Applied Intelligence (2024) 54:4689–4712
https://doi.org/10.1007/s10489-024-05403-3

R-CCF: region-aware continual contrastive fusion for weakly supervised object detection

Yongqiang Zhang[1] · Rui Tian[1] · Yin Zhang[1] · Zian Zhang[1] · Yancheng Bai[2] · Mingli Ding[1] · Wangmeng Zuo[3]

Accepted: 12 March 2024 / Published online: 3 April 2024

Abstract

Weakly-supervised learning has emerged as a compelling method for object detection by reducing the fully annotated labels requirement in the training procedure. Recently, some works have treated the detection task as a classification task, resulting in highlighting only discriminative object parts. Moreover, fully-supervised object detectors use specific modules (e.g. feature pyramid networks (FPN) and region proposal network (RPN)) to accurately localize target objects, while weakly-supervised object detectors, such as a well-designed module for object localization, rarely exist. To address the above challenges and gaps, we propose a region-aware continual contrastive fusion (R-CCF) module, which can be plugged into any off-the-shelf weak detector to improve detection performance by refining object location. Specifically, a novel region association (RA) algorithm is proposed to automatically query similarities of the most discriminative regions with their surrounding regions and then to form new rough object locations. Furthermore, we introduce an effective object integration (OI) constraint, including a class sub-constraint and a distance sub-constraint, to refine the rough object locations from the RA algorithm further and achieve accurate object regions. By integrating our R-CCF module into weakly supervised detector architectures and training end-to-end, we can continually refine object locations by contrastively fusing the discriminative regions with surrounding patches. Extensive experiments demonstrate the effectiveness of the proposed method in weakly supervised object detection and show that integrating R-CCF into the state-of-the-art MIST [1] achieves 58.3% in mAP on the PASCAL VOC2007 benchmark, surpassing MIST by 0.2% absolutely. Moreover, R-CCF based on OICR [2] and WSDDN [3] achieve 42.5% and 32.5% in mAP on the PASCAL VOC2007, which is 1.3% and 2.1% higher than the baseline detectors, respectively. We also test the robustness of R-CCF on the PASCAL VOC 2012 dataset, and R-CCF outperforms the baseline methods clearly.

Keywords Object detection · Weakly supervised learning · Contrastive learning · Region refining

1 Introduction

Weakly Supervised Object Detection(WSOD) [1, 3–8] is a task of learning an object detector (i.e. classification and localization) with only image-level labels that indicate the presence or absence of target object classes in an image. More recently, WSOD has attracted considerable attention because it is unnecessary for tight bounding box annotations, which are usually non-existent or unavailable in many practical applications (e.g. specific type airplanes or ships detection). Compared with the fully supervised object detection paradigm [9, 10] which requires object-level annotations (i.e. object bounding boxes) in the training stage, WSOD is more promising but also more challenging to obtain satisfactory performance.

Challenge In WSOD, most of the previous methods follow the multiple instance learning (MIL) pipeline [3, 11–15]to detect objects with only image-level labels. First, they treat the detection task as a region classification task, resulting in the highlighting only discriminative object parts [16–20]. Additionally, to pursue a high recall rate in the absence of ground-truth bounding boxes, region proposals in WSOD methods have to be redundant, so some methods [6, 7, 17, 21] attempt to design the precise region proposals generation strategy for covering the whole object region or at least containing a larger portion of objects. Furthermore, some works [1, 2, 22, 23] are proposed to refine the instance classifier by a set of proposal scores, which guide the weakly

Yongqiang Zhang and Rui Tian contributed equally to this work

✉ Yongqiang Zhang
zhangyongqiang@hit.edu.cn;
yongqiang.zhang.hit@gmail.com

Extended author information available on the last page of the article

supervised detector to focus on the true positive regions. However, to the best of our knowledge, the performance of weakly supervised detectors remains far behind that of fully supervised detectors. Given a complex image that contains multiple instances of objects, especially in the presence of partial occlusion or non-rigid objects, fully supervised object detectors can use specific modules [9, 24–27] to achieve accurate localization, while weakly supervised detectors rarely have a well-designed module to improve the localization performance.

Main idea Can we design a module to bridge the gap between fully-supervised and weakly-supervised detection methods and alleviate the shortcomings of weakly-supervised object detectors (i.e. the problem of highlighting only discriminative object parts)? Thus, in this paper, we propose a region-aware continual contrastive fusion (R-CCF) framework for weakly supervised object detection, which includes two main components: a RA algorithm and an OI constraint. As shown in Fig. 1, for an input image, we first employ an off-the-shelf weakly supervised detector and a region extractor to obtain query regions (the red bounding boxes) and key regions (the yellow bounding box). Then, a RA algorithm is proposed to automatically query the similarities of the most discriminative regions (i.e. query regions) with their surrounding regions (i.e. key regions), and form new rough object regions. Moreover, an OI constraint is further proposed to refine the rough object region and achieve an accurate object region.

Specifically, the RA algorithm examines two kinds of regions: 1) the predicted bounding boxes from baseline detectors, and 2) the surrounding regions. We combine the contrastive learning [28–31] and cluster process [32–36] method to automatically query the similarities of the most discriminative regions (i.e. the predicted bounding boxes from baseline detectors, noted as query regions) with their surrounding regions (i.e. key regions). First, RA algorithm continually improves the feature embedding ability of the vision transformer (ViT) based on the contrastive learning paradigm and optimization strategy as described in Section 3.5. Then, a clustering process with the above trained ViT is introduced to calculate the similarities score of query region and key regions and identify the positive key regions(k+ regions as shown in Fig. 1). Here, these positive key regions and query regions are used to form a new rough object region.

In terms of OI constraint, it refines the rough object regions from the RA algorithm by removing some positive keys in low confidence, which includes a double-check strategy: a class sub-constraint and a distance sub-constraint (described in Section 3.4). The class sub-constraint and distance sub-constraint are designed to measure the correctness of clustering results. Only the positive key regions satisfy class sub-constraint and distance sub-constraint simultaneously, they are selected as the candidates of object regions to form the final accurate object regions.

In Section 4, ablation study experiments prove that the RA algorithm and OI constraint play an important role in our pipeline, and R-CCF is effective for solving the problem of highlighting only the most discriminative areas in WSOD. Note that our region-aware continual contrastive fusion(R-CCF) mines accurate object regions and achieves good results for the following reasons. With the help of contrastive learning and unsupervised cluster in the RA algorithm, we train two encoders (i.e. base encoder and momentum encoder) to extract the high-dimensional features of query and key regions and find high-similarity key regions corresponding to

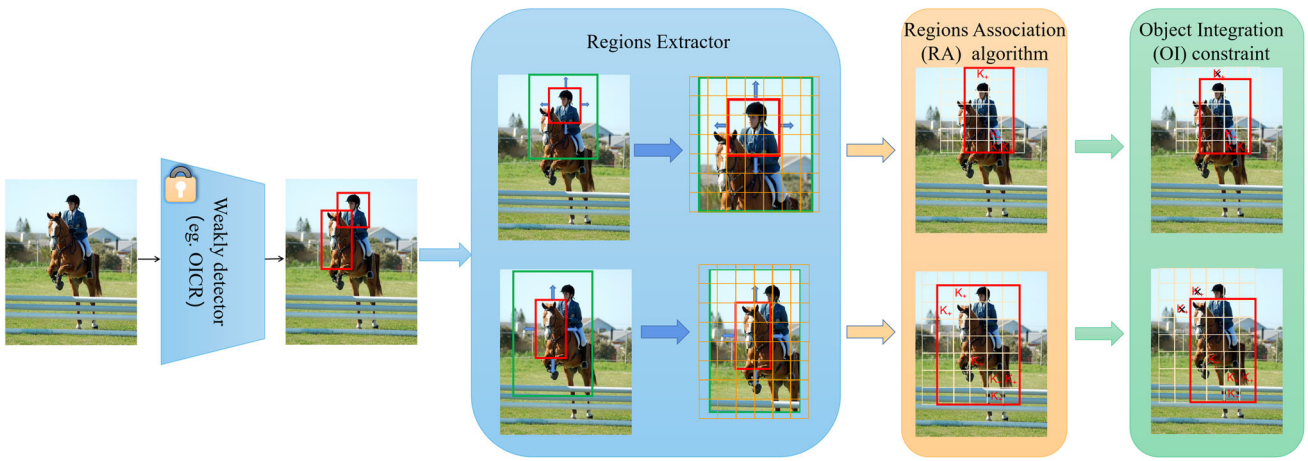


Fig. 1 An illustration of our proposed R-CCF. The query regions and key regions are denoted by red bounding boxes and yellow bounding boxes respectively. We regard clustered positive key regions as k+ regions, and the process of removing some extraneous or background regions is shown by black crosses. Best seen on a computer, in color and zoomed in

query regions. Then, the query regions and key regions with high similarity are fused into a new object region by the similarity score. Moreover, the proposed OI constraint, including the class sub-constraint and distance sub-constraint, is used to obtain complete object regions by preventing the unsteady cluster process and removing false positive key regions. Benefiting from the feature embedding ability of ViT, the clustering process, and the constraints, we also designed a training strategy to improve feature embedding ability and model robustness further.

In summary, we make the following four contributions for weakly supervised object detection in this work:

- A region-aware continual contrastive fusion (R-CCF) framework is proposed to solve the problem of highlighting only discriminative object parts in WSOD, which is used to bridge the performance gap between fully-supervised and weakly-supervised detectors. Our R-CCF can be integrated into any weakly detector architectures, and trained with the results of existing weakly supervised detectors in an end-to-end manner.
- A novel region association (RA) algorithm is proposed to automatically query the similarities of the most discriminative regions and their surrounding regions and to form new rough object regions based on the clustering results. Here, a new multi-region predicted loss with an optimization strategy is designed to improve the embedding feature ability of ViT for query and key regions.
- We introduce an effective constraint to further refine the rough object regions from the RA algorithm, named Object Integration(OI) constraint. Here, OI constraint is proposed as a double-check strategy, i.e. class sub-constraint and distance sub-constraint, to ensure the correctness of clustering results, where some positive key regions with low confidence are removed and the accurate object regions are achieved.
- Extensive experiments demonstrate the effectiveness of the proposed method in WSOD and show that R-CCF achieves 58.3% in mAP on the PASCAL VOC 2007 dataset, which is a state-of-the-art detection performance. Moreover, plugging R-CCF into some popular and classical weakly supervised object detectors further demonstrates the effectiveness of our proposed method, i.e. we outperform OICR by 1.3%. Surprisingly, based on a simple weakly supervised object detector WSDDN, our R-CCF increases the mAP from 30.4% to 32.5% without an iterative refinement strategy. We also test the robustness of R-CCF on PASCAL VOC 2012 dataset, and R-CCF outperforms the baseline methods clearly.

The rest of the paper is organized as follows. We review the related recent literature in Section 2. In Section 3, the detailed methods of our region-aware continual contrastive fusion (R-CCF) for WSOD are presented, which include region extractor, RA algorithm, and OI constraint. Moreover, the multi-region predicted loss function and optimization strategy are described in detail. In Section 4, we first conduct some ablation study analysis to validate the effectiveness of each proposed component in our pipeline, and then we show some experiments compared with previous state-of-the-art on the widely used WSOD datasets(i.e. PASCAL VOC2007 and VOC2012). This is followed by the conclusions and future work in Section 5.

2 Related work

Driven by the great success of contrastive learning and clustering for exploring the similarity and dissimilarity of instance features, a series of methods have been proposed for unsupervised learning in recent years. In this paper, we focus on the problem of highlighting only discriminative parts of objects in WSOD and further solve this intractable problem by combining advanced contrastive learning and clustering techniques with weakly object detection methods. In this section, we review the methods specifically relate to the WSOD task.

2.1 Weakly supervised object detection (WSOD)

Weakly supervised object detection (WSOD) is a task of learning classification and localization information from image-level labels and does not need bounding box annotations. Recently, numerous WSOD methods have been proposed, and competitive performance has been achieved. Here we divide the existing methods into the following categories: initial high-quality region proposal-based methods [6, 7, 38–40], refining instance classifier-based methods [3, 22, 23, 41, 42], mining completely object regions-based methods [16, 17, 43–45] and weakly-supervised to fully-supervised conversion methods [4, 5, 14, 46].

Initial high-quality region proposals-based methods usually improve detection recall while reducing the redundancy before feeding region proposals into weakly supervised object detectors. For example, Gong et al. [7] combine selective search and a gradient-weighted class activation mapping (Grad-CAM) to obtain region proposals that can better envelop the entire objects. Mao et al. [38] employ weighted gradient features on the multiple convolutional layers to generate a gradient-pyramid feature for object localization. The refining instance classifier-based methods are designed to iteratively guide the detector and help it to focus on the true positive regions. For instance, Zeng et al. [23] propose a novel WSOD method, i.e. using bottom-up and top-down objectness distillation to improve the deep

objectness representation of CNN, which could measure the probability of a bounding box including a complete object. Ren et al. [41] propose an instance dual-optimization framework, which optimizes the feature of noisy instances existing in the training samples and focuses on global salient information. Mining completely object regions-based methods is concerned with the relationship between the predicted and surrounding regions. This relationship is then used to obtain a complete region for each target object. Cai et al. [45] propose a novel selective weakly supervised detection to progressively localize the complete object even though the object region is partially occluded and very tiny. In terms of weakly-supervised to fully-supervised conversion methods, they combine the advantages of fully-supervised (e.g. regression ability) and weakly-supervised (e.g. inexpensive training annotation) learning to perform object detection in a weakly manner. Zhang et al. [5] present a novel weakly-supervised to fully-supervised framework, where pseudo ground-truth excavation algorithm and pseudo ground-truth adaptation algorithm are proposed to find accurate pseudo ground-truths of each instance in an image, and the mined pseudo ground-truths are used to train a fully-supervised detector for producing the final detection results.

However, although these methods use high-quality region proposals, refined strategy, mining completely object region methods and weakly-supervised to fully-supervised conversion to detect object regions respectively, the detection results always highlighting only discriminative object parts instead of the whole object region. To solve this problem, some works focus on transferring the confidence of the same class labels from the most discriminative region to surrounding regions (e.g. PCL [22]). In this paper, we propose a region-aware continual contrastive fusion (R-CCF) module to query the similarity of the most discriminative regions and surrounding regions by continual contrastive learning. The surrounding regions with high confidence are considered as parts of the object region to form a new complete and accurate region for target objects. Note that the similarity calculated by our R-CCF denotes the consistency of distance and direction (i.e. euclidean distance and cosine similarity) between regions in the feature space while previous works only consider the spatial adjacency relationship between the most discriminative region and surrounding regions, such as OICR [2], MIST [1].

2.2 Contrastive learning

Contrastive learning is an unsupervised feature representation learning technique, and it extracts the discriminative representations for each image to attract similar (positive) samples and dispel different (negative) samples. Most approaches for contrastive learning are developed based on the instance discriminative task [28–31, 47], and significant performance has been achieved in unsupervised representation learning. For example, He et al. [30] follow a simple instance discrimination task(i.e. a query matches a key if they are encoded into different views of the same image) to build a dynamic dictionary with a queue and a moving-averaged encoder, and design an unsupervised pre-train framework for learning the feature representation. Chen et al. [29] propose a novel self-supervised contrastive learning framework that incrementally detects and explicitly removes the false negative samples. Li et al. [31] bridge contrastive learning with clustering to learn low-level features for the task of instance discrimination, and encode semantic structures discovered by clustering into the learned embedding space. Chen et al. [28] propose a new unsupervised framework for vision Transformer [37] by investigating the effects of several fundamental components and achieve state-of-the-art in various downstream tasks.

Existing contrastive learning methods first set pretext task and train robust feature extractor under unsupervised paradigms, and then the robust feature extractor is widely applied to various downstream tasks. In fact, the above process(i.e. unsupervised contrastive learning and downstream task fine-tuning) is not implemented in an end-to-end manner. Compared with these methods, we introduce contrastive learning into weakly supervised object detection and directly learn the feature representations of the query regions, key regions and fused regions in end-to-end manner. Note that our feature extractor(i.e. ViT) is not fine-tuned or pre-trained. Then, we continually compare the similarity of query and key regions in the clustering process to find similar surrounding regions with the most discriminative regions. The mined similar surrounding regions are further used to form the accurate object regions.

2.3 Deep clustering

Unsupervised clustering is the task of partitioning an unlabeled dataset into different semantic categories where the prior knowledge of a labeled set is not available. With the development of deep learning, some clustering methods [32–36] based on deep learning have been proposed in recent years, which can be roughly divided into two categories. The first category exploits the pairwise similarity of the samples to generate pseudo-labels for clustering [32, 34, 36]. For example, Mathilde et al. [34] propose a novel deep clustering method that jointly learns the parameters of a neural network and the assignments of cluster results. They use the subsequent assignments as supervision to update the network weights. Whereas, the second category [33, 35] uses neighborhood aggregation of feature embedding to bring the similar instances closer and simultaneously push away the dissimilar instances. Wouter et al. [35] propose a two-step approach for the unsupervised clustering process, where

a self-supervised task is employed to obtain semantically meaningful features and further use the obtained features as a prior to bring closer or push away each instance.

Compared with the above two category methods which utilize image-level features to assign images into different clusters, our R-CCF combined with the clustering process assigns the positive key regions to the corresponding object region using region-level features. Our R-CCF queries the similarity of the most discriminative regions and surrounding regions automatically, then the positive surrounding regions are assigned into the same cluster with the most discriminative region based on the clustering results. Finally, similar query regions and key regions are merged into complete object regions as the final detection results.

3 Proposed method

In this section, we describe the details of the proposed region-aware continual contrastive fusion(R-CCF) for WSOD. First, we introduce the pipeline of the proposed R-CCF module, as shown in Fig. 2, which intuitively presents the details of the Region Association(RA) algorithm and the Object Integration(OI) constraint. Then, we describe the process of generating query regions and key regions using region extractor. Additionally, the RA algorithm combines contrastive learning with a clustering process to mine rough object regions. After that, we show that the OI constraint refines the rough object region from the RA algorithm by removing positive keys in low confidence and achieves accurate object regions. Finally, an effective optimization strategy is used to train our R-CCF module in an end-to-end fashion base on the off-the-shelf weakly supervised object detectors.

3.1 Overview of R-CCF module

As shown in Fig. 2, there are three components in the proposed R-CCF framework. The first component is the region extractor, and input images are fed into the off-the-shelf weakly supervised object detector to obtain the most discriminative regions and surrounding regions. The second component is the RA algorithm, which combines contrastive learning in RA algorithm(a) and clustering process in RA algorithm(b). The RA algorithm(a) follows the MoCo v3 paradigm to improve the feature embedding ability of ViT for the most discriminative regions and surrounding regions. Meanwhile, a multi-region predicted loss and an optimization strategy are well-designed to optimize the ViT

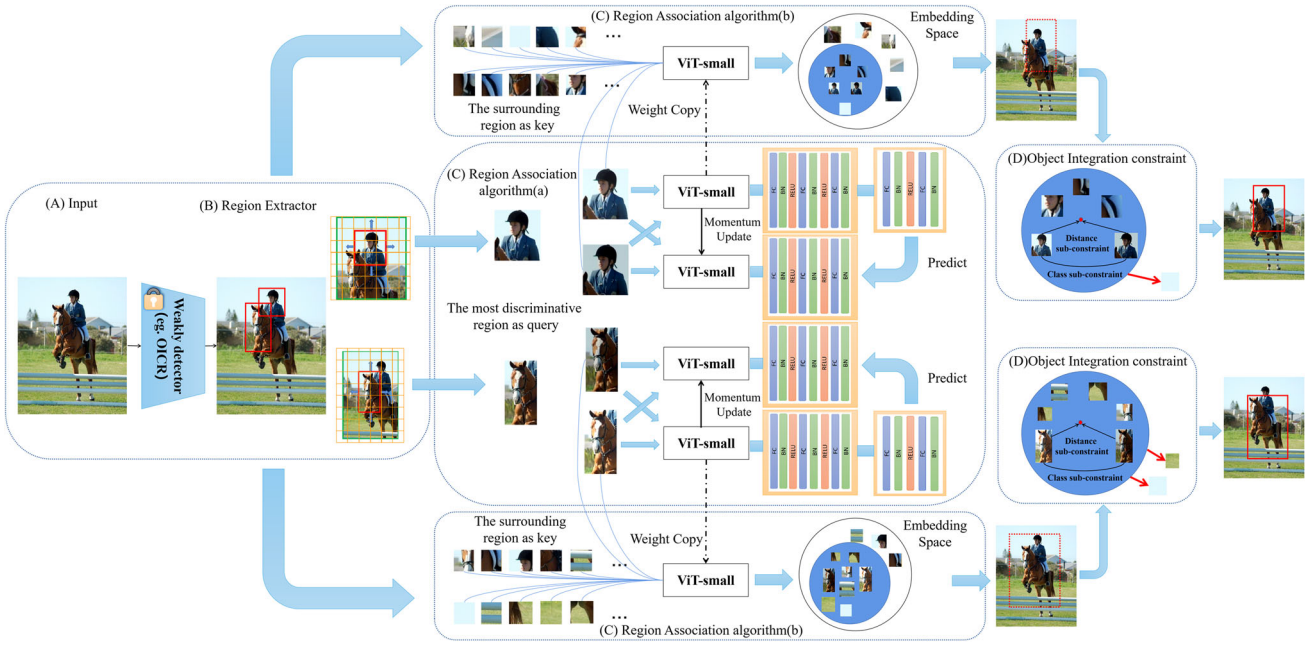


Fig. 2 The pipeline of the region-aware continual contrastive fusion (R-CCF) for WSOD. (A) The input images are fed into an off-the-shelf weakly supervised detector, and the initial red bounding boxes denote the predicted object location regions. (B) The query regions(red bounding boxes) and key regions(yellow bounding boxes) are generated by our region extractor. (C) The region association algorithm(a) follows the MoCo v3 paradigm to improve the feature embedding ability of ViT for query regions and key regions, and the RA algorithm (b) continually queries the similarity of the query region and key region to further form new rough object regions(i.e. dash red boxes). (D) Object integration constraint, including class sub-constraint and distance sub-constraint, removes some unsteady and noisy positive key regions with low confidence (i.e. red arrow in OI constraint) and an accurate object region (solid red boxes) is achieved. Best seen on a computer, in color and zoomed in

architecture. The RA algorithm(b) is proposed to continually query similarities of the most discriminative regions and their surrounding regions. Based on the similarity results in the clustering process, a rough object region that contains the whole object or almost the complete region of an object is mined. The third component is the OI constraint, including class sub-constraint and distance sub-constraint. It is used to remove some unsteady and noisy regions with low confidence and refine the rough object region from the RA algorithm to form a final complete object region.

3.2 Region extractor

As shown in Fig. 2, given an image, the R-CCF module first employs an off-the-shelf weakly supervised detector to predict the object location regions and treats these predicted locations as query regions. Then, we extend the predicted bounding boxes by a certain ratio α to obtain surrounding regions and crop them into patches as key regions. The coordinates of key regions are obtained from query regions (x, y, w, h) to $(x, y, \alpha w, \alpha h)$ and $\alpha > 1$. Note that x and y denote the center coordinate of the query region. The query regions can be expressed as (1):

$$B_d = \{b_r\}_{r=1}^{R}, \tag{1}$$

where b_r denotes the object regions predicted by the weakly supervised detector in an image, i.e. known as the most discriminative region, and R is the total number of predicted regions. The subscript d in B_d denotes the most discriminative object region(i.e. 'd' is the acronym of **d**iscriminative object region.)

Similarly, the key regions can be formulated as (2):

$$B_s = \{b_{rn}\}_{n=1}^{N_r}, \tag{2}$$

where b_{rn} denotes the n-th key regions corresponding to the r-th query region, i.e. known as the surrounding regions, and N_r is the total number of key regions corresponding to b_r. The subscript s in B_s denotes the surrounding region(i.e. 's' is the acronym of **s**urrounding region.)

3.3 Region association algorithm

For each input image I_i, query regions are generated by the baseline detector in the region extractor. In the RA algorithm(a), we sequentially apply four strong augmentations [28], including Random Color Jitter, Random Gray Scale, Random Gaussian Blur and Random Solarize, to transform an inputted query region into two correlated views $b_r^{'}$ and $b_r^{''}$. Then, we use a vision transformer as the base encoder $f_b(*)$ and momentum encoder $f_m(*)$ to extract embedding features of the augmented query regions. Specifically, we adopt ViT-small [37] to learn representation $h_i^{'} = f_b(b_r^{'})$ and $h_i^{''} = f_m(b_r^{''})$. Furthermore, a neural network projection head $g(*)$ maps embedding features from a low-dimensional space to a high-dimensional space. A prediction head $pr(*)$ after the projection head is used to optimize the multi-region predicted loss $L_{predict}$ in the base encoder branch. The above process can be expressed as minimizing the $L_{predict}$ loss function, and it can be formulated as (3)(4)(5):

$$L_{predict} = ctr(z_{r1}^{'}, z_{r2}^{''}) + ctr(z_{r2}^{'}, z_{r1}^{''}) \tag{3}$$

$$z_{r1}^{'} = pr(g(f_b(b_r^{'}))), z_{r1}^{''} = g(f_m(b_r^{'})) \tag{4}$$

$$z_{r2}^{'} = pr(g(f_b(b_r^{''}))), z_{r2}^{''} = g(f_m(b_r^{''})) \tag{5}$$

where $ctr(*)$ denotes a function of predicted loss that predicts A to B and B to A based on contrastive learning [28]. $z_{r1}^{'}$(or $z_{r2}^{'}$) and $z_{r1}^{''}$(or $z_{r2}^{''}$) denote the high-dimension and non-linear representation mapped by the prediction head $pr(*)$ and projection head $g(*)$ from the feature of query regions. Equations 4 and 5 represent different feature embedding branches in the RA algorithms(a).

Note that our base encoder branch in the RA algorithms(a) consists of a backbone (e.g. vision transformer), a projection head and an extra prediction head. The momentum encoder branch includes a backbone and a projection head, which is updated by the moving-average base encoder. However, the solution to the multi-region predicted loss optimization considers only the query regions and ViT lacks the ability to embed key region features at the early training stage, which may result in a queried ambiguous problem in the RA algorithm(b).

Thus, in the process of RA algorithm(b), we treat the RA algorithm(a) as a pre-training process and use the pre-training ViT to extract the embedding features of query and key regions for realizing semantic clustering. Specifically, we follow the same procedure in PCL [31] to perform K-means clustering based on the $f_b(b_r)$ and $f_b(b_{rn})$ features of the query region and key regions corresponding to b_r. With the continuous clustering process, we denote the K-th cluster centroid as c_K, and each key region is assigned as a positive or a negative key. If a key region is assigned to the same cluster as the query region, it will be treated as a positive key region. Otherwise, the key region will be considered as a negative key region. Based on the mined positive keys $k+$ and negative keys $k-$, we fuse the query region with the positive key regions and form a new object region.

Nonetheless, after the RA algorithm(b), we observe that some unsteady and noisy key regions with low confidence are classified as positive key regions, which inevitably disrupt the clustering process. Based on the above phenomenon, we conduct a qualitative analysis of the queried ambiguous problem. We found that, in the early training stage, our RA

algorithm(a) considers only the feature representation of the query regions and ignores the key regions, which is sub-optimal for the clustering process in the RA algorithm(b). In this case, those unsteady and noisy key regions usually obtain a low confidence value, while true positive key regions can achieve a high confidence.

Intuitively, to solve this sub-optimal problem, we select the high-confidence key regions to form new object regions as in paper [29]. For each query region b_i, we generate $b_i^{'}$ and $b_i^{''}$ by data augmentation, feed $b_i^{'}$, $b_i^{''}$ and b_{ij} into the RA algorithm, compute cluster centroids, and select the positive key regions whose confidence are larger than T_{score}. Here, the confidence of the positive key is calculated by (6):

$$T = \frac{sim(h_i, c_k)}{\sum_{k=1}^{K} sim(h_i, c_k)} \tag{6}$$

In (6), it is a softmax of cosine similarities between the regions representation h_i and each centroid c_k. We believe that the positive key regions with high confidence can be at least a part of an object, which can relieve the noisy phenomenon to some extent.

3.4 Object integration constraint

In the RA algorithm, although we can mine positive key regions with high confidence and obtain a new object region, it ignores a situation that augmented $b_i^{'}$ and $b_i^{''}$ may be assigned into different clusters. To solve this problem, we propose an enhanced constraint during the clustering process, named Object Integration(OI) constraint, which includes class sub-constraint and distance sub-constraint as shown in Algorithm 1.

More specifically, the class sub-constraint and distance sub-constraint are designed to measure the correctness of clustering results. In the class sub-constraint, we perform different data augmentation on query regions b_i for obtaining $b_i^{'}$ and $b_i^{''}$. Then, we conduct a cluster operation between $b_i^{'}$, $b_i^{''}$ and each key regions b_{ij}. If the clustering process in the RA algorithm classifies $b_i^{'}$ and $b_i^{''}$ into the same cluster, which is treated as a successful clustering, otherwise we ignore the clustering results. After the above class sub-constraint is met, the distance sub-constraint is further used to calculate distances between $b_i^{'}$, $b_i^{''}$, and their cluster centroids. If the difference between these two distances is greater than a threshold (e.g. 0.1), similarly we will discard the cluster results in this clustering procedure. When class sub-constraint and distance sub-constraint are satisfied simultaneously, we employ cosine similarity to represent the direction similarity of the positive key region b_{ij} and query regions ($b_i^{'}$ or $b_i^{''}$). Finally, if cosine similarity is larger than a threshold (e.g. 0.95), we will save those remaining positive keys, which will be merged as the final object regions.

Algorithm 1 Mine accurate object region based on object integration constraint.

Input: $B_d = \{b_1, ..., b_R\}$, Distance Threshold $T_{distance}$
$B_s = \{b_{11}, ..., b_{1N_1}\} \bigcup \{b_{21}, ..., b_{2N_2}\} ... \bigcup \{b_{R1}, ..., b_{RN_R}\}$
$k+$ confidence threshold T_{score}
for i to R **do**
  Initialize $G_d = \phi$, Initialize $G_s = \phi$ // init two lists and start RA and OI R times
  G_d.append(b_i) // append query region to the list
  Generate $b_i^{'}$ and $b_i^{''}$ by data augmentation
  for j to N_i **do**
    Feed the $b_i^{'}$, $b_i^{''}$ and b_{ij} into the RA algorithm
    Extract cluster centroid $C_{b_{ij}}$, $C_{b_i^{'}}$, $C_{b_i^{''}}$
    Compute distances between $b_i^{'}$/ $b_i^{''}$ and their cluster centroid as $T^{'}$ and $T^{''}$
    if $C_{b_i^{'}} = C_{b_i^{''}}$ and $\left|T^{'} - T^{''}\right| \leq T_{distance}$ **then**
      // satisfy class sub-constraint and distance sub-constraint
      Select the key region b_{ij} with the same label $C_{b_i^{'}}$ or $C_{b_i^{''}}$ as $k+$
      Compute cosine similarity T of $k+$ and $b_i^{'}$ or $b_i^{''}$
      if $T > T_{score}$ **then**
        G_s.append($k+$)// append the region whose confidence is larger than a threshold
      end if
    else
      Neglect the clustering result
    end if
  end for
  Update object regions G by fusing G_d and G_s
end for
Output: New object regions G

3.5 Optimization strategy

Based on the contrastive learning and clustering in the RA algorithm and OI constraint, the R-CCF module can automatically query similarities of the most discriminative regions with their surrounding regions, and the regions with high similarity will be merged to form the accurate object regions. However, at the early training stage, the feature extraction ability of ViT is insufficient for describing key regions. Therefore, we have developed a detailed training strategy to solve the problem of weak feature extraction ability, and the detailed training strategy is described as follows.

We train the R-CCF framework for 100 epochs, where the training process is divided into three stages, including the warm-up ViT stage, the query and clustering stage, and the accurate object mining stage. At the warm-up ViT stage (i.e. 0-29 epoch), we train the ViT to capture high-level semantic knowledge of query regions. Then, the ViT architecture is further used to extract the representation of query regions and key regions in RA algorithm(b) in the following stage. At the query and clustering stage, we conduct a

Algorithm 2 Train Strategy of R-CCF module

All the trainable parameters are initialized as random values which are denoted by W_t, W_g, W_p in vision Transformer, project head and prediction head respectively.

Input: W_t, W_g, W_p, and the total number of epoch is 100
β, learning rate decay coefficient
$z_{d'}$, $z_{d''}$, the feature of query region in different data augment
$z_{G'}$, $z_{G''}$, the feature of updated region in different data augment
$z_{\bar{d}'}$, $z_{\bar{d}''}$, the feature of not fused query region in different data augment
for *epoch* to 100 **do**
if $epoch < 30$ **then**
$W_t, W_g, W_p \leftarrow ctr(z_{d'}, z_{d''}) + ctr(z_{d''}, z_{d'})$, the warm-up stage only calculates the loss of query regions
else if $epoch \geq 30$ and $epoch\%5 == 0$ **then**
$W_t, W_g, W_p \leftarrow \beta * [ctr(z_{G'}, z_{G''}) + ctr(z_{G''}, z_{G'})]$, the query and clustering stage calculates the loss of updated object regions
else
$W_t, W_g, W_p \leftarrow \beta * [ctr(z_{G'}, z_{G''}) + ctr(z_{G''}, z_{G'}) + ctr(z_{\bar{d}'}, z_{\bar{d}''}) + ctr(z_{\bar{d}''}, z_{\bar{d}'})]$, the accurate object mining stage calculates the loss of updated object regions and the not fused query regions
end if
end for
Output: W_t, W_g, W_p

clustering operation in each 5th epoch (e.g. 30, 35, ..., 100), i.e. feed query regions and key regions into the RA algorithm to automatically find positive key regions and merge them into rough object regions. After the RA algorithm, the OI constraint is proposed to discard the incorrect clustering results and refine the rough object regions for obtaining the whole object region (i.e. updated object regions). At the accurate object mining stage (e.g. $31-34$, ..., $96-99$), we continually train ViT using updated object regions to learn the semantic knowledge of query regions and key regions, which can further improve the feature embedding ability of ViT. Moreover, unfused query regions are still fed into ViT to further query their surrounding regions.

Note that the multi-region predicted loss in RA algorithm is optimized by different inputs at three training stages, and the details about the training strategy and optimization can be found in Algorithm 2.

4 Experiments

In this section, we experimentally validate our R-CCF module and analyze each of its components for WSOD, which includes some ablation studies and comparisons against other state-of-the-art detectors on the PASCAL VOC 2007 and VOC2012 datasets. Moreover, we further conduct experiments in a cross-dataset setting from VOC2007 to COCO2017 to validate the generalization ability of our method. Finally, some qualitative results of R-CCF module are shown on the VOC 2007 dataset and VOC 2012 dataset.

4.1 Datasets and evaluation metrics

All of the baseline models, i.e. MIST [1], OICR [2], and WSDDN [3], use PASCAL VOC 2007 dataset and VOC 2012 dataset to validate their effectiveness. For a fair comparison, following these baselines, we choose the database(i.e. PASCAL VOC 2007 dataset and VOC 2012 dataset) to validate the effectiveness of our method, which have 9,963 and 22,531 images from 20 object categories, respectively. The greatest advantage of WSOD methods is that their training process does not require instance-level annotated data, and a large quantity of image-level annotated data is available. Therefore, we are interested in studying whether more training data can further improve the detection performance in WSOD. To this end, we train our proposed R-CCF framework on VOC 2007 trainval (5011 images) and 0712 trainval (16,551 images) and evaluate it on the VOC 2007 test set to demonstrate the effectiveness of our method. We follow the standard metrics for WSOD and use mean average precision (mAP) [49] as the evaluation metric for evaluating our model on the testing set of VOC 2007 and VOC 2012. Furthermore, correct location (CorLoc) [50] is used to evaluate the localization accuracy of our model on the VOC 2007 and VOC 2012 training sets. The metrics comply with the PASCAL criterion, where a positive detection has an IoU>0.5 with the ground-truth.

Moreover, we report the performance of our R-CCF in a cross-dataset setting from VOC2007 to COCO2017 to demonstrate the generalization ability of our method, where the VOC2007 trainval set(5011 images) serves as the training data and the COCO2017 val set(5000 images) is used as the testing data. Since there is a difference in the number of categories between the COCO dataset and VOC datasets, we only report the performance of the same class as the VOC dataset on the COCO dataset following the above evaluation criterion.

4.2 Implementation details

In the region extractor, we set $\alpha = 1.2$ to define the scope of the surrounding regions, which we crop these surrounding regions into 32*32 patches as key regions and resize these regions to fixed 224*224 as inputs of the contrastive learning architecture. In the RA algorithm, the best performance of the clustering process is obtained with cluster centroids K=20, so we set $K = 20$ in all of our experiments. In the OI constraint, we set the distance threshold $T_{distance} = 0.1$ to measure the correctness of clustering results for the distance sub-constraint. We set the score threshold $T_{score} = 0.95$ to

represent the direction similarity of the positive key regions and query regions.

In the optimization strategy, we set the learning rate decay coefficient $\beta = 0.5$ at the query with the clustering stage and the accurate object regions mining stage. The other training settings in the R-CCF module are adopted from the setups in MoCo v3 [28]. Specifically, the R-CCF module uses AdamW optimizer, and the batch size is set to 96. We search for lr(initial lr=1.5e-4) and weight decay wd(initial wd=1e-6) based on 100-epoch results. We adopt the learning rate warm-up for the first 30 epochs and the lr follows a cosine decay schedule after warm-up. Our augmentation policy includes Random Color Jitter, Random Grayscale, Random Gaussian Blur, and Random Solarize, which is used to transform an inputted query region into two correlated views $b_r^{'}$ and $b_r^{''}$ as in MoCo v3. In all our experiments, we run the publicly available deep learning framework PyTorch1.7.1 on an NVIDIA TESLA P100 GPU with CPU Xeon E5-1620 v4@3.50GHz.

Note that the test set was never involved in the tuning of any parameters or hyperparameters in all experiments (i.e. the comparisons with other state-of-the-art detectors and the ablation study).

4.3 Main results

The R-CCF module only needs to calculate the multi-region predicted loss on the predicted object regions from off-the-shelf weakly supervised detectors, so it can be easily applied to various weakly methods. To validate the effectiveness of our R-CCF, we conduct experiments based on three classic and popular detectors trained on the PASCAL VOC 2007 and VOC 2012 datasets, including WSDDN [3], OICR [2], MIST [1]. We plug our R-CCF module into the above three weakly supervised detectors and train them in an end-to-end manner.

We stress that our module does not use any label during training (no bounding box and class annotations are used), the only thing we need is the predicted regions by the baseline weakly supervised detectors. The reason is that we use the contrastive learning and clustering methods to automatically enlarge the most discriminative object regions generated by any existing weakly supervised detector. Thus, any off-the-shelf weakly detectors can be used as our baseline detector.

As shown in Table 1, we can see that the R-CCF module achieves 32.5%, 42.5% and 55.2% in mAP on the PASCAL VOC 2007 test set, which are 2.1%, 1.3% and 0.3% higher than the baseline detectors WSDDN [3], OICR [2], MIST [1] respectively. All the baseline detectors gain mAP improvements with the R-CCF module, which demonstrates the effectiveness of our proposed method in WSOD. Interestingly, based on the predicted results of a simple weakly-supervised object detector WSDDN-L [3], our R-CCF increases the mAP by a large margin, i.e. from 30.4% to 32.5% without an iterative refinement strategy, which proves our method can handle some complex scenarios. Moreover, based on the state-of-the-art weakly supervised detector [1] with refine strategy, our proposed method still improves the mAP performance from 54.9% to 55.2%. The above experiments prove that our proposed method is not only suitable for the method based on refined strategy but also for the simplest weakly-supervised object detector without the refinement iteration.

Table 2 shows the CorLoc on the VOC 2007 trainval set, similarly, we can see that our proposed method surpasses all baseline detectors clearly. For example, R-CCF based on OICR achieves 62.4% in CorLoc, exceeding the baseline by a large margin (i.e. 1.8%). Furthermore, R-CCF based on other weakly detectors (e.g. WSDDN and MIST) also improves in CorLoc, i.e. 49.7% *vs.* 49.4, and 68.9% *vs.* 68.8%. The comparison of CorLoc performance demonstrates that our method can solve the problem of highlighting only the most discriminative regions in WSOD.

To validate the effectiveness and robustness of our proposed method, we also conduct experiments on the VOC 2012 dataset, i.e. training our R-CCF framework on the VOC 2012 trainval set and evaluating on the VOC 2012 test set. Table 3 shows that the R-CCF module achieves 41.5% and 54.6% in mAP, which are 3.6% and 0.2% higher than the baseline detectors OICR [2] and MIST [1] respectively.

Moreover, R-CCF based on these two baseline detectors also improves in CorLoc(i.e. 62.1% vs. 63.2%, and 71.2% vs. 71.4%). According to the improvement in mAP and CorLoc, we can conclude that our proposed method tends to focus on a complete object region and, indeed, refines the predicted results of baseline detectors. Meanwhile, to analyze the complexity of the proposed method, we test the parameters and giga floating-point operations per second(GFLOPs) of the model that affect training and inference time and the required computational resources. The last two columns in Table 3 denote the number of network parameters and the GFLOPs results by testing a batch of the first 100 color images and their corresponding region proposals on the VOC2012 test set with a TESLA P100 GPU. As shown in Table 3, compared with the baseline (i.e. OICR), our module brings 44 M parameters and acceptable computational cost (362.28 *vs.* 498.91 in GFLOPs), but the mAP and CorLoc performance is improved, i.e. +3.6% and + 1.1% based on OICR, which further verifies the effectiveness of our R-CCF module for WSOD.

Furthermore, to validate whether more training data can further improve the detection performance, we follow the experimental setting (i.e. using an extra training set) of MIST [1] to train our R-CFF module, i.e. we use the predicted results of MIST on the VOC0712train/val sets as the input(i.e. query regions and key regions) of our R-CCF module. Similarly, we plug our R-CCF module into the MIST [1] weakly supervised detector and train it in an end-to-end

Table 1 Average precision(AP) (%) of our proposed R-CCF framework based on different baseline detectors (i.e. WSDDN [3], OICR [2] and MIST [1]) on the VOC2007 test set

Method	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	TV	mAP
WSDDN [3]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	30.4
WSDDN+R-CCF(Ours)	39.4	48.4	30.1	18.6	10.2	48.8	47.6	37.6	9.1	25.9	28.4	30.3	42.1	46.8	15.1	12.4	28.3	37.4	48.4	44.8	32.5(+2.1)
OICR [2]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
OICR+R-CCF(Ours)	61.2	67.9	43.6	14.3	12.5	67.2	66.7	40.1	20.3	49.5	35.3	30.1	33.8	67.4	5.7	20.5	41.7	42.6	62.1	67.3	42.5(+1.3)
MIST [1]	68.8	77.7	57.0	27.7	28.9	69.1	74.5	67.0	32.1	73.2	48.1	45.2	54.4	73.7	25.0	29.3	64.1	53.8	65.3	65.2	54.9
MIST+R-CCF(Ours)	68.4	77.8	57.5	27.7	29.5	69.2	74.6	67.5	32.2	73.2	47.9	46.0	54.4	74.2	25.0	29.4	63.8	54.2	65.5	65.3	55.2(+0.3)

Table 2 CorLoc (%) of our proposed R-CCF framework based on different baseline detectors (i.e. WSDDN [3], OICR [2] and MIST [1]) on the VOC2007 trainval set

Method	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	TV	Corloc
WSDDN [3]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	49.4
WSDDN+R-CCF(Ours)	70.8	63.5	54.7	34.0	39.7	68.0	67.7	49.1	35.5	54.8	33.8	39.3	37.4	77.5	7.5	43.2	60.8	25.0	68.8	63.1	49.7(+0.3)
OICR [2]	81.7	80.4	48.7	49.5	32.8	81.7	85.4	40.1	40.6	79.5	35.7	33.7	60.5	88.8	21.8	57.9	76.3	59.9	75.3	81.4	60.6
OICR+R-CCF(Ours)	80.4	82.7	67.6	42.6	41.2	80.2	85.5	53.5	42.7	80.8	43.7	40.5	52.0	88.8	16.2	57.1	81.4	53.2	74.1	82.8	62.4(+1.8)
MIST [1]	87.5	82.4	76.0	58.0	44.7	82.2	87.5	71.2	49.1	81.5	51.7	53.3	71.4	92.8	38.2	52.8	79.4	61.0	78.3	76.0	68.8
MIST+R-CCF(Ours)	87.5	82.8	76.6	58.0	44.7	82.8	87.5	71.5	49.1	81.5	51.7	53.3	71.8	92.8	38.2	52.8	80.4	61.5	78.3	76.0	68.9(+0.1)

Table 3 The mAP(%), Corloc(%), Parameters(M) and GFLOPs of our proposed R-CCF framework based on different baseline detectors (i.e. OICR [2] and MIST [1]) on the VOC2012 test set and VOC2012 trainval set

Methods	mAP(%)	CorLoc(%)	Parameters(M)	GFLOPs
OICR [2]	37.9	62.1	134	362.28
OICR+R-CCF(Ours)	41.5(+3.6)	63.2(+1.1)	178	498.91
MIST [1]	54.4	71.2	136	503.14
MIST+R-CCF(Ours)	54.6(+0.2)	71.4(+0.2)	180	604.42

The Parameters denote the number of network parameters and the CFLOPs (lower is better) is computed on the first 100 images of VOC2012 test set with a TESLA P100 GPU

manner. Table 4 shows that R-CCF achieves state-of-the-art performance (i.e. 58.3% in mAP) on the PASCAL VOC 2007 test set, which is 0.2% higher than the baseline detector. This improvement further demonstrates the effectiveness of our proposed method in WSOD.

Finally, we report the performance of our R-CCF in a cross-dataset setting from VOC to COCO to demonstrate the generalization ability of our method, i.e. we use a model trained on the VOC dataset to conduct inference tests directly on the COCO dataset. Since there is a difference in the number of categories between the COCO dataset and the VOC datasets, we only report the performance of the same class as the VOC dataset on the COCO dataset. As shown in Table 5, compared to the MIST baseline trained on the COCO dataset, our method without training on the COCO dataset achieves a remarkable result of 28.1% in mAP, which demonstrates the generalization ability of our proposed R-CCF across different datasets.

4.4 Ablation studies

To validate the contribution of each component in our R-CCF module including the RA algorithm, OI constraint, and other hyper-parameters, we conduct ablation experiments to prove their effectiveness on the VOC 2007 dataset. Note that all ablation experiments are based on MIST [1] and the training is terminated at 80 epochs unless otherwise stated.

Table 4 The mAP performance of our proposed R-CCF module trained with different train sets based on the MIST detector on the VOC 2007 test set

Methods	07trian/val	0712trian/val
MIST [1]	54.9	58.1
MIST+R-CCF(Ours)	55.2	58.3
↑	+0.3	+0.2

07train/val means the models are trained on the VOC 2007 trainval set, and 0712train/val denotes the models are trained on the VOC 2007 and 2012 trainval sets

Effectiveness of the RA algorithm To verify the effectiveness of the RA algorithm in our framework, as shown in Table 6, we compare the proposed R-CCF module with the RA algorithm against the baseline detector (without RA algorithm) on the VOC 2007 test set. Although the baseline detector (i.e. MIST [1]) already performs well on this dataset, we can observe that the performance of R-CCF with RA algorithm still outperforms the baseline detector by 0.21% in mAP (from 54.9% to 55.11%). The reason for this improvement is that the baseline detector often focuses on the most discriminative parts of objects. In contrast, the RA algorithm follows the MoCo v3 paradigm and a clustering process is conducted to automatically query similarities of the most discriminative regions and their surrounding regions, the selected surrounding regions with the most discriminative regions are further used to merge as the candidates of final accurate object regions. Although the most discriminative object regions from the baseline detector contain limited object information, R-CCF module with the RA algorithm can identify the similarity of surrounding regions and form new object regions using contrastive learning between query regions and key regions. Moreover, we continually train the vision transformer using our well-designed optimization strategy to learn the semantic knowledge of query regions and key regions, which further improves the feature embedding ability of the ViT. Based on the above reasons, the RA algorithm solves the problem of highlighting only discriminative object parts in WSOD, which also demonstrates the effectiveness of our proposed RA algorithm in WSOD.

Effectiveness of the OI constraint To further validate the effectiveness of the OI constraint, we show the performance of our proposed R-CCF module with and without OI constraint on the VOC 2007 test set. As shown in the 2nd and 3rd row of Table 6, we can see that the R-CCF module with OI constraint obtains 55.17% in mAP, surpassing the baseline clearly (from 54.9% to 55.17%). As mentioned earlier, the reason is that the class sub-constraint and distance sub-constraints in the OI constraint measure the correctness of the clustering process and hinder harmful clustering processes, which are important for removing positive key regions with low confidence at the early training stage. By introducing the OI constraint, the detection performance increases from 55.11% to 55.17% compared with using only the RA algorithm. Clearly, this improvement validates the claim that the OI constraint can remove positive key regions with low confidence. The remaining positive key regions and query regions are used to generate the final object regions. The above experiments and analysis demonstrate the effectiveness of our proposed OI constraint.

Table 5 Comparison results in mAP(%) for testing generalization capability of the proposed R-CCF on COCO2017 val

Method	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	TV	mAP
MIST [1]	61.8	32.1	26.8	13.2	21.0	66.5	29.7	31.6	14.6	45.7	11.3	38.8	52.4	46.9	3.5	19.4	20.3	10.8	62.1	61.8	33.5
MIST+R-CCF	54.7	25.4	20.4	12.7	14.6	52.7	22.3	51.4	12.0	25.6	21.6	31.5	35.4	37.3	21.5	7.2	24.5	26.9	50.8	40.8	28.1

Table 6 Ablation study results of RA algorithm and OI constraint on VOC2007 test set

Method	RA algorithm	OI constraint	mAP(%)
R-CCF module	✗	✗	54.9(baseline)
	✓	✗	55.11(+0.21)
	✓	✓	55.17(+0.27)

The performance of R-CCF module with or without RA algorithm and OI constraint are reported, which are based on the MIST detector

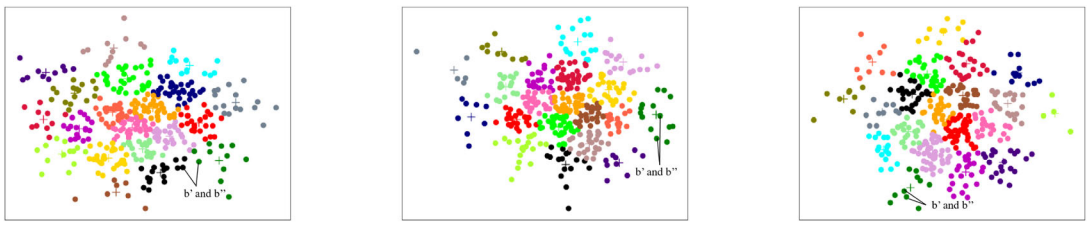
Table 7 The performance of different learning rate decay coefficients β in the optimization strategy on the VOC2007 test set, where β=0 denotes the well-designed optimization strategy is not used

Learning rate decay coefficient β	0	0.2	0.5	0.8	1.0
mAP(%)	55.13	55.15	55.17	55.17	55.16

Furthermore, a class sub-constraint and a distance sub-constraint in the OI constraint are proposed to measure the correctness of clustering results and further refine the rough object region from the RA algorithm. Specifically, the class sub-constraint judges whether query regions from different views(i.e. $b_i^{'}$ and $b_i^{''}$) are assigned to the same cluster. When the class sub-constraint is met, the distance sub-constraint further is used to calculate distances between $b_i^{'}$, $b_i^{''}$ and their cluster centroid, if the difference between these two distances is less than a threshold, we think that this clustering process is valid and correct, and the final regions generated by R-CCF can cover objects completely. Here, we qualitatively show the influence of these two sub-constraints, i.e. class sub-constraint and distance sub-constraint in the OI constraint.

Figure 3 shows three typical clustering results based on the class sub-constraint and the distance sub-constraint, including mismatching class sub-constraint, mismatching distance sub-constraint, and matching OI constraint (i.e. both the class sub-constraint and distance sub-constraint are satisfied). First, the class sub-constraint judges whether $b_i^{'}$ and $b_i^{''}$ are assigned into the same cluster, and Fig. 3(a) shows the case of mismatching class sub-constraint, where $b_i^{'}$ and $b_i^{''}$ are assigned into different clusters. Then, the distance sub-constraint is used to calculate distances between $b_i^{'}$ and $b_i^{''}$ and the cluster centroid, we can observe that the difference between these two distances is greater than a threshold $T_{distance}$ will result in a mismatching distance sub-constraint, as shown in Fig. 3(b). Finally, we visualize the cases of class sub-constraint and distance sub-constraint are satisfied simultaneously, where all latent representations of query regions and positive key regions are spatially adjacent and associate with the same object in embedding space, as shown in Fig. 3(c), thus the clustering process is correct and the final regions generated by R-CCF can cover objects accurately.

Effectiveness of the optimization strategy To validate the effectiveness of our optimization strategy in Section 3.5, we compare the performance of R-CCF with and without the optimization strategy(i.e. $\beta \neq 0$ and $\beta = 0$) on the VOC2007 test set. Table 7 shows that the detection performance increases from 55.13% to 55.17% when R-CCF uses the well-designed optimization strategy. Based on the above experimental results, we can conclude that the optimization strategy can further improve the feature embedding ability of ViT, which also demonstrates the effectiveness of our optimization strategy. We also use different learning rate decay coefficients β to analyze the sensitivity of this parameter in ViT feature embedding. Thus, we conduct several experiments to investigate the influence of β. As shown in Table 7, when $\beta \in (0, 1)$, the performance is almost unchanged, which indicates our optimization strategy is not sensitive to the learning rate decay coefficient β. We obtain the best performance (i.e. 55.17%) in mAP when $\beta = 0.5$, so we set $\beta = 0.5$ in all experiments in this paper.



(a) Mismatch Class sub-Constraint (b) Mismatch Dis. sub-Constraint (c) Matched OI Constraint

Fig. 3 Three typical clustering results based on the class sub-constraint and distance sub-constraint, where + represents the centroids of each cluster, · represents embedding features of query regions or key regions. (a) Mismatching class sub-constraint, $b_i^{'}$ and $b_i^{''}$ are assigned into different clusters. (b) Mismatching distance sub-constraint, the distance between $b_i^{'}$ and $b_i^{''}$ and the cluster centroid is large than a distance threshold (i.e. $T_{distance}$). (c) Matching class sub-constraint and distance sub-constraint simultaneously, all latent representations of query regions and key regions in the same cluster are spatially adjacent and associate with the same object. Best viewed in color and on computer

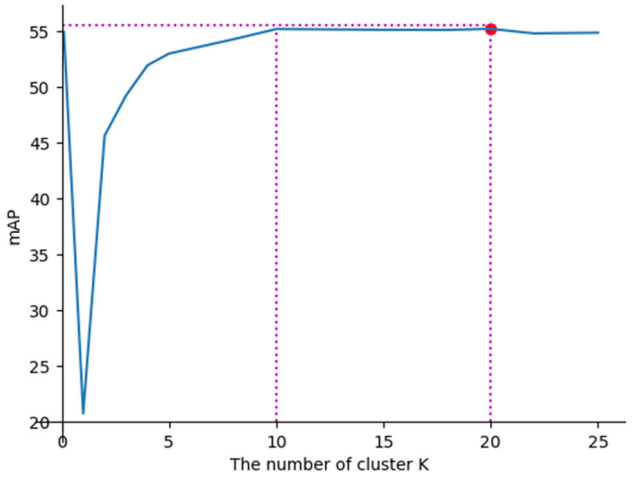


Fig. 4 The mAP performance for various choices of K on the VOC 2007 test set, where K denotes the number of cluster centroids. The red solid dot indicates the setting K=20 with the highest mAP

Analysis of cluster number K In this subsection, we conduct ablation experiments to find a suitable cluster number K for generating high-quality rough object regions during the clustering process in the RA algorithm. Figure 4 shows the comparison of different cluster numbers K on the VOC 2007 test set. It can be observed that the initial performance in mAP(i.e. MIST detector [1] as baseline) is 54.9%. By increasing the clustering number K, the mAP first reduced substantially compared with K=0, where K=1 denotes that a query region and its key regions are assigned to the same cluster. Then, the cluster number K gradually increases, and the best performance 55.17% in mAP(i.e. K=20) is achieved, which indicates that the suitable cluster number K improves the mAP performance with an ascending trend($10 \leq K \leq 20$). We think the reason is that the RA algorithm finds the positive key regions with high confidence and provides high-quality rough object regions for the OI constraint. However, by continuing to increase the cluster number K (e.g. $K > 20$), the detection network converges to the baseline mAP performance around 54.9%. This is because each region may become an independent cluster in these cases, and query regions and key regions are not fused, i.e. the number of cluster centroids is greater than the number of regions. For quantitatively understanding the sensitivity of cluster number K in our R-CCF module, Table 8 shows the comparison of mAP performance utilizing the different cluster number K (K=0, 1, 2, 3,..., 25) to train our module based on the baseline on the VOC 2007 test set. We can observe that the detection performance improves to 55.17% from 54.9% as K increases in the clustering process since the RA algorithm provides accurate object regions for the OI constraint. However, after setting the cluster number $K > 20$, the mAP converges to the baseline performance(i.e. 54.9%), the reason is that the detection network converges to one point as explained in the above. Thus, we set the cluster number K to 20 (i.e. K=20) for all the experiments in our paper.

Analysis of the scope of surrounding regions α We conduct ablation experiments to discuss the importance of the scope of surrounding regions α, which directly affects the performance(i.e. mAP) and computational cost(i.e. GFLOPs) of our proposed method. Figure 5 and Table 9 show the comparison results utilizing different α (1.0, 1.2, 1.5, and 2.0) to train our module on VOC 2007 test set. It can be observed that the initial performance in mAP(i.e. MIST baseline [1]) is 54.9%. By increasing α, the mAP first increases and reaches the best performance 55.17% in mAP(i.e. α=1.2). Then, the performance gradually decreases, which indicates that a large range of key regions is associated with adjacent instances of the same category. Meanwhile, the computational cost continually increases, as shown in Fig. 5. The reason is that a large number of query regions and key regions are used in the continual contrastive fusion process. Therefore, to balance detection performance(i.e. mAP) and computational costs(i.e. GFLOPs), we set α to 1.2 for for all the experiments in our paper.

Analysis of the ratio of surrounding regions As stated in the region extractor (Section 3.2), 60% of surrounding regions are randomly used in the RA algorithm and the OI constraint. To make clear how to tune the ratio of surrounding regions in the region extractor, we compare the performance of R-CCF on the VOC2007 test set under the setting of using 60% and 100% surrounding regions as the key regions, respectively. Table 10 shows that the best detection performance (i.e. 55.17%) is achieved when the ratio of surrounding regions is set to 60%. Meanwhile, the clustering computation time decreases by 202 ms (from 444 ms to 242 ms). There are two reasons for this improvement: the 60% ratio

Table 8 The mAP performance in different cluster numbers on the VOC 2007 test set, where K denotes the number of clusters and K=0 represents the performance without the RA algorithm

K	0	1	2	3	4	5	8
mAP(%)	54.9	20.71	45.62	49.19	51.90	52.95	54.23
↑	0	−34.19	−9.28	−5.71	−3.00	−1.95	−0.67
K	10	15	18	**20**	22	25	
mAP(%)	55.15	55.08	55.07	**55.17**	54.96	55.02	
↑	+0.25	+0.18	+0.17	**+0.27**	+0.06	+0.12	

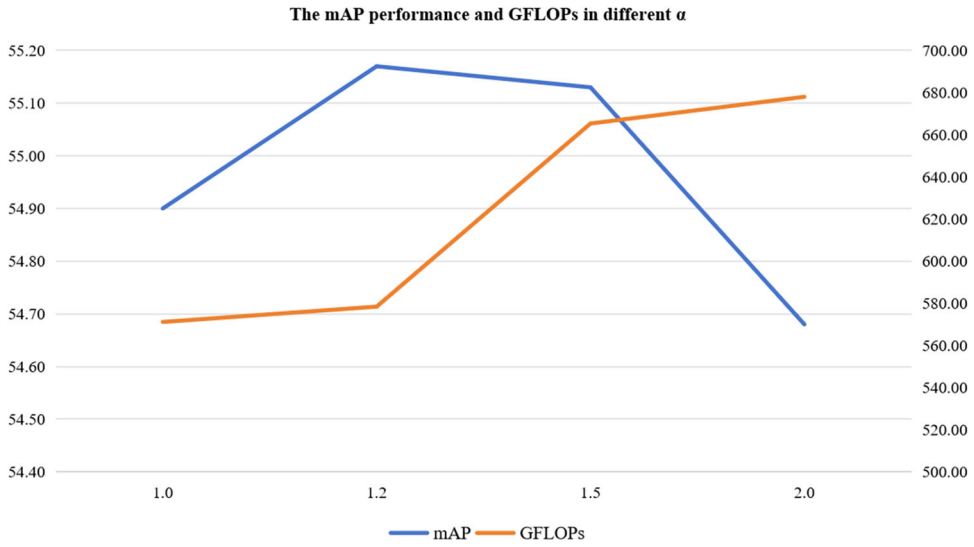


Fig. 5 The performance(i.e. mAP) and computational costs(i.e. GFLOPs) for different α on VOC 2007 test set, where mAP(larger is better) and CFLOPs (lower is better) are computed on the first 100 images of VOC2007 test set with a TESLA P100 GPU

of surrounding regions can prominently reduce the time of querying and clustering and 60% of surrounding regions still efficiently mine accurate object localization regions(i.e. positive key regions can be clearly detected using only a few key regions in a row or column). Based on the above experimental results and analysis, we prove the effectiveness of setting the 60% ratio of the surrounding regions as key regions.

4.5 State-of-the-art comparison

We compare our proposed R-CCF module with several state-of-the-art weakly supervised detectors on the VOC 2007 test set, including SDCN [51], C-MIL [52], Yang [53], PCL++ [22], SLV [54], WSOD2 [23], PredNet [55], C-MIDN [56], Zhang [5], SDCN+FRCNN [51], SLV+FRCNN [54], IDO [41], MIST [1], C-MIL+FRCNN [52], C-MIDN+FRCNN [56], PredNet+FRCNN [55], Yang+FRCNN [53] and IDO+FRCNN [41]. The upper part in Table 11 shows the detection performance of each weakly supervised detector. Additionally, we can observe that the proposed method surpasses all previous approaches on the VOC2007 test set. For example, the R-CCF module exceeds the previous state-of-the-art method (i.e. MIST) 0.3%. Moreover, though we do not train a fully-supervised detector like Faster-RCNN [60] by using pseudo ground-truths, such as SDCN+FRCNN [51], SLV+FRCNN [54], Zhang et al. [5], MIST [1], C-MIL+FRCNN [52], C-MIDN+FRCNN [56], PredNet+FRCNN [55], and Yang+FRCNN [53], R-CCF module can still achieve the highest performance of 55.2% in mAP, which further demonstrates the effectiveness of our proposed method in the task of WSOD.

Furthermore, to validate whether more training data can further improve detection performance, we compare our proposed method with other detectors using additional training sets on the VOC 2007 test set. As shown in the lower part of Table 11, our method also achieves the best performance (i.e. 58.3%), which is a new state-of-the-art performance in WSOD, surpassing all previous methods [1, 23, 43, 57–59] clearly. This improvement further demonstrates the robustness and generalization of our proposed method in weakly supervised object detection.

We also show the CorLoc performance of our R-CCF based on MIST compared with other state-of-the-art methods on the VOC 2007 trainval set. As shown in Table 12, we can see that our R-CCF (i.e. 68.9%) surpasses the baseline MIST (i.e. 68.8%) by 0.1%, which proves our method can refine the target object locations. In addition, we achieve a competitive performance in CorLoc compared with other SOTA methods, such as WSOD2 [23], PredNet [55], and SLV [54]. Since our R-CCF is orthometric with the baselines and it can

Table 9 The mAP and GFLOPs performance in different α, where mAP(larger is better) and CFLOPs (lower is better) are computed on the first 100 images of VOC2007 test set with a TESLA P100 GPU

α	1.0	1.2	1.5	2.0
mAP(%)	54.90	55.17	55.13	54.68
GFLOPs	571.14	578.42	665.34	678.01

Table 10 The mAP performance in different ratios of surrounding regions, where clustering time represents the time of performing R-CCF module once at the query and clustering stage

Ratio of the surrounding regions	mAP(%)	Clustering time(ms)
60%	55.17	242
100%	55.11	444

Table 11 Comparison results in AP and mAP (%) for different weakly supervised methods on the PASCAL VOC 2007 test set

Method	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	TV	mAP
SDCN [51]	59.8	67.1	32.0	34.7	22.8	67.1	63.8	67.9	22.5	48.9	47.8	60.5	51.7	65.2	11.8	20.6	42.1	54.7	60.8	64.3	48.3
C-MIL [52]	62.5	58.4	49.5	32.1	19.8	70.5	66.1	63.4	20.0	60.5	52.9	53.5	57.4	68.9	8.4	24.6	51.8	58.7	66.7	63.5	50.5
Yang [53]	57.6	70.8	50.7	28.3	27.2	72.5	69.1	65.0	26.9	64.5	47.4	47.7	53.5	66.9	13.7	29.3	56.0	54.9	63.4	65.2	51.5
PCL++ [22]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	52.9
SLV [54]	65.6	71.4	49.0	37.1	24.6	69.6	70.3	70.6	30.8	63.1	36.0	61.4	65.3	68.4	12.4	29.9	52.4	60.0	67.6	64.5	53.5
WSOD2 [23]	65.1	64.8	57.2	39.2	24.3	69.8	66.2	61.0	29.8	64.6	42.5	60.1	71.2	70.7	21.9	28.1	58.6	59.7	52.2	64.8	53.6
PredNet [55]	66.7	69.5	52.8	31.4	24.7	74.5	74.1	67.3	14.6	53.0	46.1	52.9	69.9	70.8	18.5	28.4	54.6	60.7	67.1	60.4	52.9
C-MIDN [56]	53.3	71.5	49.8	26.1	20.3	70.3	69.9	68.3	28.7	65.3	45.1	64.6	58.0	71.2	20.0	27.5	54.9	54.9	69.4	63.5	52.6
MIST [1]	68.8	77.7	57.0	27.7	28.9	69.1	74.5	67.0	32.1	73.2	48.1	45.2	54.4	73.7	35.0	29.3	64.1	53.8	65.3	65.2	54.9
Zhang [5]	64.3	67.1	49.6	34.0	13.4	71.3	68.7	70.9	24.8	56.5	51.8	70.2	61.3	65.8	27.0	24.6	51.7	62.3	65.8	61.2	53.1
IDO [41]	65.4	70.5	50.7	24.6	24.0	73.1	71.6	65.1	28.6	63.9	49.3	61.1	56.3	70.9	10.1	29.2	53.7	59.2	69.8	68.2	53.3
SDCN+FRCNN [51]	61.1	70.6	40.2	32.8	23.9	63.4	68.9	68.2	18.3	60.2	53.5	63.6	53.6	66.1	14.6	21.8	50.5	56.9	62.4	67.9	51.0
SLV+FRCNN [54]	62.1	72.1	54.1	34.5	25.6	66.7	67.4	77.2	24.2	61.6	47.5	71.6	72.0	67.2	12.1	24.6	51.7	61.1	65.3	60.1	53.9
C-MIL+FRCNN [52]	61.8	60.9	56.2	28.9	18.9	68.2	69.6	71.4	18.5	64.3	57.2	66.9	65.9	65.7	13.8	22.9	54.1	61.9	68.2	66.1	53.1
C-MIDN+FRCNN [56]	54.1	74.5	56.9	26.4	22.2	68.7	68.9	74.8	25.2	64.8	46.4	70.3	66.3	67.5	21.6	24.4	53.0	59.7	68.7	58.9	53.6
PredNet+FRCNN [55]	67.7	70.4	52.9	31.3	26.1	75.5	73.7	68.6	14.9	54.0	47.3	53.7	70.8	70.2	19.7	29.2	54.9	61.3	67.6	61.2	53.6
Yang+FRCNN [53]	59.8	72.8	54.4	35.6	30.2	74.4	70.6	74.5	27.7	68.0	51.7	46.3	63.7	68.6	14.8	27.8	54.9	60.9	65.1	67.4	54.5
IDO+FRCNN [41]	61.3	77.9	59.3	36.1	22.5	75.0	74.3	75.7	30.6	66.6	50.6	51.5	51.9	70.9	19.5	28.7	57.2	58.2	64.1	69.1	55.1
Ours	68.4	77.8	57.5	27.7	29.5	69.2	74.6	67.5	32.2	73.2	47.9	46.0	54.4	74.2	25.0	29.4	63.8	54.2	65.5	65.3	**55.2**
WebReIETH* [57]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	46.0
WSOD2* [23]	68.2	70.7	61.5	42.3	28.0	73.4	69.3	52.3	32.7	71.9	42.8	57.9	73.8	71.4	25.5	29.2	61.6	60.9	56.5	70.7	56.0
zhong(R50-C4)* [43]	64.4	45.0	62.1	42.8	42.4	73.1	73.2	76.0	28.2	78.6	28.5	75.1	74.6	67.7	57.5	11.6	65.6	55.4	72.2	61.3	57.77
MIST* [1]	70.11	78.39	65.41	38.88	35.03	75.05	69.76	60.31	35.64	76.01	44.68	50.69	60.70	78.12	21.65	31.71	64.84	70.05	61.81	73.42	58.11
Boosting* [58]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	56.5
WSOD-CAT* [59]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	58.0
Ours*	70.1	78.4	65.7	39.3	35.0	75.1	69.8	60.3	35.6	76.0	44.7	52.3	60.8	78.1	21.6	32.0	64.8	70.1	61.8	73.4	**58.3**

The upper part shows the results without auxiliary training data, and the lower part shows the results when using auxiliary training data. * means using auxiliary training sets, where zhong(R50-C4)* [43], Boosting [58] and WSOD-CAT [59] use COCO dataset as auxiliary training set and other methods use VOC 2012train/val set as the auxiliary training set

Table 12 Comparison results in CorLoc (%) for different weakly supervised methods on the PASCAL VOC 2007 trainval set

Method	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	TV	CorLoc
Cinbis [61]	56.6	58.3	28.4	20.7	6.8	54.9	69.1	20.8	9.2	50.5	10.2	29.0	58.0	64.9	36.7	18.7	56.5	13.2	54.9	59.4	38.8
Bilen [62]	66.4	59.3	42.7	20.4	21.3	63.4	74.3	59.6	21.1	58.2	14.0	38.5	49.5	60.0	19.8	39.2	41.7	30.1	50.2	44.1	43.7
Wang [63]	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	48.5
Li [64]	78.2	67.1	61.8	38.1	36.1	61.8	78.8	55.2	28.5	68.8	18.5	49.2	64.1	73.5	21.4	47.4	64.6	22.3	60.9	52.3	52.4
WSDDN [3]	65.1	58.8	58.5	33.1	39.8	68.3	60.2	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
Teh [65]	84.0	64.6	70.0	62.4	25.8	80.6	73.9	71.5	35.7	81.6	46.5	71.3	79.1	78.8	56.7	34.3	69.8	56.7	77.0	72.7	64.6
ContextLocNet [66]	83.3	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	47.1	59.5	67.0	83.5	35.3	39.9	67.0	49.7	63.5	65.2	55.1
OICR [2]	81.7	80.4	48.7	49.5	32.8	81.7	85.4	40.1	40.6	79.5	35.7	33.7	60.5	88.8	21.8	57.9	76.3	59.9	75.3	81.4	60.6
Jie [67]	72.7	55.3	53.0	27.8	35.2	68.6	81.9	60.7	11.6	71.6	29.7	54.3	64.3	88.2	22.2	53.7	72.2	52.6	68.9	75.5	56.1
Diba [68]	83.9	72.8	64.5	44.1	40.1	65.7	82.5	58.9	33.7	72.5	25.6	53.7	67.4	77.4	26.8	49.1	68.1	27.9	64.5	55.7	56.7
Wei [16]	84.2	74.1	61.3	52.1	32.1	76.7	82.9	66.6	42.3	70.6	39.5	57.0	61.2	88.4	9.3	54.6	72.2	60.0	65.0	70.3	61.0
Wan [69]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	61.4
PCL [22]	79.6	85.5	62.2	47.9	37.0	83.8	83.4	43.0	38.3	80.1	50.6	30.9	57.8	90.8	27.0	58.2	75.3	68.5	75.7	78.9	62.7
PCL++ [22]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	67.2
Tang [70]	77.5	81.2	55.3	19.7	44.3	80.2	86.6	69.5	10.1	87.7	68.4	52.1	84.4	91.6	57.4	63.4	77.3	58.1	57.0	53.8	63.8
SDCN [51]	85.0	83.9	58.9	59.6	43.1	79.7	85.2	77.9	31.3	78.1	50.6	75.6	76.2	88.4	49.7	56.4	73.2	62.6	77.2	79.9	68.6
Shen [71]	82.9	74.0	73.4	47.1	60.9	80.4	77.5	78.8	18.6	70.0	56.7	67.0	64.5	84.0	47.0	50.1	71.9	57.6	83.3	43.5	64.5
C-MIL [52]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	65.0
IDO [41]	83.3	81.2	67.9	50.0	46.2	85.3	88.7	70.6	42.8	78.8	51.0	74.2	76.9	90.4	15.5	54.2	74.2	62.9	81.0	82.4	67.9
Yang [53]	80.0	83.9	74.2	53.2	48.5	82.7	86.2	69.5	39.3	82.9	53.6	61.4	72.4	91.2	22.4	57.5	83.5	64.8	75.7	77.1	68.0
MIST [1]	87.5	82.4	76.0	58.0	44.7	82.2	87.5	71.2	49.1	81.5	51.7	53.3	71.4	92.8	38.2	52.8	79.4	61.0	78.3	76.0	68.8
WSOD2 [23]	87.1	80.0	74.8	60.1	36.6	79.2	83.8	70.6	43.5	88.4	46.0	74.7	87.4	90.8	44.2	52.4	81.4	61.8	67.7	79.9	69.5
PredNet [55]	88.6	86.3	71.8	53.4	51.2	87.6	89.0	65.3	33.2	86.6	58.8	65.9	87.7	93.3	30.9	58.9	83.4	67.8	78.7	80.2	70.9
SLV [54]	84.6	84.3	73.3	58.5	49.2	80.2	87.0	79.4	46.8	83.6	41.8	79.3	88.8	90.4	19.5	59.7	79.4	67.7	82.9	83.2	71.0
Ours	87.5	82.8	76.6	58.0	44.7	82.2	87.6	71.5	49.1	81.5	51.7	53.3	71.8	92.8	38.2	52.8	80.4	61.5	78.3	76.1	68.9

The upper/lower part are results from single and multiple models respectively

plug into any existing weakly supervised detector, we believe our R-CCF can further achieve a higher CorLoc performance if based on a better baseline detector (e.g. SLV), which is left in the future.

Finally, we compare R-CCF module with several state-of-the-art weakly detectors on the VOC 2012 test set, including OICR [2],C-WSL [72], WSRPN [70],C-MIL [52], Yang [53], C-MIDN [56], WSOD2 [23], OIM [73], PredNet [55], IDO [41], SLV [54], IDO+FRCNN [41], MIST [1] and CASD [74], as shown in Table 13. Table 13 shows that the proposed method clearly surpasses all previous models on the VOC 2012 test set. Regarding the mAP metric, our proposed method exceeds the previous state-of-the-art method (i.e. MIST) by 0.2% (from 54.4% to 54.6%) absolutely. In terms of the CorLoc metric, our proposed method achieves the competitive results (i.e. 71.4% vs. 72.3%) on the VOC2012 trainval set. The reason for these improvements is that our proposed method can achieve a more accurate object region than other weakly supervised detection methods. The comparison on the VOC 2012 dataset also proves that R-CCF is effective in WSOD.

4.6 Qualitative results

Figure 6 shows some detection results predicted by our proposed R-CCF module on the VOC2007 and VOC2012 test sets. In the first four rows, we show some detection results in images that include a single or multiple objects. From these qualitative results, we observe that our method can successfully mine the whole object region even though some object instances are occluded and non-rigid. This demonstrates the effectiveness of our proposed method in solving the problem of highlighting only the most discriminative regions. Moreover, we also visualize some failure cases in the last two rows in Fig. 6, i.e. missed objects, and grouped instances. The main reason is that our R-CCF is influenced by the predictions from the baseline weakly supervised detector. Our R-CCF can not solve these two problems in that the regions of objects predicted by the baseline detector are missed and grouped. However, for the main problem of highlighting only discriminative object parts in WSOD, our proposed R-CCF can detect the complete object region well, as shown in the first four rows in Fig. 6.

Table 13 Comparison results in mAP(%) and CorLoc(%) for different weakly supervised methods on the PASCAL VOC 2012 test set

Method	mAP(%)	CorLoc(%)
OICR [2]	37.9	62.1
C-WSL [72]	43.0	64.9
WSRPN [70]	43.4	64.9
C-MIL [52]	46.7	67.4
Yang [53]	46.8	69.5
C-MIDN [56]	50.2	71.2
WSOD2 [23]	47.2	71.9
OIM [73]	45.3	67.1
PredNet [55]	48.4	69.5
IDO [41]	49.0	67.7
SLV [54]	49.2	69.2
IDO+FRCNN [41]	51.7	71.2
CASD [74]	53.6	**72.3**
MIST [1]	54.4	71.2
Ours	**54.6**	71.4

Clearly, our method obtains the best performance in mAP(%) and gets a comparable performance in CorLoc(%)

5 Conclusion

In this paper, we propose an end-to-end R-CCF module for weakly supervised object detection (WSOD). To solve the problem of highlighting only discriminative object parts in WSOD, a novel RA algorithm is proposed to identify similarities between query regions and key regions, and new rough object regions are formed based on the clustering results. In the RA algorithm, a multi-region predicted loss and an optimization strategy are designed to optimize ViT for better extracting the representation of query regions and key regions. However, some unsteady and noisy key regions with low confidence are classified as positive key regions, which will inevitably disrupt the RA algorithm and become a sub-optimal problem. To overcome this problem, OI constraint, including class sub-constraint and distance sub-constraint, is proposed to refine rough object regions from the RA algorithm, and accurate object regions are achieved by removing positive key regions with low confidence. Meanwhile, a well-designed training strategy is proposed to overcome the sub-optimal problem further by improving the feature embedding ability of ViT. Extensive experiments on the widely used WSOD dataset demonstrate the effectiveness of our proposed R-CCF module for detecting complete object regions, and show that we outperform previous state-of-the-art methods by a large margin, where the largest improvement is registered for non-grid objects.

Detecting completely object regions in WSOD is still an open challenge, in the future, we plan to design more effective methods for learning the stronger feature maps of query regions and key regions, and then further explore their relationship for accurate fusing. Moreover, encouraged by the

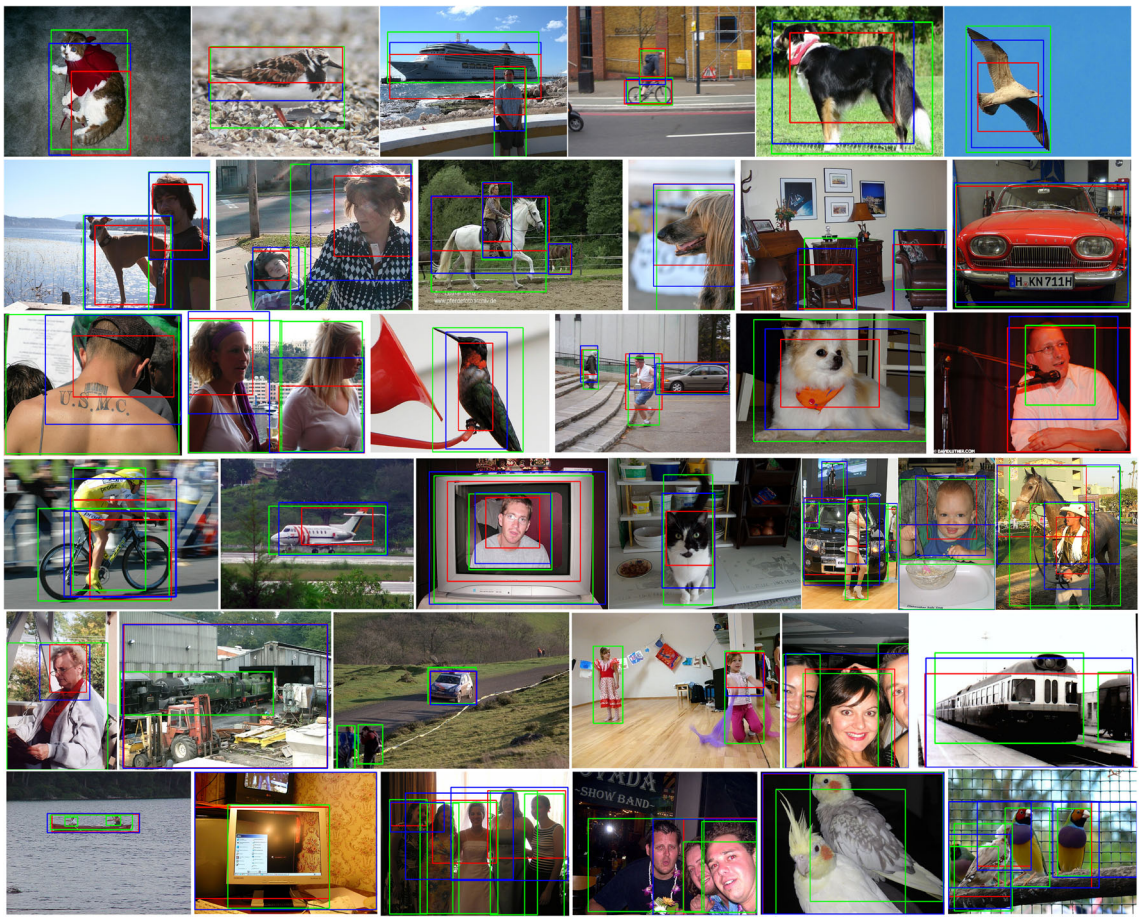


Fig. 6 Qualitative results of our proposed R-CCF module and the baseline detector on the VOC2007 and VOC2012 test sets. Blue and red bounding boxes denote our predictions and baseline predictions, respectively. Green bounding boxes denote the ground truths. In the first four rows, we show that some object instances are successfully detected by our proposed method. We also show some failure cases, where the final predicted object regions(i.e. blue bounding box) hard cover the complete object in the fifth and sixth rows. Best seen on a computer in color

achieved results, we plan to extend our proposed method to other kinds of weakly supervised tasks. Finally, with the development of the vision-language model(e.g. CLIP), we will also introduce CLIP to the weakly supervised tasks, which fully utilizes the image-level label semantic information of NLP to save high-quality region proposals and reduce expensive computational costs.

Acknowledgements This work is supported by the China Postdoctoral Science Foundation (Grant No. 259822), the National Postdoctoral program for Innovative Talents (Grant No. BX20200108), the National Science Foundation of China (Grant No. 62206077 and No. 61976070), and the Science Foundation of Heilongjiang Province (LH2021F024).

Author Contributions Yongqiang Zhang and Rui Tian conceived of the presented idea. Yongqiang Zhang, Rui Tian, Yin Zhang and Zian Zhang carried out the experiment and wrote the manuscript with support from Wangmeng Zuo and Mingli Ding. MingliDing supervised the project. Mingli Ding and Yancheng Bai contributed to the final version of the manuscript.

Data availability and access The datasets used during the current study are available as follows: 1) PASCAL VOC 2007: http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html 2) PASCAL VOC 2012: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html 3) COCO 2017:https://cocodataset.org/.

Declarations

Conflict of interest We would like to note that in the manuscript entitled "R-CCF: Region-aware Continual Contrastive Fusion for Weakly Supervised Object Detection", no conflict of interest exists in the submission of this manuscript, and the manuscript is approved by all authors for publication.

Ethical and informed consent We confirm that the used data does not include any Ethical consent.

References

1. Ren Z, Yu Z, Yang X, Liu MY, Lee YJ, Schwing AG, Kautz J (2020) Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10598–10607
2. Tang P, Wang X, Bai X, Liu W (2017) Multiple instance detection network with online instance classifier refinement. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2843–2851
3. Bilen H, Vedaldi A (2016) Weakly supervised deep detection networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2846–2854
4. Zhang Y, Bai Y, Ding M, Li Y, Ghanem B (2018) W2f: a weakly-supervised to fully-supervised framework for object detection. In: CVPR. IEEE, pp 928–936
5. Zhang Y, Bai Y, Ding M, Li Y, Ghanem B (2018) Weakly-supervised object detection via mining pseudo ground truth bounding-boxes. Pattern Recognit 84:68–81
6. Zhang Y, Ding M, Bai Y, Xu M, Ghanem B (2019) Beyond weakly supervised: pseudo ground truths mining for missing bounding-boxes object detection. IEEE Trans Circuits Syst Video Technol 30(4):983–997
7. Cheng G, Yang J, Gao D, Guo L, Han J (2020) High-quality proposals for weakly supervised object detection. IEEE Trans Image Process 29:5794–5804
8. Peng J, Wang H, Yue S, Zhang Z (2022) Context-aware co-supervision for accurate object detection. Pattern Recognit 121:108199
9. Dai X, Chen Y, Yang J, Zhang P, Yuan L, Zhang L (2021) Dynamic detr: end-to-end object detection with dynamic attention. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2988–2997
10. Li F, Zhang H, Liu S, Guo J, Ni LM, Zhang L (2022) Dn-detr: accelerate detr training by introducing query denoising. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13619–13627
11. Wang Y, Ilic V, Li J, Kisačanin B, Pavlovic V (2023a) Alwod: active learning for weakly-supervised object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6459–6469
12. Wang Y, Guerrero R, Pavlovic V (2023b) D2f2wod: learning object proposals for weakly-supervised object detection via progressive domain adaptation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 22–31
13. Feng X, Yao X, Shen H, Cheng G, Xiao B, Han J (2023) Learning an invariant and equivariant network for weakly supervised object detection. IEEE Trans Pattern Anal Mach Intell
14. Sui L, Zhang CL, Wu J (2022) Salvage of supervision in weakly supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14227–14236
15. Gao W, Wan F, Yue J, Xu S, Ye Q (2022) Discrepant multiple instance learning for weakly supervised object detection. Pattern Recognit 122:108233
16. Wei Y, Shen Z, Cheng B, Shi H, Xiong J, Feng J, Huang T (2018) Ts2c: tight box mining with surrounding segmentation context for weakly supervised object detection. In: Proceedings of the European conference on computer vision (ECCV), pp 434–450
17. Choe J, Han D, Yun S, Ha JW, Oh SJ, Shim H (2021) Region-based dropout with attention prior for weakly supervised object localization. Pattern Recognit 116:107949
18. Murtaza S, Belharbi S, Pedersoli M, Sarraf A, Granger E (2023) Discriminative sampling of proposals in self-supervised transformers for weakly supervised object localization. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 155–165
19. Shao F, Chen L, Shao J, Ji W, Xiao S, Ye L, Zhuang Y, Xiao J (2022) Deep learning for weakly-supervised object detection and localization: a survey. Neurocomputing 496:192–207
20. Bai J, Ren J, Xiao Z, Chen Z, Gao C, Ali TAA, Jiao L (2023) Localizing from classification: self-directed weakly supervised object localization for remote sensing images. IEEE Trans Neural Netw Learn Syst
21. Hui W, Tan C, Gu G, Zhao Y (2022) Gradient-based refined class activation map for weakly supervised object localization. Pattern Recognit 128:108664
22. Tang P, Wang X, Bai S, Shen W, Bai X, Liu W, Yuille A (2018) Pcl: proposal cluster learning for weakly supervised object detection. IEEE Trans Pattern Anal Mach Intell 42(1):176–191
23. Zeng Z, Liu B, Fu J, Chao H, Zhang L (2019) Wsod2: learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8292–8300
24. Wang J, Chen Y, Dong Z, Gao M (2023) Improved yolov5 network for real-time multi-scale traffic sign detection. Neural Comput Appl 35(10):7853–7865
25. Piao Z, Wang J, Tanga L, Zhao B, Wang W (2022) Accloc: anchor-free and two-stage detector for accurate object localization. Pattern Recognit 126:108523
26. Wang J, Zhao C, Huo Z, Qiao Y, Sima H (2022) High quality proposal feature generation for crowded pedestrian detection. Pattern Recognit 128:108605
27. Shao Z, Su Y, Zhou Y, Meng F, Zhu H, Liu B, Yao R (2023) Ct-net: arbitrary-shaped text detection via contour transformer. IEEE Trans Circuits Syst Video Technol
28. Chen X, Xie S, He K (2021) An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9640–9649
29. Chen TS, Hung WC, Tseng HY, Chien SY, Yang MH (2021b) Incremental false negative detection for contrastive learning. Preprint arXiv:2106.03719
30. 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
31. Li J, Zhou P, Xiong C, Hoi SC (2020) Prototypical contrastive learning of unsupervised representations. arXiv:2005.04966
32. Lim S, Park J, Lee M, Lee H (2023) Unsupervised object discovery with pseudo label generated using k-means and self-supervised transformer. Neurocomputing 545:126326
33. Zhuang C, Zhai AL, Yamins D (2019) Local aggregation for unsupervised learning of visual embeddings. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6002–6012
34. Caron M, Bojanowski P, Joulin A, Douze M (2018) Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV), pp 132–149
35. Van Gansbeke W, Vandenhende S, Georgoulis S, Proesmans M, Van Gool L (2020) Scan: learning to classify images without labels. In: European conference on computer vision, Springer, pp 268–285
36. Niu C, Shan H, Wang G (2022) Spice: semantic pseudo-labeling for image clustering. IEEE Trans Image Process 31:7264–7278
37. 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

38. Mao Z, Zhou Y, Sun J, Wu H, Pan F, Ahmad B (2023) Weakly-supervised object localization with gradient-pyramid feature. Appl Intell 53(3):2923–2935
39. Ramaswamy HG et al (2020) Ablation-cam: visual explanations for deep convolutional network via gradient-free localization. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 983–991
40. Jia Q, Wei S, Ruan T, Zhao Y, Zhao Y (2021) Gradingnet: towards providing reliable supervisions for weakly supervised object detection by grading the box candidates. Proceedings of the AAAI Conference on Artificial Intelligence, vol 35, pp 1682–1690
41. Ren Z, Tang Y, Zhang W (2023) Ido: instance dual-optimization for weakly supervised object detection. Appl Intell 1–18
42. Feng X, Han J, Yao X, Cheng G (2020) Tcanet: triple context-aware network for weakly supervised object detection in remote sensing images. IEEE Trans Geosci Remote Sens 59(8):6946–6955
43. Zhong Y, Wang J, Peng J, Zhang L (2020) Boosting weakly supervised object detection with progressive knowledge transfer. In: European conference on computer vision, Springer, pp 615–631
44. Hou L, Zhang Y, Fu K, Li J (2021) Informative and consistent correspondence mining for cross-domain weakly supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9929–9938
45. Cai Y, Tan X, Tan X (2017) Selective weakly supervised human detection under arbitrary poses. Pattern Recognit 65:223–237
46. Huang Z, Bao Y, Dong B, Zhou E, Zuo W (2022) W2n: switching from weak supervision to noisy supervision for object detection. In: European conference on computer vision, Springer, pp 708–724
47. Chen T, Kornblith S, Norouzi M, Hinton G (2020a) A simple framework for contrastive learning of visual representations. In: International conference on machine learning, PMLR, pp 1597–1607
48. Chen T, Kornblith S, Swersky K, Norouzi M, Hinton GE (2020) Big self-supervised models are strong semi-supervised learners. Adv Neural Inf Process Syst 33:22243–22255
49. Deselaers T, Alexe B, Ferrari V (2012) Weakly supervised localization and learning with generic knowledge. Int J Comput Vis 100(3):275–293
50. Deselaers T, Alexe B, Ferrari V (2012) Weakly supervised localization and learning with generic knowledge. Int J Comput Vis 100(3):275–293
51. Li X, Kan M, Shan S, Chen X (2019) Weakly supervised object detection with segmentation collaboration. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9735–9744
52. Wan F, Liu C, Ke W, Ji X, Jiao J, Ye Q (2019) C-mil: continuation multiple instance learning for weakly supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2199–2208
53. Yang K, Li D, Dou Y (2019) Towards precise end-to-end weakly supervised object detection network. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8372–8381
54. Chen Z, Fu Z, Jiang R, Chen Y, Hua XS (2020) Slv: spatial likelihood voting for weakly supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12995–13004
55. Arun A, Jawahar CV, Kumar MP (2019) Dissimilarity coefficient based weakly supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9432–9441
56. Gao Y, Liu B, Guo N, Ye X, Wan F, You H, Fan D (2019) C-midn: coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9834–9843
57. Tao Q, Yang H, Cai J (2018) Exploiting web images for weakly supervised object detection. IEEE Trans Multimed 21(5):1135–1146
58. Dong B, Huang Z, Guo Y, Wang Q, Niu Z, Zuo W (2021) Boosting weakly supervised object detection via learning bounding box adjusters. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2876–2885
59. Cao T, Du L, Zhang X, Chen S, Zhang Y, Wang YF (2021) Cat: weakly supervised object detection with category transfer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3070–3079
60. Ren S, He K, Girshick R, Sun J (2017) Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
61. Gokberk Cinbis R, Verbeek J, Schmid C (2014) Multi-fold mil training for weakly supervised object localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2409–2416
62. Bilen H, Pedersoli M, Tuytelaars T (2015) Weakly supervised object detection with convex clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1081–1089
63. Wang C, Ren W, Huang K, Tan T (2014) Weakly supervised object localization with latent category learning. In: European conference on computer vision, Springer, pp 431–445
64. Li D, Huang JB, Li Y, Wang S, Yang MH (2016) Weakly supervised object localization with progressive domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3512–3520
65. Teh EW, Rochan M, Wang Y (2016) Attention networks for weakly supervised object localization. In: BMVC, pp 1–11
66. Kantorov V, Oquab M, Cho M, Laptev I (2016) Contextlocnet: context-aware deep network models for weakly supervised localization. In: European conference on computer vision, Springer, pp 350–365
67. Jie Z, Wei Y, Jin X, Feng J, Liu W (2017) Deep self-taught learning for weakly supervised object localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1377–1385
68. Diba A, Sharma V, Pazandeh A, Pirsiavash H, Van Gool L (2017) Weakly supervised cascaded convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 914–922
69. Wan F, Wei P, Jiao J, Han Z, Ye Q (2018) Min-entropy latent model for weakly supervised object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1297–1306
70. Tang P, Wang X, Wang A, Yan Y, Liu W, Huang J, Yuille A (2018) Weakly supervised region proposal network and object detection. In: Proceedings of the European conference on computer vision (ECCV), pp 352–368
71. Shen Y, Ji R, Wang Y, Wu Y, Cao L (2019) Cyclic guidance for weakly supervised joint detection and segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 697–707
72. Gao M, Li A, Yu R, Morariu VI, Davis LS (2018) C-wsl: count-guided weakly supervised localization. In: Proceedings of the European conference on computer vision (ECCV), pp 152–168
73. Sun G, Wang W, Dai J, Van Gool L (2020) Mining cross-image semantics for weakly supervised semantic segmentation. In: European conference on computer vision, Springer, pp 347–365
74. Huang Z, Zou Y, Kumar B, Huang D (2020) Comprehensive attention self-distillation for weakly-supervised object detection. Adv Neural Inf Process Syst 33:16797–16807

Yongqiang Zhang received M.S. and Ph.D. degrees in instrument science and technology from the Harbin Institute of Technology (HIT), Harbin, China, in 2015 and 2020, respectively. He worked at King Abdullah University of Science and Technology (KAUST) as a visiting student from 2017 to 2018. Currently, he is an assistant professor in the School of Instrumentation Science and Engineering at the Harbin Institute of Technology. His research areas are computer vision, pattern recognition, machine learning, and deep learning. His research interests mainly include face detection, weakly/fully supervised object detection, activity detection, and image and video understanding in the real world.

Rui Tian received the MS degree in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 2023. He is currently a PhD student from the Harbin Institute of Technology (HIT). His research areas are computer vision, pattern recognition, machine learning, and deep learning. His research interests mainly include weakly supervised object detection, depth estimation and contrastive learning.

Yin Zhang received the MS degree in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 2021. He is currently a PhD student from the Harbin Institute of Technology (HIT). His research areas are computer vision, pattern recognition, machine learning, and deep learning. His research interests mainly include object detection, knowledge distillation, image super-resolution.

Zian Zhang received the B.S. and M.S. degrees in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 2018 and 2020, respectively. Currently, he is a doctoral student in the School of Instrumentation Science and Engineering at Harbin Institute of Technology. His research areas are computer vision, pattern recognition, machine learning, and deep learning. His research interests mainly include human pose estimation, object detection, action recognition, gait recognition, image and video understanding in the monitoring system.

Yancheng Bai is an associate professor in Pattern Recognition and Intelligent System at Institute of Software, Chinese Academy of Sciences, Beijing, China. He received the MS degree from Sichuan University, Sichuan, China, in 2009. He received the PhD degree from Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014. He worked as a post-doctoral at King Abdullah University of Science and Technology (KAUST) from 2016 to 2018. His research interests include computer vision, pattern recognition, artificial intelligence, machine learning, and deep learning.

Mingli Ding received B.S., M.S. and Ph.D. degrees in instrument science and technology from the Harbin Institute of Technology (HIT), Harbin, China, in 1996, 1997 and 2001, respectively. He worked as a visiting scholar in France from 2009 to 2010. Currently, he is a professor in the School of Instrumentation Science and Engineering at the Harbin Institute of Technology. Prof. Ding's research interests are intelligence tests and information processing, automation test technology, computer vision, and machine learning. He has published over 40 papers in peer-reviewed journals and conferences.

Wangmeng Zuo received his Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007. From 2004 to 2006, he was a research assistant with the Department of Computing, the Hong Kong Polytechnic University, Hong Kong. From 2009 to 2010, he was a visiting professor with Microsoft Research Asia. He is currently a professor with the School of Computer Science and Technology, the Harbin Institute of Technology. He has published over 100 articles in top-tier academic journals and conferences. His current research interests include image enhancement and restoration, image/video generation, and image classification. He has served as an associate editor for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and a senior editor of the Journal of Electronic Imaging.

Authors and Affiliations

Yongqiang Zhang[1] · Rui Tian[1] · Yin Zhang[1] · Zian Zhang[1] · Yancheng Bai[2] · Mingli Ding[1] · Wangmeng Zuo[3]

Rui Tian
rui.tian.hit@gmail.com

Yin Zhang
yin.zhang.hit@gmail.com

Zian Zhang
20b901007@stu.hit.edu.cn

Yancheng Bai
yancheng.bai.1987@gmail.com

Mingli Ding
dingml@hit.edu.cn

Wangmeng Zuo
cswmzuo@gmail.com

[1] School of Instrumentation Science and Engineering, Harbin Institute of Technology, Harbin, China

[2] Institute of Software, Chinese Academy of Sciences, Beijing, China

[3] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China