CLIP-Driven Transformer for Weakly Supervised Object Localization

Zhiwei Chen, Yunhang Shen, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji, *Senior Member, IEEE*

***Abstract*—Weakly supervised object localization (WSOL) aims to localize objects using only image-level labels as supervision. Despite recent advancements incorporating transformers into WSOL have resulted in improvements, these methods often rely on category-agnostic attention maps, leading to suboptimal object localization. This paper presents a novel CLIP-Driven TRansformer (CDTR) that learns category-aware representations for accurate object localization. Specifically, we initially propose a Category-aware Stimulation Module (CSM) that embeds learnable category biases into self-attention maps, enhancing the learning process with auxiliary supervision. Additionally, an Object Constraint Module (OCM) is designed to refine object regions in a self-supervised manner, leveraging the discriminative potential of the self-attention maps provided by CSM. To create a synergistic connection between CSM and OCM, we further develop a Semantic Kernel Integrator (SKI), which generates a semantic kernel for self-attention maps. Meanwhile, we explore the CLIP model and design a Semantic Boost Adapter (SBA) to enrich object representations by integrating semantic-specific image and text representations into self-attention maps. Extensive experimental evaluations on benchmark datasets, such as `CUB-200-2011` and `ILSVRC` highlight the superior performance of our CDTR framework.**

***Index Terms*—Weakly supervised learning, object localization, vision transformer.**

I. Introduction

WEAKLY supervised learning utilizes minimal or coarse supervision during training, providing a cost-effective alternative to fully supervised methods. Specifically, weakly supervised object localization (WSOL) concentrates on localizing objects using solely image-level annotations. This task is increasingly recognized in diverse applications, attributed to its capacity to circumvent the need for labor-intensive bounding box annotations [1], [2], [3], [4], [5], [6]. Furthermore, it is applicable to various downstream tasks, including weakly supervised object detection (WSOD) [1], [7], [8], [9], [10], weakly supervised semantic segmentation (WSSS) [11], [12], [13], [14], weakly supervised instance segmentation (WSIS) [15], [15], [16], [17].

Traditional WSOL techniques [18], [19], [20], [21], [22], [23], [24], [25] primarily utilize Class Activation Maps (CAM) [26] to identify discriminative image regions for object localization, relying solely on image-level labels. However, these methods tend to underestimate the actual object regions, frequently leading to bounding boxes significantly smaller than the true object extent. To address this limitation, several approaches have been developed to augment CAM, including graph propagation [27], data augmentation [28], [29], adversarial erasing [22], [23], [30], and spatial relation activation [21], [25], [31]. Despite their effectiveness, these techniques remain limited in comprehensively capturing the entire extent of objects, attributable to the intrinsic limitations of convolutional neural networks (CNNs) [32], [33] in global feature relation exploration.

The emergence of visual transformers [34] represents a substantial paradigm shift in computer vision, characterized by their capability to match CNNs in feature extraction through purely transformer-based architectures. Pioneering studies such as those by Gao et al. [3] initiated the fusion of semantic-aware tokens and semantic-agnostic attention maps for localization purposes. Subsequent research by Chen et al. [33] and Bai et al. [5] investigated the use of transformers for object localization, concentrating on local-continuous visual patterns and semantic similarities, respectively. The feature maps for a specific category are derived from the ground-truth image-level class label. However, the attention maps generated from class tokens capture long-range feature dependencies in a category-agnostic manner [3], which means it does not differentiate between object classes (i.e., not foreground-focused). Hence, this lack of distinction introduces category-agnostic noise into the localization maps, which in turn affects the accuracy and precision of the bounding boxes generated for object localization.

In recent years, the CLIP model [35] has gained significant attention due to its impressive performance in various vision-language tasks. Trained on a massive dataset of image-text pairs, CLIP leverages pseudo-supervision to achieve high accuracy in tasks such as segmentation [36], [37]. However, this reliance on a large-scale dataset poses limitations. Specifically, in scenarios

Received 25 January 2024; revised 31 December 2024; accepted 22 February 2025. Date of publication 14 March 2025; date of current version 7 May 2025. This work was supported in part by the National Science Fund for Distinguished Young Scholars under Grant 62025603, in part by the National Natural Science Foundation of China under Grant U21B2037, Grant U22B2051, Grant U23A20383, Grant 62176222, Grant 62176223, Grant 62176226, Grant 62072386, Grant 62072387, Grant 62072389, Grant 62002305, and Grant 62272401, and in part by the Natural Science Foundation of Fujian Province of China under Grant 2021J06003 and Grant 2022J06001. Recommended for acceptance by J. Verbeek. *(Corresponding author: Liujuan Cao.)*

Zhiwei Chen, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji are with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen 361005, China (e-mail: zhiweichen.cn@gmail.com; caoliujuan@xmu.edu.cn; zsc_2016@xmu.edu.cn; rrji@xmu.edu.cn).

Yunhang Shen is with Youtu Laboratory, Tencent, Shanghai 200233, China (e-mail: odysseyshen@tencent.com).

The code and models for this study are available at github.com/zhiweichen0012/CDTR.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2025.3548704, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2025.3548704

0162-8828 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

where such extensive datasets are unavailable or where CLIP cannot be utilized, the applicability of this approach diminishes.

Drawing on insights from the above analysis, this work introduces the **C**LIP-**D**riven **TR**ansformer (CDTR), a meticulously designed framework that leverages category information to enhance transformer-based attention in weakly supervised object localization. As illustrated in Fig. 1, our approach focuses on generating category-aware attention maps, which can effectively learn the discriminative representation of specific object classes. In particular, a Category-aware Stimulation Module (CSM) is introduced into the transformer attention mechanism, inducing a learning bias that associates self-attention maps with specific categories. CSM serves as auxiliary supervision, guiding the learning of more effective transformer representations and establishing a robust one-to-one relationship between self-attention maps and their corresponding classes. Complementing this module, an Object Constraint Module (OCM) is devised to precisely refine the object regions within the category-aware attention maps, utilizing a self-supervised approach. OCM plays a crucial role in reducing background distractions and generating pixel-level pseudo labels from self-attention maps, thereby accurately pinpointing the precise object regions. Additionally, a Semantic Kernel Integrator (SKI) is proposed to supplement and reinforce the image-level supervision. SKI generates a semantic kernel for self-attention maps, facilitating a synergistic connection between CSM and OCM throughout the training phase. Furthermore, we design a Semantic Boost Adapter (SBA) to fine-tune the WSOL model by infusing more comprehensive semantic-specific image and text representations into the attention maps. SBA is specifically developed based on the Contrastive Language-Image Pre-training (CLIP) model [35], infusing more comprehensive category-specific information into the attention maps. Finally, an innovative automatic weighted loss mechanism [38] is incorporated to dynamically adjust the loss weights during the training phase, thereby enhancing the model's performance. To validate the effectiveness of the proposed CDTR, comprehensive experiments are conducted on challenging WSOL benchmarks. The contributions of this work can be summarized as follows:

1) We introduce a CLIP-Driven TRansformer (CDTR), a novel framework for weakly supervised object localization, which significantly enhances category awareness in self-attention maps and deepens the understanding of long-range feature dependencies.
2) A Category-aware Stimulation Module (CSM) is innovatively designed to embed category-specific insights into self-attention maps across transformer blocks.
3) An Object Constraint Module (OCM) is introduced to refine the object regions within the category-aware attention maps in a self-supervised manner.
4) A Semantic Kernel Integrator (SKI) is proposed to synergistically connect CSM and OCM, enhancing category-aware learning within the CDTR framework.
5) A Semantic Boost Adapter (SBA) is developed to introduce stronger priors, leveraging the CLIP model to further augment the semantic perception capabilities of the CDTR framework.

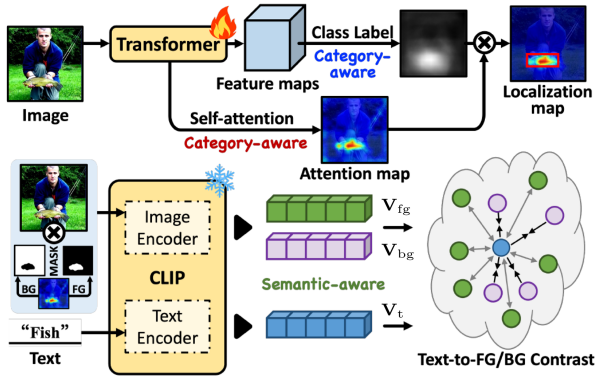


Fig. 1. *Illustration of the proposed CDTR.* Category awareness is bolstered by integrating insights from semantic prior information into self-attention maps. Meanwhile, we transfer the semantic knowledge of the CLIP model from the image level to the pixel level in self-attention maps during the training phase. Lastly, the generated category-aware attention maps are coupled with category-aware feature maps to yield more accurate and comprehensive localization maps. This approach significantly reduces irrelevant noise and enhances the accuracy of object localization. Predicted bounding boxes are depicted in red.

6) The proposed CDTR achieves significant and consistent performance improvements on `CUB-200-2011` and `ILSVRC` with 81.33% and 58.20% Top-1 localization accuracy, respectively.

This paper represents an expanded and more comprehensive version of our conference publication [39], with several significant enhancements and novel contributions: **1)** An in-depth analysis of the category-agnostic nature of transformer-based WSOL methods is presented, providing a detailed and visually enhanced examination, as elaborated in Section III-A. **2)** The introduction of a Semantic Kernel Integrator (SKI) as a novel module effectively bridges the CSM and OCM, enriching category-aware learning within the CDTR framework, as detailed in Section III-E. **3)** We introduce a Semantic Boost Adapter (SBA), leveraging the CLIP model to refine semantic processes in category-aware object localization, discussed in Section III-F. **4)** Substantial improvements have been implemented in the paper's presentation, encompassing enhanced motivation, more illustrative diagrams, refined formulations, comprehensive experimental analysis, and key results. Furthermore, various sections have been restructured to enhance readability and offer more detailed insights into the motivations, quantitative and qualitative comparisons, and discussions.

The remainder of this paper is structured as follows: Section II provides a review of existing works pertinent to WSOL and the CLIP model. The detailed methodology is elaborated in Section III. Experimental results validating the proposed CDTR are detailed in Section IV. Discussions on the limitations and future research directions are provided in Section V. The paper concludes with a summary of the findings in Section VI.

II. Related Work

A. Weakly Supervised Object Localization

CNN-based Methods for WSOL: The cornerstone of CNN-based WSOL methods is the Class Activation Maps

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

(CAM) [26] technique. This approach generates object bounding boxes from class activation maps derived from classification networks. However, CAM often focuses narrowly on the most discriminative object parts, neglecting less conspicuous yet crucial regions. To overcome this, a plethora of methods have emerged. Image and feature augmentation strategies, for instance, have been employed to emphasize non-discriminative object parts. Singh et al. [29] experimented with hiding random image patches during training, while Yun et al. [28] introduced the innovative cut-and-paste technique among training images. Additionally, Zhang et al. [20] and Mai et al. [22] mined different discriminative regions by two adversary classifiers. Junsk et al. [30] erased discriminative spatial positions on the feature map to capture the integral extent of the object. Chen et al. [23] 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. [21] utilized a discrepant activation method to learn complementary and discriminative visual patterns. Zhang et al. [31] adopted constraints to prompt the consistency of object features within the same categories. Pan et al. [24] leveraged structural information incorporated into convolutional features to distill the structure-preserving ability of features. Apart from the above works, Xie et al. [40] proposed a new paradigm that learns a foreground prediction map to achieve localization. Wu et al. [41] proposed a background activation suppression strategy to learn foreground prediction maps for object localization. Wang et al. [19] explored the intrinsic discrimination and consistency in the image classification task pipeline. Zhu et al. [42] 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. [43], Guo et al. [25] and Wei et al. [44], divided WSOL into two independent sub-tasks, 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.

The majority of the methods mentioned utilize convolutional neural networks, which inherently lack the ability to capture global information, rendering them prone to focusing on local discriminative object regions [3], [32]. To address this limitation, a transformer-based method is proposed for WSOL.

Transformer-based Methods for WSOL: Recent advancements in computer vision have shown the great potential of transformers for various tasks [45], [46], [47], [48], [49], [50], [51], [52], as they excel in capturing long-range dependencies and perform better than convolutional neural networks (CNNs). Dosovitskiy et al. [46] 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. [3] 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. [33] highlighted local details of global representations using learnable kernels and cross-patch information guided by the class-token attention map. Gupta et al. [53] improved localization maps by incorporating a patch-based attention dropout layer into the transformer attention blocks. Bai et al. [5] considered the semantic similarities of patch tokens and their spatial relationships for WSOL.

Nevertheless, transformer-based approaches frequently employ category-agnostic attention maps, which results in the incorporation of background noise into object localization. This work endeavors to overcome this limitation by integrating category-aware information into these maps, consequently enhancing the specificity and accuracy of object localization.

B. Weakly Supervised Object Detection

Weakly supervised object detection (WSOD) seeks to simultaneously perform image classification and instance localization using training data annotated solely with image-level labels. Unlike WSOL, WSOD assumes each image contains at least one object of the specified category. Most approaches [54], [55], [56], [57], [58] employ multiple-instance learning (MIL) to recast WSOD as a multi-label classification problem.

For one-stage training methods, given the off-the-shelf proposals [59], [60], [61], the model selects high-scored instances from bags during training. Bilen et al. [62] first introduced a two-stream weakly supervised deep detection network, selecting positive samples by multiplying classification and detection scores. Henceforth, Tang et al. [63] combined WSDDN and multi-stage instance classifiers into an online instance classifier refinement framework to improve proposal classification ability. Tang et al. [57] further improved OICR with a robust proposal generation module based on proposal cluster learning. Zeng et al. [58] and Ren et al. [64] introduced bounding box regression into the WSOD network, where proposals with highest scores are selected as pseudo ground-truths to train the regression branch. Besides, Wu et al. [65] proposed to capture the occurrence of misclassification and mitigate its adverse effect on training. Wu et al. [66] sorted proposals by the score from high to low and chose the proposals preceding the one with the largest drop. For multi-stage training methods [67], [68], [69], an additional detector is trained using proposals with the highest scores as pseudo ground truths, based on outputs from the WSOD network of the previous stage. Algorithms like pseudo ground truth adaptation [70] and zigzag learning [71] have been proposed to refine these pseudo labels, critically impacting detector performance.

However, most WSOD methods rely on a large number of candidate proposals and high-resolution inputs, imposing substantial computational burdens [43], [72]. Consequently, they are challenging to apply to large-scale datasets for weakly supervised object localization.

C. Semi-Supervised Segmentation

Recent years have witnessed remarkable advancements in image-level semi-supervised learning, largely propelled by two predominant paradigms: self-training [73], [74] and consistency regularization [75], [76]. For semi-supervised segmentation, primary localization maps are generated through class

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

activation mapping (CAM) [26], which are then employed to train a segmentation network in conjunction with the available pixel-level annotations within a semi-supervised framework. For instance, Lai et al. [77] proposed a context-aware consistency mechanism to ensure alignment between features representing the same identity across diverse contexts, thereby bolstering the robustness of representations against environmental variability. Similarly, Ouali et al. [78] introduced cross-consistency training, enforcing invariance in predictions under diverse perturbations applied to encoder outputs. Sohn et al. [79] enhanced pseudo masks by generating synthetic labels through a synergistic application of consistency regularization and pseudo-labeling. In this work, we delve into the potential of semantic-aware consistency within the attention maps of transformers to advance the field of WSOL.

D. Few-Shot Learning

The dependency of deep learning models on extensive datasets, which are often unavailable in practical applications, has underscored the necessity of few-shot learning techniques. Few-shot learning aims to recognize novel classes using only a limited number of training samples. Traditional few-shot learning methods, such as fine-tuning [80], data and feature augmentation [81], and meta-learning [82], have been widely explored. More recently, vision-language pre-trained models [35], [83], [84], [85] have been utilized to transfer their robust representational capabilities into few-shot learning scenarios. For instance, Zhou et al. [86] improved image-text alignment through continuous prompt refinement. Qiu et al. [87] proposed a cross-modal module to effectively identify informative regions in images and aggregate visual features. Zhang et al. [88] introduced a training-free approach that directly configures adapter weights using a cache model, circumventing conventional SGD-based fine-tuning. Besides, Zhou et al. [84] extended the existing CoOp framework by incorporating a lightweight neural network that generates input-conditional tokens tailored to each image. In contrast to these methods, which primarily focus on image-text contrastive learning, our work leverages the pre-trained CLIP model to derive more discriminative visual representations for WSOL.

E. Contrastive Language-Image Pretraining

The advent of Contrastive Language-Image Pretraining (CLIP) [35] has marked a significant breakthrough in bridging the semantic gap between visual perception and linguistic representation. Pre-trained on an extensive collection of image-text pairs, CLIP has demonstrated an unparalleled capability to comprehend a broad spectrum of visual concepts through a linguistic lens. Central to the CLIP model is its unique architecture comprising an image encoder and a text encoder. These components collaboratively learn to generate corresponding embeddings, thereby measuring the congruence between visual and textual entities.

CLIP's versatile nature has seen its application across a diverse array of computer vision tasks [13], [36], [37], [89]. In particular, its ability to perform zero-shot learning and understand images in relation to textual descriptions has been leveraged in various contexts. Chen et al. [37] employed the pretrained CLIP vision encoder to pioneer open-vocabulary zero-shot semantic segmentation models. Similarly, Liang et al. [36] fine-tuned the CLIP model on masked image regions and their respective text descriptions, enhancing its applicability in open-vocabulary semantic segmentation. Lin et al. [13] demonstrated the model's remarkable proficiency in semantic segmentation, even with only image-level labels and without the need for additional training.

In this study, the focus is on harnessing semantic cues from the CLIP model to fine-tune the WSOL model. By leveraging the extensive, pre-trained semantic knowledge of CLIP, this research aims to tackle the category-agnostic challenge inherent in transformer-based methods for WSOL.

III. Methodology

This section begins with a critical analysis of the self-attention mechanism within the transformer, particularly focusing on why these attention maps tend to be category-agnostic. A comprehensive overview of the proposed CLIP-Driven TRansformer (CDTR) is then provided. Subsequently, the paper delves into the intricate details of the four novel and pivotal modules comprising the core of CDTR: Category-aware Stimulation Module (CSM), Object Constraint Module (OCM), Semantic Kernel Integrator (SKI), and Semantic Boost Adapter (SBA). Finally, the modules are incorporated with the transformer structure into a joint optimization framework, as illustrated in Fig. 3.

A. Qualitative Analysis and Findings

Following Dosovitskiy et al. [46], pure transformers have demonstrated exceptional performance in the image classification task. Consider an input image denoted as $I \in \mathbb{R}^{H\times W\times M}$, where H, W, and M represent its height, width, and the number of channels, respectively. The input sequence, $\mathbf{X} \in \mathbb{R}^{(N+1)\times D}$, is propagated through a transformer encoder consisting of L successive transformer blocks. Within the multi-head attention module of each transformer block, the attention matrix is denoted as $\mathbf{A}_{\text{am}} \in \mathbb{R}^{L\times S\times(N+1)\times(N+1)}$, where L and S signify the number of blocks and heads, respectively. The class-token attention vector is then extracted and reshaped from $\mathbf{A}_{\text{am}}$, yielding an attention map $\bar{\mathbf{A}} \in \mathbb{R}^{L\times S\times w\times h}$. However, this attention map acquires information from various representation subspaces at different positions in the input image, covering all regions of interest. This makes it difficult for the model to locate a specific object, as illustrated in Fig. 2. For example, in the top sample of the second column, it is revealed that both the *Fish* and *Person* are activated, despite the ground-truth label being assigned only to the *Fish*.

The analysis attributes the category-agnostic nature of these attention maps primarily to two types of noise: **1)** Base noise, which arises from the model's prior knowledge of currently irrelevant classes. For example, in an image labeled as *Stingray* (the top first-column sample), the attention map unexpectedly activates regions corresponding not only to *Stingray* but also to *Person*. This phenomenon indicates that the model's prior

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

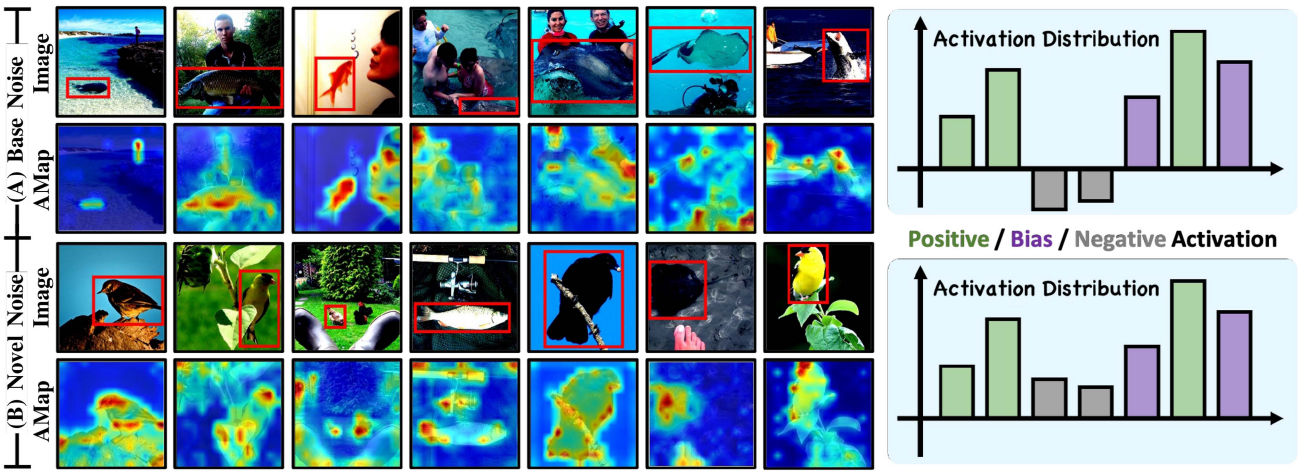


Fig. 2. *Exploration of attention maps (AMaps) within the transformer framework.* (a) Base Noise: Activations corresponding to prior learned but currently irrelevant class categories, leading to misaligned context activations. (b) Novel Noise: Activations corresponding to the regions beyond the network's prior knowledge, bringing in pure background clusters. The ground-truth bounding boxes are in Red.

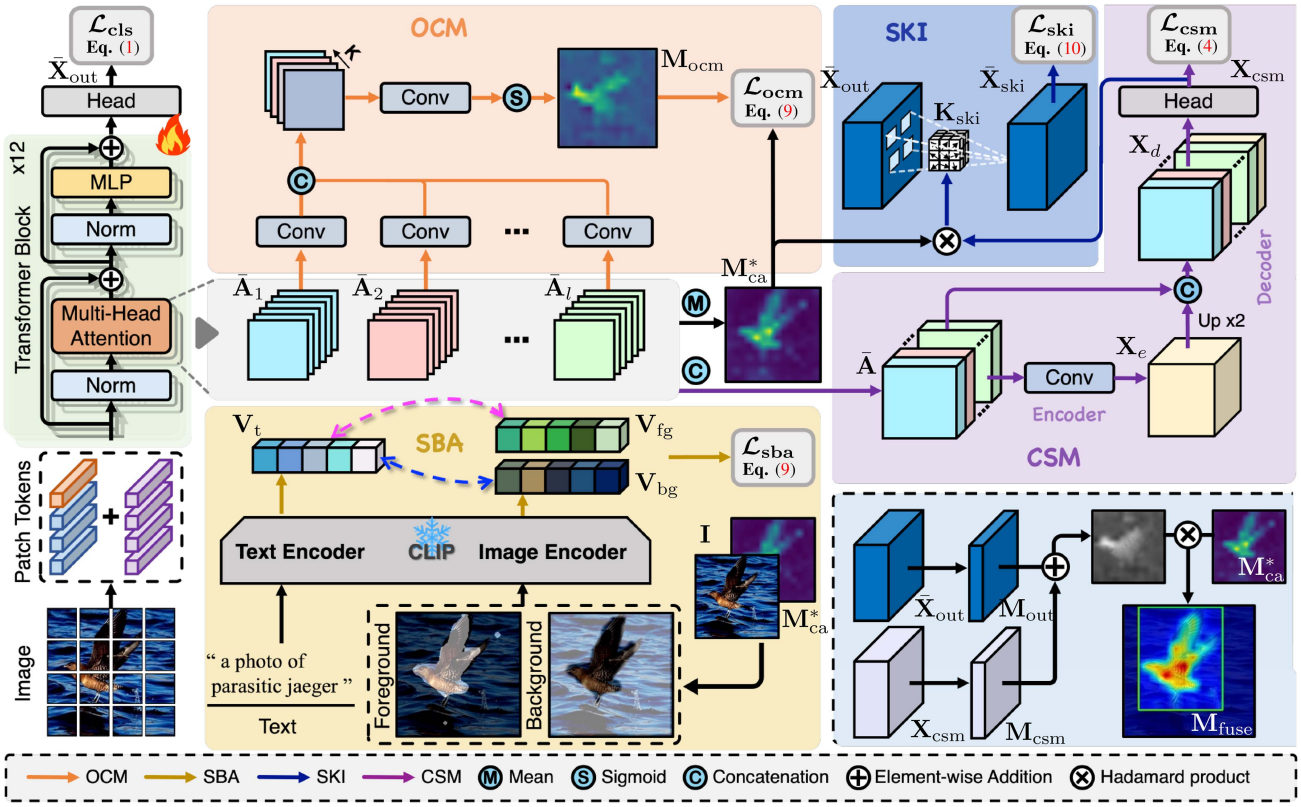


Fig. 3. *The architecture of the proposed CLIP-Driven TRansformer (CDTR).* It consists of a vision transformer backbone, a CLIP model, an object constraint module (OCM), a semantic boost adapter (SBA), a semantic kernel integrator (SKI), and a category-aware stimulation module (CSM). In the inference 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.

exposure to *Person* features inadvertently intrudes into the localization process for *Stingray*. That is, *Person* is a related category on the ILSVRC dataset. It highlights the challenge of segregating current relevant features from the prior knowledge of the model. **2)** Novel noise, which refers to background clusters that are completely outside the scope of the model's prior knowledge. These noise activations dilute the model's focus, leading it to allocate attention to irrelevant background clusters. For example, in images labeled as *Bird* (the bottom first-column sample), the attention map not only highlights the *Bird* but also extends to background clusters like *Wood*. These unlearned, novel background clusters introduce a scattered pattern of activation,

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

diluting the model's focus and erroneously directing attention to irrelevant background regions.

This paper aims to address these challenges by refining the attention mechanism to more accurately isolate and highlight features relevant to the target object. This refinement is essential for minimizing the influence of both base and novel noise in the object localization task. By making attention maps foreground-focused, the proposed approach facilitates more accurate, category-aware object localization in weakly supervised settings, thereby increasing the model's interpretative accuracy and relevance.

B. Framework Overview

This section provides a comprehensive overview of the proposed CLIP-driven transformer (CDTR), which mainly comprises four modules: a category-aware stimulation module (CSM) to infuse category awareness into self-attention maps, an object constraint module (OCM) to refine these attention maps in a self-supervised manner, a semantic kernel integrator (SKI) to connect CSM and OCM with a shift kernel, and a semantic boost adapter (SBA) to enrich object representations by leveraging the CLIP model.

Consider an input image $I \in \mathbb{R}^{H\times W\times M}$, where H, W, and M denote its height, width, and the number of channels, respectively. The image is first divided 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 prepended to the tokens, together with a position embedding $\mathbf{T}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$, forming the input sequence $\mathbf{X} \in \mathbb{R}^{(N+1)\times D}$, which is then fed into the transformer encoder with L consecutive transformer blocks.

To allocate semantic information, CSM is applied to all 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 added following $\mathbf{X}_{\text{csm}}$ to enhance category awareness. Subsequently, OCM is applied to the self-attention maps to refine object regions, leading to the derivation of the object constraint loss $\mathcal{L}_{\text{ocm}}$. This process entails a meticulous examination and adjustment of the attention maps to sharpen the focus on the relevant object areas, simultaneously suppressing background noise. Furthermore, a Semantic Kernel Integrator (SKI) is proposed to build a semantic kernel shift for the self-attention maps, resulting in a shift loss $\mathcal{L}_{\text{ski}}$. It connects CSM with OCM during the training phase, thus harmonizing the interplay between category-aware learning and object localization. Additionally, apart from the coarse-grained image-level label supervision, we design a Semantic Boost Adapter (SBA) to infuse image-text pair priors into the localization model, fine-tuning the category feature distribution of self-attention maps. We then obtain a semantic boost loss $\mathcal{L}_{\text{sba}}$.

Let $\mathbf{X}_{\text{out}} \in \mathbb{R}^{(N+1)\times D}$ represent the output feature map obtained from the transformer encoder. In accordance with the methodology in Gao et al. [3], we discard the class token from $\mathbf{X}_{\text{out}}$ and apply a convolutional layer to transform the feature map into $\bar{\mathbf{X}}_{\text{out}} \in \mathbb{R}^{C\times w\times h}$. Subsequently, a global average pooling layer is employed on $\bar{\mathbf{X}}_{\text{out}}$ to generate the class probability distribution $\hat{y}$ for classification prediction. With the corresponding image-level one-hot encoding label y, the classification loss function is defined as follows:

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

Note that C is the number of classes. During the training phase, our approach integrates an automatic weighted mechanism [38]. This mechanism is designed to dynamically balance the contributions of various modules, thereby optimizing the overall learning process. Specifically, OCM is employed on the self-attention maps with the objective of refining and delineating the object regions more accurately, yielding the object constraint loss $\mathcal{L}_{\text{ocm}}$, which comprises a noise suppression loss $\mathcal{L}^{\text{s}}_{\text{ocm}}$ and an object awakening loss $\mathcal{L}^{\text{a}}_{\text{ocm}}$. SKI then fosters a synergistic connection between CSM and OCM, resulting in a semantic kernel loss $\mathcal{L}_{\text{ski}}$. Moreover, SBA is utilized to infuse image-text pair priors into the localization model, fine-tuning the category feature distribution of self-attention maps. This process yields the semantic boost loss $\mathcal{L}_{\text{sba}}$. The overall training loss effectively integrates the distinct aspects of our model in a harmonious and balanced manner:

$$\begin{aligned}\mathcal{L} = {} & \frac{1}{2\lambda_1^2}\mathcal{L}_{\text{cls}} + \log\left(1+\lambda_1^2\right) + \frac{1}{2\lambda_2^2}\mathcal{L}_{\text{csm}} + \log\left(1+\lambda_2^2\right) \\ & + \frac{1}{2\lambda_3^2}\mathcal{L}^{\text{s}}_{\text{ocm}} + \log\left(1+\lambda_3^2\right) + \frac{1}{2\lambda_4^2}\mathcal{L}^{\text{a}}_{\text{ocm}} + \log\left(1+\lambda_4^2\right) \\ & + \frac{1}{2\lambda_5^2}\mathcal{L}_{\text{ski}} + \log\left(1+\lambda_5^2\right) + \alpha\mathcal{L}_{\text{sba}},\end{aligned} \tag{2}$$

where learnable parameters λ_1-λ_5 are initialized at 1, and α is set to 2.4 on `CUB-200-2011` [90] and 0.01 on `ILSVRC` [91].

During the testing phase, an object map $\mathbf{M}_{\text{out}} \in \mathbb{R}^{w\times h}$ is derived from the transformed feature map $\bar{\mathbf{X}}_{\text{out}}$, based on the class predicted by the model. This map delineates the spatial distribution of the predicted class within the image, thereby providing an initial localization of the object. Simultaneously, we also generate a secondary object map $\mathbf{M}_{\text{csm}} \in \mathbb{R}^{w\times h}$ from the category-aware feature map $\mathbf{X}_{\text{csm}}$. Originating from CSM, this map provides an additional perspective on the object's location, highlighting various aspects of the object's presence in the image. To achieve a comprehensive and precise localization map, the two object maps are fused with a category-aware attention map $\mathbf{M}^*_{\text{ca}}$, which is generated from the self-attention maps. The process can be formulated as follows:

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

where $\otimes$ denotes Hadamard product. Following the fusion, $\mathbf{M}_{\text{fuse}}$ is resized to match the dimensions of the original input image through linear interpolation. This resizing step is essential for aligning the localization map with the original image's scale, ensuring that the final object localization is accurately represented. Finally, the predicted box is obtained by computing the tightest bounding box that encompasses the largest connected area of foreground pixels in $\mathbf{M}_{\text{fuse}}$, following the approach used in prior works like Zhou et al. [26] and Gao et al. [3].

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

C. Category-Aware Stimulation Module

In order 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. As described in Section III-A, the attention map $\bar{\mathbf{A}}$, derived from the transformer's multi-head attention mechanism, aggregates information from varied representation subspaces across different image positions. Although this comprehensive aggregation is robust in feature capturing, it inadvertently encompasses all regions of interest within the image. This can be attributed to the absence of category supervision in $\bar{\mathbf{A}}$ during the training process. To address this issue, the Category-aware Stimulation Module (CSM) is introduced to optimize the attention map $\bar{\mathbf{A}}$.

Specifically, $\bar{\mathbf{A}}$ is transformed into $\mathbf{X}_A \in \mathbb{R}^{(LS)\times w\times h}$ by vector transformation, with L and S denoting the number of transformer blocks and heads, respectively. The transformed map, $\mathbf{X}_A$, functions as the input for our carefully designed encoder-decoder layer, which aims to extract more precise semantic information to distinguish various regions of interest. The encoder, including a max-pooling operation with a stride of 2 and two 3×3 convolutional layers, efficiently condenses the information from $\mathbf{X}_A$ into a downsampled feature map $\mathbf{X}_e \in \mathbb{R}^{G\times \frac{w}{2}\times \frac{h}{2}}$. The number of feature channels, indicated by G, represents a crucial parameter established to optimize the encoding process. The decoder processes this feature map, up-sampling it back to the original scale and subsequently concatenating it with the input feature map $\mathbf{X}_A$. This procedure guarantees the integration of detailed local features with broader contextual information. Subsequently, the concatenated feature map is processed through two additional 3×3 convolutional layers, resulting in the output feature map $\mathbf{X}_d$. In order to enhance category awareness within our model, an auxiliary classification head is attached to $\mathbf{X}_d$. This head, comprising a 1×1 convolutional layer followed by a global average pooling layer, is designed to produce 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 function for CSM is formulated as follows:

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

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

D. Object Constraint Module

CSM incorporates category information into the category-agnostic attention maps, thereby reducing the background interference of the localization maps. Then, we propose the Object Constraint Module (OCM), designed to further refine the attention maps for precise object localization. OCM capitalizes on the improved attention maps provided by CSM, leveraging their discriminative potential to generate pseudo labels. These pseudo labels serve in a self-supervised manner, directing the attention mechanism to concentrate more accurately on the object regions.

OCM comprises two mechanisms: noise suppression and object awakening. The noise suppression mechanism is specifically designed to restrict the influence of background clutter within the attention maps, effectively filtering out irrelevant spatial information. Conversely, the object awakening mechanism concentrates on enhancing the regions corresponding to the objects of interest. This dual approach ensures a balanced focus, simultaneously minimizing background interference while accentuating object features.

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 $\mathbf{M}^*_{\text{ca}} \in \mathbb{R}^{w\times h}$, which is formulated as:

$$\mathbf{M}^*_{\text{ca}} = \frac{1}{LS}\sum_{l}^{L}\sum_{s}^{S}\bar{\mathbf{A}}_{(l,s)}. \tag{5}$$

Note that L and S represent the number of transformer blocks and heads, respectively. Since long-range dependency is preserved in $\mathbf{M}^*_{\text{ca}}$, we treat values below the p-percentile as background clutters, and suppress them to zero. The loss function for noise suppression is defined as:

$$\bar{\mathbf{M}}^*_{\text{ca}(i,j)} = \begin{cases} 0, & \text{if } \mathbf{M}^*_{\text{ca}(i,j)} > \mathbf{\Phi}_p\left(\mathbf{M}^*_{\text{ca}}\right), \\ \mathbf{M}^*_{\text{ca}(i,j)}, & \text{otherwise}, \end{cases} \tag{6}$$

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

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

Object awakening mechanism: In the pursuit of identifying and learning optimal object regions, a critical step is the generation of reliable attention pseudo labels to serve as supervision. Initially, the attention map, denoted as $\bar{\mathbf{A}}$, is split along the block dimension to yield $\bar{\mathbf{A}}_l \in \mathbb{R}^{S\times w\times h}$ for each l-th block. Subsequently, a 1×1 convolutional layer is employed to transform this into a new feature map $\mathbf{X}_l \in \mathbb{R}^{1\times w\times h}$, which is designed to closely approximate the spatial distribution of the object within 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 $\mathbf{M}_{\text{ocm}} \in \mathbb{R}^{w\times h}$. This pseudo map plays a pivotal role in supervising and refining the category-aware attention map throughout the training phase. The activation loss is formulated as:

$$\mathcal{L}^{\text{a}}_{\text{ocm}} = \frac{1}{hw}\sum_{i}^{h}\sum_{j}^{w}\left(\mathbf{M}_{\text{ocm}(i,j)} - \mathbf{M}^*_{\text{ca}(i,j)}\right)^2. \tag{8}$$

To optimize the training of OCM, we formulate a joint objective function incorporating two distinct loss terms: the noise suppression loss $\mathcal{L}^{\text{s}}_{\text{ocm}}$ and the object awakening loss $\mathcal{L}^{\text{a}}_{\text{ocm}}$. The overall loss for OCM is defined as follows:

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

$$\mathcal{L}_{\text{ocm}} = \frac{1}{2\lambda_3^2}\mathcal{L}_{\text{ocm}}^{\text{s}} + \log\left(1+\lambda_3^2\right) + \frac{1}{2\lambda_4^2}\mathcal{L}_{\text{ocm}}^{\text{a}} + \log\left(1+\lambda_4^2\right), \tag{9}$$

where λ_3 and λ_4 are learnable parameters.

E. Semantic Kernel Integrator

With CSM and OCM, the self-attention maps are trained under the guidance of category information. However, it is observed that the two modules are independent, which limits the optimization of the model in the training phase. To address this issue, we introduce the Semantic Kernel Integrator (SKI), a novel component designed to establish a synergistic link between CSM and OCM during the training process.

The operational principle of SKI centers around adapting to the spatial variations within the feature maps produced by CSM and OCM. Consider the feature map $\mathbf{X}_{\text{csm}} \in \mathbb{R}^{C\times w\times h}$, derived from CSM, and the map $\mathbf{M}_{\text{ca}}^{*} \in \mathbb{R}^{w\times h}$, obtained from OCM. First, $\mathbf{M}_{\text{ca}}^{*}$ is applied to $\mathbf{X}_{\text{csm}}$ via Hadamard product, effectively merging the semantic and attentional aspects within CSM and OCM. By performing this multiplication, we integrate the category-specific insights from CSM with the object-centric attention from OCM. Subsequently, a 3×3 convolutional layer is applied to this combined feature map to produce semantic offsets. Based on these offsets, we design a semantic kernel $\mathbf{K}_{\text{ski}}$, which is then used in an additional convolutional layer on the feature map $\bar{\mathbf{X}}_{\text{out}}$. This process allows the model to recognize semantic-aware spatial patterns and variations of the objects. Consequently, a new feature map $\mathbf{X}_{\text{ski}} \in \mathbb{R}^{C\times w\times h}$ is obtained, where C signifies the number of classes. To establish a link to category awareness, a global average pooling layer is utilized, culminating in the class probability distribution $\hat{y}$. The loss function for SKI is formulated as follows:

$$\mathcal{L}_{\text{ski}} = -\sum_{i}^{C} y_i \log\left(\frac{e^{\hat{y}_i}}{\sum_j^C e^{\hat{y}_j}}\right). \tag{10}$$

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

F. Semantic Boost Adapter

Due to the difficulty in learning object semantic information solely from coarse-grained image-level label supervision, we propose a Semantic Boost Adapter (SBA) to augment the semantic perception of objects under the guidance of natural language supervision.

Consider $\mathcal{F}_t$ and $\mathcal{F}_i$ as the text and image encoders, respectively, in the frozen contrastive language-image pretraining (CLIP) model [35]. As depicted in Fig. 3, the input image I is multiplied by the category-aware attention map $\mathbf{M}_{\text{ca}}^{*}$ and its complement $(1-\mathbf{M}_{\text{ca}}^{*})$ to isolate foreground objects and background pixels. These modified images are then processed through the image encoder, resulting in the extraction of image representation vectors $\mathbf{v}_f$ (foreground) and $\mathbf{v}_b$ (background), defined as:

$$\mathbf{v}_f = \mathcal{F}_i(I\cdot\mathbf{M}_{\text{ca}}^{*}), \quad \mathbf{v}_b = \mathcal{F}_i\left(I\cdot(1-\mathbf{M}_{\text{ca}}^{*})\right). \tag{11}$$

Following CLIP [35], the text prompt $\mathbf{t}$ for image I is structured as "a photo of {}", e.g., "a photo of Stingray". Subsequently, the text representations are obtained as follows:

$$\mathbf{v}_t = \mathcal{F}_t(\mathbf{t}). \tag{12}$$

To enhance the precision in identifying activated object regions, the object-text pair-matching loss $\mathcal{L}_{\text{sba}}$ is introduced. This loss function is designed to facilitate the learning of more distinct foreground and background patterns, calculated as:

$$\mathcal{L}_{\text{sba}} = -\log\left(\frac{\mathbf{s}_f}{\mathbf{s}_f+\mathbf{s}_b+\xi}\right), \tag{13}$$

$$\mathbf{s}_f = \texttt{sim}(\mathbf{v}_t,\mathbf{v}_f), \quad \mathbf{s}_b = \texttt{sim}(\mathbf{v}_t,\mathbf{v}_b), \tag{14}$$

where $\texttt{sim}(\cdot)$ denotes the cosine similarity function. The value of ξ is very small, ensuring that the equation makes sense. By minimizing $\mathcal{L}_{\text{sba}}$, our approach achieves a more effective activation of target object regions in $\mathbf{M}_{\text{ca}}^{*}$. With the alignment of text and image pairs in CLIP, this attention map learns category awareness through text features, which helps in capturing the whole object. Additionally, SBA pushes $\mathbf{v}_b$ away from $\mathbf{v}_t$, reducing the effect of background noise. Consequently, this map acquires category awareness from text feature integration, thereby effectively reducing the presence of irrelevant background clusters.

IV. Experiments

A. Experimental Settings

Datasets: We evaluate the effectiveness of our proposed method on two widely-used WSOL (Weakly Supervised Object Localization) benchmarks: `CUB-200-2011` [90], `ILSVRC` [91] and `OpenImages` [18]. The `CUB-200-2011` dataset, focusing on avian species, encompasses a collection of 200 distinct bird categories, comprising 5,994 images for training and 5,794 for testing. The `OpenImages` dataset contains 100 categories, including 29,819 and 5,000 images in training and test set, respectively. In a more expansive scope, the `ILSVRC` dataset (i.e., the subset of the ImageNet dataset [97]) includes an extensive array of 1,000 classes, featuring a substantial corpus of 1,281,197 training images and 50,000 images in its validation set. We train our model using only the training set and evaluate it on the validation set for `ILSVRC` (the test set for `CUB-200-2011`), where the bounding box annotations are solely used for evaluation purposes.

Evaluation Metrics: Following previous methods [26], [33], [41], 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

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

TABLE I
LOCALIZATION COMPARISON WITH STATE-OF-THE-ART METHODS

Methods	Venue	Backbone	CUB-200-2011 Loc Acc.			ILSVRC Loc Acc.		
			Top-1	Top-5	GT-known	Top-1	Top-5	GT-known
CAM [26]	CVPR16	VGG16	41.06	50.66	55.10	42.80	54.86	59.00
ACoL [20]	CVPR18	VGG16	45.92	56.51	62.96	45.83	59.43	62.96
ADL [92]	TPAMI20	VGG16	52.36	–	75.41	44.92	–	–
I2C [31]	ECCV20	VGG16	55.99	68.34	–	47.41	58.51	63.90
MEIL [22]	CVPR20	VGG16	57.46	–	73.84	46.81	–	–
SPA [24]	CVPR21	VGG16	60.27	72.50	77.29	49.56	61.32	65.05
FAM [93]	ICCV21	VGG16	69.26	–	89.26	51.96	–	–
Kim et al. [4]	CVPR22	VGG16	70.83	88.07	93.17	49.94	63.25	68.92
TAFormer [94]	TPAMI23	VGG16	72.02	85.94	90.84	53.42	67.73	**74.02**
CAM [26]	CVPR16	InceptionV3	41.06	50.66	55.10	46.29	58.19	62.68
I2C [31]	ECCV20	InceptionV3	55.99	68.34	72.60	53.11	64.13	68.50
GCNet [95]	ECCV20	InceptionV3	58.58	71.00	75.30	49.06	58.09	–
SPA [24]	CVPR21	InceptionV3	53.59	66.50	72.14	52.73	64.27	68.33
FAM [93]	ICCV21	InceptionV3	70.67	–	87.25	55.24	–	68.62
TAFormer [94]	TPAMI23	InceptionV3	73.32	84.08	88.66	56.00	66.49	69.69
TS-CAM [3]	ICCV21	Deit-S	71.30	83.80	87.70	53.40	64.30	67.60
LCTR [33]	AAAI22	Deit-S	79.20	89.90	92.40	56.10	65.80	68.70
SCM [5]	ECCV22	Deit-S	76.40	91.60	96.60	56.10	66.40	68.80
LCAR [96]	AAAI23	Deit-S	77.40	–	95.90	57.10	–	70.70
CATR [39]	ICCV23	Deit-S	79.62	92.08	94.94	56.90	66.64	69.25
CDTR (ours)	This Work	Deit-S	**81.33**	**94.06**	**96.89**	**58.20**	**68.05**	70.72

The best results are highlighted in bold, second are underlined.

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. For the mask, we use the pixel average precision (PxAP) [18].

Implementation Details: We construct our proposed CDTR using the Deit-S backbone [98] pre-trained on ILSVRC [91] and adopt TS-CAM [3] as our baseline method. Specifically, we replace the MLP head with a convolutional layer and append a global average pooling layer on top of it. The input images are resized to 256×256 and then randomly cropped to 224×224. For the optimization of CDTR, we use an AdamW optimizer [99] 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 110 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 64 and a learning rate of 5e-4. We use the grid search to select the optimal values for the parameters.

B. Comparison With State-of-the-Arts

Quantitative Comparison: We compare the proposed CDTR with state-of-the methods on CUB-200-2011 [90] and ILSVRC [91]. As shown in Table I, CDTR achieves stable and superior localization performance. On CUB-200-2011 [90], CDTR exhibits a significant enhancement over the baseline TS-CAM [3] in terms of Top-1/Top-5/GT-known Loc metrics, achieving Top-1 Loc accuracy of 81.33% and GT-known Loc accuracy of 96.89% . Furthermore, compared with the CNN-based state-of-the-art methods (TAFormer [94]), CDTR respectively achieves improvements of 8.01% in terms of Top-1 Loc. Additionally, CDTR exhibits 1.98% improvement in Top-5 Loc compared to the transformer-based state-of-the-art method CATR [39], which further highlights the potential of transformers for WSOL. On ILSVRC [91], the proposed CDTR surpasses the baseline TS-CAM [3] by margins of 4.80% and 3.12% in terms of Top-1 Loc and GT-knwon Loc, respectively. Remarkably, CDTR achieves a Top-1 Loc accuracy of 58.20% , outperforming all transformer-based methods. Compared to the CNN-based methods, the proposed CDTR achieves state-of-the-art performance, with only a slightly lower GT-Known Loc than TAFormer [94]. Note that CDTR adopts the CLIP model trained on a vast image-caption dataset. While benefiting from the large-scale supervision provided by CLIP, CDTR may encounter challenges in some constrained scenarios where CLIP is either not useful or unavailable (please see Section V for details).

Table II further presents a comparison between CDTR and other methods in terms of classification performance. The results demonstrate that CDTR achieves superior classification accuracy. Specifically, on CUB-200-2011 [90], CDTR manifests a remarkable 83.87% Top-1 Cls and a 97.03% Top-5 Cls. In comparison to the benchmark TS-CAM [3], CDTR notches up a noteworthy enhancement, achieving a 3.57% and 2.23% uptick in Top-1 Cls and Top-5 Cls, respectively. On ILSVRC [91], the proposed CDTR achieves the best 93.67% Top-5 Cls and is slightly lower than TAFormer [94] in Top-1 Cls (77.66% vs 77.76%). Despite a lesser count of accurately classified images, CDTR still trumps TAFormer [94] (58.20% vs 53.42%) in localization performance. Note that Top-1 and Top-5 consider both the localization and classification accuracy, i.e., a prediction is correct only if both localization and classification are correct.

Table III presents a comparison of CDTR with methods that adopt a separate localization-classification pipeline. It is worth noting that these multi-stage approaches achieve remarkable results, but require separate networks for localization and classification that must undergo distinct training phases. For instance, SPOL [100] employs three distinct networks for WSOL. The first two separate

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

TABLE II
CLASSIFICATION COMPARISON WITH STATE-OF-THE-ART METHODS

Methods	Venue	Backbone	CUB-200-2011 Cls Acc.		ILSVRC Cls Acc.	
			Top-1	Top-5	Top-1	Top-5
CAM [26]	CVPR16	VGG16	76.60	92.50	66.60	88.60
ACoL [20]	CVPR18	VGG16	71.90	–	67.50	88.00
ADL [92]	TPAMI20	VGG16	65.27	–	–	–
MEIL [22]	CVPR20	VGG16	74.77	–	70.27	–
SPA [24]	CVPR21	VGG16	76.11	92.15	–	–
FAM [93]	ICCV21	VGG16	77.26	–	70.90	–
TAFormer [94]	TPAMI23	VGG16	79.13	–	70.67	–
CAM [26]	CVPR16	InceptionV3	73.80	91.50	73.30	91.80
I2C [31]	ECCV20	InceptionV3	–	–	73.30	91.60
SPA [24]	CVPR21	InceptionV3	73.51	91.39	73.26	91.81
FAM [93]	ICCV21	InceptionV3	81.25	–	77.63	–
TAFormer [94]	TPAMI23	InceptionV3	82.22	–	**77.76**	–
TS-CAM [3]	ICCV21	Deit-S	80.30	94.80	74.30	92.10
LCTR [33]	AAAI22	Deit-S	**85.00**	**97.10**	77.10	93.40
SCM [5]	ECCV22	Deit-S	78.50	94.50	76.70	93.00
LCAR [96]	AAAI23	Deit-S	80.60	–	75.90	–
CATR [39]	ICCV23	Deit-S	83.72	96.82	77.25	93.64
CDTR (ours)	This Work	Deit-S	83.87	97.03	77.66	**93.67**

The best results are highlighted in bold, second are underlined.

TABLE III
COMPARISON WITH THE METHODS BASED ON A SEPARATE LOCALIZATION-CLASSIFICATION PIPELINE.

Methods	Backbone		CUB-200-2011 Loc Acc.			ILSVRC Loc Acc.		
	Localization	Classification	Top-1	Top-5	GT-known	Top-1	Top-5	GT-known
PSOL [43]	InceptionV3	InceptionV3	65.51	83.44	-	54.82	63.25	65.21
PSOL [43]	ResNet50	ResNet50	70.68	86.64	90.00	53.98	63.08	65.44
PSOL [43]	DenseNet161	DenseNet161	74.97	89.12	93.01	55.31	64.18	66.28
PSOL [43]	DenseNet161	EfficientNet-B7	77.44	89.51	93.01	58.00	65.02	66.28
SLT-Net [25]	VGG16	VGG16	67.80	–	87.60	51.20	62.40	67.20
SLT-Net [25]	InceptionV3	InceptionV3	66.10	–	86.50	55.70	65.40	67.60
SPOL [100]	ResNet50†	DenseNet161	79.74	93.69	96.46	56.40	66.48	69.02
SPOL [100]	ResNet50†	EfficientNet-B7	80.12	93.44	96.46	59.14	67.15	69.02
ISIC [44]	ResNet50	ResNet50	80.68	**94.08**	**97.32**	**59.61**	67.84	70.01
CATR [39]	Deit-S		79.62	92.08	94.94	56.90	66.64	69.25
CDTR (ours)	Deit-S		**81.33**	94.06	96.89	58.20	**68.05**	**70.72**

The best results are highlighted in bold, second are underlined. '†' indicates the backbone is modified.

modified ResNet50 are used for generating class activation maps and foreground segmentation, respectively. An additional DenseNet161/EffcientNet-B7 is then employed for classification. Contrasting this multi-stage paradigm, the proposed CDTR introduces an integrated framework utilizing a singular network architecture, thereby significantly enhancing computational efficiency. Furthermore, CDTR not only achieves the best Top-1 Loc (81.33%) on CUB-200-2011 [90] but also demonstrates competitive Top-1 Loc (58.20% vs 59.61%) on ILSVRC [91], further proving the effectiveness of the proposed method.

Visual Comparison: We compare the localization results of the proposed CDTR, CATR [39] and TS-CAM [3] on CUB-200-2011 [90] and ILSVRC [91] in Fig. 4. Our proposed CDTR consistently generates more accurate and refined localization maps that encompass category-aware object regions, exhibiting sharper and more compact boundaries compared to TS-CAM [3] and CATR [39]. The notable instances of this superiority can be observable in Fig. 4, particularly the fifth-column sample in the bottom part. Here, TS-CAM [3]'s representation is marred by category-agnostic noise that inadvertently highlights *Person* in addition to the target object *Shovel*. While CATR [39] addresses this issue to some extent by integrating category information - thereby focusing the network's attention more acutely on the target object region relative to TS-CAM [3] - it falls short of the precision exhibited by the proposed CDTR in this paper. Additionally, Fig. 6 offers further visualizations of localization maps generated by our CDTR method. These visualizations reveal the method's effectiveness in preserving long-range feature dependencies and encompassing the entire extent of the objects under consideration.

Localization Quality: We present a statistical analysis of the Intersection over Union (IoU) between predicted bounding boxes and ground-truth bounding boxes, as depicted in Fig. 5. On CUB-200-2011 [90], our method attains a median IoU of 82.52% , signifying a marked enhancement over the existing baseline established by TS-CAM [3], with an improvement margin of 12.80% . Similarly, on ILSVRC [91], our method exhibits an increment of 1.90% in IoU compared to the baseline.

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

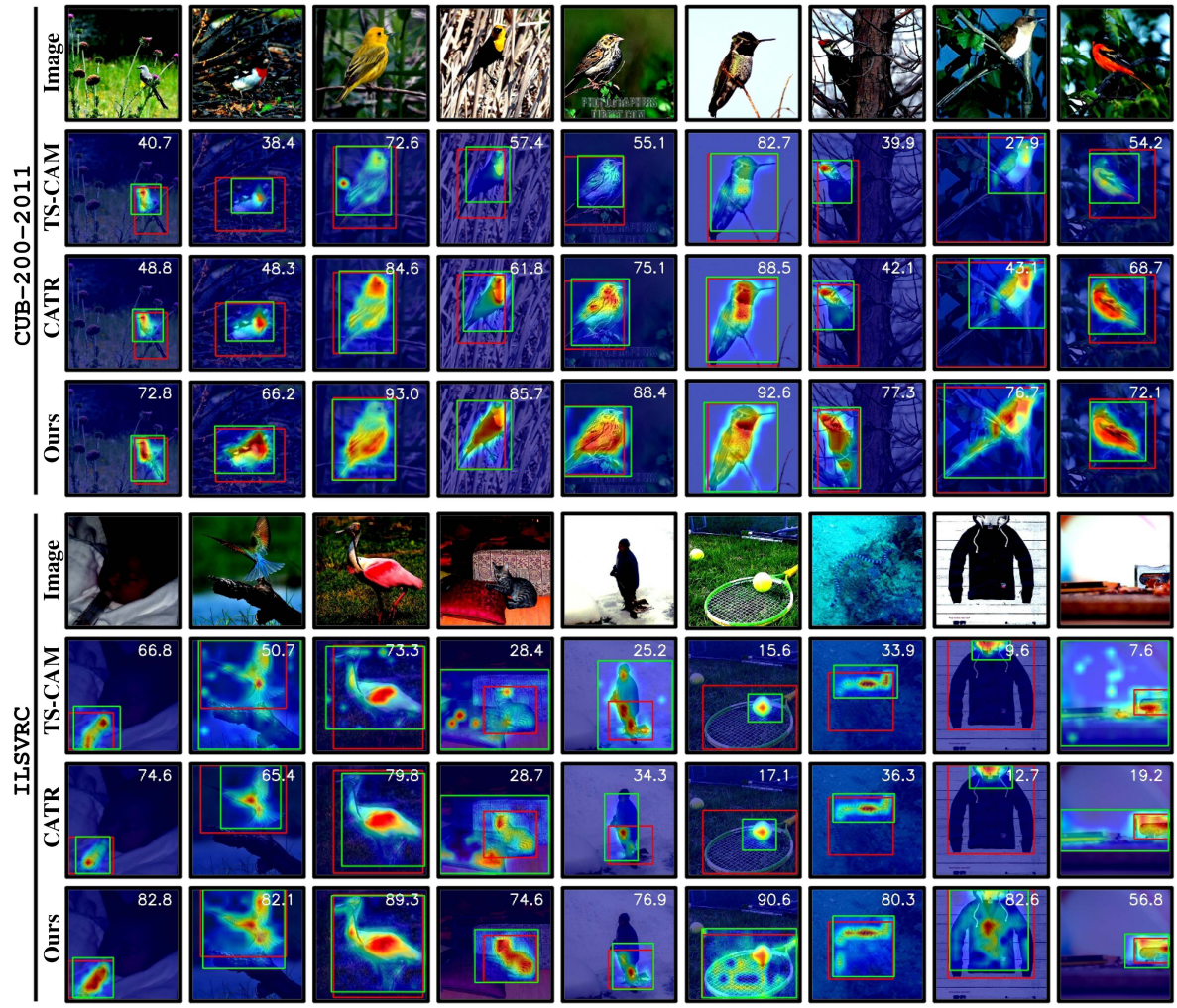


Fig. 4. *Visualization comparison with the baseline TS-CAM [3] and CATR [39] method on* `CUB-200-2011` *and* `ILSVRC`. The ground-truth bounding boxes are in Red, the predictions are in Green and the corresponding IoU values (%) are displayed in white text.

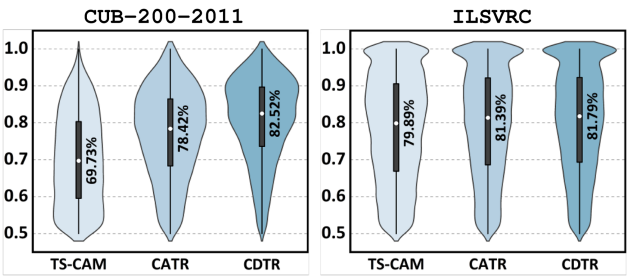


Fig. 5. Statistical analysis about localization quality.

These findings collectively indicate that the proposed CDTR method significantly augments the localization quality across both datasets.

Segmentation Quality: We present the segmentation performance of CDTR on `OpenImages` [18]. As depicted in Table IV, the proposed CDTR performs best in the PxAP metric. Compared with the baseline method CATR [39], we achieve an increase of 2.28% (67.35% vs 69.63%). The visualizations of the localization maps are presented in Fig. 6. We can observe that the proposed CDTR demonstrates a strong ability to accurately activate the target objects. For instance, in the example shown in the last column, the localization map precisely identifies the target *Ball* without interference from surrounding background elements, such as *Person*.

TABLE IV
SEGMENTATION QUALITY OF THE LOCALIZATION MAP

Method	Venue	Backbone	PxAp
TS-CAM [3]	ICCV21	Deit-S	54.26
CATR [39]	ICCV23	Deit-S	67.35
CDTR (ours)	This Work	Deit-S	**69.63**

C. Ablation Study

In this subsection, an extensive series of experiments are conducted to empirically assess the efficacy of CDTR.

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

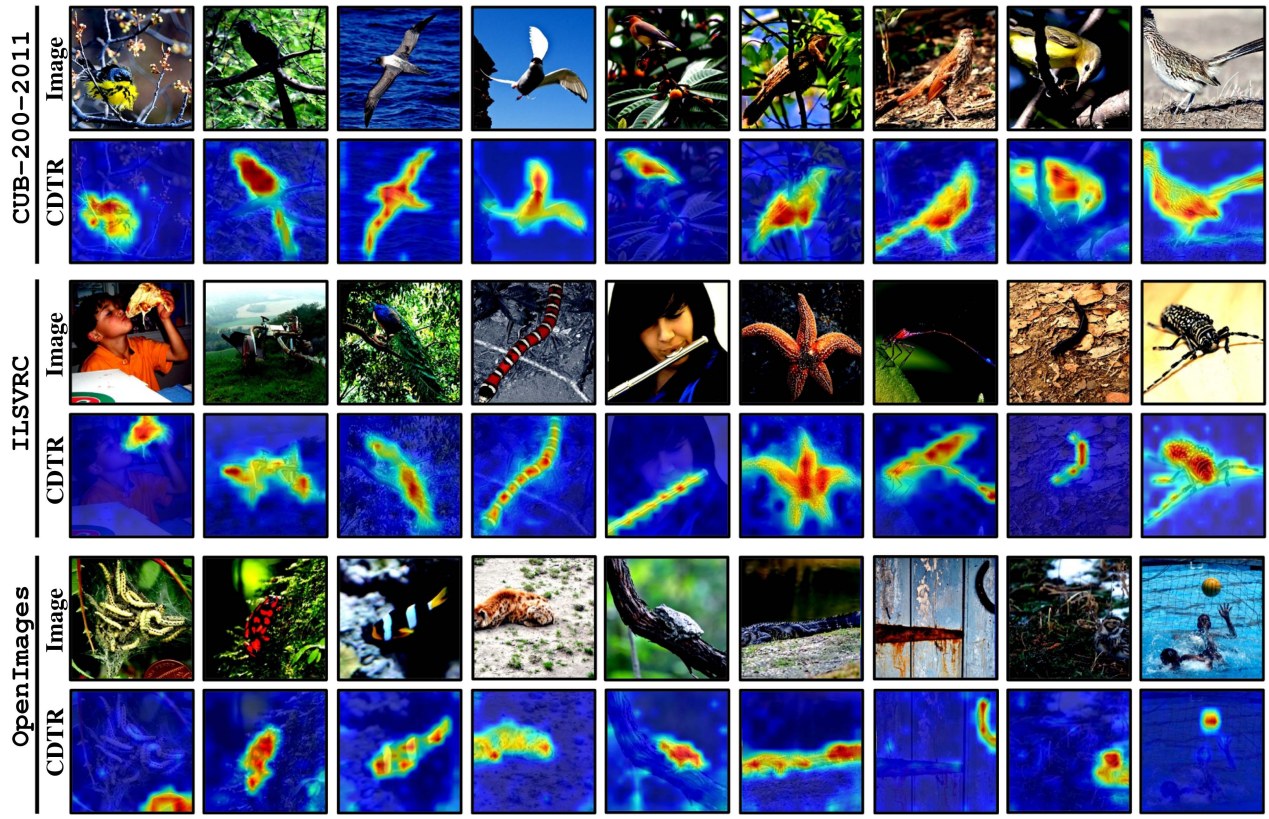


Fig. 6. Visualizaiton of localizaiton maps on CUB-200-2011, ILSVRC and OpenImages.

TABLE V
ABLATION STUDY OF CDTR

	CSM	OCM	SKI	SBA	Top-1	Top-5	GT-k.
CATR [39]					74.16	86.42	89.28
	✓				77.59	90.35	93.22
		✓			76.10	88.89	91.58
	✓	✓			79.62	92.08	94.94
CDTR	✓	✓	✓		80.77	93.42	96.30
	✓	✓		✓	80.42	92.90	95.88
	✓	✓	✓	✓	**81.33**	**94.06**	**96.89**

TABLE VI
COMPARISON OF PARAMETERS AND MACS W.R.T. SKI AND SBA

Method	Input Size	#Param. (M)	MACs (G)	GT-k. (%)
CATR [39]	224^2	22.92	4.49	94.94
+ SKI	224^2	23.32 (+0.40)	4.49 (0)	**96.30** (+1.36)
+ SBA	224^2	22.92 (+0.00)	4.49 (0)	**95.88** (+0.94)
+ SKI + SBA	224^2	23.32 (+0.40)	4.49 (0)	**96.89** (+1.95)

Ablation studies of CDTR components: We compare the performance of the proposed approach using different modules. We re-implement TS-CAM [3] and take it as our baseline. Table V shows that the CSM, which builds a connection between the attention map and the category information, contributes to enhancements of 3.43% and 3.94% in terms of Top-1 Loc and GT-known Loc, respectively. Remarkably, the integration of both CSM and OCM culminates in an eminent Top-1 Loc accuracy, peaking at 79.62% . Moreover, the proposed SKI builds a connection between CSM and OCM, which further improves the GT-known Loc (from 94.94% to 96.30%). The model achieves the best localization accuracy when the four modules are activated, with Top-1 Loc and GT-known Loc accuracies reaching 81.33% and 96.89% , respectively.

Parameters of SKI and SBA: In Table VI, we compare the parameters of the proposed methods with the baseline method, i.e., CATR [39] (the initial conference version). The framework integrating the proposed SKI, which has 23.32 M parameters and 4.49 MACs, achieves a 1.36% improvement in performance compared to CATR [39], which has 22.92 M parameters and 4.49 MACs (94.94% vs 96.30%). Additionally, we introduce SBA, which does not add any parameters. By constraining the semantic information of the objects, SBA enhances localization performance by 0.94% compared to the baseline method [39]. These results underscore the effectiveness of SKI and SBA in improving localization performance with minimal additional computational overhead.

Hyperparameter G in CSM: We investigate the effect of G in terms of Top-1/Top-5/GT-known Loc. Here, G represents the number of channels in the encoder part of CSM, which aims to reduce the channel dimension while maintaining essential spatial information of features. Referencing Fig. 7, we vary G values

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

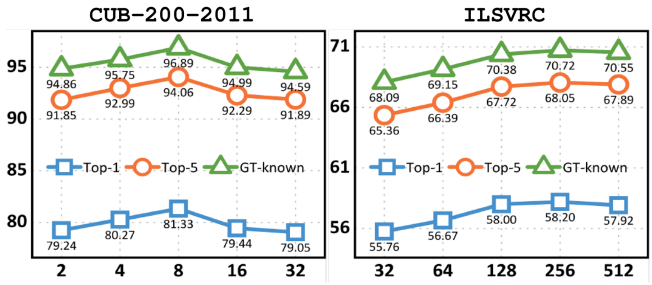


Fig. 7. Performance analysis w.r.t. hyperparameter G in CSM.

TABLE VII
PERFORMANCE ANALYSIS W.R.T. HYPERPARAMETER p IN OCM

p	0.10	0.25	0.30	0.35	0.40
Top-1	80.53	**81.33**	81.17	80.89	79.85
Top-5	93.47	**94.06**	96.01	93.88	92.91
GT-k.	96.22	**96.89**	83.11	95.98	95.58

p	0.05	0.10	0.15	0.20	0.25
Top-1	57.82	**58.20**	57.36	57.12	56.99
Top-5	67.80	**68.05**	67.78	67.69	67.50
GT-k.	70.19	**70.72**	70.10	70.00	69.78

Top: CUB-200-2011, bottom: ILSVRC.

TABLE VIII
PERFORMANCE ANALYSIS W.R.T. THE HYPERPARAMETER K IN OCM

K	3	5	7	9	11
Top-1	79.03	80.75	81.01	**81.33**	80.79
Top-5	91.85	93.54	93.90	**94.06**	93.71
GT-k.	94.51	96.31	96.77	**96.89**	96.53

K	4	6	8	10	12
Top-1	56.37	57.04	57.84	**58.20**	57.82
Top-5	66.07	66.93	67.73	**68.05**	67.84
GT-k.	68.76	69.60	70.45	**70.72**	70.57

Top: CUB-200-2011, bottom: ILSVRC.

to study the localization performance changes. From these results, we observe that optimal localization accuracy is attained at $G = 8$ for CUB-200-2011 and $G = 256$ for ILSVRC. Nevertheless, an increase in G beyond these thresholds tends to result in diminished performance, a phenomenon we attribute to the emergence of parameter redundancy.

Hyperparameter p in OCM: We investigate the impact of the hyperparameter p, which determines the extent of background noise suppression. The experimental results are summarized in Table VII. It is observed that the localization accuracy peaks at $p = 0.25$, followed by a degradation in performance on CUB-200-2011. This occurs because as p increases, the model increasingly suppresses more regions, thereby inadvertently affecting the foreground as well. The results in Table VII on ILSVRC also corroborate this observation, with performance dropping when $p \geq 0.1$.

Hyperparameter K in OCM: We demonstrate the effect of the hyperparameter K, which represents the number of attention maps aggregated from top-K transformer blocks. The experimental results are summarized in Table VIII. Optimal localization accuracy is achieved when K is set to 9 and 6 on CUB-200-2011 and ILSVRC, respectively. Increasing K results in a greater number of self-attention maps, thus offering enhanced insights for generating the pixel-level pseudo map. However, a noticeable decline in performance is observed when K exceeds its optimal value, likely due to overfitting.

Quality of pixel-level pseudo map in OCM: We visualize the pixel-level pseudo map (i.e., $\mathbf{M}_{\text{ocm}}$) of OCM in Fig. 8. It is observed that $\mathbf{M}_{\text{ocm}}$ encompasses the class-specific features, effectively highlighting 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. Consequently, OCM effectively refines the object regions for the category-aware attention map in a self-supervised manner, leading to the precise activation of object regions.

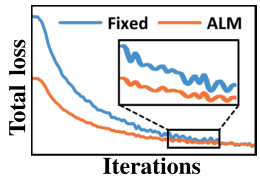
Hyperparameter α in total loss: Fig. 9 illustrates the sensitivity of localization quality to the hyperparameters α in $\mathcal{L}$ (2). α is defined as the factor controlling the semantic knowledge from the CLIP model during the training phase. From the results, we can observe that optimal localization accuracy occurs when $\alpha = 2.4$ on CUB-200-2011 and $\alpha = 0.01$ on ILSVRC. This finding also supports the argument that incorporating prior knowledge of the CLIP model is highly beneficial for fine-grained features.

Effects of the automatic weighted loss mechanism: We investigate the effects of the automatic weighted loss mechanism (ALM) [38] on our losses from two perspectives. First, we examine the performance when the ALM is not used, and λ_1-λ_5 are set to 1 for training. The results in Table IX indicate that using the ALM contributes to a slight improvement in localization performance. Second, we analyze the changes in five learnable parameters (i.e., λ_1-λ_5) in (2) during the training process. It should be noted that these learnable parameters are all initialized to 1. As shown in Fig. 10, it is observed that λ_2 and λ_5 are consistently larger during training, even though the losses weighted by λ_1, λ_2 and λ_5 are the same classification loss (i.e., the cross-entropy loss). This can be attributed to the category-aware information provided by CSM and SKI, which is beneficial to classification. This observation supports the argument that the modules facilitate the connection between the attention maps and the specific classes. Furthermore, the results indicate that OCM plays a supporting role in refining the object regions, as the values of λ_3 and λ_4 tend to decrease during the training phase.

Effects of the CLIP model: To provide a fair comparison with other WSOL methods that do not use CLIP, we modify representative WSOL methods (i.e., TS-CAM [3] and LCTR [33])

TABLE IX
PERFORMANCE ANALYSIS W.R.T. LOSS WEIGHTING MECHANISM

Weighting Mechanism	Loc Acc. Top-1	Top-5	GT-k.
Fixed	80.39	93.29	96.07
ALM	**81.33**	**94.06**	**96.89**
+Δ	+0.94	+0.77	+0.82



Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

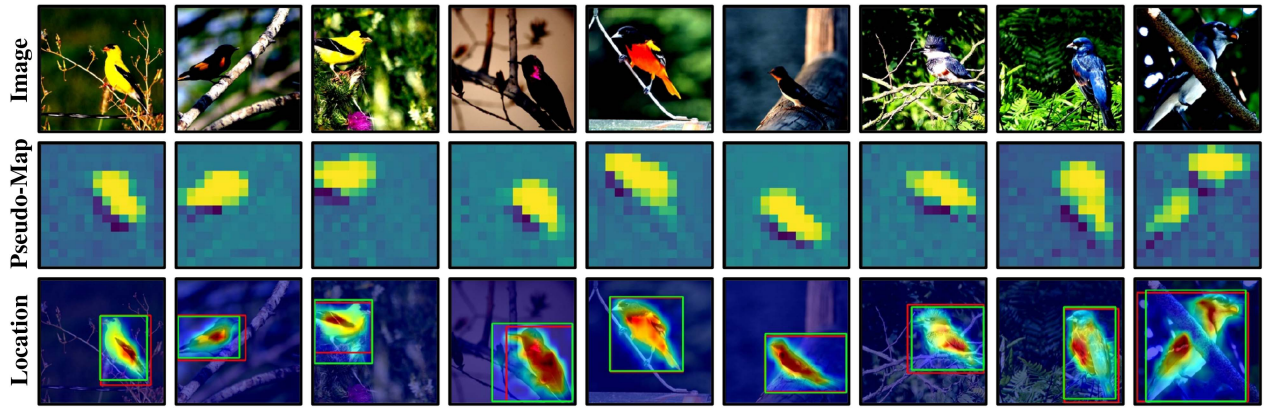


Fig. 8. *Visualization of the pixel-level pseudo map* M_{ocm}. The ground-truth bounding boxes are in Red, and the predictions are in Green.

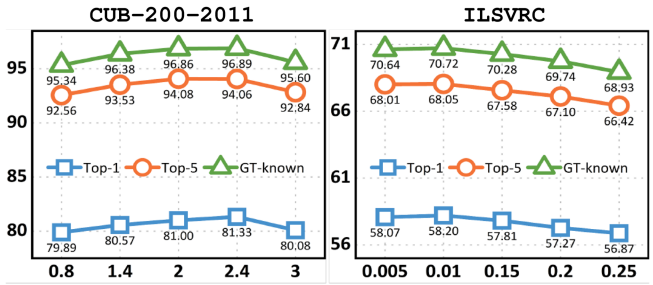


Fig. 9. Performance analysis w.r.t. the hyperparameter α in total loss.

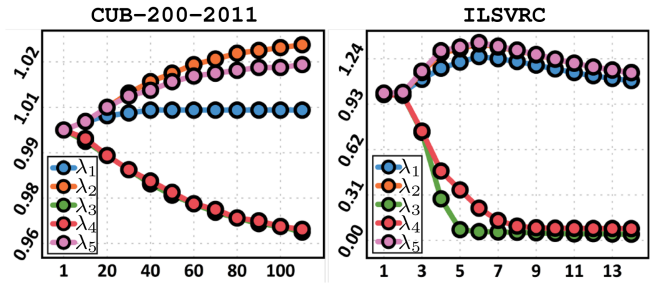


Fig. 10. Analysis about loss weights in the training phase.

TABLE X
PERFORMANCE ANALYSIS W.R.T. CLIP

Method	CUB-200-2011 Top-1 Loc Acc.	ILSVRC Top-1 Loc Acc.
TS-CAM [3]	74.16	53.40
TS-CAM ♣	75.20	53.92
LCTR [33]	79.20	56.10
LCTR ♣	80.00	56.33
CDTR ♢ (ours)	80.42	56.90
CDTR (ours)	**81.33**	**58.20**

♣ indicates adopting the CLIP knowledge in the training phase. ♢ indicates the using only SBA.

by integrating CLIP prior information into the training process. Specifically, we introduce SBA into these methods while maintaining their original experimental settings. The experimental results are presented in Table X. The results show that integrating the CLIP model improves the localization performance of both TS-CAM [3] and LCTR [33]. For instance, the modified TS-CAM [3] with CLIP achieved an accuracy improvement of 1.04% on CUB-200-2011 compared to the original TS-CAM. In addition, we can observe that using only SBA, i.e., applying only CLIP to CDTR, achieves the best Top-1 Loc. accuracy of 80.42% and 56.90% on CUB-200-2011 and ILSVRC, compared to other WSOL methods. This superior performance also demonstrates SBA's effectiveness in both leveraging visual and textual information for the WSOL task. Lastly, the proposed CDTR achieves the best localization performance, with accuracies of 81.33% and 58.20% on CUB-200-2011 and ILSVRC, respectively.

TABLE XI
COMPARISON OF MODEL COMPLEXITY

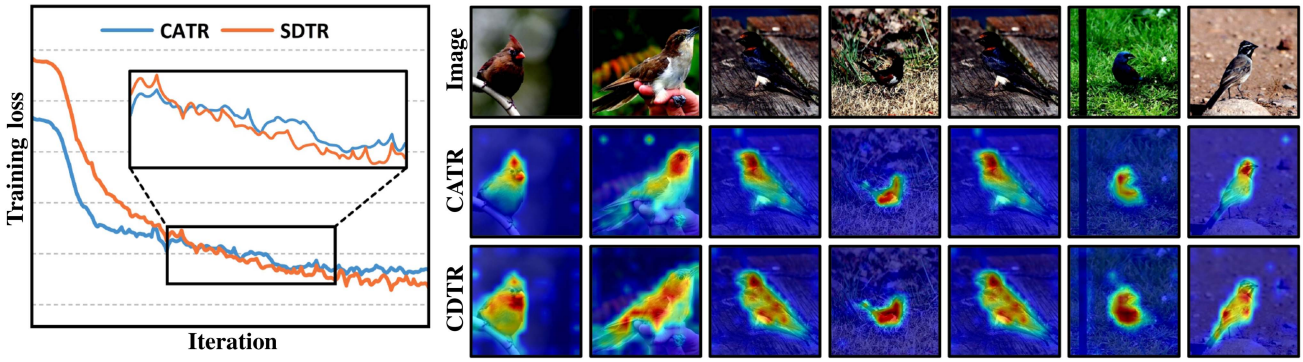
Method	Input Size	#Param. (M)	RunTime (ms/img)	Memory (M)	Top-1 Loc. (%)
TS-CAM∗ [3]	224^2	22.36	0.544	171.19	74.16
LCTR∗ [33]	224^2	25.76	0.842	197.93	79.20
CATR [39]	224^2	22.92	0.642	175.76	79.62
CDTR (ours)	224^2	23.32	0.685	178.17	**81.33**

*indicates the re-implement method.

Model Complexity: In Table XI, a comparison of model complexity is presented with other methods. From this comparison, it is clear that the proposed CDTR achieves the highest Top-1 Loc at 81.33% , while also maintaining a competitive balance in runtime and memory usage. Compared to CATR [39], CDTR demonstrates a 1.71% increase in Top-1 Loc with minimal additional computational overhead. Additionally, compared to LCTR [33], CDTR achieves a 2.13% improvement in Top-1 Loc with a 2.44 M reduction in parameters.

Comparison with previous conference version: We compare the improvement over the conference version in both quantitative and qualitative aspects. As shown in Fig. 11(a), the curve of the training loss is displayed with the training iterations for both CATR (the previous conference version) and CDTR (this work). It is observed that the loss curve of CDTR converges to a lower point and demonstrates a more stable convergence trend. This suggests that the semantic communication between CSM and

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.



(a) The training loss curves (b) Visualization of the localization maps

Fig. 11. Experimental comparison between CATR (the previous conference version) and CDTR (this work).

TABLE XII
COMPARISON WITH SAT [52]

Method	CUB-200-2011 Top-1 Loc Acc.	ILSVRC Top-1 Loc Acc.
SAT [52]	80.96	**60.15**
CDTR (ours)	81.33	58.20
CDTR♠(ours)	**82.20**	60.00

♠ indicates using the same localization map generation mechanism as in SAT [52].

TABLE XIII
LOCALIZATION PERFORMANCE ON test SPLIT OF CUB-200-2011-V2

Method	Loc. Acc		
	Top-1	Top-5	GT-known
TS-CAM [3]	74.08	85.95	88.71
CATR [39]	79.13	90.58	93.29
CDTR (ours)	**81.12**	**93.75**	**96.58**

Hyperparameters have been optimized over the train-fullsup split.

OCM enables the classification network to better learn the category regions, thus leading to faster convergence. Fig. 11(b) also illustrates that more explicit semantic information contributes to the complete activation of the target object. In comparison with the previous CATR, CDTR exhibits more robustness in the learning of the object regions and thereby improves localization accuracy. From the visualization comparison in Fig. 4, it is observed that CDTR can accurately activate and locate objects with poor CATR localization. From the localization quality analysis in Fig. 5, it is evident that the proposed method improves by 4.1% and 0.4% on CUB-200-2011 and ILSVRC, respectively, compared to CATR. From the localization accuracy comparison in Table I, CDTR achieves gains of 1.95% and 1.47% in GT-known Loc on CUB-200-2011 and ILSVRC, respectively.

Comparison with SAT [52]: The proposed CDTR employs the cls-token attention map for object localization, similar to existing WSOL methods such as TS-CAM [3], LCTR [33], SCM [5]. In contrast, SAT [52] proposes to use the spatial-token attention map for object localization, leveraging different experimental benchmarks. To ensure a fair comparison, we conduct experiments within the CDTR framework using the same localization map generation mechanism as SAT [52]. As shown in Table XII, CDTR demonstrates improved localization performance on both CUB-200-2011 and ILSVRC when using the spatial-token attention map. In particular, CDTR achieves a Top-1 localization accuracy of 82.20% on the CUB-200-2011 dataset, marking a 1.24% improvement over SAT [52], and a Top-1 localization accuracy of 60.00% on the ILSVRC dataset, which is slightly lower than SAT [52]. It is noteworthy that the baseline performances of our method and SAT are 53.40% and 56.86% , respectively, with relative performance improvements of 6.60% and 3.29% , underscoring the effectiveness of our approach in object localization. However, the proposed CDTR employs a more intricate pipeline compared to SAT [52]. This observation highlights a valuable avenue for future research: simplifying and optimizing our design to improve efficiency.

Ablation studies on different dataset configurations: To further evaluate the effectiveness of our proposed method, we conduct evaluations of CDTR on the CUB-200-2011-v2 dataset, following the experimental protocols outlined by Choe et al. [18]. The dataset is divided into three disjoint subsets: train-weaksup, train-fullsup, and test. In contrast to the original CUB-200-2011 dataset [90], CUB-200-2011-v2 [18] includes additional annotated images intended for hyperparameter tuning, designated as train-fullsup. Accordingly, we train CDTR using the train-weaksup subset, perform hyperparameter optimization on train-fullsup, and subsequently evaluate the model on the test subset. For comparative analysis, we also implement the recent TS-CAM [3] and CATR [39] methods under identical experimental conditions. The experiments are presented in Table XIII. It can be observed that the proposed CDTR demonstrates superior localization performance, achieving 81.12% in Top-1 localization accuracy. Moreover, we note that light discrepancies when compared to the results reported under the original settings in Table I. These differences can be attributed to the hyperparameter tuning being performed on a disjoint dataset split.

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

TABLE XIV
LOCALIZATION PERFORMANCE FOR DIFFERENT DOMAIN DATASETS

Setting			Brain Tumor	NEU-DET
CATR [39]	SKI	SBA	Top-1 Loc.	Top-1 Loc.
✓			46.83	19.82
✓	✓		48.26	20.77
✓		✓	46.85	19.80
✓	✓	✓	**48.38**	**21.00**

V. DISCUSSION

Expansion: To verify the generalizability of the CDTR, we conduct experiments on two datasets from different domains, namely the Brain Tumor dataset [101] and the NEU-DET dataset [102]. The Brain Tumor dataset consists of Magnetic Resonance Imaging (MRI) images of brain tumors. The dataset contains a total of 5,249 images, with 4,737 images for training and 512 images for validation. The NEU-DET dataset is a surface defect database containing six types of typical surface defects in hot-rolled steel strips. The dataset includes 1,800 grayscale images, with 1,440 images for training and 360 images for validation. Table XIV summarizes the experimental results. This demonstrates that the proposed CDTR achieves an improvement of 1.55% and 1.18% in localization accuracy compared to the baseline CATR [39] on the Brain Tumor dataset [101] and the NEU-DET dataset [102], respectively. Additionally, we observe that both CATR [39] and CDTR exhibit limited localization capabilities on both datasets. We attribute this to the inherent limitations of the weakly supervised object localization pipeline. Specifically, the Transformer's patch-based processing mechanism poses challenges, making it difficult to avoid blurred edges and grid-like artifacts when generating masks from attention maps [103]. Furthermore, we find that SBA has minimal impact on performance improvement. These results support the notion that CLIP [35] is suboptimal in cross-domain scenarios. On the other hand, it validates the robustness of the other components of our approach. Additionally, we visualize the localization maps on the Brain Tumor dataset [101] and the NEU-DET dataset [102] in Fig. 12. The results suggest that CDTR has the potential to identify targets in images from different domains. We believe these insights provide a deeper understanding of the strengths and weaknesses of our method and suggest valuable directions for future research.

Limitation: Failure localization maps are presented in Fig. 13. It is found that the CDTR model demonstrates sub-optimal performance in localizing objects within complex environments (i.e., camouflage-prone situations). For instance, in the first column of Fig. 13, the *Bittern* is obscured by numerous weeds, hindering the distinction of its complete area. In the second column, this sample depicts a chainsaw that is sawing wood, with the chainsaw blade embedded in the wood being a challenging region to separate. This limitation is attributed to the lack of instance-level supervision, rendering it challenging for the weakly supervised object localization task to highlight "camouflage" objects.

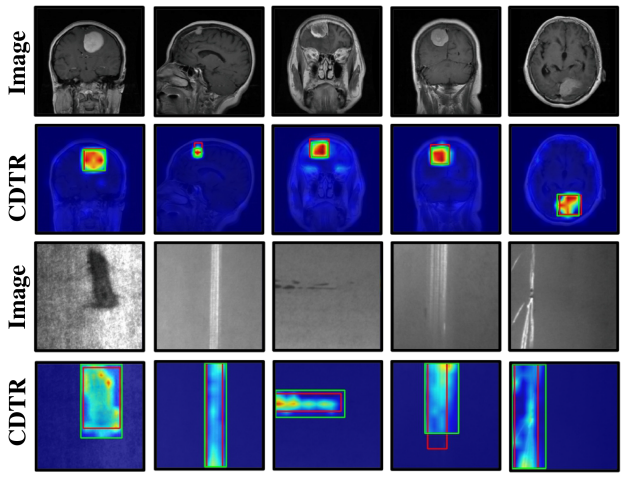


Fig. 12. Visualization of localization results on the Brain Tumor dataset (TOP) and the NEU-DET dataset (Bottom). The ground-truth bounding boxes are in Red, the predictions are in Green.

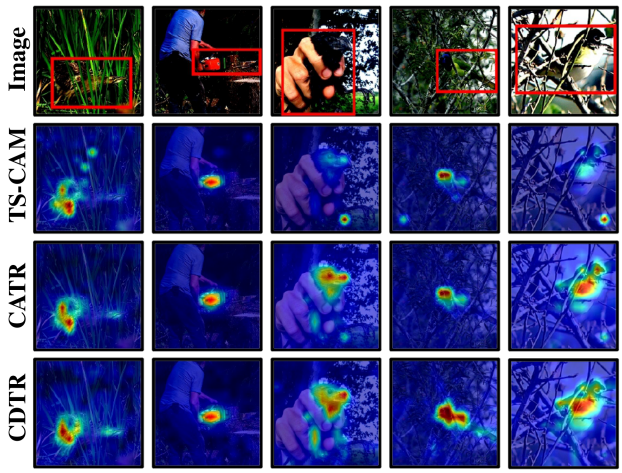


Fig. 13. Localization error analysis in TS-CAM, CATR, and our proposed CDTR. The ground-truth bounding boxes are in Red.

Future works: Future work will focus on two primary areas: **1)** enhancing the localization performance for "camouflage" objects and **2)** expanding the application scope of CDTR.

To address the challenges in localizing "camouflage" objects, the following research directions are considered promising: **1)** Enhancing the category-aware constraint mechanism by implementing varying tolerances for regions at distinct pixel levels. **2)** Integrating multi-scale unified learning from the domain of camouflage object detection to emphasize object regions and boundary features.

VI. CONCLUSION

This paper identifies that previous transformer-based works rely on category-agnostic attention maps to generate localization maps, making them prone to background noise. To address this issue, we propose the CLIP-Driven TRansformer (CDTR), which aims to improve localization map generation by enhancing the category awareness of attention maps to focus on

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

foreground objects. First, a category-aware stimulation module (CSM) is introduced that learns category information for the self-attention maps. Furthermore, an object constraint module (OCM) is proposed to refine object regions for category-aware attention maps in a self-supervised manner. Third, a semantic kernel integrator (SKI) is designed to establish a connection between CSM and OCM, facilitating enhanced communication of category awareness. Additionally, a semantic boost adapter (SBA) is introduced to further refine category-aware maps using prior knowledge. Extensive experiments on `CUB-200-2011` and `ILSVRC` benchmarks validate the effectiveness of the proposed CDTR, which outperforms state-of-the-art methods.

REFERENCES

[1] Z. Chen, R. Ji, J. Wu, and Y. Shen, "Multi-scale features for weakly supervised lesion detection of cerebral hemorrhage with collaborative learning," in *Proc. 1st ACM Int. Conf. Multimedia Asia*, 2019, pp. 1–7.

[2] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5866–5885, Sep. 2022.

[3] W. Gao et al., "TS-CAM: Token semantic coupled attention map for weakly supervised object localization," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 2886–2895.

[4] E. Kim, S. Kim, J. Lee, H. Kim, and S. Yoon, "Bridging the gap between classification and localization for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14258–14267.

[5] H. Bai, R. Zhang, J. Wang, and X. Wan, "Weakly supervised object localization via transformer with implicit spatial calibration," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 612–628.

[6] L. Zhu et al., "Background-aware classification activation map for weakly supervised object localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14175–14191, Dec. 2023.

[7] Y. Shen et al., "Enabling deep residual networks for weakly supervised object detection," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 118–136.

[8] Y. Shen, R. Ji, Z. Chen, Y. Wu, and F. Huang, "UWSOD: Toward fully-supervised-level capacity weakly supervised object detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 7005–7019.

[9] Y. Shen et al., "Noise-aware fully webly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11326–11335.

[10] O. Veksler, "Test time adaptation with regularized loss for weakly supervised salient object detection," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 7360–7369.

[11] L. Ru, H. Zheng, Y. Zhan, and B. Du, "Token contrast for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3093–3102.

[12] S. Rong, B. Tu, Z. Wang, and J. Li, "Boundary-enhanced co-training for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19574–19584.

[13] Y. Lin et al., "Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15305–15314.

[14] J. Hanna, M. Mommert, and D. Borth, "Sparse multimodal vision transformer for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2144–2153.

[15] T. Cheng, X. Wang, S. Chen, Q. Zhang, and W. Liu, "BoxTeacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3145–3154.

[16] A. Arun, C. Jawahar, and M. P. Kumar, "Weakly supervised instance segmentation by learning annotation consistent instances," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 254–270.

[17] Y. Shen et al., "Parallel detection-and-segmentation learning for weakly supervised instance segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 8198–8208.

[18] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim, "Evaluating weakly supervised object localization methods right," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3133–3142.

[19] C. Wang, R. Xu, S. Xu, W. Meng, R. Wang, and X. Zhang, "Exploring intrinsic discrimination and consistency for weakly supervised object localization," *IEEE Trans. Image Process.*, vol. 33, pp. 1045–1058, 2024.

[20] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1325–1334.

[21] H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji, and Q. Ye, "DANet: Divergent activation for weakly supervised object localization," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6589–6598.

[22] J. Mai, M. Yang, and W. Luo, "Erasing integrated learning: A simple yet effective approach for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8766–8775.

[23] Z. Chen, L. Cao, Y. Shen, F. Lian, Y. Wu, and R. Ji, "E2Net: Excitative-expansile learning for weakly supervised object localization," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 573–581.

[24] X. Pan et al., "Unveiling the potential of structure preserving for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11642–11651.

[25] G. Guo, J. Han, F. Wan, and D. Zhang, "Strengthen learning tolerance for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7399–7408.

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[27] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Soft proposal networks for weakly supervised object localization," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1841–1850.

[28] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6023–6032.

[29] K. Kumar Singh and Y. Jae Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 3524–3533.

[30] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2219–2228.

[31] X. Zhang, Y. Wei, and Y. Yang, "Inter-image communication for weakly supervised localization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 271–287.

[32] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 367–376.

[33] Z. Chen et al., "LCTR: On awakening the local continuity of transformer for weakly supervised object localization," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2022, pp. 410–418.

[34] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[35] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 8748–8763.

[36] F. Liang et al., "Open-vocabulary semantic segmentation with mask-adapted clip," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7061–7070.

[37] J. Chen et al., "Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 699–710.

[38] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," 2018, *arXiv: 1805.06334*.

[39] Z. Chen et al., "Category-aware allocation transformer for weakly supervised object localization," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 6643–6652.

[40] J. Xie, C. Luo, X. Zhu, Z. Jin, W. Lu, and L. Shen, "Online refinement of low-level feature based activation map for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 132–141.

[41] P. Wu, W. Zhai, and Y. Cao, "Background activation suppression for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14228–14237.

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

[42] L. Zhu, Q. She, Q. Chen, Y. You, B. Wang, and Y. Lu, "Weakly supervised object localization as domain adaption," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14637–14646.
[43] C.-L. Zhang, Y.-H. Cao, and J. Wu, "Rethinking the route towards weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13460–13469.
[44] J. Wei, S. Wang, S. K. Zhou, S. Cui, and Z. Li, "Weakly supervised object localization through inter-class feature similarity and intra-class appearance consistency," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 195–210.
[45] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, 2022.
[46] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–21.
[47] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Multi-class token transformer for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4310–4319.
[48] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 15013–15022.
[49] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 357–366.
[50] J. Hu et al., "ISTR: End-to-end instance segmentation with transformers," 2021, *arXiv:2105.00637*.
[51] L. Tan, P. Dai, R. Ji, and Y. Wu, "Dynamic prototype mask for occluded person re-identification," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 531–540.
[52] P. Wu, W. Zhai, Y. Cao, J. Luo, and Z.-J. Zha, "Spatial-aware token for weakly supervised object localization," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 1844–1854.
[53] S. Gupta, S. Lakhotia, A. Rawat, and R. Tallamraju, "ViToL: Vision transformer for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4101–4110.
[54] Y. Shen, R. Ji, C. Wang, X. Li, and X. Li, "Weakly supervised object detection via object-specific pixel gradient," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5960–5970, Dec. 2018.
[55] B. Lai and X. Gong, "Saliency guided end-to-end learning for weakly supervised object detection," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2051–2059.
[56] Y. Wei et al., "TS2C: Tight box mining with surrounding segmentation context for weakly supervised object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 434–450.
[57] P. Tang et al., "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.
[58] Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, "WSOD2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 8292–8300.
[59] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 328–335.
[60] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.
[61] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, pp. 154–171, 2013.
[62] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2846–2854.
[63] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2843–2851.
[64] Z. Ren et al., "Instance-aware, context-focused, and memory-efficient weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10598–10607.
[65] Z. Wu, Y. Xu, J. Yang, and X. Li, "Misclassification in weakly supervised object detection," *IEEE Trans. Image Process.*, vol. 33, pp. 3413–3427, 2024.
[66] Z. Wu, J. Wen, Y. Xu, J. Yang, and D. Zhang, "Multiple instance detection networks with adaptive instance refinement," *IEEE Trans. Multimedia*, vol. 25, pp. 267–279, 2023.
[67] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, "C-mil: Continuation multiple instance learning for weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2199–2208.
[68] Y. Gao et al., "C-MIDN: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9834–9843.
[69] X. Li, M. Kan, S. Shan, and X. Chen, "Weakly supervised object detection with segmentation collaboration," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9735–9744.
[70] Y. Zhang, Y. Bai, M. Ding, Y. Li, and B. Ghanem, "W2F: A weakly-supervised to fully-supervised framework for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 928–936.
[71] X. Zhang, J. Feng, H. Xiong, and Q. Tian, "Zigzag learning for weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4262–4270.
[72] M. Meng, T. Zhang, W. Yang, J. Zhao, Y. Zhang, and F. Wu, "Diverse complementary part mining for weakly supervised object localization," *IEEE Trans. Image Process.*, vol. 31, pp. 1774–1788, 2022.
[73] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, Art. no. 896.
[74] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 529–536.
[75] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3365–3373.
[76] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1171–1179.
[77] X. Lai et al., "Semi-supervised semantic segmentation with directional context-aware consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1205–1214.
[78] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12674–12684.
[79] K. Sohn et al., "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 596–608.
[80] Z. Shen, Z. Liu, J. Qin, M. Savvides, and K.-T. Cheng, "Partial is better than all: Revisiting fine-tuning strategy for few-shot learning," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2021, pp. 9594–9602.
[81] S. W. Yoon, J. Seo, and J. Moon, "TapNet: Neural network augmented with task-adaptive projection for few-shot learning," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 7115–7123.
[82] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.
[83] Y. Li et al., "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," 2021, *arXiv:2110.05208*.
[84] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16 816–16 825.
[85] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625–5644, Aug. 2024.
[86] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.
[87] L. Qiu et al., "VT-clip: Enhancing vision-language models with visual-guided texts," 2021, *arXiv:2112.02399*.
[88] R. Zhang et al., "Tip-adapter: Training-free clip-adapter for better vision-language modeling," 2021, *arXiv:2111.03930*.
[89] J. Xie, X. Hou, K. Ye, and L. Shen, "CLIMS: Cross language image matching for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4483–4492.
[90] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200–2011 dataset," California Institute of Technology, Aug. 2011.

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.

[91] O. Russakovsky et al., *ImageNet Large Scale Visual Recognition Challenge*, vol. 115. Berlin, Germany: Springer, 2015, pp. 211–252.
[92] J. Choe, S. Lee, and H. Shim, "Attention-based dropout layer for weakly supervised single object localization and semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4256–4271, Dec. 2021.
[93] M. Meng, T. Zhang, Q. Tian, Y. Zhang, and F. Wu, "Foreground activation maps for weakly supervised object localization," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 3385–3395.
[94] M. Meng, T. Zhang, Z. Zhang, Y. Zhang, and F. Wu, "Task-aware weakly supervised object localization with transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9109–9121, Jul. 2023.
[95] W. Lu, X. Jia, W. Xie, L. Shen, Y. Zhou, and J. Duan, "Geometry constrained weakly supervised object localization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 481–496.
[96] Y. Pan, Y. Yao, Y. Cao, C. Chen, and X. Lu, "Coarse2fine: Local consistency aware re-prediction for weakly supervised object localization," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2023, pp. 2002–2010.
[97] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[98] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 10347–10357.
[99] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–8.
[100] J. Wei, Q. Wang, Z. Li, S. Wang, S. K. Zhou, and S. Cui, "Shallow feature matters for weakly supervised object localization," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 5993–6001.
[101] Kaggle - Brain Tumor MRI Dataset, 2024. [Online]. Available: https://www.kaggle.com/datasets/ahmedsorour1/mri-for-brain-tumor-with-bounding-boxes
[102] K. Song and Y. Yan, "A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects," *Appl. Surf. Sci.*, vol. 285, pp. 858–864, 2013.
[103] H. Chen, J. An, B. Jiang, L. Xia, Y. Bai, and Z. Gao, "WS-MTST: Weakly supervised multi-label brain tumor segmentation with transformers," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 12, pp. 5914–5925, Dec. 2023.

Zhiwei Chen is currently working toward the PhD degree with Xiamen University, China. His publications on top-tier journals and conferences include *IEEE Transactions on Neural Networks and Learning Systems*, ICCV, CVPR, ECCV, AAAI, and so on. His research interests include computer vision and machine learning, especially weakly supervised object localization and weakly supervised semantic segmentation.

Yunhang Shen received the BS, MS, and PhD degrees from the Department of Artificial Intelligence in Xiamen University, China, in 2014, 2017, and 2021, respectively. He has published more than ten papers as the first author in international journals and conferences including *IEEE Transactions on Image Processing*, *IEEE Transactions on Neural Networks and Learning Systems*, ICCV, CVPR, NeurIPS, and so on. His current research interests include computer vision, pattern recognition, and machine learning.

Liujuan Cao is a professor with Xiamen University. She was awarded the Joint Funds of the National Natural Science Foundation of China (2022), the Excellent Youth Foundation of Fujian Province (2022), and the First Prize for Scientific and Technological Progress in Fujian Province (2020). Her research falls in the field of computer vision, deep learning, weakly supervised learning, and small object detection. She has published 60+ papers in ACM/IEEE Transactions and top-tier conferences, such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Neural Networks and Learning Systems*, CVPR, and ICCV.

Shengchuan Zhang received the BEng degree in electronic information engineering from Southwest University, Chongqing, China, in 2011 and the PhD degree in information and telecommunications engineering, School of Electronic Engineering, Xidian University, Xi'an, China, in 2016. He is currently an assistant professor in School of Informatics Xiamen University. His current research interests include computer vision and pattern recognition. He has published some scientific papers in leading journals like *IEEE Transactions on Image Processing*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Circuits and Systems for Video Technology*, *Pattern Recognition*, *etc*.

Rongrong Ji (Senior Member, IEEE) is a Nanqiang distinguished professor with Xiamen University, the deputy director with the Office of Science and Technology, Xiamen University, and the director of Media Analytics and Computing Lab. He was awarded as the National Science Foundation for Excellent Young Scholars (2014), the National Ten Thousand Plan for Young Top Talents (2017), and the National Science Foundation for Distinguished Young Scholars (2020). His research falls in the field of computer vision, multimedia analysis, and machine learning. He has published 50+ papers in ACM/IEEE Transactions, including *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *International Journal of Computer Vision*, and 100+ full papers on top-tier conferences, such as CVPR and NeurIPS. His publications have got more than 10K citations in Google Scholar. He was the recipient of the Best Paper Award of ACM Multimedia 2011. He has served as area chairs in top-tier conferences such as CVPR and ACM Multimedia. He is also an advisory member for Artificial Intelligence Construction in the Electronic Information Education Committee of the National Ministry of Education.

Authorized licensed use limited to: Southern University of Science and Technology. Downloaded on January 07,2026 at 05:18:02 UTC from IEEE Xplore. Restrictions apply.