Applied Intelligence (2025) 55:618
https://doi.org/10.1007/s10489-025-06514-1

Dual-branch contrastive learning for weakly supervised object localization

Zebin Guo[1,2] · Dong Li[1,2] · Zhengjun Du[1,2] · Bingfeng Seng[1,2]

Accepted: 25 March 2025

Abstract
The weakly supervised object localization task uses image-level labels to train object localization models. Traditional convolutional neural network (CNN)-based methods usually localize objects using a class activation map. However, the class activation map usually suffers from the problem of activating a small part of the object that is most discriminative. Meanwhile, the methods based on the Vision Transformer can capture long-range feature dependencies but tend to ignore local feature details. In this paper, we innovatively propose a dual-branch contrastive learning (DBC) method that consists of a Transformer and a CNN branch. The method can effectively separate the background and foreground of an image and fuse the features of Transformer and CNN through contrastive learning. Specifically, the method separates the background and foreground representations of the image using the initially generated class-agnostic activation maps. Then, the representations of the same image from different branches form positive pairs for contrastive learning. The background and foreground representations from the same branch form negative pairs. Finally, the DBC method forces the model to separate the background and foreground representations through negative contrastive loss and makes the model fuse the features of two branches through positive contrastive loss. Experiments on the ILSVRC benchmark show that the proposed method can achieve a Top-1 localization accuracy of 59.9% and a GT-known localization accuracy of 71.7%, which are better metrics than those of the state-of-the-art methods with the same parameter complexity.

Keywords Deep learning · Computer vision · Weakly supervised object localization · Dual-branch network · Contrastive learning

1 Introduction

The object localization models based on convolutional neural networks (CNNs) require a large amount of accurate annotations (i.e., bounding box labels) for training. However, obtaining many complex annotations is time-consuming and labor-intensive. To solve this problem, weakly supervised object localization (WSOL) has received increasing attention because it enables the training of localization models using only image-level labels [1]. The key to weakly supervised object localization is enabling the framework to learn the complete target region from image-level supervision signals during the learning process [2–4]. The current methods are primarily based on either CNNs or vision transformer architectures, each having its own advantages and disadvantages.

CNN-based WSOL methods typically utilize a class activation map (CAM) [5] to generate object localization maps for estimating bounding boxes. However, the activated object region generated by CAM is generally smaller than the actual object region because the model trained for classification tends to focus on the most discriminative regions [6, 7]. Many studies have been conducted to address this problem, e.g., adversarial erasing [8–10], divergent activation [11–13] and gradient-based CAM [14–16]. However, the inherent inability of CNNs to capture the dependency of long-range features is complicated and has not been fundamentally solved [7]. CNN extracts features through convolution operations, which enables it to extract local features effectively, but it has difficulty capturing global cues (Fig. 1).

✉ Dong Li
lidong@qhu.edu.cn

[1] School of Computer Technology and Application, Qinghai University, Xining 810000, China

[2] Intelligent Computing and Application Laboratory of Qinghai Province, Xining, China

Published online: 03 April 2025

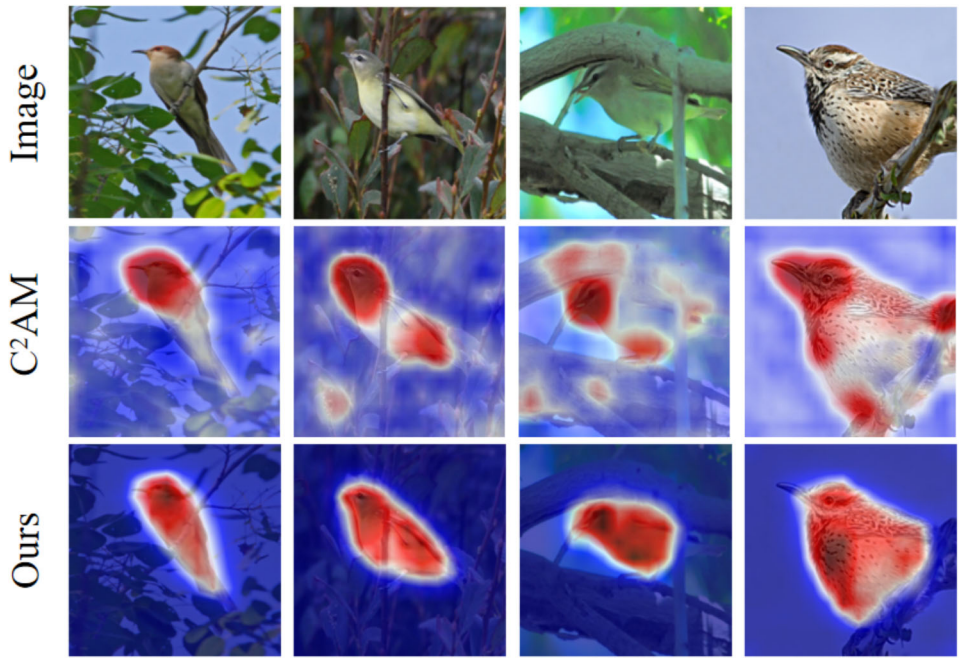

Fig. 1 Comparison of activation maps between our method and C^2AM method

With the success of the vision transformer (ViT) [17] in the field of computer vision, researchers have realized its ability to capture global cues. ViT divides the image into patch tokens with positional embeddings and then processes them through a cascaded block sequence containing a self-attention mechanism [18] and a multilayer perceptron (MLP). Benefiting from the self-attention mechanism, ViT can learn long-range semantic correlations adaptively. Gao et al. [7] proposed a token semantic coupled attention map (TS-CAM) to introduce ViT into WSOL. TS-CAM generates semantic-aware localization by integrating the semantic-agnostic attention map of ViT with semantic-aware CAM, resulting in a comprehensive activation domain. Unfortunately, ViT easily ignores local feature details, making it susceptible to irrelevant background interference. Since CNNs and ViT each have strengths and weaknesses, some studies have proposed leveraging the advantages of both CNNs and ViT. Peng et al. [19] proposed Conformer, which is a hybrid concurrent dual-branch structure with a Transformer branch and a CNN branch for target classification. As CNNs and Transformers have the ability to capture features at different levels, the Conformer designs a feature coupling unit that enables the fusion of the CNN and Transformer features. However, the Conformer lacks fusion from the perspective of the object localization activation map, as it is a classification network.

In summary, CNNs prioritize local features while often neglecting long-range feature associations, whereas Transformers effectively capture long-range features but may overlook local details. By integrating the local feature details of CNNs with the global feature associations of ViTs, we can achieve more precise object localization. However, owing to the inherent structural differences between CNNs and ViTs, simple feature fusion methods may not effectively leverage the advantages of both architectures.

In recent years, contrastive learning methods [20–26] have achieved excellent results in the field of unsupervised classification and have been applied in other fields. Some studies [27–29] on dual-branch contrastive learning indicate that pulling positive sample representations closer in high-dimensional feature space not only facilitates feature fusion but also preserves the distinct characteristics of the features [30]. This is crucial for integrating the features of CNNs and ViTs while highlighting the advantages of both architectures. Furthermore, the approach of pushing negative sample pairs apart effectively enhances feature separation between the foreground and background [31]. Xie et al. [31] proposed cross-image foreground background contrastive learning of class-agnostic activation map (C^2AM) for WSOL. C^2AM uses a similar background or foreground to form a positive pair and uses the background and foreground to form a negative example pair. C^2AM separates the background and foreground via contrastive learning to extract accurate bounding boxes. However, C^2AM is limited by the inherent shortcomings of CNNs. In addition, C^2AM does not consider semantic information when positive pairs are constructed on the basis of similar foregrounds or backgrounds. Although there are several comparative learning methods, there is no comparative learning method using two-branch networks in the WSOL domain.

In this paper, we innovatively propose a dual-branch contrastive learning (DBC) method that uses a hybrid dual-branch network comprising a Transformer branch and a CNN branch. The innovation of DBC lies primarily in employing contrastive learning to achieve a more comprehensive fusion of CNN and Transformer features while simultaneously separating the foreground from the background, thereby enabling accurate localization of the target. Through this approach, we can leverage the efficiency of CNNs in processing local features and spatial information, while also harnessing the advantages of ViTs in capturing the global context and long-range dependencies. Our method provides a novel perspective for the integration of CNNs and ViTs within the framework of WOSL, and it has the potential to achieve enhanced localization performance and improved generalization capabilities in complex scenarios.

Specifically, we generate the background and foreground representations of the image using class-agnostic activation maps extracted independently from the Transformer and CNN branches. Since the Transformer branch and the CNN branch tend to capture global and local features respectively, the same image generates different foreground representations in the Transformer and CNN branches. However, the foreground representations of different branches should contain the same semantic information. Therefore, the representations of the same image from different branches form a positive pair. The background and foreground representations of an image form a negative pair for contrastive learning because they contain different semantic information. The representations of negative pairs are pushed apart by negative contrastive loss to separate the background and foreground regions in the class-agnostic activation map. We design a positive contrastive loss to pull close the representations of positive pairs which can comprehensively fuse local features (CNN) and global features (Transformer). In this method, the attention map of the Transformer branch can learn local feature clues from the activation map of the CNN branch to mitigate the background interference caused by the disruption of spatial topology. In addition, the activation map of the CNN branch can learn global cues from the attention map of the Transformer branch. Comprehensive experiments on the CUB-200-2011 and ILSVRC datasets confirmed the effectiveness of our method.

The contributions of our work are as follows:

- We propose a novel dual-branch contrastive learning (DBC) for WSOL based on the hybrid concurrent dual-branch network. This approach integrates the long-range feature dependency ability of Transformers with the local feature perception ability of CNNs.
- We design a novel positive contrastive loss function that facilitates a more comprehensive integration of CNN and Transformer features by pulling close the representations of positive pairs.
- DBC achieves 59.9% Top-1 localization accuracy and 71.7% GT-known localization accuracy performance on the ILSVRC dataset, which are better than the state-of-the-art methods with the same parameter complexity.

2 Related work

CNN-based approaches for WSOL The objective of the weakly supervised object localization (WSOL) [1] task is to learn object localizations solely on the basis of image-level annotations, without bounding box annotations. Zhou et al. [5] first introduce the CAM into WSOL. CAM produces the object activation map through the last fully connected layer, which includes semantic-aware localization. Nouman Ahmad et al. [32] propose Ghost-UNET++ for automatic CT image dataset segmentation. However, the activated object region generated by CAM is generally smaller than the actual object region because the model trained for classification tends to focus on the most discriminative regions [6, 7]. Numerous studies have been proposed to address the localized activation of CAM.

Some works [8–13] employ an erasing approach, in which a portion of a picture is erased to push the models to focus on expanded object sections. HaS [11] and CutMix [13] randomly erase grid patches from input images. ACoL [9] and ADL [10] adopt adversarial erasing and utilize adversarially trained classifiers to reconstruct missing regions. DANet [12] optimizes object localization through divergent activation. SPG [33] and I2C [34] enforce classification networks to learn pixel correlations from multiple layers through constraints. Some works [14–16] improve the generation of CAM by leveraging backpropagated gradients specific to a particular class. Grad-CAM [15] summarizes the gradient as the importance of neurons in aggregating feature maps. Nouman and Öfverstedt et al. [35] evaluate three-layer CT imaging using Grad-CAM for saliency analysis. Grad-CAM++ [36] improves Grad-CAM by applying pixelwise weighting to the output gradients. BagCAMs [16] proposes the regional localizer generation (RLG) to form the final localization map via effective weighting. Unlike other methods, PSOL [6], SPOL [37], SLT-Net [38] and C^2AM [31] divide WSOL into two independent subtasks: class-agnostic object localization and object classification. C^2AM [31] adopts contrastive learning to disentangle the background and foreground in the class-agnostic object localization task. Many other approaches have also contributed. FAM [39] proposes a module to emphasize foreground objects and a module to explore discriminative parts. Zhu et al. [40] regard WSOL as a domain adaption (DA) task and design a DA-WSOL pipeline to enhance localization performance. Chen

et al. [41] propose a computational method for improving CAM using k-means clustering.

The aforementioned studies have proposed several methods to extend local activation to global activation. However, the inherent defect of CNNs, which tend to capture partial semantic features with local receptive fields, has not been fundamentally resolved. Moreover, it is necessary to explore how to balance image classification and target localization.

Transformer-based approaches for WSOL The vision transformer (ViT) [17] divides the image into patch tokens and sends them to a cascaded block sequence containing the self-attention mechanism [18]. Owing to the self-attention mechanism, ViT can learn long-range semantic correlations adaptively. However, it fails to capture local feature details and neglects spatial coherence. Numerous studies have been conducted to address these issues. TS-CAM [7] first introduces ViT into WSOL and proposes a semantic coupling strategy using Deit [42] as the backbone. TS-CAM generates semantic-aware localization by integrating the semantic-agnostic attention map of ViT and semantic-aware CAM. Although TS-CAM obtains a larger activation domain, it suffers from background interference. To address this issue, TRT [43] proposes a re-attention mechanism to inhibit the effects of background interference. The Token Priority Scoring Module (TPSM) within the TRT framework generates context-aware features through cumulative importance sampling. The context-aware module is also utilized to generate discriminative features. LCTR [44] considers cross-patch information and designs a cue-digging module to improve the local perception ability of global features in the presence of long-range feature dependencies. SCM [45] implicitly optimizes attention representations of the Transformer and generates accurate activation maps on the basis of spatial and contextual coherence.

Although the Vision Transformer addresses the global activation issue, it still faces several challenges, such as capturing local feature details and addressing irrelevant background interference caused by the destruction of spatial topology.

Contrastive learning Contrastive learning operates on the principle of pulling close samples from the positive pair and pushing apart samples from the negative pair in feature space [20–26]. On the basis of object class labels, a positive pair is formed by samples from the same class, whereas a negative pair is formed by samples with different class labels [46–48]. Wang et al. [48] propose two crucial properties of contrastive loss, i.e., the alignment of features from positive pairs and the uniformity of the induced distribution of the normalized features on the hypersphere. C^2AM [31] constructs negative pairs using cross-image foreground and background representations. The positive pair is formed by similar background and background or foreground and foreground representations from two different images. However, C^2AM considers only the similarity of the foreground representations and ignores the semantic information.

In other fields, methods that employ results generated by two network branches for contrastive learning exit. Xiang and Chen [27] achieve credible underwater image enhancement results by constraining semantic information consistency between the clear image domain and the degraded image domain through contrastive learning. Wang et al. [49] utilize a contrastive learning branch to enhance the extraction of discriminative features in the identification of abnormal fasteners. Tian and Sun [50] introduce a cluster-based dual-branch contrastive learning framework that synergistically combines cluster-based unsupervised domain adaptation with contrastive learning to mitigate the effects of upper-body clothing color on person re-identification. Zhang et al. [28] used a dual-branch contrastive method to compare each of the two generated views with the original graph and then jointly optimized them for graph data classification. Mansoor Hayat and Supavadee Aramvith [51] enhance endoscopic image super-resolution by promoting complex interactions between features extracted from the left and right views of endoscopic images. Chen et al. [29] propose a two-branch long-tail recognition method consisting of an imbalanced learning branch and a contrastive learning branch for long-tailed dataset classification. However, relatively few methods have been proposed specifically for the WSOL field. Moreover, the aforementioned approaches primarily employ the same network architecture for the branch networks, which does not adequately capitalize on the strengths of CNNs and ViTs.

3 Methodology

In this section, we first present the preliminaries for the hybrid dual-branch network Conformer. Then, we provide a detailed description of our DBC method, as shown in Fig. 2.

3.1 Hybrid concurrent dual-branch network

For the Conformer [19], an Image $I \in \mathbb{R}^{3\times W\times H}$ is fed into the stem module, which consists of convolution and max pooling layers. The stem module extracts initial local features F^0 and feeds them to the dual branches. The Transformer branch, similar to ViT [17], constructs N patch tokens $T_N \in \mathbb{R}^{N\times D}$ by projecting and flattening F^0 and a class token $T_{cla} \in \mathbb{R}^{1\times D}$, where D denotes the dimension of the tokens. The tokens are fed into a sequence of L cascaded Transformer blocks. Finally, the class token T_{cla}^L is taken out and provided as input to the MLP classifier. l indicates the

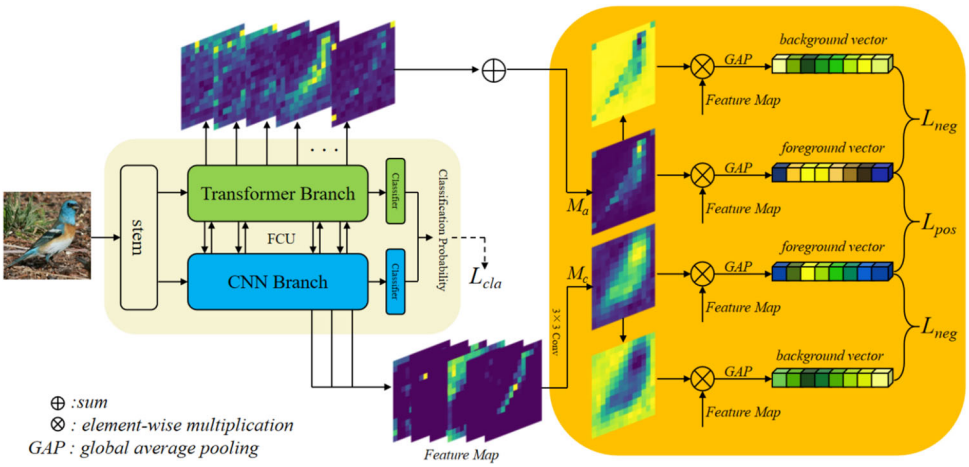

Fig. 2 Overview of the proposed DBC, which consists of a Conformer backbone for feature extraction and a part of dual-branch contrastive learning

class token of the l-th Transformer block. The classification probability of the Transformer branch is calculated as

$$p_t = \text{softmax}(\text{MLP}(T_{cla}^L)), \tag{1}$$

where $p_t \in \mathbb{R}^{1\times S}$ and S denotes the number of classes. MLP$(\cdot)$ denotes the classification function implemented by the MLP block.

The CNN branch comprises L cascaded convolution blocks. Owing to the misalignment of features between the CNN and Transformer blocks, each CNN block establishes interactions with the Transformer block through a feature coupling unit (FCU) to effectively couple the local features and global representations. For the output of the l-th convolution block $F^l \in \mathbb{R}^{c\times w\times h}$, FCU first applies a 1×1 convolution to align the channel dimensions of the patch embeddings for F^l, followed by downsampling to form the patch tokens. Finally, the class token T_{cla}^L from the Transformer branch is spliced and the patch embedding is added. For the patch tokens from the Transformer block, FCU also aligns their dimensions with the feature maps using a 1×1 convolution. The tokens are subsequently upsampled and added to the feature maps. By facilitating the interaction between the feature maps and patch embeddings in this way, the FCU enables the integration of features from both CNNs and ViTs. Finally, the CNN features F^L are pooled and fed to another classifier. The classification probability of the CNN branch is calculated as

$$p_c = \text{softmax}(\text{MLP}(\text{GAP}(F^L))), \tag{2}$$

where $p_c \in \mathbb{R}^{1\times S}$. GAP$(\cdot)$ denotes the global average pooling function. The prediction results are a simple summary of the outputs from the two classifiers, and the classification loss function is defined as

$$\mathcal{L}_{cla} = \text{CrossEntropyLoss}(\frac{p_t + p_c}{2}, y), \tag{3}$$

where $y \in \mathbb{R}^{1\times S}$ denotes the ground truth label and CrossEntropyLoss$(\cdot)$ denotes the cross-entropy loss function.

3.2 Dual-branch contrastive learning

Although the Conformer uses the FCU to couple the CNN and Transformer features, it lacks fusion from the perspective of the object localization activation maps. The proposed dual-branch contrastive learning (DBC) method combines the localization maps of the dual branches by constructing foreground features as positive pairs. Specifically, it first extracts attention maps from the Transformer branch. Moreover, it extracts foreground feature maps from the CNN branch and generates foreground activation maps through convolution. Decoupling the attention and activation maps yields the background and foreground vectors for the CNN and VIT branches, respectively. The foreground vectors of the different branches form a positive pair, whereas the background and foreground vectors form negative pairs. Contrastive learning by pushing apart the negative pairs and pulling close the positive pairs in the feature space.

Attention map We denote $t^l \in \mathbb{R}^{(N+1)\times D}$ as the input of the l-th Transformer block. The attention matrix $M^l \in \mathbb{R}^{h\times(N+1)\times(N+1)}$ of the multihead self-attention module for the l-th Transformer block is calculated as

$$M^l = \text{softmax}\left(\frac{Q^l K^{l\top}}{\sqrt{D/h}}\right), \tag{4}$$

where Q^l and K^l denote the queries and keys projected from the input t^l, respectively. $\top$ is the transpose operator and h indicates the number of self-attention heads.Then, we extract the class token vector $A^l \in \mathbb{R}^{1\times N}$ from the attention matrix M^l. Since the CNN and ViT branches have different feature scales, we normalize the attention map. The final class-agnostic attention map is defined as

$$M_a = \text{norm}(\Gamma^{w\times h}(\frac{1}{L}\sum_l A^l)), \tag{5}$$

where $\Gamma^{w\times h}$ denotes the reshape operation which converts the attention vector to the attention map $M_a \in \mathbb{R}^{w\times h}$. norm$(\cdot)$ denotes the normalization function.

Foreground activation map The output of the last convolution block F^L is applied to classification. To separate the classification and location tasks, we extract high-level feature maps $F^* \in \mathbb{R}^{D\times w\times h}$ from the previous convolution blocks. The class-agnostic foreground activation map is defined by convolution as

$$M_c = \text{norm}(\sum_d F_d^* * k_{1,d}), \tag{6}$$

where $d \in \{1, 2, ..., D\}$ denotes the d-th feature map. $k_{1,d} \in \mathbb{R}^{1\times D\times 3\times 3}$ denotes the convolution kernel weights. Since M_c characterizes the foreground region, the class-agnostic background activation map can be calculated as $(1 - M_c)$. The final output class-agnostic activation map, which is the localization of the object, is calculated as

$$M = M_a + M_c. \tag{7}$$

We utilize a special pooling operation to disentangle the feature maps F^* into background and foreground feature vectors. The foreground activation vector $\mathbf{f}^c \in \mathbb{R}^{1\times D}$ and background activation vector $\mathbf{b}^c \in \mathbb{R}^{1\times D}$ of the CNN branch are formulated as

$$\begin{aligned} \mathbf{f}^c &= \text{GAP}(F^* \otimes M_c), \\ \mathbf{b}^c &= \text{GAP}(F^* \otimes (1 - M_c)), \end{aligned} \tag{8}$$

where $\otimes$ denotes an element-wise multiplication operation. The attention map can also be disentangled into a foreground vector $\mathbf{f}^t \in \mathbb{R}^{1\times D}$ and background vector $\mathbf{b}^t \in \mathbb{R}^{1\times D}$ as follows:

$$\begin{aligned} \mathbf{f}^t &= \text{GAP}(F^* \otimes M_a), \\ \mathbf{b}^t &= \text{GAP}(F^* \otimes (1 - M_a)). \end{aligned} \tag{9}$$

Contrastive loss There is a significant disparity in semantic information between the foreground and background. This contrast holds true for the foreground and background of different images as well. Therefore, for both the within-image and cross-image, the distance between the foreground and background features in a high-dimensional space should be greater. The representations of the foreground and background from any image form a negative pair, i.e., $(\mathbf{f}^c, \mathbf{b}^c)$. Consequently, there should be a greater separation between the background representations and foreground representations in the feature space. The negative contrastive loss is designed as

$$\mathcal{L}_{neg} = \mathcal{L}_{neg}^c + \mathcal{L}_{neg}^t, \tag{10}$$

$$\begin{aligned} \mathcal{L}_{neg}^c &= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\log(1 - \text{sim}(\mathbf{f}_i^c, \mathbf{b}_j^c)), \\ \mathcal{L}_{neg}^t &= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\log(1 - \text{sim}(\mathbf{f}_i^t, \mathbf{b}_j^t)), \end{aligned} \tag{11}$$

where sim$(\cdot)$ is a function used to calculate cosine similarity. $\mathcal{L}_{neg}$ is composed of the loss function of the CNN branch $\mathcal{L}_{neg}^c$ and Transformer branch $\mathcal{L}_{neg}^t$. When $i = j$ represents the within-image, when $i \neq j$ represents the cross-image.

Traditional contrastive learning methods create positive pairs by applying different image augmentation techniques to the same image. In the dual-branch network, the CNN and Transformer branches possess inherent characteristics that predispose them to capture local and global features, respectively. As a result, these two branches produce distinct foreground representations for the same image. Nevertheless, the semantic information contained in the foreground representations of identical images should be consistent. Therefore, the foreground features from different branches of the same image form positive pairs with the same semantics. The positive contrastive loss is formulated as

$$\mathcal{L}_{pos} = -\frac{1}{n}\sum_{i=1}^{n}\log(\text{sim}(\mathbf{f}^c, \mathbf{f}^t)). \tag{12}$$

By employing this approach to construct positive pairs, the integration of two branch features can be enhanced comprehensively. The attention map of the Transformer branch can learn local feature clues from the CNN branch, whereas the activation map of the CNN branch can learn global clues from the Transformer branch.

Algorithm 1 Workflow of DBC on conformer.

Require: Input image I, Backbone $conformer(\cdot)$
1: Calculating classification probability p with backbone $conformer(I)$.
2: Generating attention map M_a by (5).
3: Obtaining high-level feature maps F^* with convolution blocks.
4: Generating foreground activation map M_c by (6).
5: Generating final localization map M by (7).
6: Generating foreground vectors and background vectors by (8) and (9).
7: Calculating positive contrastive loss $\mathcal{L}_{pos}$ and negative contrastive loss $\mathcal{L}_{neg}$ by (10) and (12).
8: Calculating final loss $\mathcal{L}$ and backward propagating.
Ensure: Localization Map M, Classification Probability p.

The final loss function $\mathcal{L}$ is defined as

$$\mathcal{L} = \mathcal{L}_{cla} + \lambda_{neg} * \mathcal{L}_{neg} + \lambda_{pos} * \mathcal{L}_{pos}, \tag{13}$$

where λ_{neg} and λ_{pos} represent the weights of loss function. The negative contrastive loss $\mathcal{L}_{neg}$ is applied to direct the class-agnostic activation map from the dual branch to separate the background and foreground regions. Additionally, the final class-agnostic activation map fuses the features of the CNN and ViT branches using positive contrastive loss $\mathcal{L}_{pos}$. Algorithm 1 illustrates the workflow of DBC for object localization based on a Conformer model.

4 Experiments

4.1 Experimental settings

Dataset We evaluate the proposed method on two commonly used benchmarks, i.e., CUB-200-2011 [52] and ILSVRC [53]. CUB-200-2011 is a fine-grained bird dataset consisting of images of 200 bird species. It includes a training set and a test set. The training set and test set contain 5,994 and 5,794 images, respectively. ILSVRC consists of a training set of over 1.2 million images and a validation set of 5,000 images, encompassing 1,000 categories. The model is fine-tuned using only category labels from the training set and evaluated on the validation set.

Evaluation metrics Following [5, 54], we adopt the Top-1/Top-5 localization accuracy (Top-1/Top-5 $Loc.Acc$), GT-known localization accuracy (GT-known $Loc.Acc$) and maximal box accuracy (MaxBoxAccV2) [55] as evaluation metrics. For GT-known localization, a prediction is considered correct when the intersection over union (IoU) between the predicted box and the ground-truth box is greater than 50%. Top-1/Top-5 localization is correct when the predicted Top-1/Top-5 classification is correct and when the GT-known localization is correct.

Implementation details The Conformer [19] is adopted as the backbone and pre-trained on ILSVRC. The input images are resized to 256×256 pixels and randomly cropped to 224×224 pixels. We apply AdamW [56] with ϵ=1e-8, β_1=0.9, β_2=0.99 and a weight decay of 5e-4 during training. For CUB-200-2011, we use a batch size of 64 and an initial learning rate of 5e-4 to train the model for 60 epochs with one Nvidia Tesla V100 GPU. For ILSVRC, the training procedure lasts 20 epochs with a learning rate of 1e-6 and batch size of 128 on two Nvidia Tesla V100 GPUs. We evaluate the performance of the model on the validation set in each epoch. The model parameters of the best Top-1 $Loc.Acc$ performance will be saved.

4.2 Performance

Localization performance As shown in Table 1, we compare our method using Conformer-Ti and Conformer-S [19] as the backbones with the state-of-the-art (SOTA) methods on the CUB-200-2011 dataset. Our method employing Conformer-S as a backbone achieves Top-1, Top-5, and Gt-Known localization accuracy of 80.9%, 94.1% and 97.3%, respectively. Notably, our method also achieves 90% Gt-Known $Loc.Acc$ on the Comformer-Ti backbone. Compared with the baseline method TS-CAM [7] on the Conformer-S backbone, our method outperforms it by 4.0%, 3.3% and 3.3% in terms of Top-1 $Loc.Acc$, Top-5 $Loc.Acc$ and Gt-Known $Loc.Acc$. Furthermore, our method is superior to the single-stage CNN-based methods and ViT-based methods in all the metrics. DBC outperforms the two-stage SOTA method (C^2AM) in terms of the Gt-Known $Loc.Acc$ and Top-5 $Loc.Acc$.

Table 2 compares DBC with other methods on the ILSVRC. Our method achieves Top-1 $Loc.Acc$ of 59.9% and 59.4% are obtained for the Conformer-Ti and Conformer-S backbones, respectively. In addition, DBC using Conformer-Ti as the backbone achieves the GT-known $Loc.Acc$ of 71.7% and Top-5 $Loc.Acc$ of 69.2%. Compared with TS-CAM, DBC achieves performance gains of 6.5% and 4.1% in terms of Top-1 and Gt-Known $Loc.Acc$. Compared with the SOTA single-stage CNN-based methods and ViT-based methods, DBC outperforms them by 3.8% and 1.1%, respectively, in terms of Top-1 $Loc.Acc$. Compared with the two-stage CNN-based SOTA methods, DBC achieves performance gains of 0.6% for the Top-1 $Loc.Acc$ and 2.7% for the Gt-Known $Loc.Acc$.

In addition, we use the MaxBoxAccV2 metric [55] for a comparison with the other SOTA methods in Table 3. Our method has a significant advantage over other methods, achieving a 4.5% improvement.

Table 1 Localization accuracy on the CUB-200-2011 test set

Method	Backbone	CUB-200-2011		
		Top-1 *Loc*	Top-5 *Loc*	GT-known *Loc*
CAM [5]	VGG16	44.2	52.2	56.0
ACoL [9]	VGG16	45.9	61.0	–
ORNet [57]	VGG16	67.7	80.8	86.2
BAS [58]	VGG16	71.3	85.3	91.1
SPG [33]	InceptionV3	46.6	59.4	–
I^2C [34]	InceptionV3	55.9	68.3	72.6
FAM [39]	InceptionV3	70.7	–	87.3
CREAM [59]	InceptionV3	71.8	86.4	90.4
BagCAMs [16]	InceptionV3	60.1	–	89.8
ADL [10]	ResNet50-SE	62.3	80.3	–
DA-WSOL [40]	ResNet50	66.7	–	81.8
BGC [60]	ResNet50	53.8	65.8	69.9
PSOL [6]	DenseNet161+EfficientNet-B7	77.4	89.5	93.0
SPOL [37]	ResNet50+EfficientNet-B7	80.1	93.4	96.5
C^2AM [31]	DenseNet161+EfficientNet-B7	**83.3**	92.7	94.5
TS-CAM [7]	Deit-S	71.3	83.8	87.7
SCM [45]	Deit-S	76.4	91.6	96.6
TRT [43]	Deit-B	76.5	88.0	91.1
LCTR [44]	Deit-S	79.2	89.9	92.4
Ours	Conformer-Ti	77.2	92.4	97.0
TS-CAM [7]	Conformer-S	77.2	90.9	94.1
Ours	Conformer-S	80.9	**94.1**	**97.3**

The best performance is shown in bold

Table 2 Localization accuracy on the ImageNet-1K validation set

Method	Backbone	ILSVRC		
		Top-1 *Loc*	Top-5 *Loc*	GT-known *Loc*
CAM [5]	VGG16	42.8	54.9	59.0
ACoL [9]	VGG16	45.8	63.3	–
ORNet [57]	VGG16	52.1	63.9	68.3
BAS [58]	VGG16	53.0	65.4	69.6
SPG [33]	InceptionV3	48.6	60.0	64.7
I^2C [34]	InceptionV3	53.1	64.1	68.5
FAM [39]	InceptionV3	55.2	–	68.6
CREAM [59]	InceptionV3	56.1	66.2	69.0
BagCAMs [16]	InceptionV3	53.9	–	71.0
ADL [10]	ResNet50-SE	48.5	–	–
DA-WSOL [40]	ResNet50	55.8	–	70.3
BGC [60]	ResNet50	53.8	65.8	69.9
PSOL [6]	DenseNet161+EfficientNet-B7	56.4	66.5	69.0
SPOL [37]	ResNet50+EfficientNet-B7	59.1	67.2	69.0
C^2AM [31]	DenseNet161+EfficientNet-B7	59.3	66.7	68.2
TS-CAM [7]	Deit-S	53.4	64.3	67.6
TRT [43]	Deit-B	58.8	68.3	70.7
SCM [45]	Deit-S	56.1	66.4	68.8
ViTOL [54]	Deit-B	57.6	–	71.3
LCTR [44]	Deit-S	56.1	65.8	68.7
Ours	Conformer-Ti	**59.9**	**69.2**	**71.7**
TS-CAM [7]	Conformer-S	57.6	65.3	67.1
Ours	Conformer-S	59.4	67.3	69.3

The best performance is shown in bold

Table 3 MaxBoxAccV2 on the CUB-200-2011 dataset

Method	Backbone	MaxBoxAccV2
BagCAMs	InceptionV3	76.9
BGC	ResNet50	75.9
C^2AM	ResNet50	83.8
ViTOL	Deit-B	69.2
TRT	Deit-B	82.1
Ours	Conformer-Ti	86.6

Parameter complexity Table 4 shows the parameters of different methods on the CUB-200-2011 dataset. Compared with the elaborate methods, our method is simpler and only requires an additional convolutional layer. This indicates that only a few extra parameters are required for superior results. Under similar parameter complexity, our method outperforms TS-CAM. Compared with the two-stage methods (e.g., PSOL [6] and C^2AM [31]), our method employs only a single backbone and has significantly fewer parameters. Specifically, DBC achieves better GT-known $Loc.Acc$ than C^2AM, which uses only approximately 25% of the parameters.

Visualization In Fig. 3, we visualize the final localization maps of CAM [5], TS-CAM [7] and our method on the CUB-200-2011 and ILSVRC datasets which are based on Conformer-S [19]. As shown in the figure, although the Conformer fuses the features of the Transformer and CNN via the FCU module, CAM also exclusively focuses on the most discriminative region of the object. This finding indicates that relying solely on the FCU to fuse features is insufficient for the WSOL on the Conformer. In contrast to CAM, due to the destruction of spatial topology, TS-CAM is susceptible to background interference despite its ability to capture long-range features. Furthermore, since TS-CAM combines semantic-aware CAM in a simple way, the problem of local activation is not completely solved. Compared with the former methods, our method integrates local feature details to achieve precise localization while also capturing long-distance cues more effectively. This approach not only activates the entire object region but also prevents background interference. In addition, through comparative learning, our method can significantly differentiate between the background and foreground compared with TS-CAM.

We visualize the activation maps from the CNN branch and the attention maps from the Transformer branch at different epochs, as shown in Fig. 4. The CNN branch gradually learns long-range feature associations from the Transformer branch (upper row of first image). In the second image, the attention map of the Transformer branch shows a gradual decrease in activation for irrelevant foreground elements (such as straw), whereas in the third image, the activation level for the foreground targets increases. This indicates that the attention maps of the Transformer branch learn local feature details from the CNN branch and mitigate background interference. This demonstrates that our foreground contrastive learning approach facilitates mutual learning of the necessary long-range and local features between the CNN and Transformer branches.

4.3 Ablation study

In this section, we conduct ablation studies to evaluate the effectiveness of our method, utilizing Conformer-S as the backbone. In Table 5, we compare our methods under different hybrid strategies and architectures, including a vertical hybrid network that first extracts features using ResNet before feeding them to ViT, as well as a straightforward architecture that employs ResNet and ViT branches without any feature fusion. Our method achieves good localization performance with the simple vertical hybrid network, achieving a 94.7% GT-known $Loc.Acc$ on CUB and 66.8% on ILSVRC; however, it is constrained by the classification performance. In the straightforward parallel network using ResNet and ViT, the lack of feature fusion results in significant misalignment between the features from the two branches.

In Table 6, we explore the effects of the loss functions. Compared with only using the classification loss $\mathcal{L}_{cla}$, the employment of positive contrastive loss $\mathcal{L}_{pos}$ or negative contrastive loss $\mathcal{L}_{neg}$ can significantly improve the localization performance of class-agnostic activation maps. A better per-

Table 4 Comparison of parameters on the CUB-200-2011 dataset

Method	Backbone	#Params(M)	Top-1 *Loc*	GT-known *Loc*
CAM	VGG-16	19.6	44.2	56.0
TS-CAM	Deit-S	25.1	71.3	87.7
C^2AM	DenseNet161+ EfficientNet-B7	94.8	83.3	94.5
Ours	Conformer-Ti	23.0	80.9	97.3

Table 5 Comparison of different architectures

Backbone	CUB-200-2011		ILSVRC	
	Top-1 *Loc*	GT *Loc*	Top-1 *Loc*	GT *Loc*
Resnet50+ViT-S (Vertical Hybrid)	72.2	94.7	55.4	66.8
Resnet50+ViT-S (Dual Branch)	69.0	91.6	51.7	61.4
Conformer-Ti	77.2	97.0	59.9	71.7
Conformer-S	80.9	97.3	59.4	69.3

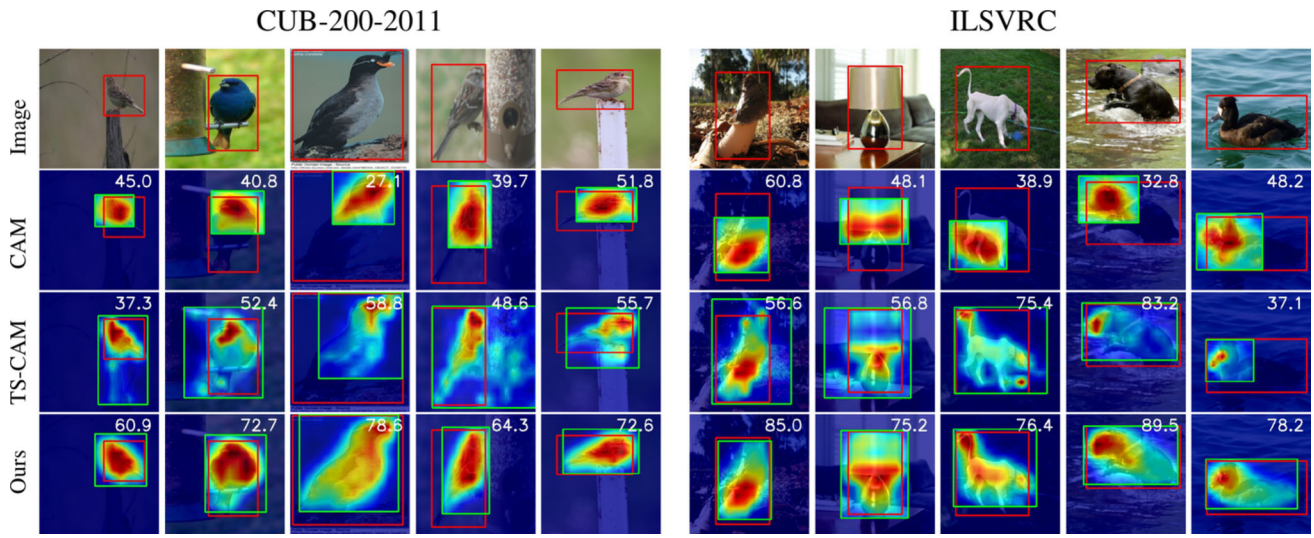

Fig. 3 Visualization of different methods on CUB-200-2011 and ILSVRC datasets

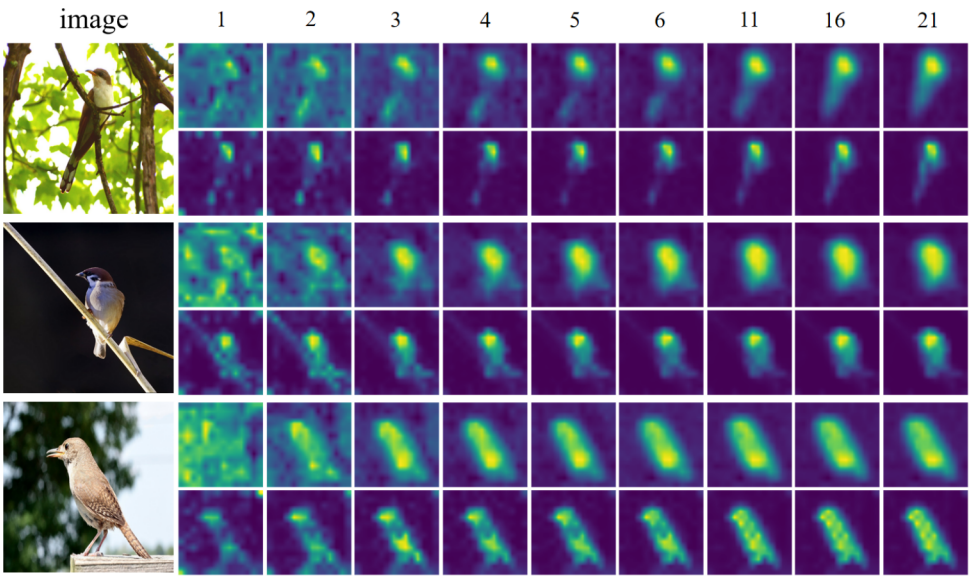

Fig. 4 Visualization of the activation maps of the CNN branch and the attention maps of the Transformer branch at different epochs

Fig. 5 Visualization of the effects of loss functions on CUB-200-2011 dataset

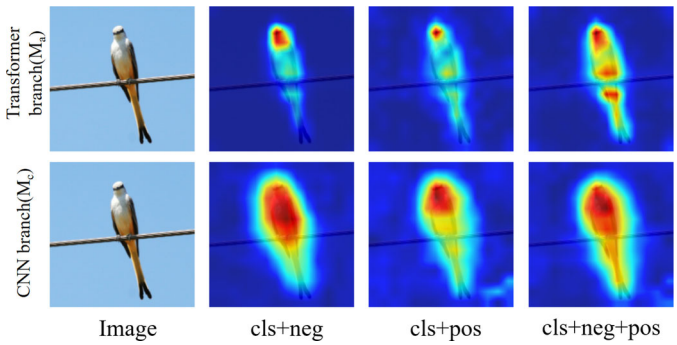

Table 6 The effects of loss functions on CUB-200-2011 test set

$\mathcal{L}_{cla}$	$\mathcal{L}_{pos}$	$\mathcal{L}_{neg}$	Top-1 *Loc*	Top-5 *Loc*	GT-known *Loc*
√			79.2	91.9	94.8
√		√	80.4	93.5	96.7
√	√		80.1	93.4	96.5
√	√	√	**80.9**	**94.1**	**97.3**

The bold text indicates the best performance

formance can be achieved by using both positive contrastive loss and negative contrastive loss. This suggests that the two losses independently improve model localization accuracy through different mechanisms. As illustrated in the second column of Fig. 5, without the positive contrastive loss, the activation patterns between the foreground activation map of CNN branch and the attention map of Transformer branch exhibit significant divergence. This demonstrates that the positive contrastive loss critically guides the model to coherently integrate long-range dependencies (captured by the Transformer) with localized details (extracted by the CNN). When the negative contrastive loss is removed (third column), the CNN and Transformer branches exhibit similar activation patterns. However, this configuration demonstrates incomplete discrimination between foreground and background regions(e.g., background interference observed in the tail region). This observation suggests that the negative contrastive loss plays a critical role in guiding the model to disentangle foreground features from background by explicitly enforcing divergence between foreground and background representations. By utilizing both positive contrastive loss $\mathcal{L}_{pos}$ and negative contrastive loss $\mathcal{L}_{neg}$, it is possible to integrate long-range features and local details while emphasizing the foreground object.

We constructed negative pairs using foreground and background samples from all images. In Table 7, we compare this approach (cross-image) to one that uses only the foreground and background from the same image to form negative pairs (within-image). When negative pairs are constructed using only the foreground and background from the same image, the ability of the model to distinguish between the foreground and background weakens due to a lack of sufficient negative samples for comparison.

Table 7 Performance of different negative pair construction methods for contrastive learning using Conforme-S as the backbone

Negative pair	Top-1 *Cls*	Top-1 *Loc*	GT-known *Loc*
Cross-image	82.7	80.9	97.3
Within-image	82.3	75.4	91.4

Table 8 Localization accuracy of class-agnostic activation maps from different branches on CUB-200-2011 test set

Branch	Top-1 *Loc*	Top-5 *Loc*	GT-known *Loc*
Transformer(M_a)	81.1	94.3	97.6
CNN(M_c)	80.4	93.5	96.9
Transformer+CNN($M_a + M_c$)	80.9	94.1	97.3

Table 8 presents a comparison of the localization accuracy achieved by the class-agnostic activation maps obtained from the different branches. From the experimental results, there is little difference in using the class-agnostic activation maps from the Transformer or CNN branch alone as the output. This indicates that both branches learn features from another branch. The output combines the localization results of the dual branches by adding the category-agnostic activation maps together. This allows the output to consider the features of both the Transformer and CNN branches.

We evaluate the effects of hyperparameters λ_{pos} and λ_{neg} in Fig. 6. λ_{pos} and λ_{neg} are the weights of the contrastive losses $\mathcal{L}_{pos}$ and $\mathcal{L}_{neg}$, respectively. Our method achieves the best performance when $\lambda_{pos} = 1.0$ and $\lambda_{neg} = 0.1$. DBC is sensitive to changes in the weight of the loss function. Furthermore, we compare the performance under different convolution kernel sizes in Table 9. The experimental results show that a convolution kernel size of 3×3 achieves better performance.

4.4 Discussion of details

Computational efficiency We present the parameter count, MACs, memory usage, and training speed of the different backbones in the Table 10. Our method employs Conformer-Ti and Conformer-S as the backbones, with parameter counts of 23.0 million and 36.6 million, respectively. The number of multiply accumulate operations (MACs) is 5.2 billion for Conformer-Ti and 10.6 billion for Conformer-S. The memory required for inference is 161.8 MB for Conformer-Ti and 309.7 MB for Conformer-S. We trained the Conformer-Ti and Conformer-S models with a batch size of 64, requiring approximately 34 seconds and 55 seconds per epoch, respectively. Compared with the two-stage SOTA method C2AM that employs DenseNet and EfficientNet backbones, our approach achieves a reduced total parameter count (23.0 vs. 28.5+66.3 = 94.8 M). Crucially, C2AM's requirement for two independent networks to perform distinct tasks necessitates concurrent memory allocation for both architectures, whereas our unified framework demonstrates 73% lower memory consumption (161.8 vs. 156.6+448.6 = 605.2 MB). These advantages collectively

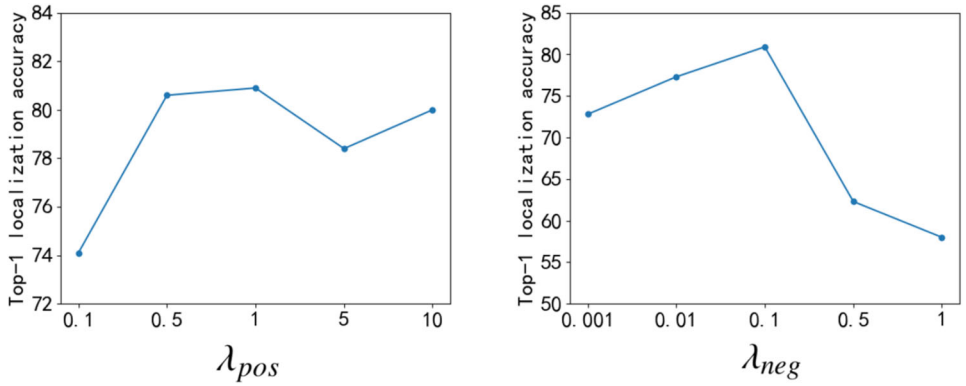

Fig. 6 Evaluation of the hyperparameters λ_{pos} and λ_{neg} on CUB-200-2011 dataset

enable deployment on resource-constrained devices without compromising performance. Also in terms of time to train an epoch our method is shorter overall (34s vs. 101s+55s), which means our method is less expensive to train. The combined parameter/memory/training-time advantages position our method as a cost-effective solution for practical deployment scenarios.

Foreground-background separation results We conduct tests on the augmented PASCAL VOC 2012 [61] dataset provided by SBD [62]. This is a general semantic segmentation dataset that includes 1 background class and 20 foreground object categories. We also trained using only class labels to generate background activation maps. In Table 11, we calculate the intersection over union (IoU), precision, and recall for the foreground and visualize the results in Fig. 7. The baseline method C^2AM constructs positive and negative pairs solely from the features in the CNN. This limitation results in insufficient attention to long-range features. Additionally, C^2AM lacks labels for the foreground regions, preventing it from determining whether the extracted feature vectors represent the foreground or the background. Our method uses attention maps, which typically highlight the activation of the foreground, serving as a guide for distinguishing between the foreground and background in separated features. This enables a clear differentiation between the foreground and background.

Hyperparameter selection Combined with the ablation experiments on the loss function, increasing λ_{pos} can guide the model to improve the integration of the local and global features, whereas increasing λ_{neg} directs the model to focus more on distinguishing between the foreground and background. λ_{pos} and λ_{neg} can be adjusted on the basis of the characteristics of the dataset to accommodate different data distributions. For samples where distinguishing between the foreground and background is more difficult, the value of λ_{neg} can be increased. Conversely, for samples with long-range semantic relationships, the value of λ_{pos} can be increased. Through systematic grid search experiments for hyperparameter optimization, we empirically identified that configurations maintaining a a setting of λ_{pos}:λ_{neg} = 10:1 consistently deliver superior localization performance. This phenomenon can be attributed to the asymmetric sample in contrastive pair construction: The negative contrastive loss engages all foreground-background feature pairs (N^2), while the positive loss utilizes only cross-branch foreground pairs (N). Consequently, the negative loss necessitating down-weighting to prevent dominance in the multi-task learning framework.

Table 9 Performance under different convolution kernel sizes on CUB-200-2011 test set

Kernel size	Top-1 *Loc*	Top-5 *Loc*	GT-known *Loc*
1 × 1	79.1	93.5	96.7
3 × 3	**80.9**	**94.1**	**97.3**
5 × 5	79.2	92.6	95.6

The bold text indicates the best performance

Table 10 Parameter count, MACs, memory usage, and training time of different backbones

Backbone	#Params (M)	MACs (G)	Memory (MB)	Training Time(64/s)
DenseNet161	28.5	7.8	156.6	101
EfficientNet-B7	66.3	5.3	448.6	55
Conformer-Ti	23.0	5.2	161.8	34
Conformer-S	36.6	10.6	309.7	55

Table 11 Background segmentation performance on the VOC dataset

Method	IoU	Prec.	Recall
C^2AM	85.5	92.6	91.8
Ours	85.9	90.7	94.2

4.5 Error analysis

We analyzed all 127 images with localization errors in the CUB-200 test set(5794 images) in Fig. 8. We categorized the causes of errors as follows: object occlusion, water reflection, partial activation, irrelevant interference, low resolution, multiple instances, and label errors. The occlusion of objects, such as branches, divides the target into two parts, making it difficult to activate the non-discriminative regions of the tail. The reflections on the water surface share the same features as the object, which is difficult to address by class labeling alone. Partial activation is usually caused by omissions for narrow tails. Irrelevant interference is due to misrecognition from co-occurring foregrounds. Because the output localization map is a low-resolution image of size 14×14, it leads to larger activation maps for smaller objects in some of the images when they are interpolated to high resolution. Images with multiple instances and label errors have annotation issues; however, the localization of the objects is correct. In summary, subsequent research directions can be directed toward improving the semantic discontinuity caused by occlusion.

5 Conclusion

In this work, we propose dual-branch contrastive learning on hybrid concurrent dual-branch networks to merge the strengths of Transformer and CNNs for weakly supervised object localization. DBC exploits the consistency of features across different branches. The foregrounds from different branches of the same picture form positive pairs and the background and foreground from different images form negative pairs. The background and foreground regions of the class-agnostic activation map are distinguished by pushing apart the representations of foreground and background in the feature space. DBC effectively integrates the long-range feature dependency and local feature details by pulling close the representations of dual branches to generate accurate localization. Extensive experiments conducted on the CUB-200-2011 and ILSVRC2012 datasets demonstrate that our method presents a viable approach for enhancing localization performance.

Our method has several hyperparameters and is sensitive to the hyperparameters of the loss function. Therefore, future work needs to focus on designing a more effective method

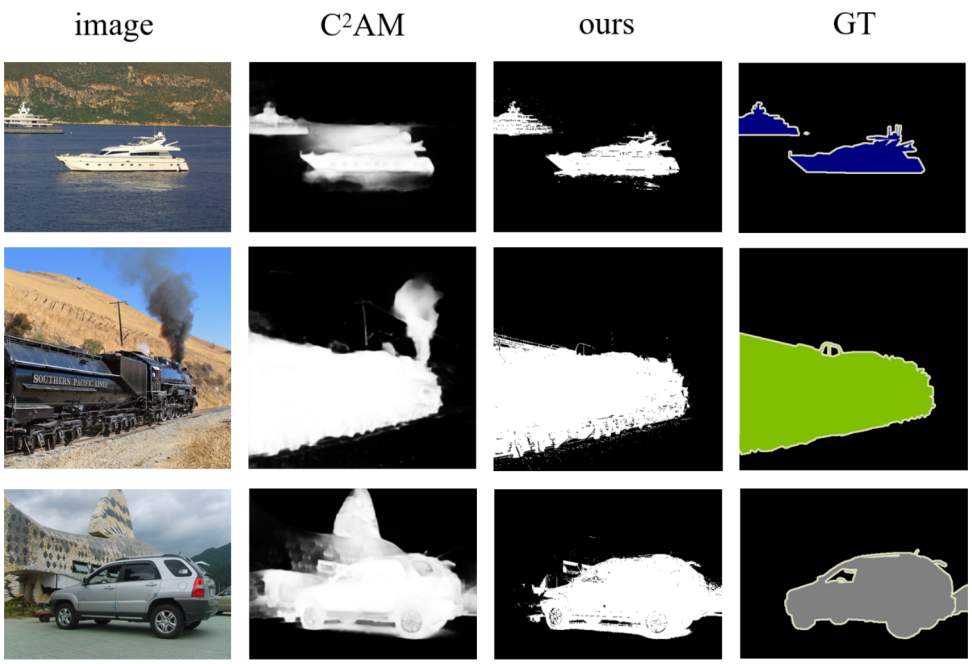

Fig. 7 Visualization of background cues on the VOC dataset

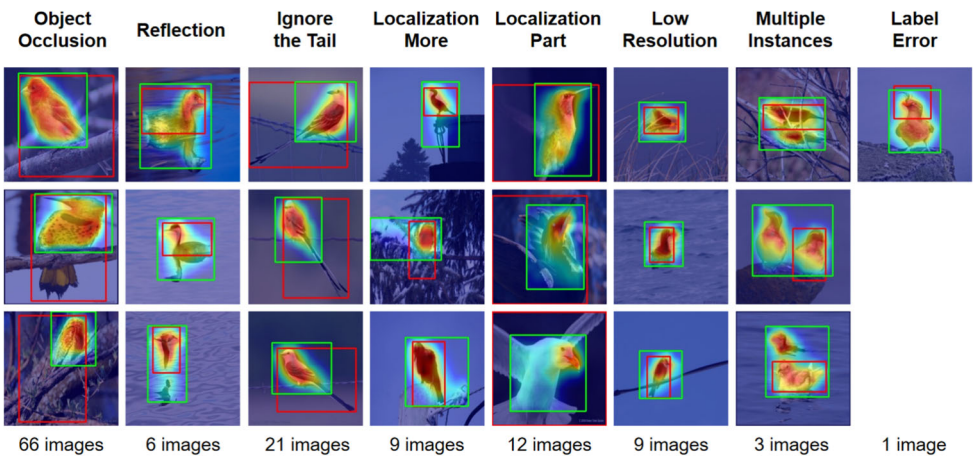

Fig. 8 Visualization of error analysis on CUB on CUB-200-2011

for integrating the loss function. In addition, it is also an important work to improve the semantic discontinuity caused by occlusion.

Author Contributions Zebin Guo: Methodology, Software, Investigation, Formal Analysis, Writing-Original Draft; Dong Li: Conceptualization, Funding Acquisition, Supervision, Writing - Review & Editing; Zhengjun Du: Writing-Review & Editing; Bingfeng Seng: Writing-Review & Editing.

Funding This research was supported by the National Natural Science Foundation of China (No. 62366043); the Subtopic of Qinghai Top Ten Innovation Platform Project (No. ZYYSDPT-2023-02).

Data Availability Not applicable.

Materials Availability Not applicable.

Code Availability Code generated or used during the study are available from the corresponding author by request.

Declarations

Conflict of Interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

1. Zhang D, Han J, Cheng G, Yang M-H (2021) Weakly supervised object localization and detection: A survey. IEEE Trans Pattern Anal Mach Intell 44(9):5866–5885
2. Wu C, Li M, Gao Y, Xie X, Ng WW, Musyafa A (2024) Weakly supervised object localization with background suppression erasing for art authentication and copyright protection. Mach Intell Res 21(1):89–103
3. Zhai W, Wu P, Zhu K, Cao Y, Wu F, Zha Z-J (2024) Background activation suppression for weakly supervised object localization and semantic segmentation. Int J Comput Vis 132(3):750–775
4. Hui W, Gu G, Wang B (2023) Shallow feature-driven dual-edges localization network for weakly supervised localization. Mach Intell Res 20(6):923–936
5. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929
6. Zhang C-L, Cao Y-H, Wu J (2020) Rethinking the route towards weakly supervised object localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13460–13469
7. Gao W, Wan F, Pan X, Peng Z, Tian Q, Han Z, Zhou B, Ye Q (2021) Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2886–2895
8. Mai J, Yang M, Luo W (2020) Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8766–8775
9. Zhang X, Wei Y, Feng J, Yang Y, Huang TS (2018) Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1325–1334
10. Choe J, Shim H (2019) Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2219–2228
11. Kumar Singh K, Jae Lee Y (2017) Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: Proceedings of the IEEE international conference on computer vision, pp 3524–3533
12. Xue H, Liu C, Wan F, Jiao J, Ji X, Ye Q (2019) Danet: Divergent activation for weakly supervised object localization. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6589–6598
13. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y (2019) Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6023–6032
14. Tan C, Gu G, Ruan T, Wei S, Zhao Y (2020) Dual-gradients localization framework for weakly supervised object localization. In: Proceedings of the 28th ACM international conference on multimedia, pp 1976–1984
15. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via

gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
16. Zhu L, Chen Q, Jin L, You Y, Lu Y (2022) Bagging regional classification activation maps for weakly supervised object localization. In: European conference on computer vision, pp 176–192
17. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst. 30
19. Peng Z, Huang W, Gu S, Xie L, Wang Y, Jiao J, Ye Q (2021) Conformer: Local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 367–376
20. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International conference on machine learning, pp 1597–1607
21. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729–9738
22. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, Krishnan D (2020) Supervised contrastive learning. Adv Neural Inf Process Syst 33:18661–18673
23. Wu Z, Xiong Y, Yu SX, Lin D (2018) Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3733–3742
24. Ye M, Zhang X, Yuen PC, Chang S-F (2019) Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6210–6219
25. Tian Y, Krishnan D, Isola P (2020) Contrastive multiview coding. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, 23–28 August, 2020, Proceedings, Part XI 16, pp 776–794
26. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A (2021) Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9650–9660
27. Xiang Y, Chen Z (2023) Dual-branch contrastive learning for image enhancement of underwater internet of things. In: 2023 8th International Conference on Image, Vision and Computing (ICIVC), pp 245–250
28. Zhang H, Cao J, Li K, Wang Y, Li R (2023) Dual-branch contrastive learning for network representation learning. In: International conference on neural information processing, pp 185–197
29. Chen Q, Huang T, Zhu G, Lin E (2023) A dual-branch model with inter-and intra-branch contrastive loss for long-tailed recognition. Neural Netw 168:214–222
30. Ke G, Hong Z, Zeng Z, Liu Z, Sun Y, Xie Y (2021) Conan: contrastive fusion networks for multi-view clustering. In: 2021 IEEE International conference on big data (Big Data), pp 653–660
31. Xie J, Xiang J, Chen J, Hou X, Zhao X, Shen L (2022) C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 989–998
32. Ahmad N, Strand R, Sparresäter B, Tarai S, Lundström E, Bergström G, Ahlström H, Kullberg J (2023) Automatic segmentation of large-scale ct image datasets for detailed body composition analysis. BMC Bioinf. 24
33. Zhang X, Wei Y, Kang G, Yang Y, Huang T (2018) Self-produced guidance for weakly-supervised object localization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 597–613
34. Zhang X, Wei Y, Yang Y (2020) Inter-image communication for weakly supervised localization. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, 23–28 August, 2020, Proceedings, Part XIX 16, pp 271–287
35. Ahmad N, Öfverstedt J, Tarai S, Bergström G, Ahlström H, Kullberg J (2024) Interpretable uncertainty-aware deep regression with cohort saliency analysis for three-slice ct imaging studies. In: Medical imaging with deep learning
36. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp 839–847
37. Wei J, Wang Q, Li Z, Wang S, Zhou SK, Cui S (2021) Shallow feature matters for weakly supervised object localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5993–6001
38. Guo G, Han J, Wan F, Zhang D (2021) Strengthen learning tolerance for weakly supervised object localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7403–7412
39. Meng M, Zhang T, Tian Q, Zhang Y, Wu F (2021) Foreground activation maps for weakly supervised object localization. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3385–3395
40. Zhu L, She Q, Chen Q, You Y, Wang B, Lu Y (2022) Weakly supervised object localization as domain adaption. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14637–14646
41. Chen Z, Sun Q (2023) Extracting class activation maps from non-discriminative features as well. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3135–3144
42. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H (2021) Training data-efficient image transformers & distillation through attention. In: International conference on machine learning, pp 10347–10357
43. Su H, Ye Y, Chen Z, Song M, Cheng L (2022) Re-attention transformer for weakly supervised object localization. In: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, 21-24 November, 2022
44. Chen Z, Wang C, Wang Y, Jiang G, Shen Y, Tai Y, Wang C, Zhang W, Cao L (2022) Lctr: On awakening the local continuity of transformer for weakly supervised object localization. In: Proceedings of the AAAI conference on artificial intelligence, vol 36, pp 410–418
45. Bai H, Zhang R, Wang J, Wan X (2022) Weakly supervised object localization via transformer with implicit spatial calibration. In: European conference on computer vision, pp 612–628
46. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol 1, pp 539–546
47. Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol 2, pp 1735–1742
48. Wang T, Isola P (2020) Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning, pp 9929–9939
49. Wang J, Wang S, Zhao X, Wu J, Li Q (2024) Abnormal fastener recognition via dual-branch supervised contrastive learning network with hard feature synthesis. IEEE Sens J

50. Tian Q, Sun J (2023) Cluster-based dual-branch contrastive learning for unsupervised domain adaptation person re-identification. Knowl-Based Syst. 280:111026
51. Hayat M, Aramvith S (2024) Saliency-aware deep learning approach for enhanced endoscopic image super-resolution. IEEE Access. 12:83452–83465
52. Wah C, Branson S, Welinder P, Perona P, Belongie S (2011) The caltech-ucsd birds-200-2011 dataset
53. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115:211–252
54. Gupta S, Lakhotia S, Rawat A, Tallamraju R (2022) Vitol: Vision transformer for weakly supervised object localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4101–4110
55. Choe J, Oh SJ, Lee S, Chun S, Akata Z, Shim H (2020) Evaluating weakly supervised object localization methods right. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
56. Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv:1711.05101
57. Xie J, Luo C, Zhu X, Jin Z, Lu W, Shen L (2021) Online refinement of low-level feature based activation map for weakly supervised object localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 132–141
58. Wu P, Zhai W, Cao Y (2022) Background activation suppression for weakly supervised object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 14248–14257
59. Xu J, Hou J, Zhang Y, Feng R, Zhao R-W, Zhang T, Lu X, Gao S (2022) Cream: Weakly supervised object localization via class re-activation mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9437–9446
60. Kim E, Kim S, Lee J, Kim H, Yoon S (2022) Bridging the gap between classification and localization for weakly supervised object localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14258–14267
61. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88:303–338
62. Hariharan B, Arbeláez P, Bourdev L, Maji S, Malik J (2011) Semantic contours from inverse detectors. In: 2011 International conference on computer vision, pp 991–998