

Self-Classification Enhancement and Correction for Weakly Supervised Object Detection

Yufei Yin¹, Lechao Cheng^{2*}, Wengang Zhou³, Jiajun Deng⁴, Zhou Yu^{1*} and Houqiang Li³

¹School of Computer Science, Hangzhou Dianzi University

²School of Computer Science and Information Engineering, Hefei University of Technology

³EEIS Department, University of Science and Technology of China

⁴Australian Institute for Machine Learning, University of Adelaide

yinyf@hdu.edu.cn, chenglc@hfut.edu.cn, zhgw@ustc.edu.cn,
jiajun.deng@adelaide.edu.au, yuz@hdu.edu.cn, lihq@ustc.edu.cn

Abstract

In recent years, weakly supervised object detection (WSOD) has attracted much attention due to its low labeling cost. The success of recent WSOD models is often ascribed to the two-stage multi-class classification (MCC) task, *i.e.*, multiple instance learning and online classification refinement. Despite achieving non-trivial progresses, these methods overlook potential classification ambiguities between these two MCC tasks and fail to leverage their unique strengths. In this work, we introduce a novel WSOD framework to ameliorate these two issues. For one thing, we propose a self-classification enhancement module that integrates intra-class binary classification (ICBC) to bridge the gap between the two distinct MCC tasks. The ICBC task enhances the network’s discrimination between positive and mis-located samples in a class-wise manner and forges a mutually reinforcing relationship with the MCC task. For another, we propose a self-classification correction algorithm during inference, which combines the results of both MCC tasks to effectively reduce the misclassified predictions. Extensive experiments on the prevalent VOC 2007 & 2012 datasets demonstrate the superior performance of our framework.

1 Introduction

Object detection aims to localize objects of interest and classify them, which is a fundamental task in the field of computer vision. The recent decade has witnessed rapid progress [Girshick, 2015; Ren *et al.*, 2015; Liu *et al.*, 2016; Redmon *et al.*, 2016] in various object detection scenarios [Nie *et al.*, 2023; Wang *et al.*, 2023; Jiao *et al.*, 2024; Wang *et al.*, 2024b], benefiting from the development of convolutional neural networks (CNN). In spite of the remarkable advances, current fine-grained instance-level annotations are labor-intensive and time-consuming to obtain. This paper focuses on the domain of weakly supervised object detection

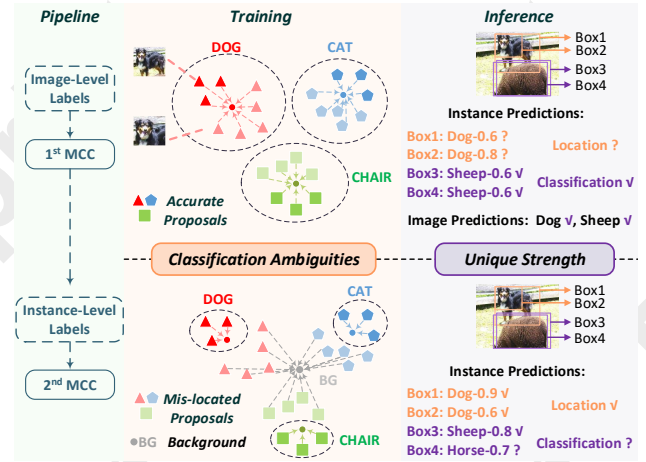


Figure 1: Comparison between the two distinct MCC tasks.

(WSOD) [Su *et al.*, 2022], which requires only image-level annotations, *i.e.*, existing object categories in a given image, to achieve the object detection task.

Recent WSOD approaches [Tang *et al.*, 2017; Wei *et al.*, 2018; Ren *et al.*, 2020; Huang *et al.*, 2020] generally convert WSOD into a two-stage multi-class classification (MCC) pipeline. In the first stage, a multiple instance detection network (MIDN) [Bilen and Vedaldi, 2016] is constructed, leveraging multiple instance learning to introduce competition both among object classes and proposals (denoted as 1st MCC). This process effectively identifies candidate regions that contain significant class-specific patterns. However, MIDN often suffers from partial-located issues, wherein high scores are assigned to the detections localizing only the most discriminative parts. To overcome this challenge, in the second stage, cascaded online multi-class classifiers [Tang *et al.*, 2017; Chen *et al.*, 2020] are integrated to refine the classification scores of MIDN, and assorted strategies [Zeng *et al.*, 2019; Ren *et al.*, 2020] are designed to generate pseudo instance-level labels for training these classifiers (denoted as 2nd MCC). Despite the promising results achieved by these methods, as shown in Figure 1, they overlook the potential classification ambiguities and the unique strengths of the two

*Corresponding authors: Lechao Cheng and Zhou Yu

distinct multi-class classification tasks across the two stages:

- (i) During the 1st MCC task, some mis-located proposals, especially those only containing discriminative parts of an object, are classified as the corresponding object class and leveraged to generate its class-specific features. However, after multiple refinements in the 2nd MCC task, these proposals will be classified as the background class, whose features are instead pushed toward those of mis-located proposals from other classes. These ambiguities compromise the quality of the class-specific features produced by the detector.
- (ii) These two MCC tasks are guided by image-level and pseudo-instance-level labels, respectively. As a result, the first MCC task excels at identifying the classes present in the image, while the second one focuses on more accurate instance-level location. However, previous methods rely solely on the second MCC task during inference, overlooking the classification benefits provided by the first task.

In this paper, We present a novel self-classification enhancement and correction (SCEC) framework to overcome these two limitations. To alleviate the *classification ambiguities*, we introduce a self-classification enhancement module during the second stage, which integrates an extra intra-class binary classification (ICBC) task to bridge the gap between the two distinct MCC tasks. ICBC task aims to enhance the network’s discrimination between positive and mis-located samples in a class-specific manner, rather than directly grouping them together with background samples into a single ‘background’ class, as done in the 2nd MCC. To sufficiently optimize the ICBC classifiers, we generate various types of mis-located samples based on the 2nd MCC results. Furthermore, the ICBC results are utilized in reverse to refine the pseudo labels for the 2nd MCC task, allowing the two tasks to complement each other. To harness the *unique strengths* of the two distinct MCC tasks, we introduce a self-classification correction algorithm, which leverages the 1st MCC results to rectify mis-classifications in the detections produced by the second one.

Our primary contributions are summarized as follows:

- We propose a self-classification enhancement module that incorporates both the base multi-class classification and an intra-class binary classification to alleviate the classification ambiguities. These two tasks are carried out in a mutually reinforcing manner.
- We introduce a self-classification correction algorithm to alleviate the mis-classification problem of detections, leveraging the image-level classification strength of the first multi-class classification results during inference.
- Extensive experiments on the prevalent PASCAL VOC 2007 and 2012 datasets demonstrate the superior performance of our framework.

2 Related Work

Weakly Supervised Object Detection. Weakly Supervised Object Detection (WSOD) has been widely studied in recent years. The pioneering work WSDDN [Bilen and Vedaldi,

2016] first integrates multiple instance learning into the CNN architecture by designing a two-stream network, *i.e.* classification branch and detection branch. By combining the results from these two branches, WSDDN converts the WSOD task into a multi-class classification problem for proposals. However, such a solution often struggles to generate accurate detections. To alleviate this problem, OICR [Tang *et al.*, 2017] proposes a two-stage pipeline where WSDDN is utilized as a basic detector, and its results are utilized to generate pseudo seed boxes for training several subsequent online instance classifiers for further refinement. Most recent WSOD approaches are developed based on this pipeline. Some methods improve the detection capability of the basic detector, *e.g.*, adding extra supplement modules for more complete detections [Yan *et al.*, 2019; Yin *et al.*, 2021; Yin *et al.*, 2022], and enhancing the generated image or proposal feature [Ren *et al.*, 2020; Huang *et al.*, 2020]. Some other methods design various strategies to improve the quality of pseudo seed boxes, *e.g.*, constructing spatial graphs [Tang *et al.*, 2018] or appearance graphs [Lin *et al.*, 2020] for top-scoring proposals, bringing top-down objectness [Zeng *et al.*, 2019], and applying spatial likelihood voting [Chen *et al.*, 2020]. Otherwise, [Yang *et al.*, 2019] add online regression branches to refine the initial proposals.

Different from them, our method extends the widely used online classifier into a self-classification enhancement module, which brings intra-class binary classification to enhance the network’s discrimination between class-specific positive and mis-located samples, thus improving the detection capability of the network.

3 Method

3.1 Overview

The overview of the proposed model is illustrated in Figure 2. First, an image and the generated region proposals are fed to the RoI feature extractor, *i.e.*, CNN backbone and an RoI pooling layer followed by two FC layers, to obtain proposal feature vectors. Next, the feature factors are fed into MIDN module to produce instance-level scores, which are summed for the training of MIDN with image-level labels. Meanwhile, the feature factors are fed into several subsequent self-classification enhancement (SCE) module to obtain MCC and ICBC scores. For each SCE module, the ICBC branch is trained using MCC scores, while the pseudo labels of MCC branch are generated from ICBC Guided Seed Mining (IGSM) algorithm. After that, an R-CNN head is constructed to produce classification scores and regression coordinates. During inference, self-classification correction (SCC) algorithm is applied to generate prediction results.

3.2 Basic WSOD Framework

In the weakly supervised setting, distinguishing between positive and negative proposals directly during training becomes challenging, given the absence of instance-level labels. To overcome this problem, the pioneering work WSDDN [Bilen and Vedaldi, 2016] adopts multiple instance learning into the CNN architecture to convert the WSOD task into the

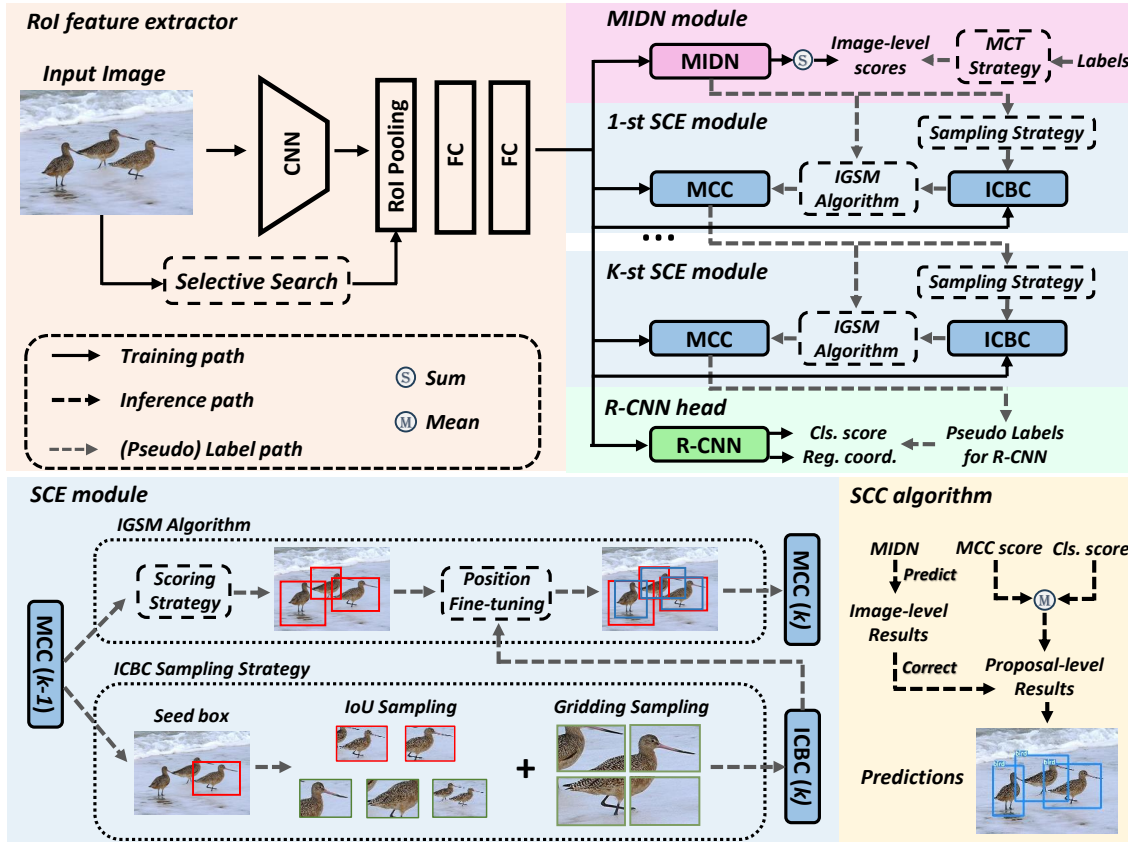


Figure 2: An overview of our self-classification enhancement and correction (SCEC) framework.

multi-class classification task for proposals. Following recent works [Tang *et al.*, 2017; Ren *et al.*, 2020], we apply WSDN as our basic detector, referred as Multiple Instance Detector Network (MIDN).

Given an image I , its image-level labels $Y = [y_1, y_2, \dots, y_C] \in \mathbb{R}^{C \times 1}$ is available according to the WSOD setting, where $y_c \in \{0, 1\}$ indicates the presence or absence of class c . Its proposals $R = \{R_1, R_2, \dots, R_N\}$ are pre-generated from Selective Search [Uijlings *et al.*, 2013] before training. First, the proposal feature vectors are generated through a CNN backbone, an RoI pooling layer [Girshick, 2015], and two FC layers. Next, these vectors are fed into two parallel branches, *i.e.*, classification and detection branches, to obtain proposal scores. For each branch, a scoring matrix $x^{\text{cls}}(x^{\text{det}}) \in \mathbb{R}^{C \times |R|}$ is first obtained by an FC layer, where $|R|$ and C represents the number of proposals and categories, respectively. The two scoring matrices are then normalized by the softmax operation through orthogonal directions, *i.e.* $\sigma(x^{\text{cls}})$ for category direction and $\sigma(x^{\text{det}})$ for proposal direction. After that, the proposal scores are generated by the element-wise product of these two matrices: $x^{\text{box}} = \sigma(x^{\text{cls}}) \odot \sigma(x^{\text{det}})$. Finally, the image-level scores are obtained by aggregating over the proposal direction of x^{box} : $x_c^{\text{img}} = \sum_{i=1}^{|R|} x_{c,i}^{\text{box}}$. In this way, the image-level labels can be utilized for supervision through binary cross-entropy loss: $\mathcal{L}_{\text{MIDN}} = -\sum_{c=1}^C [y_c \log x_c^{\text{img}} + (1 - y_c) \log (1 - x_c^{\text{img}})]$.

Following OICR [Tang *et al.*, 2017], the basic WSOD framework adds several online instance classification (OIC) branches after the basic detector to generate more accurate detections. Each branch contains an FC layer and a softmax operation and outputs a scoring matrix $x^{\text{oic}} \in \mathbb{R}^{(C+1) \times |R|}$. The top-scoring proposals from the C -th branch are utilized to generate pseudo labels y^{oic} to train the $C + 1$ -th branch, through the cross-entropy loss: $\mathcal{L}_{\text{OIC}} = -\frac{1}{|R|} \sum_{i=1}^{|R|} \sum_{c=1}^{C+1} w_i y_{c,i}^{\text{oic}} \log x_{c,i}^{\text{oic}}$. The loss weight w_i , which acts as a confidence score, is obtained from the score of the seed box which has the highest overlaps with R_i . Additionally, following [Yang *et al.*, 2019; Yin *et al.*, 2021], we construct an R-CNN branch subsequently, which contains a classification sub-branch and a regression sub-branch. The weighted cross-entropy loss and smooth-L1 loss are applied to train these two sub-branches, respectively.

3.3 Self-Classification Enhancement

The basic WSOD framework refines the initial detection results by applying OIC branches with multi-class classification. However, as illustrated in Sec. 1, this approach lead to classification ambiguities, where the features of mis-located samples from all classes are pushed together, despite that they are utilized to generate class-specific features during MIDN’s training process. Furthermore, such a solution will weaken the model’s ability to distinguish the mis-located samples from their closed intra-class positive ones. To this end, we

introduce Self-Classification Enhancement (SEC) module to tackle this problem.

The SCE module comprises two parallel branches: one for the base multi-class classification and the other for an enhanced intra-class binary classification. These branches work harmoniously during online training, supplementing each other’s strengths.

Multi-Class Classification. The multi-class classification (MCC) layer shares the same structure with the original OIC layer, containing an FC layer followed by a softmax.

Intra-Class Binary Classification. We incorporate the intra-class binary classification (ICBC) task to enhance the network’s discrimination between intra-class positive and mis-located samples. To maintain consistency with the MCC layer, we adopt a simple yet effective design to achieve the ICBC task. Specifically, given the proposal feature vector f , the ICBC branch consists of an FC layer to generate score matrices and a sigmoid function for normalization:

$$x_{c,i}^{\text{icbc}} = \sigma(FC(f)), \quad x^{\text{icbc}} \in \mathbb{R}^{C \times |R|}, \quad (1)$$

where a higher $x_{c,i}^{\text{icbc}}$ indicates that the proposal R_i is more likely to be a positive sample for class c , while a lower value suggests the opposite.

Sampling strategy for ICBC task. We adopt MCC results to select training samples U for the ICBC task. Specifically, we first choose a set of top-scoring proposals as the pseudo seed boxes $S = \{S_1, S_2, \dots, S_N\}$ according to MCC results. After that, we propose different strategies to select positive and mis-located samples based on these seed boxes.

A straightforward way is to apply the *IoU sampling strategy*. Specifically, for each proposal, we calculate its Intersection over Union (IoU) with all seed boxes, and take the maximum value IoU_i . Next, we denote the positive samples as $R_{\text{pos}} = \{R_i | IoU_i \geq \tau_h\}$ and the mis-located samples as $R_{\text{neg}} = \{R_i | \tau_l \leq IoU_i < \tau_h\}$, where IoU_i is the overlaps between R_i and its closest seed box S_j . Correspondingly, R_i shares the same category C_i with S_j . Then, we generate the pseudo label $Y_i^{\text{icbc}} = [y_{1,i}^{\text{icbc}}, y_{2,i}^{\text{icbc}}, \dots, y_{C,i}^{\text{icbc}}]$ of proposal i according to the divisions:

$$y_{c,i}^{\text{icbc}} = \begin{cases} 1, & R_i \in R_{\text{pos}} \text{ and } C_i = c, \\ 0, & \text{else.} \end{cases} \quad (2)$$

However, the samples selected from IoU sampling strategy are insufficient for achieving the ICBC task. Given the limited number of training samples, we further apply a *gridding sampling strategy* to generate additional mis-located samples for augmentation. Specifically, for each seed box with class c , we first apply box scaling by a scaling factor $\theta = 0.5$. In this way, the width and height of the seed box are randomly sampled in $[(1-\theta)w, (1+\theta)w]$ and $[(1-\theta)h, (1+\theta)h]$, respectively. The center of the box remains unchanged. Afterward, we generate an $n \times n$ grid on the scaled seed box. We treat each grid as a potential mis-located sample that may only contain a part of an object. It is worth noting that although some grids containing only background may inevitably be selected, their impact is minimal due to their limited quantity. These selected grids $G = \{G_1, G_2, \dots, G_N\}$ are then fed to

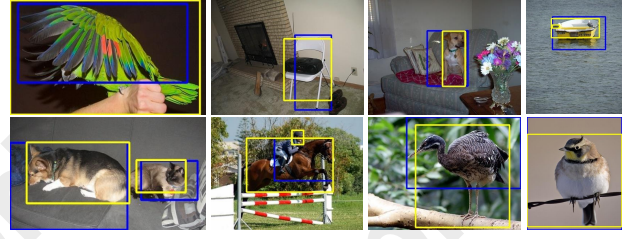


Figure 3: Comparison between the seed boxes selected from MCC scores (blue) and ICBC scores (yellow).

the RoI pooling layer to generate region features, which are subsequently passed to the ICBC layer. Their pseudo labels are assigned as follows: $y_{c,G_i}^{\text{icbc}} = 0, G_i \in G$.

We denote all the selected samples as $U = R_{\text{pos}} \cup R_{\text{neg}} \cup G$. Considering the division of positive and mis-located samples is class-specific, we utilize sample weight to ensure these samples only participate in the losses of their corresponding categories:

$$w_{c,i}^{\text{icbc}} = \begin{cases} p_i, & U_i \in R_{\text{pos}} \cup R_{\text{neg}} \text{ and } C_i = c, \\ q_i, & U_i \in G \text{ and } C_i = c, \\ 0, & \text{else,} \end{cases} \quad (3)$$

where p_i is the confidence of R_i , obtained by the MCC score of its closest seed box S_j , and q_i is set to 1.5. Finally, we adopt weighted binary cross-entropy loss for training ICBC layer:

$$\mathcal{L}_{ICBC} = -\frac{1}{|U|} \sum_{i=1}^{|U|} \sum_{c=1}^C w_{c,i}^{\text{icbc}} BCE(x_{c,i}^{\text{icbc}}, y_{c,i}^{\text{icbc}}). \quad (4)$$

ICBC Guided Seed Mining. Seed box mining plays a significant role in training the MCC branch. Some methods pursue the accuracy of seed boxes by using top-scoring strategies [Tang *et al.*, 2017], while some others focus on the recall of seed boxes by relaxing the top criteria and adopting non-maximum suppression (NMS) algorithm to remove redundant ones [Ren *et al.*, 2020]. Different from them, we adopt a soft scoring threshold to mine accurate seed boxes and utilize ICBC results for further fine-tuning.

Specifically, for each existing class c ($y_c = 1$), we first find the proposal R_k with the top score $x_{c,k}^{\text{mcc}}$. Here, we adopt the scores from the previous MCC branch following [Tang *et al.*, 2017]. Next, instead of selecting top- K proposals, we set a soft scoring threshold according to the top score in this class: $\tau_{\text{score}} = \alpha x_{c,k}^{\text{mcc}}$. We select the proposals whose MCC scores are higher than the threshold τ_{score} . After that, we apply NMS algorithm on the selected proposals to remove redundant boxes, obtaining the base seed boxes S_{base} .

Compared to the MCC, ICBC demonstrates superior proficiency in discerning the positive samples from their closed but mis-located ones in a class-specific manner. To this end, we apply ICBC results to fine-tune the obtained base seed boxes S_{base} . Briefly, for each seed box S_i in S_{base} , we first obtain its surrounding proposals by setting an overlap threshold ($\tau_{\text{sur}} = 0.5$). Next, among all the surrounding proposals, we select the one S_j that has the maximum ICBC score in

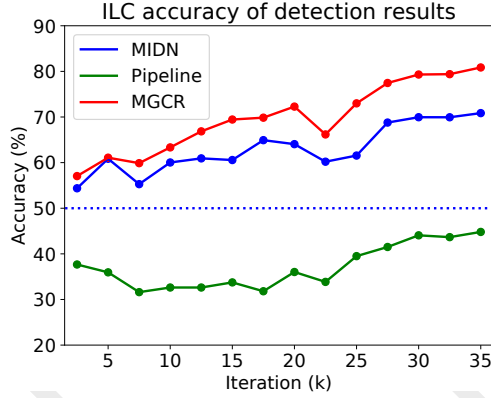


Figure 4: Comparison between image-level classification (ILC) accuracy of detections from the common pipeline (green), MIDN (blue) and SCC algorithm (red) under different training iterations.

the class of S_i , and add it to the seed boxes. According to the previous comparison between ICBC and MCC, S_i can be regarded as a potential refinement of S_i in location, which is shown in Figure 3.

Finally, the fine-tuned seed boxes S_{ft} , along with the base seed boxes S_{base} , are utilized to train the MCC layer. We generate pseudo labels of all the proposals according to their max overlaps with seed boxes. The overlap thresholds are set the same with those in ICBC training, while the mis-located samples are labeled as $C + 1$. After that, we use these pseudo labels to train MCC layer with weighted cross-entropy loss:

$$\mathcal{L}_{MCC} = -\frac{1}{|R|} \sum_{i=1}^{|R|} \sum_{c=1}^{C+1} w_i^{\text{mcc}} y_{c,i}^{\text{mcc}} \log x_{c,i}^{\text{mcc}}, \quad (5)$$

where $y_{c,i}^{\text{mcc}}$ and $x_{c,i}^{\text{mcc}}$ represent the pseudo labels and MCC score of proposal R_i in class c , respectively. The loss weight w_i^{mcc} is the score of the seed box which has the highest overlaps with R_i .

Finally, we replace the original OIC branches with our SEC modules, and we train the network end-to-end by combining all the losses mentioned before:

$$\mathcal{L}_{total} = \mathcal{L}_{MIDN} + \sum_{t=1}^T (\mathcal{L}_{MCC} + \gamma \mathcal{L}_{ICBC}) + \mathcal{L}_{R-CNN}, \quad (6)$$

where \mathcal{L}_{R-CNN} is the loss for R-CNN branch, T is the number of online enhanced instance classification modules, and γ keeps the balance between \mathcal{L}_{MCC} and \mathcal{L}_{ICBC} .

3.4 Self-Classification Correction

A common pipeline for WSOD inference involves two main steps: 1) aggregating the online MCC scores (e.g., OICs & classification scores in R-CNN branch) as the final results; 2) refining the boxes through the regression outputs. In contrast, the MIDN result is empirically discarded due to its inadequate detection performance. We often observe a common flaw in such a solution: Some non-exist categories are incorrectly assigned high scores in the final results, leading to many mis-classification samples. We conduct a toy experiment to empirically show the suboptimal classification of final results. We perform statistical analysis on the occurrences

of all detected boxes' predicted categories within the image. In other words, if the category appears in the image, the box will be considered a positive sample; otherwise, it is deemed a negative sample. This methodology allows us to calculate the image-level classification (ILC) accuracy of the bounding boxes. We evaluate multiple models at different training iterations, as shown in Figure 4, and compare the detection results from the conventional pipeline (green line) with those from MIDN (blue line).

Intrigued, two observations stand out: 1) the ILC accuracy of pipeline result falls short of the 50% mark, and 2) the MIDN result exhibits notably higher ILC accuracy compared to the pipeline one. We ascribe this disparity to two main factors. On one hand, only parts of confident proposals participate in the training of the OICs, thus the penalty for non-existent categories has been attenuated. On the other hand, MIDN employs binary cross-entropy loss for the summation of the instance-level scores, which exerts stronger constraints on the image-level classification.

Based on this observation, we propose a simple yet effective strategy to refine the pipeline results with MIDN, termed as Self-Classification Correction (SCC). Briefly, we first predict the existing image-level categories according to MIDN results through confident proposal selection and an argmax operation, and then obtain the absent categories C_N :

$$\begin{aligned} IND &\leftarrow \left\{ i \mid \max_c \mathcal{M}_{i,c} > \tau_{midn} \right\}, \\ C_P &\leftarrow \arg \max_c \mathcal{M}_{IND,c}, \\ C_I &\leftarrow \text{unique}(C_P), \\ C_N &\leftarrow \mathcal{C} - C_I, \end{aligned} \quad (7)$$

where \mathcal{M} is MIDN results, \mathcal{C} is the set of classes, and τ_{midn} is the scoring threshold for MIDN. Subsequently, we reduce the pipeline scores of these absent categories by multiplying a scoring factor λ .

As shown in Figure 4, SCC algorithm (red line) brings about substantial enhancements in ILC accuracy compared to the pipeline (green line), showcasing improvements ranging from 20% to 35%. Remarkably, pipeline utilizing SCC even achieves superior ILC accuracy when compared to MIDN. Consequently, this refinement strategy contributes to a notable reduction in instances of mis-classification, thus enhancing detector performance without introducing additional training parameters.

Additionally, to make the SCC algorithm more effective, we adopt the misclassification tolerance (MCT) strategy [Wu et al., 2024] to further improve the classification ability of MIDN. The motivation behind the MCT strategy is to introduce tolerance for unrepresentative samples with high semantic similarity to an incorrect class, thereby preventing these samples from dominating the training process and forcing the model to memorize them. Suppose N_p classes are present in the image, we identify the misclassified classes containing unrepresentative samples when their score rankings fall within the range of $[N_p, N_p + T_n]$, and assign their class weights to a when calculating \mathcal{L}_{MIDN} . Similarly, the corresponding incorrect classes, whose score rankings fall within the range of $[0, N_p]$, are also assigned the same class weight.

Methods	VOC 2007	VOC 2012
OICR [Tang <i>et al.</i> , 2017]	41.2	37.9
WS-JDS [Shen <i>et al.</i> , 2019]	45.6	39.1
C-MIL [Wan <i>et al.</i> , 2019]	50.5	46.7
Yang <i>et al.</i> [Yang <i>et al.</i> , 2019]	51.5	46.8
C-MIDN [Yan <i>et al.</i> , 2019]	52.6	50.2
Pred Net [Arun <i>et al.</i> , 2019]	52.9	48.4
SLV [Chen <i>et al.</i> , 2020]	53.5	49.2
WSOD ² [Zeng <i>et al.</i> , 2019]	53.6	47.2
CASD [Huang <i>et al.</i> , 2020]	56.8	53.6
MIST [Ren <i>et al.</i> , 2020]	54.9	52.1
IM-CFB [Yin <i>et al.</i> , 2021]	54.3	49.4
SPE [Liao <i>et al.</i> , 2022]	51.0	-
ODCL [Seo <i>et al.</i> , 2022]	56.1	54.6
CBL [Yin <i>et al.</i> , 2023]	57.4	53.5
NDI-MIL [Wang <i>et al.</i> , 2024a]	56.8	53.9
Ours	58.2	55.5

Table 1: Performance comparison among the state-of-the-art methods on PASCAL VOC 2007 and 2012. These models are evaluated in terms of mAP (%). We highlight the best and second best performance in the **red** and **blue** colors.

4 Experiments and Analysis

4.1 Datasets

Following previous works, we evaluate our proposed method on two popular object detection datasets Pascal VOC 2007 and Pascal VOC 2012 [Everingham *et al.*, 2010], which contain 20 categories. For both two datasets, we train on *trainval* splits (5,011 images in VOC 2007 and 11,540 images in VOC 2012) and applies two kinds of metrics for evaluation: (1) The mean of average precision (mAP) on the *test* split (4,951 images in VOC 2007 and 10,991 images in VOC 2012); (2) Correct localization (CorLoc) on the *trainval* split. Only image-level labels are utilized during training.

4.2 Implementation Details

Following a widely-used setting, we adopt VGG16 [Simonyan and Zisserman, 2014] pre-trained on ImageNet [Deng *et al.*, 2009] as the backbone and Selective Search [Uijlings *et al.*, 2013] for proposal generation. The whole framework is end-to-end optimized using stochastic gradient descent (SGD), and the momentum, weight decay, and batch size are set as 0.9, 5×10^{-4} , and 4, respectively. The initial learning rate is set as 1×10^{-3} for the first 70k, 170k iterations, and it is dropped by a factor of 10 for the following 20k, 40k iterations for VOC 2007 and VOC 2012, respectively. We set α to 0.9 and the NMS threshold τ_{nms} to 0.1 in the SEC module. λ and τ_{midn} in SCC algorithm are set to 0.01 and 0.001, respectively. The hyperparameters of MCT strategy are set as the same with [Wu *et al.*, 2024], *i.e.*, $T_n = 1$, $a = 0.4$. The loss weight γ is set to 0.1 for the training balance. Following the previous WSOD works, τ_l , τ_h , and T are set to 0.1, 0.5, and 3, and the images are multi-scaled to {480, 576, 688, 864, 1000, 1200} for both training and inference.

Methods	VOC 2007	VOC 2012
OICR [Tang <i>et al.</i> , 2017]	60.6	52.1
C-MIL [Wan <i>et al.</i> , 2019]	65.0	67.4
Yang <i>et al.</i> [Yang <i>et al.</i> , 2019]	68.0	69.5
C-MIDN [Yan <i>et al.</i> , 2019]	68.7	71.2
WSOD ² [Zeng <i>et al.</i> , 2019]	69.5	71.9
SLV [Chen <i>et al.</i> , 2020]	71.0	69.2
MIST [Ren <i>et al.</i> , 2020]	68.8	70.9
CASD [Huang <i>et al.</i> , 2020]	70.4	72.3
IM-CFB [Yin <i>et al.</i> , 2021]	70.7	69.6
SPE [Liao <i>et al.</i> , 2022]	70.4	-
ODCL [Seo <i>et al.</i> , 2022]	69.8	71.2
CBL [Yin <i>et al.</i> , 2023]	71.8	72.6
NDI-MIL [Wang <i>et al.</i> , 2024a]	71.0	72.2
Ours	71.9	73.4

Table 2: Performance comparison among the state-of-the-art methods on PASCAL VOC 2007 and 2012. These models are evaluated in terms of CorLoc (%). We highlight the best and second best performance in the **red** and **blue** colors.

4.3 Comparison with State-of-the-art Methods

In Table 1, we compare the performance of the state-of-art methods with single model on Pascal VOC 2007 and VOC 2012 datasets in terms of mAP. Our method achieves state-of-the-art performance with 58.2% mAP on VOC 2007 and 55.5% mAP on VOC 2012, surpassing other methods by at least 0.8% and 0.9%, respectively. Our method also achieves outstanding performance in terms of CorLoc, setting new state-of-the-art benchmarks with 71.9% on VOC 2007 and 73.4% on VOC 2012. Our method outperforms recent works [Ren *et al.*, 2020] and [Chen *et al.*, 2020] that improve the seed-box mining but remain reliant on the multi-class classification scores. On one hand, our method considers the scoring variance among different categories and images, thus making a reasonable balance between the accuracy and recall of seed boxes. On the other hand, we use the ICBC scores for further refinement, thus improving the localization accuracy of these seed boxes. Some other methods [Lin *et al.*, 2020; Zeng *et al.*, 2019] also apply other information for assistance to improve the seed-box mining, however, their improvement limits on the online training procedure. Instead, our method brings intra-class binary classification (ICBC) task to directly enhances the network’s discrimination between intra-class positive and negative samples, thus benefiting the feature representation of the whole network.

4.4 Ablation Study

Effect of Each Component. We present experimental results on VOC 2007 to validate the effectiveness of each component, as summarized in Table 3. Starting with the basic WSOD framework, referred to as the “baseline”, we achieve an initial mAP of 54.6%. Incorporating the ICBC task into the WSOD framework leads to a notable improvement of 0.6%, underscoring the effectiveness of the ICBC branches in enhancing the network’s ability to distinguish between positive and mislocated samples in a class-wise manner. Subsequently, integrating the IGSM algorithm further improves

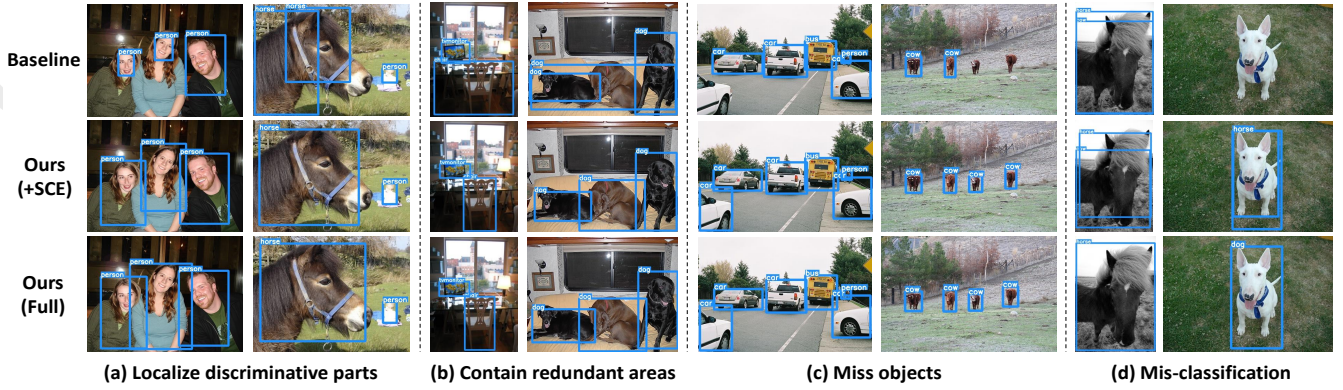


Figure 5: Qualitative results of the baseline model (1st row), the model only adding SEC module (2nd row), and our whole framework (3rd row). We show the cases including four typical challenges in WSOD: (a) Localizing only discriminative parts; (b) Containing redundant areas; (c) Missing objects; (d) Mis-classification.

Method	mAP (%)
Baseline	54.6
<i>Self-Classification Enhancement</i>	
+ Intra-class binary classification	55.2 $\uparrow_{0.6}$
+ ICBC guided seed mining	56.7 $\uparrow_{1.5}$
<i>Self-Classification Correction</i>	
+ Self-classification correction	57.5 $\uparrow_{0.8}$
+ Misclassification tolerance	58.2 $\uparrow_{0.7}$

Table 3: Ablation study of different components of our method on VOC 2007 in terms of mAP (%).

seed box quality, boosting performance to an mAP of 56.7%. Overall, the self-classification enhancement module delivers a significant mAP gain of 2.1%.

To evaluate the self-classification correction module, we first apply the self-classification correction algorithm during inference, resulting in a clear 0.8% mAP improvement. This gain can be attributed to the algorithm’s capability to effectively reduce high-scoring misclassified samples. By introducing the misclassification tolerance strategy, we achieve the highest performance of 58.2% mAP. These results demonstrate that enhancing the classification performance of MIDN can further expand the potential upper limit of the self-classification correction algorithm.

Figure 5 shows the detection results of different models on VOC 2007 test set. Compared with the baseline model (1st row), the integration of the OEIC module (2nd row) largely alleviates the mis-location problem in different cases, including localizing only discriminative parts (a) and containing redundant areas (including background and other objects) (b). Furthermore, more missing objects are detected (c) due to the design of the IGSM algorithm in SEC module. Lastly, as shown in (d), the utilization of the MSCR algorithm can further alleviate the mis-classification problem.

Effect of ICBC sampling strategy. We conduct experiments by employing different sampling strategies for the training of ICBC task, as shown in the upper part of the Table 4. Among all the settings, the combination of the IoU sampling strategy and the gridding strategy achieves the best performance and

Method	mAP (%)
<i>Intra-class binary classification</i>	
only with IoU sampling	57.3 $\downarrow_{0.9}$
with gridding sampling ($n=2$)	58.2 $\uparrow_{0.0}$
with gridding sampling ($n=3$)	57.9 $\downarrow_{0.3}$
<i>ICBC guided seed mining</i>	
without scoring strategy	57.2 $\downarrow_{1.0}$
without ICBC-guided finetuning	57.0 $\downarrow_{1.2}$

Table 4: Ablation study of self-classification enhancement module for the ICBC task on VOC 2007 in terms of mAP (%).

is insensitive to variations in grid size.

Effect of IGSM algorithm. We conduct experiments to evaluate each component of IGSM algorithm, as shown in the lower part of the Table 4. On one hand, we first replace our scoring strategy with original top-1 strategy, which only selects the proposals with highest scores as seed boxes, resulting in a 1.0% mAP drop, as our scoring strategy more effectively locates various objects using class-wise soft thresholds. Additionally, removing ICBC-guided fine-tuning causes a 1.2% mAP drop, highlighting the superior capability of ICBC in distinguishing positive samples from closely related but mislocated ones.

5 Conclusion

We propose a novel framework for weakly supervised object detection to address the limitations of the two-stage multi-class classification pipeline. On one hand, we introduce a self-classification enhancement module that enhances the network’s discrimination between intra-class positive and mis-located samples, and leverage it to enhance the quality of seed boxes. On the other hand, we introduce a self-classification correction algorithm to fine-tune the online classification scores, significantly reducing mis-classification detections. Extensive experiments on the widely used VOC datasets demonstrate the effectiveness of our framework.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants (62422204, 62472139) and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LQN25F030014, LDT23F02025F02. This work is also supported by the Open Project Program of the State Key Laboratory of CAD&CG (Grant No. A2403), Zhejiang University.

References

- [Arun *et al.*, 2019] Aditya Arun, CV Jawahar, and M Pawan Kumar. Dissimilarity coefficient based weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9432–9441, 2019.
- [Bilen and Vedaldi, 2016] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2854, 2016.
- [Chen *et al.*, 2020] Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Slv: Spatial likelihood voting for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12995–13004, 2020.
- [Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.
- [Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.
- [Girshick, 2015] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.
- [Huang *et al.*, 2020] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:16797–16807, 2020.
- [Jiao *et al.*, 2024] Yang Jiao, Zequn Jie, Shaoxiang Chen, Lechao Cheng, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Instance-aware multi-camera 3d object detection with structural priors mining and self-boosting learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2598–2606, 2024.
- [Liao *et al.*, 2022] Mingxiang Liao, Fang Wan, Yuan Yao, Zhenjun Han, Jialing Zou, Yuze Wang, Bailan Feng, Peng Yuan, and Qixiang Ye. End-to-end weakly supervised object detection with sparse proposal evolution. In *European Conference on Computer Vision*, pages 210–226. Springer, 2022.
- [Lin *et al.*, 2020] Chenhao Lin, Siwen Wang, Dongqi Xu, Yu Lu, and Wayne Zhang. Object instance mining for weakly supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11482–11489, 2020.
- [Liu *et al.*, 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2016.
- [Nie *et al.*, 2023] Yuxiang Nie, Chaowei Fang, Lechao Cheng, Liang Lin, and Guanbin Li. Adapting object size variance and class imbalance for semi-supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1966–1974, 2023.
- [Redmon *et al.*, 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.
- [Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.
- [Ren *et al.*, 2020] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10598–10607, 2020.
- [Seo *et al.*, 2022] Jinhwan Seo, Wonho Bae, Danica J Sutherland, Junhyug Noh, and Daijin Kim. Object discovery via contrastive learning for weakly supervised object detection. In *European Conference on Computer Vision*, pages 312–329. Springer, 2022.
- [Shen *et al.*, 2019] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 697–707, 2019.
- [Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [Su *et al.*, 2022] Hui Su, Yue Ye, Zhiwei Chen, Mingli Song, and Lechao Cheng. Re-attention transformer for weakly supervised object localization. *The 33rd British Machine Vision Conference (BMVC)*, 2022.
- [Tang *et al.*, 2017] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2843–2851, 2017.

- [Tang *et al.*, 2018] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Loddon Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(1):176–191, 2018.
- [Uijlings *et al.*, 2013] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013.
- [Wan *et al.*, 2019] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2199–2208, 2019.
- [Wang *et al.*, 2023] Kuo Wang, Jingyu Zhuang, Guanbin Li, Chaowei Fang, Lechao Cheng, Liang Lin, and Fan Zhou. De-biased teacher: Rethinking iou matching for semi-supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2573–2580, 2023.
- [Wang *et al.*, 2024a] Guanchun Wang, Xiangrong Zhang, Zelin Peng, Tianyang Zhang, Xu Tang, Huiyu Zhou, and Licheng Jiao. Negative deterministic information-based multiple instance learning for weakly supervised object detection and segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- [Wang *et al.*, 2024b] Kuo Wang, Lechao Cheng, Weikai Chen, Pingping Zhang, Liang Lin, Fan Zhou, and Guanbin Li. Marvelovd: Marrying object recognition and vision-language models for robust open-vocabulary object detection. In *European Conference on Computer Vision*, pages 106–122. Springer, 2024.
- [Wei *et al.*, 2018] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 434–450, 2018.
- [Wu *et al.*, 2024] Zhihao Wu, Yong Xu, Jian Yang, and Xuelong Li. Misclassification in weakly supervised object detection. *IEEE Transactions on Image Processing*, 2024.
- [Yan *et al.*, 2019] G. Yan, B. Liu, N. Guo, X. Ye, F. Wan, H. You, and D. Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9833–9842, 2019.
- [Yang *et al.*, 2019] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8372–8381, 2019.
- [Yin *et al.*, 2021] Yufei Yin, Jiajun Deng, Wengang Zhou, and Houqiang Li. Instance mining with class feature banks for weakly supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.
- [Yin *et al.*, 2022] Yufei Yin, Jiajun Deng, Wengang Zhou, Li Li, and Houqiang Li. Fi-wsod: Foreground information guided weakly supervised object detection. *IEEE Transactions on Multimedia*, 25:1890–1902, 2022.
- [Yin *et al.*, 2023] Yufei Yin, Jiajun Deng, Wengang Zhou, Li Li, and Houqiang Li. Cyclic-bootstrap labeling for weakly supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7008–7018, 2023.
- [Zeng *et al.*, 2019] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8292–8300, 2019.