Net [Guo *et al.*, 2021] divided WSOL into two independent sub-tasks and strengthened the learning tolerance to semantic mistakes to improve the localization performance. ORNet [Xie *et al.*, 2021] used low-level features instead of high-level features in the classification network to learn localization results with clearer boundaries. BAS [Wu *et al.*, 2022] and C^2AM [Xie *et al.*, 2022] directly applied a generator to generate a foreground prediction map (FPM) for localization and constrain the FPM through some appropriate loss functions.

Although local features have been well studied in these CNN-based methods, CNNs are prone to capture partial semantic features and the entire object regions cannot be well activated due to the lack of global feature perception.

2.2 Transformer-based Methods for WSOL

CNN-based WSOL approaches inevitably focus on the discriminative regions of objects, while the transformer-based methods do well in extracting the global features of input images. ViT [Dosovitskiy *et al.*, 2020] adopted the self-attention mechanism to capture the long-range feature dependencies, which performed well in image classification tasks. Inspired by ViT, TS-CAM [Gao *et al.*, 2021] applied the transformer structure [Touvron *et al.*, 2021] in WSOL for the first time, which used attention maps from patches to avoid partial activation. Then based on the TS-CAM, LCTR [Chen *et al.*, 2022] incorporated cross-patch information and used local features to enhance the local perception capability to weak response regions. SCM [Bai *et al.*, 2022] mainly considered enhancing spatial and semantic correlation of patch tokens through the graph diffusion method. LCAR [Pan *et al.*, 2023] proposed a dynamic aggregation network that replaced the post-processing of threshold segmentation to get closer predictions. TAFormer [Meng *et al.*, 2023] learned the class-agnostic foreground maps to get complete object localization. CATR [Chen *et al.*, 2023] enhanced the category awareness

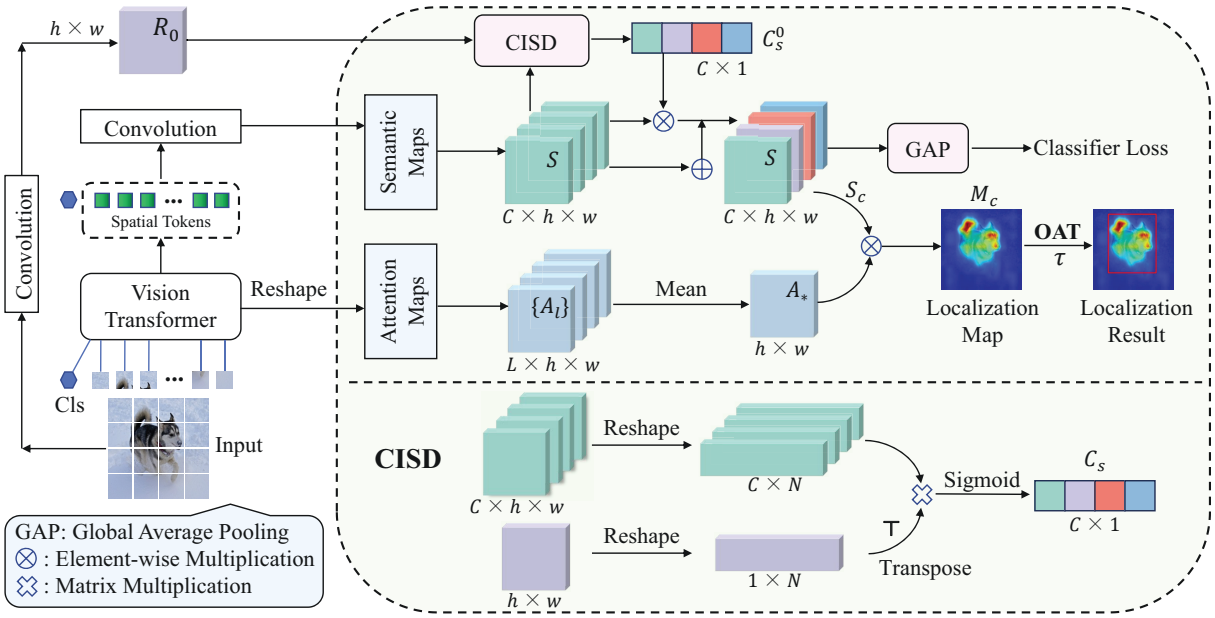


Figure 2: The object localization framework CIAT, which consists of a vision transformer for feature extraction, a consistency and integration module of shallow-deep features (CISD) and an online adaptive threshold (OAT).

of self-attention maps by learning category-aware representations for specific objects.

Despite the progress, these methods based on transformer structure ignore the shallow features in the first few layers of the neural network, which contain rich spatial information. An intuitive observation is that the shallow semantic feature has better spatial resolution while the deep semantic feature has better class discrimination ability. Our proposed method aims to force the model to explore both the detailed features and global features.

3 Methodology

In this section, we first give the overview of our proposed CIAT and then provide a detailed description of CISD and OAT. The modules are combined with the transformer structure into a joint optimization localization framework, as shown in Figure 2.

3.1 Overview

For ViT [Dosovitskiy *et al.*, 2020], an input image $I \in \mathbb{R}^{H\times W\times 3}$ is split and then linearly projected into to $N = h \times w$ patch tokens $x_p \in \mathbb{R}^{1\times D}$ of size $P \times P$, where D is the feature dimension of each token, $h = H/P$ and $w = W/P$. To extract global information from tokens, an extra class token $x_{class} \in \mathbb{R}^{1\times D}$ is added. Before these patch tokens are fed to L transformer blocks, each of which contains a multi-head attention layer and a multilayer perceptron (MLP) block, the initial input token sequence X_0 is expressed as:

$$X_0 = \{x^0_{class}; x^1_p; x^2_p; \cdots; x^N_p\} + E_{pos}, \tag{1}$$

where E_{pos} is a position embedding.

In the l-th transformer block, the output feature is denoted as $X_l \in \mathbb{R}^{(N+1)\times D}$ and self-attention mechanism is used to get an attention matrix $A_l \in \mathbb{R}^{(N+1)\times(N+1)}$ between tokens, which is formulated as:

$$A_l = \text{Softmax}(\frac{Q_l K_l^\top}{\sqrt{D/N_h}}), \tag{2}$$

where Q_l and K_l are the queries and keys obtained by linear operation. N_h is the number of heads and $\top$ is a transpose operator. We extract the attention vector $A^l_* \in \mathbb{R}^{1\times N}$ of the class token from A_l and then generate the final attention vector A_* by:

$$A_* = \frac{1}{L}\sum_{l=1}^{L} A^l_*, \tag{3}$$

which aggregates long-range feature dependency from each transformer block.

Differ from ViT, TS-CAM [Gao *et al.*, 2021] and SCM [Bai *et al.*, 2022] use the N patch tokens of the L-th transformer block for classification. The specific process is to reshape them into feature maps $F \in \mathbb{R}^{D\times h\times w}$ and to calculate a semantic map $S \in \mathbb{R}^{C\times h\times w}$ by a 3×3 convolution layer with C filters, where C is the number of categories. In the following subsection, we introduce a spatial-awareness class attention to preserve the spatial consistency in semantic feature generation process. Finally, semantic map S is fed to a global average pooling (GAP) layer [Lin *et al.*, 2014] and a softmax layer to predict the classification probability $p \in \mathbb{R}^{1\times C}$. The loss function is defined as:

$$\mathcal{L} = -\log(p_{\hat{c}}), \tag{4}$$

where $p_{\hat{c}}$ is the class probability of the correct class $\hat{c}$.

In the localization map generation phase, we first extract a semantic-agnostic map from A_* that contains long-distance dependencies and then couple it with S that contains local

information to learn the localization map M_c for each class c. The coupling procedure is formulated as:

$$M_c = \Gamma(A_*) \otimes S_c, \tag{5}$$

where $\Gamma(\cdot)$ indicates the reshape operator which coverts the vector ($\mathbb{R}^{1\times N}$) to the map ($\mathbb{R}^{h\times w}$). $\otimes$ denotes element-wise multiplication operation. The localization map M_c is adjusted to the same size as the original image by linear interpolation. Based on M_c, we model the segmentation foreground and the background as a binary classification problem and determine the threshold value based on Bayesian decision theory for final object bounding box prediction.

3.2 Consistency and Integration Module of Shallow-deep Features

CISD computes the spatial-awareness class attention through a nonlinear transform of the consistency between shallow and deep semantic features. In Figure 2, the shallow feature preserving spatial locality is first generated via a simple convolution operation with almost no increase in computational complexity. Then its consistency with the deep semantic feature processed by the multi-layer transformer incorporating global information is calculated for each class channel.

Shallow Feature Extraction. We reshape the token sequence X_l from l-th (l is a small value) transformer block to the token feature map $F_l \in \mathbb{R}^{D\times h\times w}$. Each channel of F_l focuses on different features of input images. And we reaggregate these information by one 3×3 convolutional layer with 1 filter to generate $R_l \in \mathbb{R}^{1\times h\times w}$, which contains rich image details and is calculated by:

$$R_l = \sum_d F_l^d * k_{1,d}, \tag{6}$$

where F_l^d denotes the d-th feature map, $k \in \mathbb{R}^{1\times D\times 3\times 3}$ is the convolution kernel, $k_{1,d}$ is a 3×3 kernel map indexed by 1 and d, and * is the convolution operator.

Shallow-deep Feature Consistency as Attention. We aim to fully capture channel-wise dependencies and keep spatial-semantic relations across different features with simple convolution operation. To fulfill these objectives, we calculate the similarity vector between R_l and each channel feature map of S as channel attention. Then we adopt a sigmoid activation function to learn a non-mutually-exclusive relationship as opposed to one-hot activation. The calculation procedure is formulated as:

$$C_s^l = \text{Sigmoid}(\Psi(S) \times \Psi(R_l)^\top), \tag{7}$$

where Ψ denotes the reshape function which converts the 3D vector $\mathbb{R}^{k\times h\times w}$ to the 2D vector $\mathbb{R}^{k\times N}$ ($k \in \{1, C\}$), and $\times$ denotes the matrix multiplication.

Shallow-deep Feature Integration. $C_s^l \in \mathbb{R}^{C\times 1}$ introduces the consistency information between the shallow features and the class-specific deep features. We use a channel-wise multiplication to couple S with C_s^l as well as remaining the original semantic map. Finally, S can be written as:

$$S = C_s^l \cdot S + S, \tag{8}$$

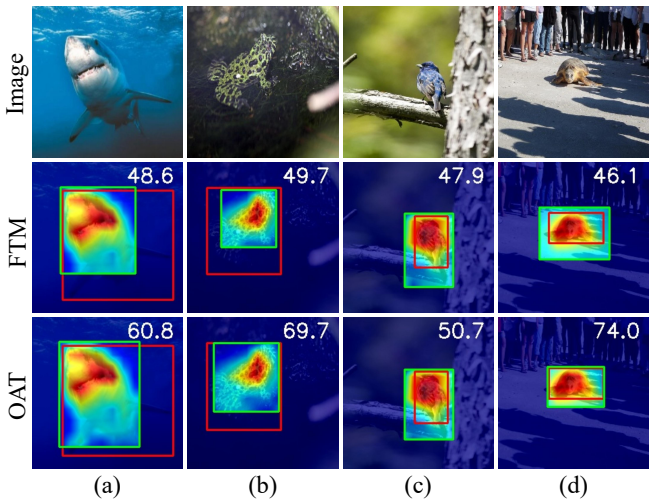


Figure 3: Comparison of localization results on FTM and OAT. Note that the ground-truth bounding boxes are in red, the predictions are in green, and IoU values (%) are shown in white text.

where $\cdot$ refers to the channel-wise multiplication. Instead of changing the structure of network or adding additional loss functions, the new semantic map integrates the class-agnostic shallow features that help extract whole object regions and the class-specific deep features that emphasize the discriminative parts. Thus, the shallow-deep feature integration makes the network focus more on features, which contribute to localization effectively.

3.3 Online Adaptive Threshold

Existing methods [Gao *et al.*, 2021; Bai *et al.*, 2022] use the same threshold to segment the localization maps of all input images. However, the objects and background values of localization maps may vary widely in different images. In the experiments, we observe that in some localization maps objects are located incorrectly, whose predictions have less than 50% IoU value with the ground-truth bounding boxes due to inappropriate thresholds. Some localization results are significantly smaller when a higher threshold is used, as presented in Figure 3(a). And the other localization results are significantly larger with a lower threshold in Figure 3(c).

The fixed threshold method (FTM) [Zhang *et al.*, 2018b] is only able to adjust the localization results of one situation above. Thus, we propose an effective online adaptive threshold (OAT) to adaptively raise the thresholds for images with larger predicted localization and lower the thresholds for images with smaller predictions. We intuitively consider following the minimum error thresholding (MET) approach [Kittler and Illingworth, 1986] to compute a theoretical threshold for each localization map.

Let M_c be divided into a foreground set M_c^{fg} and a background M_c^{bg}. The deep learning-based classification model aims to enhance the semantic map of objects and suppress the semantic map of backgrounds. Thus, it is reasonable to assume that the foreground pixels and the background pixels both follow a Gaussian distribution. Specifically, the probability density function of foreground and background values

Methods (Yr)	Backbone	Loc. Acc		
		Top-1	Top-5	GT-k.
CAM ('16)	VGG16	42.8	54.9	59.0
ACoL ('18)	VGG16	45.8	59.4	63.0
DANet ('19)	GoogLeNet	47.5	58.3	-
ORNet ('21)	VGG16	52.1	63.9	68.3
BAS ('22)	VGG16	53.0	65.4	69.6
SPG ('18)	InceptionV3	48.6	60.0	64.7
ADL ('19)	InceptionV3	48.7	-	-
MEIL ('20)	InceptionV3	49.5	-	-
GC-Net ('20)	InceptionV3	49.1	58.1	-
I^2C ('20)	InceptionV3	53.1	64.1	68.5
SLT-Net ('21)	InceptionV3	55.7	65.4	67.6
TS-CAM ('21)	Deit-S	53.4	64.3	67.6
LCTR ('21)	Deit-S	56.1	65.8	68.7
SCM ('22)	Deit-S	56.1	66.4	68.8
LCAR ('23)	Deit-S	57.1	-	70.7
TAFormer ('23)	Deit-S	56.7	66.3	70.8
CATR ('23)	Deit-S	56.9	66.6	69.3
CIAT (Ours)	Deit-S	**59.8**	**69.9**	**72.1**

Table 1: Localization accuracy on the ILSVRC validation set compared to state-of-the-art studies.

Methods (Yr)	Backbone	Cls. Acc	
		Top-1	Top-5
CAM ('16)	VGG16	68.8	88.6
ACoL ('18)	VGG16	67.5	88.0
DANet ('19)	GoogLeNet	72.5	91.4
I^2C ('20)	VGG16	69.4	89.3
ORNet ('21)	VGG16	71.6	90.4
SPG ('18)	InceptionV3	69.7	90.1
ADL ('19)	InceptionV3	72.8	-
MEIL ('20)	InceptionV3	73.3	-
GC-Net ('20)	InceptionV3	77.4	93.6
I^2C ('20)	InceptionV3	73.3	91.6
SLT-Net ('21)	InceptionV3	78.1	-
TS-CAM ('21)	Deit-S	74.3	92.1
LCTR ('21)	Deit-S	77.1	93.4
SCM ('22)	Deit-S	76.7	93.0
LCAR ('23)	Deit-S	75.9	-
TAFormer ('23)	Deit-S	77.4	-
CATR ('23)	Deit-S	77.3	93.6
CIAT (Ours)	Deit-S	**78.6**	**94.2**

Table 2: Classification accuracy on the ILSVRC validation set compared to state-of-the-art studies.

are $p(x) = N(\mu_{fg}, \sigma_{fg}^2)$ and $q(x) = N(\mu_{bg}, \sigma_{bg}^2)$, where μ_{fg} and σ_{fg} are the mean and standard deviation in M_c^{fg}, μ_{bg} and σ_{bg} are the mean and standard deviation in M_c^{bg}.

Then we define θ_{fg} is the foreground pixel ratio and the background pixel ratio is θ_{bg}, where $|\cdot|$ is the cardinal number of a set, $\theta_{fg} = |M_{fg}|/|M_c|$ and $\theta_{bg} = |M_{bg}|/|M_c|$. The probability of erroneously classifying an object point as a background point is P_1, which is defined as:

$$P_1 = \theta_{fg} \int_{-\infty}^{\tau} p(x)\, dx. \tag{9}$$

Similarly, P_2 denotes the probability that a background point is misclassified as an object point, which is defined as:

$$P_2 = \theta_{bg} \int_{\tau}^{\infty} q(x)\, dx. \tag{10}$$

Thus, the total classification error P is the sum of P_1 and P_2. Our optimization goal is to minimize P. Let $\frac{\partial P}{\partial \tau} = 0$, we can get:

$$\theta_{fg} p(\tau_t) = \theta_{bg} q(\tau_t), \tag{11}$$

where τ_t is the theoretical threshold to divide the two classes. Equation 11 can be converted to the following quadratic equation through some appropriate operations:

$$A\tau_t^2 + B\tau_t + C = 0, \tag{12}$$

where $A = \sigma_{bg}^2 - \sigma_{fg}^2$, $B = 2(\mu_{bg}\sigma_{fg}^2 - \mu_{fg}\sigma_{bg}^2)$, $C = \mu_{fg}^2\sigma_{bg}^2 - \mu_{bg}^2\sigma_{fg}^2 + 2\sigma_{fg}^2\sigma_{bg}^2 \ln(\sigma_{fg}\theta_{bg}/\sigma_{bg}\theta_{fg})$.

Due to the binary classification of foregrounds and backgrounds, we simply let $\sigma_{fg} = \sigma_{bg} = \sigma$, i.e., the foregrounds and backgrounds have the same variance. Then the theoretical threshold is defined as:

$$\tau_t = \tau_e + \tau_a, \tag{13}$$

where the first term $\tau_e = (\mu_{fg} + \mu_{bg})/2$ is seen as an empirical threshold and the second term $\tau_a = (\sigma^2/\Delta)\ln(\theta_{bg}/\theta_{fg})$ can be seen as an adjustment. Here $\Delta = \mu_{fg} - \mu_{bg}$ and σ^2/Δ is approximately independent from τ_e.

Let the mean value of τ_e for all images be τ_0. When $\theta_{bg} > \theta_{fg}$, τ_a is positive (i.e., $\tau_e < \tau_t$) and $\tau_0 - \tau_e > 0$ due to more background pixels. When $\theta_{bg} < \theta_{fg}$, τ_a is negative (i.e., $\tau_e > \tau_t$) and $\tau_0 - \tau_e < 0$ due to more foreground pixels. It can concluded that $\tau_0 - \tau_e$ and τ_a have the same sign. Thus we utilize $\lambda(\tau_0 - \tau_e)$ to approach τ_a and the theoretical threshold is realized by:

$$\tau = \tau_e + \lambda(\tau_0 - \tau_e), \tag{14}$$

where λ is a hyperparameter to control the adjustment.

4 Experiments

4.1 Experimental Settings

Datasets. We test the performance of our proposed methods on two challenging datasets: CUB-200-2011 [Wah *et al.*, 2011] and ILSVRC [Russakovsky *et al.*, 2015]. CUB-200-2011 is a small fine-grained dataset of 200 classes, with 5, 994 images used for training and 5, 794 images used for testing. For ILSVRC, we choose the subset of 1, 000 classes containing about 1.2 million training images and 50, 000 validation images. We train the model using only the image-level labels, and the bounding box labels are only used to evaluate the localization performance of the model.

Evaluation Metrics. Following some representative methods [Russakovsky *et al.*, 2015; Zhou *et al.*, 2016], we adopt Top-1/Top-5 classification accuracy (Top-1/Top-5 Cls.) for classification and adopt Top-1/Top-5 localization accuracy (Top-1/Top-5 Loc.), GT-known localization accuracy (GT-known Loc.) for localization. Specifically, Top-1/Top-5 Cls.

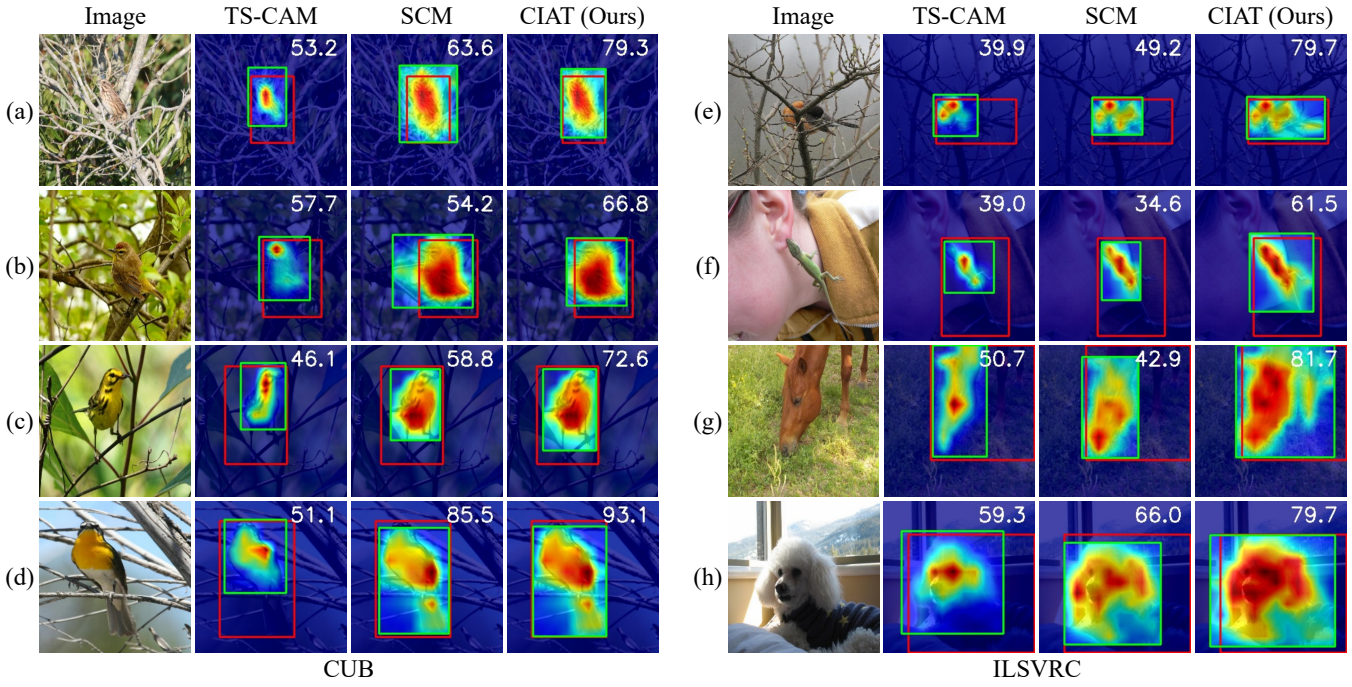


Figure 4: Visual localization results comparing with different methods on CUB-200-2011 and ILSVRC datasets. Note that the ground-truth bounding boxes are in red, the predictions are in green, and IoU values (%) are shown in white text.

is right, which means that the Top-1/Top-5 predicted category contains the correct image category label. GT-known Loc. is considered correct if only the predicted bounding boxes have over 50% IoU with at least one of the ground-truth boxes. Top-1/Top-5 Loc. is correct only when Top-1/Top-5 Cls. and GT-known Loc. are both right.

Implementation Details. We adopt SCM [Bai *et al.*, 2022] as our baseline, which uses the Deit-S backbone [Touvron *et al.*, 2021] pretrained on ILSVRC [Russakovsky *et al.*, 2015]. During the training phase, each input image is resized to 256×256 pixels and randomly cropped to 224×224 pixels. Then we use one 3×3 convolution layer initialized following He's approach [He *et al.*, 2015] instead of MLP head for classification. We choose AdamW [Loshchilov and Hutter, 2019] with $\epsilon = 1e-8$, $\beta_1 = 0.9$, $\beta_2 = 0.99$ and weight decay of 5e-4 to train the model. On CUB-200-2011, we train the model with a learning rate of 5e-5 and a batch size of 256 for 60 epochs. On ILSVRC, the training process lasts 40 epochs with a learning rate of 1e-6 and a batch size of 512.

4.2 Comparison with the State-of-the-Arts

Comparison on ILSVRC. To demonstrate the effectiveness of our methods, we first compare the performance with some state-of-the-art approaches on the more challenging million-level ILSVRC dataset. As reported in Table 1 and Table 2, CIAT achieves both the highest localization and classification accuracy on each evaluation metric, and further highlights the potential of transformers for WSOL. Notably, we surpass recent transformer-based works like CATR [Chen *et al.*, 2023] by large margins of 2.9%, 2.8%, and 1.3% on Top-1 Loc, GT-k. Loc, and Top-1 Cls, respectively.

Methods (Yr)	Backbone	Loc. Acc		
		Top-1	Top-5	GT-k.
CAM ('16)	VGG16	44.2	52.2	56.0
SPG ('18)	VGG16	48.9	57.2	58.9
ACoL ('18)	VGG16	45.9	56.5	59.3
DANet ('19)	VGG16	52.5	62.0	67.7
ADL ('19)	VGG16	52.4	-	75.4
MEIL ('20)	VGG16	57.5	-	73.8
GC-Net ('20)	VGG16	63.2	-	81.1
ORNet ('21)	VGG16	67.7	80.8	86.2
SLT-Net ('21)	VGG16	67.8	-	87.6
BAS ('22)	VGG16	71.3	85.3	91.0
TS-CAM ('21)	Deit-S	71.3	83.8	87.7
LCTR ('21)	Deit-S	79.2	89.9	92.4
SCM ('22)	Deit-S	76.4	91.6	96.6
LCAR ('23)	Deit-S	77.4	-	95.9
TAFormer ('23)	Deit-S	74.9	87.3	91.9
CATR ('23)	Deit-S	**79.6**	92.1	94.9
CIAT (Ours)	Deit-S	77.9	**92.2**	**97.1**

Table 3: Localization accuracy on the CUB-200-2011 test set compared to state-of-the-art studies.

Comparison on CUB-200-2011. Then, we compare the performance with state-of-the-art methods on CUB-200-2011 dataset. As shown in Table 3, CIAT achieves a remarkable performance of 97.1% on GT-k. Loc and Top-5 Loc of 92.2%, outperforming all transformer-based and CNN-based methods. Moreover, we achieve comparable results with state-of-the-art studies on Top-1 Loc, with a slightly lower result than LCTR [Chen *et al.*, 2022] and CATR [Chen *et al.*, 2023].

Baseline	CISD	OAT	Loc. Acc		
			Top-1	Top-5	GT-k.
✓			56.3	65.6	68.0
✓	✓		59.4	69.5	71.6
✓	✓	✓	**59.8**	**69.9**	**72.1**

Table 4: Ablation results on ILSVRC validation set when applying different configurations.

Module	Dataset	l-th	Loc. Acc		
			Top-1	Top-5	GT-k.
CISD (Ours)	ILSVRC	0	**59.4**	**69.5**	**71.6**
		1	59.3	**69.5**	**71.6**
		2	59.1	69.2	71.4
		6	58.4	68.2	70.7
		7	58.2	68.1	70.6

Table 5: Ablation results in the CISD when extracting shallow features from l-th block on ILSVRC validation set.

Methods	Resolution	FTM	OAT	GT-known
TS-CAM*	14×14	✓		67.3
			✓	68.2 (+0.9)
SCM	14×14	✓		68.8
			✓	69.5 (+0.7)
CIAT (Ours)	14×14	✓		71.6
			✓	**72.1** (+0.5)

Table 6: Comparison of different threshold segmentation methods based on some works of WSOL. Note that * indicates the re-implemented method.

Visualization Comparison. To further demonstrate the effectiveness of CIAT, we visualize the localization results on two datasets in Figure 4. Compared to TS-CAM and SCM, our approach succeeds in exploring more complete range of object regions containing both the weak and high response regions. For instance, in Figure 4(e) and Figure 4(f), our methods force the model to aggregate more detailed features, such as the tails of these animals, while others fail to capture them. Similarly, in Figure 4(a) and Figure 4(b), our methods contribute to filtering out the background noise in complex environment, leading to more accurate localization results.

4.3 Ablation Studies

First, we show the accuracy results with different configurations in Table 4. It is shown that CISD greatly increases the localization performance on the ILSVRC validation set, compared to the baseline method. The results indicate that keeping the semantic-spatial information consistency between shallow features and deep features is important for localization. During the post-processing time, OAT is used to obtain the closer prediction bounding boxes, which further brings an improvement for localization. We get the best results when applying both CISD and OAT.

Then, we explore the design details of each method to maximize their performance in a series of experiments. Table 5 shows results when extracting features from different blocks. Notably, we observe a downward trend on these metrics when using deeper features. These results indicate that CISD performs best when extracting shallow features from the 0-th block due to its more image details and spatial information. Moreover, to evaluate OAT's performance with other methods, we select TS-CAM [Gao *et al.*, 2021], SCM [Bai *et al.*, 2022], and CIAT to testify OAT. As shown in Table 6, we adopt OAT instead of the commonly used fixed threshold method (FTM) for each method and all get a 0.5% - 0.9% GT-known Loc improvement.

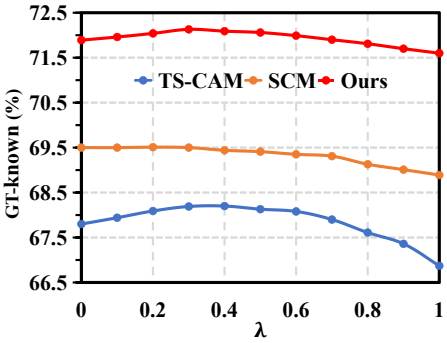


Figure 5: Performance analyses of hyperparameter λ in OAT.

Finally, we perform the sensitivity analysis of hyperparameters through extensive experiments. Our methods are simple with fewer hyperparameters and there is no need to design additional loss functions and balance each loss. Figure 5 further shows the effect of the hyperparameter λ, which is used to control the adjustment. For the same λ, our methods get much higher localization results than TS-CAM and SCM. For a method, a larger or smaller λ may lead to a decline in the results and the performance is worst when $\lambda = 1$.

5 Conclusion

In this paper, we propose the Consistency and Integration model with Adaptive Thresholds (CIAT), a novel approach for weakly supervised object localization. CIAT consists of two components, which improve the object activation and the foreground segmentation, respectively. The first component is a consistency and integration module of shallow-deep features (CISD), which significantly brings an improvement for object localization by maintaining semantic-spatial information consistency between shallow features and deep features. The other component is an online adaptive threshold (OAT) that utilizes prior information of each localization map to compute a robust threshold for a more accurate prediction. Extensive experiments demonstrate that our proposed methods effectively improve the performance of representative transformer-based methods for WSOL.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62176271) and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515011692).

References

[Bai *et al.*, 2022] Haotian Bai, Ruimao Zhang, Jiong Wang, and Xiang Wan. Weakly supervised object localization via transformer with implicit spatial calibration. In *Proceedings of the European Conference on Computer Vision*, pages 612–628. Springer, 2022.

[Bochkovskiy *et al.*, 2020] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[Chen *et al.*, 2022] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 410–418, 2022.

[Chen *et al.*, 2023] Zhiwei Chen, Jinren Ding, Liujuan Cao, Yunhang Shen, Shengchuan Zhang, Guannan Jiang, and Rongrong Ji. Category-aware allocation transformer for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6643–6652, 2023.

[Choe and Shim, 2019] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019.

[Dosovitskiy *et al.*, 2020] 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. *arXiv preprint arXiv:2010.11929*, 2020.

[Gao *et al.*, 2021] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2886–2895, 2021.

[Guo *et al.*, 2021] Guangyu Guo, Junwei Han, Fang Wan, and Dingwen Zhang. Strengthen learning tolerance for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7403–7412, 2021.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1026–1034, 2015.

[Kittler and Illingworth, 1986] Josef Kittler and John Illingworth. Minimum error thresholding. *Pattern Recognition*, 19(1):41–47, 1986.

[Kumar Singh and Jae Lee, 2017] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3524–3533, 2017.

[Lin *et al.*, 2014] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *International Conference on Learning Representations*, 2014.

[Liu *et al.*, 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

[Mai *et al.*, 2020] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8766–8775, 2020.

[Meng *et al.*, 2023] Meng Meng, Tianzhu Zhang, Zhe Zhang, Yongdong Zhang, and Feng Wu. Task-aware weakly supervised object localization with transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9109–9121, 2023.

[Pan *et al.*, 2023] Yixuan Pan, Yao Yao, Yichao Cao, Chongjin Chen, and Xiaobo Lu. Coarse2fine: Local consistency aware re-prediction for weakly supervised object localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2002–2010, 2023.

[Russakovsky *et al.*, 2015] 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. *International Journal of Computer Vision*, 115:211–252, 2015.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[Wang *et al.*, 2023] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023.

[Wu *et al.*, 2022] Pingyu Wu, Wei Zhai, and Yang Cao. Background activation suppression for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14228–14237, 2022.

[Xie *et al.*, 2021] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 132–141, 2021.

[Xie *et al.*, 2022] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–998, 2022.

[Xue *et al.*, 2019] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6589–6598, 2019.

[Zhang *et al.*, 2018a] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018.

[Zhang *et al.*, 2018b] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European Conference on Computer Vision*, pages 597–613, 2018.

[Zhang *et al.*, 2020] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *Proceedings of the European Conference on Computer Vision*, pages 271–287. Springer, 2020.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.