*electronics*

MDPI

*Article*

# Semi-Supervised Object Detection with Multi-Scale Regularization and Bounding Box Re-Prediction

**Yeqin Shao [1,*], Chang Lv [1], Ruowei Zhang [2], He Yin [3], Meiqin Che [1], Guoqing Yang [4] and Quan Jiang [1,*]**

[1] School of Transportation and Civil Engineering, Nantong University, Nantong 226019, China; dragonquest@stmail.ntu.edu.cn (C.L.); meiqin.che@ntu.edu.cn (M.C.)
[2] College of Electrical Engineering, Nantong University, Nantong 226004, China; 2111310006@stmail.ntu.edu.cn
[3] School of Information Science and Technology, Nantong University, Nantong 226019, China; 2010320026@stmail.ntu.edu.cn
[4] Suzhou Research Institute of Industrial Technology, Zhejiang University, Hangzhou 310058, China; ygq78@zju.edu.cn
* Correspondence: hnsyk@ntu.edu.cn (Y.S.); jiang.q@ntu.edu.cn (Q.J.)

**Abstract:** Semi-supervised object detection has become a hot topic in recent years, but there are still some challenges regarding false detection, duplicate detection, and inaccurate localization. This paper presents a semi-supervised object detection method with multi-scale regularization and bounding box re-prediction. Specifically, to improve the generalization of the two-stage object detector and to make consistent predictions related to the image and its down-sampled counterpart, a novel multi-scale regularization loss is proposed for the region proposal network and the region-of-interest head. Then, in addition to using the classification probabilities of the pseudo-labels to exploit the unlabeled data, this paper proposes a novel bounding box re-prediction strategy to re-predict the bounding boxes of the pseudo-labels in the unlabeled images and select the pseudo-labels with reliable bounding boxes (location coordinates) to improve the model's localization accuracy based on its unsupervised localization loss. Experiments on the public MS COCO and Pascal VOC show that our proposed method achieves a competitive detection performance compared to other state-of-the-art methods. Furthermore, our method offers a multi-scale regularization strategy and a reliably located pseudo-label screening strategy, both of which facilitate the development of semi-supervised object detection techniques and boost the object detection performance in autonomous driving, industrial inspection, and agriculture automation.

**Keywords:** object detection; semi-supervised learning; multi-scale regularization; bounding box re-prediction



**Citation:** Shao, Y.; Lv, C.; Zhang, R.; Yin, H.; Che, M.; Yang, G.; Jiang, Q. Semi-Supervised Object Detection with Multi-Scale Regularization and Bounding Box Re-Prediction. *Electronics* **2024**, *13*, 221. https://doi.org/10.3390/electronics13010221

Academic Editor: Yue Wu

Received: 29 November 2023
Revised: 21 December 2023
Accepted: 2 January 2024
Published: 3 January 2024



## 1. Introduction

Object detection is used to identify and locate objects of interest (such as humans, animals, vehicles, and other objects) in images or videos. In recent years, deep learning-based object detection methods have been rapidly developed. However, the main deep learning-based object detection methods [1–7] are supervised and heavily rely on large-scale manually annotated datasets, such as MS COCO. When the dataset samples are insufficient, the model generalization is compromised, resulting in lower detection accuracy. On the other hand, the manual annotation of a large-scale dataset is highly labor-intensive and expensive. Therefore, semi-supervised learning [8–16], which involves training with a small amount of labeled data and a large amount of unlabeled data, is being investigated by more and more researchers. By incorporating both labeled and unlabeled data, semi-supervised learning can exploit the rich information in the unlabeled data and generate a more robust and accurate model, reducing its dependence on a large amount of labeled data.

Semi-supervised object detection applies semi-supervised learning to object detection, exploiting the information from unlabeled data. The semi-supervised object detection

methods mainly fall into two categories: the methods based on consistency regularization [17–21] and the methods based on pseudo-labels [22–33]. The former uses consistency regularization to constrain the model to obtain the same prediction results for a given image and its perturbed counterpart. The latter, as a mainstream version of the semi-supervised object detection method, first utilizes a teacher model to predict pseudo-labels of unlabeled data and then employs the labeled data and unlabeled data to train a student model. During training, the teacher model's parameters are updated using the Exponential Moving Average [34] (EMA) based on the student model's parameters.

In recent years, many researchers have explored semi-supervised object detection methods. Jeong et al. [17] proposed a consistency loss to produce the same predictions for an image and its horizontally flipped counterpart. Zhou et al. [18] replaced the sparse pseudo-boxes with dense prediction as a united and straightforward form of the pseudo-label. Guo et al. [19] proposed Scale-Equivalent Distillation (SED) to alleviate the noise problem that arises from the false negative samples and inaccurate bounding box regression. Li et al. [20] proposed Multi-view Scale-invariant Learning (MSL) with mechanisms of both label- and feature-level consistency to achieve feature consistency by aligning the shifted feature pyramids in two varied scaled images. Miyato et al. [21] proposed a new regularization method based on virtual adversarial loss, which is a new measure of the local smoothness of the conditional label distribution. Sohn et al. [22] presented a self-training and augmentation-driven consistency regularization framework, which trained a detector with a limited number of annotated samples and generated pseudo-labels from unlabeled samples. To further improve the quality of pseudo-labels, Zhou et al. [23] proposed the Instant-Teaching framework, which employed a co-rectify method to rectify erroneous predictions made by two structurally identical but independently trained models. Tang et al. [24] proposed the Humble Teacher framework that utilized the EMA method and soft-label mechanism for improving the accuracy of semi-supervised models. Li et al. [25] proposed a novel self-correcting pseudo-label module and pseudo-label-guided copy–paste technology to generate more reliable predictions and enhance instance representation learning within diverse complex scenes. Liu et al. [26] introduced the Unbiased Teacher framework to employ a real-time pseudo-label generation method, which addressed the class-imbalance issue in object detection using focal loss [27] and the EMA. Xu et al. [28] proposed the Soft Teacher framework, which introduced classification loss to address the imbalanced foreground–background samples and utilized a bounding box discrepancy filter to fully leverage the bounding box regression information from the unlabeled data. Feng et al. [29] proposed a semi-supervised object detection method based on position confidence weighting and introduced a Location-Aware Head (LAH) to reduce the pseudo-label noise in the unlabeled data. Kim et al. [30] introduced a simple yet effective data augmentation method, Mix/UnMix (MUM), to unmix feature tiles from the mixed image tiles under the semi-supervised object detection framework. Cai et al. [31] proposed a semi-supervised object detection method based on teacher–student models with strong–weak heads. The strong and weak heads of the teacher model solved the quality measurement problem of pseudo-label localization. Unbiased Teacher v2 [32] and Dense Learning (DSL) [33] tried to combine semi-supervised learning methods with anchor-free detectors. The former method removed the misleading impact on bounding box regression in pseudo-labels by estimating the corresponding uncertainty, while the latter method proposed an adaptive filtering strategy to assign dense pseudo-labels to each pixel. Additionallly, the latter introduced an integrated teacher model to improve the stability and quality of pseudo-labels.

Although pseudo-label-based semi-supervised object detection methods have led to the achievement of numerous successes in recent years, these approaches ignore the importance of regularization for semi-supervised detectors. These methods have a poor generalization capability, especially if there is a lack of labeled data. Meanwhile, in previous works, based on the detection results of the teacher model, a threshold was applied to each detection's highest classification probabilities to filter out a subset of high-quality, reliable

pseudo-labels. These pseudo-labels were employed in the student model's training for the computation of classification loss. However, the classification probabilities of pseudo-labels can only reflect the classification results of the detected objects instead of their localization information; hence, unsupervised localization loss has often been ignored in previous works. To address these issues, based on the Unbiased Teacher method, we propose a semi-supervised object detection method with multi-scale regularization and bounding box re-prediction. The contributions of this paper are as follows:

(1) A novel multi-scale regularization (MSR) strategy is proposed to constrain the Faster R-CNN and to generate the same prediction results for both the input images and their corresponding down-sampled ones, thus achieving better detection accuracy with the student model.
(2) A novel bounding box re-prediction (BBRP) strategy is presented to re-predict the object's bounding box, obtaining reliably located pseudo-labels of unlabeled data and thus improving the localization capability of the student model.

Experiments using the public MS COCO and Pascal VOC, which have been widely used in previous semi-supervised object detection works [17,22–26,30], show that the proposed method achieves a promising performance.

## 2. Materials and Methods

### *2.1. Overview of Our Proposed Method*

The framework of our proposed semi-supervised object detection method based on multi-scale regularization and bounding box re-prediction is shown in Figure 1. Firstly, the dataset is divided into labeled data, $\mathcal{X}^{\mathcal{L}} = \{x_i^l|_{i=1}^{N_l}\}$, and unlabeled data, $\mathcal{X}^{\mathcal{U}} = \{x_i^u|_{i=1}^{N_u}\}$. $N_l$ and $N_u$ are the amounts of labeled and unlabeled data, respectively. $N_u \gg N_l$. The annotations of the labeled data are represented as $\mathcal{Y}^{\mathcal{L}} = \{y_i^l|_{i=1}^{N_l}\}$. Each annotation includes the center coordinates, width, height, and object classes of a bounding box. Then, based on the teacher–student framework, the teacher and student models both adopt the two-stage Faster R-CNN [6] as the object detector to predict the detection results. Specifically, in the first stage, the backbone network of the Faster R-CNN extracts the features of images and generates abundant region proposals for the foreground objects through the region proposal network (RPN). In the second stage, the features of all proposals are scaled to a fixed size using the region-of-interest (RoI) pooling operation, and then the RoI head is used to obtain the final labels, including the localization information and classification probabilities.

To alleviate the issue of false detection and duplicate detection, a multi-scale regularization strategy is integrated into the student Faster R-CNN to down-sample the strongly augmented images and constrain these images and the corresponding down-sampled ones so that both the RPN and RoI head of the student model produce consistent output. We employ $L_{scale}$ loss to constrain the student model in both the RPN and RoI head for accurate detection results.

On the other hand, for unlabeled data, based on the pseudo-labels predicted by the teacher model, our method uses thresholding technology to select reliably classified pseudo-labels, whose classification probabilities are higher than a threshold of $\tau_1$. Meanwhile, our method employs bounding box re-prediction to select reliably located pseudo-labels, whose location coordinates are steady. Using these reliably classified pseudo-labels and reliably located pseudo-labels, our method improves the student model's generalization of unlabeled data with the constraint of the $L_{\text{unsup}}$ loss, which includes the unsupervised classification loss and unsupervised localization loss from the RPN and RoI head in the student model.

Additionally, for the labeled data and their corresponding labels, we train the student model in a supervised way, with the constraint of the $L_{\text{sup}}$ loss, which includes the supervised classification loss and supervised localization loss from the RPN and RoI head in the student model.

It should be noted that our method adopts three kinds of losses: supervised loss, $L_{sup}$; unsupervised loss, $L_{unsup}$; and multi-scale regularization loss $L_{scale}$, which are all derived from both the RPN and RoI head of the student model. The student model uses the Gradient Descent method to update its parameters according to the total loss function. Different from the student model, the teacher model does not compute the loss function, and it only uses the EMA to slightly update its parameters according to those of the student model.

Moreover, during training, we first use the available labeled data to train and optimize our teacher model with the supervised loss. Then, we duplicate the trained weights from the teacher model and apply them to the student model. After that, we train the teacher–student framework using the three losses. Here, the teacher model takes the weakly augmented images as input to accurately predict the pseudo-labels, while the student model takes the strongly augmented images as input to be more robust. During the inference phase, we only use the teacher model to produce the final object detection results.

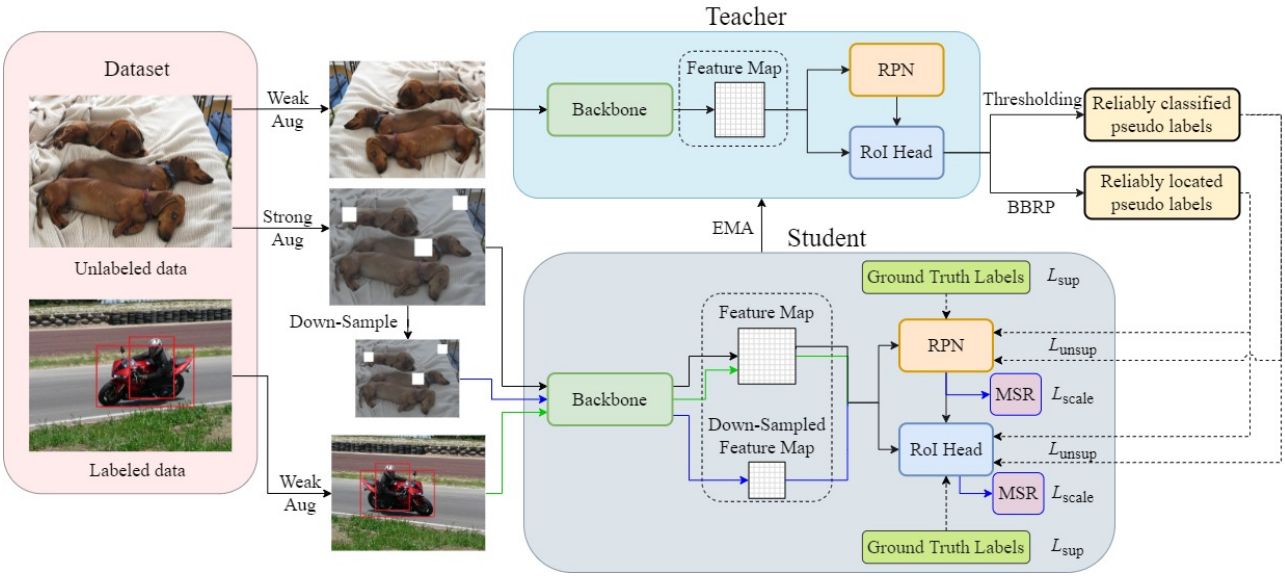


**Figure 1.** The framework of the semi-supervised object detection method based on multi-scale regularization and bounding box re-prediction. Our semi-supervised framework uses a teacher–student framework, where the teacher model predicts pseudo-labels of unlabeled data and screens the reliably classified pseudo-labels and reliably located pseudo-labels by the thresholding method and the bounding box re-prediction (BBRP) method. The student model computes supervised loss, unsupervised loss, and multi-scale regularization loss based on the ground-truth labels, reliably classified pseudo-labels, and reliably located pseudo-labels, respectively.

### *2.2. Multi-Scale Regularization*

To improve the generalization of the student model, we propose multi-scale regularization loss for the RPN and RoI head of the detector. The pseudo-code is shown in Algorithm 1.

Multi-scale regularization is illustrated in Figures 2 and 3. To increase the richness of the training samples, the Scale Jittering strategy [28] is applied to the labeled data and unlabeled data for data augmentation. Here, the images are randomly scaled within the range from −50% to 150%. These scaled images then enable the model to obtain different-sized features during training for better robustness.

**Algorithm 1** Multi-scale regularization.

**Input:** Unlabeled Image $x^{\mathrm{u}}$;
**Output:** Multi-Scale Regularization Loss $L_{\mathrm{scale}}^{\mathrm{rpn}}$, $L_{\mathrm{scale}}^{\mathrm{roihead}}$;
$x^{\mathrm{u}}$ = Scale Jittering($x^{\mathrm{u}}$);
$x_{\mathrm{d}}^{\mathrm{u}}$ = Down_Sample($x^{\mathrm{u}}$);
$F$ = Backbone_Student($x^{\mathrm{u}}$);
$F_{\mathrm{d}}$ = Backbone_Student($x_{\mathrm{d}}^{\mathrm{u}}$);
*Proposals*, $\widehat{\boldsymbol{OBJ}}$ = RPN_Student($F$);
$Proposals_{\mathrm{d}}$ = Down_Sample(*Proposals*);
_, $\widehat{\boldsymbol{OBJ}}_{\mathrm{d}}$ = RPN_Student($F_{\mathrm{d}}$);
$\widehat{\boldsymbol{OBJ}}^{*}$ = Down_Sample($\widehat{\boldsymbol{OBJ}}$);
Compute $L_{\mathrm{scale}}^{\mathrm{rpn}}$ according to Equation (1);
$\widehat{\boldsymbol{CLS}}$, $\widehat{\boldsymbol{REG}}$ = RoI_Head_Student($x^{\mathrm{u}}$, $F$, *Proposals*);
$\widehat{\boldsymbol{CLS}}_{\mathrm{d}}$, $\widehat{\boldsymbol{REG}}_{\mathrm{d}}$ = RoI_Head_Student($x_{\mathrm{d}}^{\mathrm{u}}$, $F_{\mathrm{d}}$, $Proposals_{\mathrm{d}}$);
Compute $L_{\mathrm{scale}}^{\mathrm{roihead}}$ according to Equation (2);
**return** $L_{\mathrm{scale}}^{\mathrm{rpn}}$, $L_{\mathrm{scale}}^{\mathrm{roihead}}$.

Subsequently, we input the image $x^{\mathrm{u}}$ and the corresponding down-sampled one $x_{\mathrm{d}}^{\mathrm{u}}$ into the backbone network with a Feature Pyramid Network [35] (FPN) to extract feature maps $F$ and $F^{\mathrm{d}}$ and generate region proposals using the RPN of the student model. After that, the objectness map $\widehat{\boldsymbol{OBJ}}$ of $x^{\mathrm{u}}$ generated by the RPN is down-sampled to obtain $\widehat{\boldsymbol{OBJ}}^{*}$, which has the same shape as that of the objectness map of $x_{\mathrm{d}}^{\mathrm{u}}$. The down-sample operation is called MaxPooling. Thus, the $L_{scale}$ loss for the RPN is computed as the Euclidean distance between the objectness map of $x^{\mathrm{u}}$ and that of $x_{\mathrm{d}}^{\mathrm{u}}$, as shown in Equation (1).

$$L_{\mathrm{scale}}^{\mathrm{rpn}}(\widehat{\boldsymbol{OBJ}}^{*}, \widehat{\boldsymbol{OBJ}}_{\mathrm{d}}) = \frac{1}{m}\sum_{i=1}^{m}\left\|\widehat{\boldsymbol{obj}}^{*}[i] - \widehat{\boldsymbol{obj}}_{\mathrm{d}}[i]\right\|_2^2 \tag{1}$$

where $\widehat{\boldsymbol{obj}}[i] \in \mathbb{R}^{A\times W_{\mathrm{i}}\times H_{\mathrm{i}}}$ and $\widehat{\boldsymbol{obj}}_{\mathrm{d}}[i] \in \mathbb{R}^{A\times \frac{W_{\mathrm{i}}}{2}\times \frac{H_{\mathrm{i}}}{2}}$ represent the $i$-th objectness map of $x^{\mathrm{u}}$ and $x_{\mathrm{d}}^{\mathrm{u}}$ generated by the RPN, respectively. $\widehat{\boldsymbol{obj}}^{*}[i] \in \mathbb{R}^{A\times \frac{W_{\mathrm{i}}}{2}\times \frac{H_{\mathrm{i}}}{2}}$ is the spatially down-sampled version of $\widehat{\boldsymbol{obj}}[i]$. $A$ is the number of anchors in an image. $W_{\mathrm{i}} = \frac{W}{2^{(i-1)}}$ and $H_{\mathrm{i}} = \frac{H}{2^{(i-1)}}$ present the width and height of the $i$-th objectness map $\widehat{\boldsymbol{obj}}[i]$. $W$ and $H$ are the width and height of the largest objectness map. $m$ is the number of objectness maps (for one single image, the backbone network of the detector outputs $m$ feature maps of different scales, and each feature map produces a corresponding objectness map; here, $m = 5$).

Similarly, to make the detector's prediction more robust, our method constrains the RoI head outputs of the student model on the feature map and the down-sampled ones to make them more consistent for the unlabeled data, as shown in Figure 3. Firstly, our student model employs the backbone and RPN to obtain the feature maps and region proposals of images $x^{\mathrm{u}}$ and $x_{\mathrm{d}}^{\mathrm{u}}$, respectively. Then, based on the feature maps and region proposals, RoI pooling converts the features inside the proposals into a fixed size. Finally, these converted features from the two feature maps are inputted into the RoI head to obtain the classification output and regression output. Here, $F_{\mathrm{d}}$ shares the same region proposals with $F$ due to the down-sampling operation. The RoI head output (classification output or regression output) of the two feature maps ($F$ and $F_{\mathrm{d}}$) have the same size, so the $L_{scale}$ loss for the RoI head can be computed according to Equation (2).

$$L_{\mathrm{scale}}^{\mathrm{roihead}}(\widehat{\boldsymbol{CLS}}, \widehat{\boldsymbol{CLS}}_{\mathrm{d}}, \widehat{\boldsymbol{REG}}, \widehat{\boldsymbol{REG}}_{\mathrm{d}}) = \tfrac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\boldsymbol{cls}}^{\,i} - \widehat{\boldsymbol{cls}}_{\mathrm{d}}^{\,i}\right\|_2^2 + \tfrac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\boldsymbol{reg}}^{\,i} - \widehat{\boldsymbol{reg}}_{\mathrm{d}}^{\,i}\right\|_2^2 \tag{2}$$

where $\widehat{\boldsymbol{cls}}^{i} \in \mathbb{R}^{C}$ and $\widehat{\boldsymbol{reg}}^{i} \in \mathbb{R}^{4}$ are the classification output (classification probabilities) and regression output (bounding box) of the $i$-th region proposal of image $x^{\mathrm{u}}$. $\widehat{\boldsymbol{cls}}^{i}_{\mathrm{d}} \in \mathbb{R}^{C}$ and $\widehat{\boldsymbol{reg}}^{i}_{\mathrm{d}} \in \mathbb{R}^{4}$ are the classification output and regression output of the $i$-th region proposal of image $x^{\mathrm{u}}_{\mathrm{d}}$. $C$ is the number of object classes ($C = 80$ for the COCO dataset). $n$ is the number of region proposals, which we set to be 1000 for each image in our method. Since the regression outputs of images $x^{\mathrm{u}}$ and $x^{\mathrm{u}}_{\mathrm{d}}$ are of different scales, the parameters of each regressed bounding box are normalized using Equation (3).

$$\begin{aligned} x^{\mathrm{reg}} &= (\hat{x} - x_{\mathrm{a}})/w_{\mathrm{a}}, & y^{\mathrm{reg}} &= (\hat{y} - y_{\mathrm{a}})/h_{\mathrm{a}} \\ w^{\mathrm{reg}} &= \log(\hat{w}/w_{\mathrm{a}}), & h^{\mathrm{reg}} &= \log(\hat{h}/h_{\mathrm{a}}) \\ x^{\mathrm{reg}}_{\mathrm{d}} &= (\hat{x}_{\mathrm{d}} - x_{\mathrm{d,a}})/w_{\mathrm{d,a}}, & y^{\mathrm{reg}}_{\mathrm{d}} &= (\hat{y}_{\mathrm{d}} - y_{\mathrm{d,a}})/h_{\mathrm{d,a}} \\ w^{\mathrm{reg}}_{\mathrm{d}} &= \log(\hat{w}_{\mathrm{d}}/w_{\mathrm{d,a}}), & h^{\mathrm{reg}}_{\mathrm{d}} &= \log(\hat{h}_{\mathrm{d}}/h_{\mathrm{d,a}}) \end{aligned} \tag{3}$$

where $x$, $y$, $w$, and $h$ represent the center coordinates, width, and height of the region proposal. Variables $\hat{x}$, $x_{\mathrm{a}}$, and $x^{\mathrm{reg}}$ represent the predicted box, anchor box, and normalized regressed box, respectively (as do $y$, $w$, and $h$). $\widehat{\boldsymbol{reg}}^{i} = [x^{\mathrm{reg}}, y^{\mathrm{reg}}, w^{\mathrm{reg}}, h^{\mathrm{reg}}]$ and $\widehat{\boldsymbol{reg}}^{i}_{\mathrm{d}} = [x^{\mathrm{reg}}_{\mathrm{d}}, y^{\mathrm{reg}}_{\mathrm{d}}, w^{\mathrm{reg}}_{\mathrm{d}}, h^{\mathrm{reg}}_{\mathrm{d}}]$.

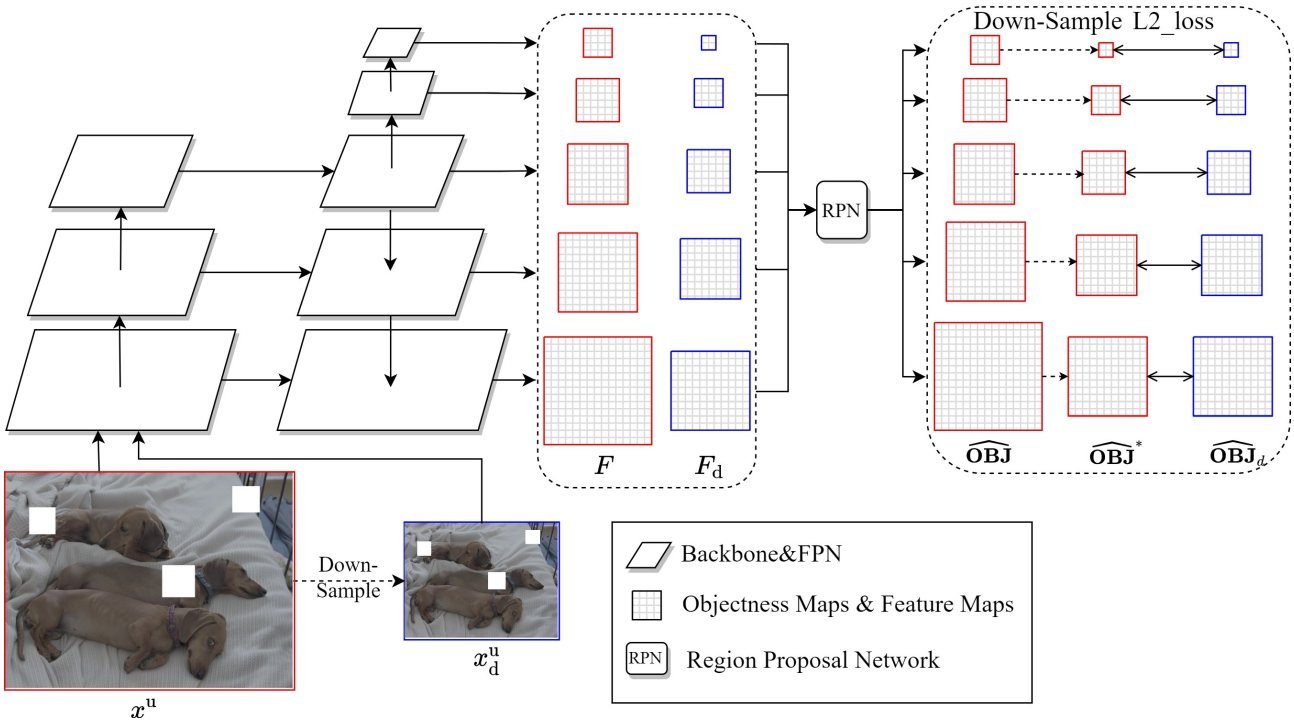


**Figure 2.** An illustration of multi-scale regularization loss in the RPN. Based on the image and its down-sampled counterpart, our method extracts the feature maps and constrains the RPN to produce consistent results for images and their down-sampled ones.

Finally, the multi-scale regularization loss from the RPN and RoI head in the student model can be summarized with Equation (4).

$$L_{\mathrm{scale}} = L^{\mathrm{rpn}}_{\mathrm{scale}} + L^{\mathrm{roihead}}_{\mathrm{scale}} \tag{4}$$

It is worth noting that the loss functions $L^{\mathrm{rpn}}_{\mathrm{scale}}$ and $L^{\mathrm{roihead}}_{\mathrm{scale}}$ can be computed with one forward propagation based on the image $x^{\mathrm{u}}$ and the corresponding down-sampled one $x^{\mathrm{u}}_{\mathrm{d}}$. To make it clear, we illustrate the two loss functions separately.

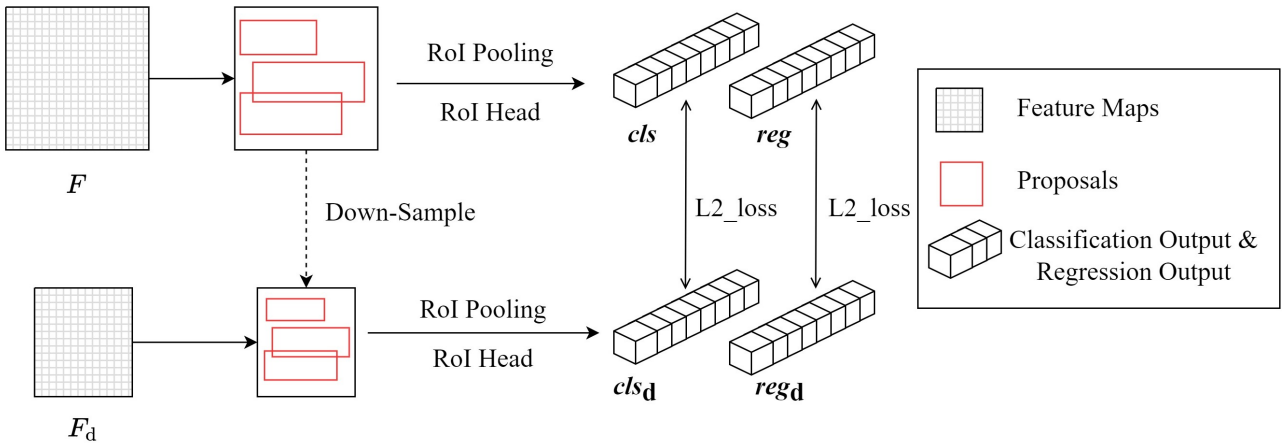


**Figure 3.** An illustration of multi-scale regularization loss in the RoI head. Based on $F$, $F_d$, and region proposals, our method constrains the RoI head to generate consistent predictions for feature maps of different scales.

### *2.3. Bounding Box Re-Prediction*

Most of the semi-supervised methods set a threshold for the predicted classification score to obtain reliable pseudo-labels. However, these pseudo-labels only ensure that the classification output of the detection is reliable and cannot guarantee the reliability of the regression output, i.e., the bounding box, for detection. Therefore, we propose the bounding box re-prediction strategy to refine the bounding box of each pseudo-label initially predicted by the teacher model for better localization accuracy.

Specifically, the unlabeled image $x^u$ is input into the teacher model, and the RPN of the teacher model predicts the region proposals. Based on these region proposals, a classification and regression operation is performed for the RoI head of the teacher model to obtain the prediction results $B_0$ (including the classification scores and location coordinates). Then, we set a threshold $\tau_1$ for the classification score to filter out the reliably classified pseudo-labels for the unsupervised classification loss computation. To further obtain reliably located pseudo-labels, based on the initial prediction results $B_0$ from the teacher model, our method employs BBRP on the RoI head of the teacher model to re-predict the bounding box of each region proposal and to obtain the refined prediction results $B_{tmp}$. The pseudo-code of BBRP is shown in Algorithm 2. Thus, our method obtains reliably located pseudo-labels for unsupervised localization loss computation on unlabeled data.

**Algorithm 2** Bounding box re-prediction.

**Input:** Feature map $F$, prediction results $B_0$;
**Output:** Reliably located pseudo-labels $y^{\mathrm{u}}_{\mathrm{loc}}$;

$B_{tmp}$=RoI_Head_Teacher($F$, $B_0$);
$y^{\mathrm{u}}_{\mathrm{loc}}$={};
**for** $b$ in $B_{tmp}$ **do**
  **if** $b$.cls_prob $> \tau_2$ **then**
    **for** $b_0$ in $B_0$ **do**
      o $= \mathrm{IoU}(b, b_0)$;
      **if** o $> \tau_3$ **then**
        $y^{\mathrm{u}}_{\mathrm{loc}} \leftarrow \mathrm{Append}(b)$;
      **end if**
    **end for**
  **end if**
**end for**
**return** $y^{\mathrm{u}}_{\mathrm{loc}}$.

### 2.4. Loss Function

The loss function of our proposed method is as follows:

$$L = L_{\text{sup}} + \alpha L_{\text{unsup}} + \beta L_{\text{scale}} \tag{5}$$

$$L_{\text{sup}} = L_{\text{cls}}^{\text{rpn}}(\hat{\mathcal{Y}}_{\overline{\text{cls}}}^{\text{l}}, \mathcal{Y}_{\overline{\text{cls}}}^{\text{l}}) + L_{\text{loc}}^{\text{rpn}}(\hat{\mathcal{Y}}_{\text{loc}}^{\text{l}}, \mathcal{Y}_{\text{loc}}^{\text{l}}) + L_{\text{cls}}^{\text{roihead}}(\hat{\mathcal{Y}}_{\text{cls}}^{\text{l}}, \mathcal{Y}_{\text{cls}}^{\text{l}}) + L_{\text{loc}}^{\text{roihead}}(\hat{\mathcal{Y}}_{\text{loc}}^{\text{l}}, \mathcal{Y}_{\text{loc}}^{\text{l}}) \tag{6}$$

$$L_{\text{unsup}} = L_{\text{cls}}^{\text{rpn}}(\hat{\mathcal{Y}}_{\overline{\text{cls}}}^{\text{u}}, \mathcal{Y}_{\overline{\text{cls}}}^{\text{u}}) + L_{\text{loc}}^{\text{rpn}}(\hat{\mathcal{Y}}_{\text{loc}}^{\text{u}}, \mathcal{Y}_{\text{loc}}^{\text{u}}) + L_{\text{cls}}^{\text{roihead}}(\hat{\mathcal{Y}}_{\text{cls}}^{\text{u}}, \mathcal{Y}_{\text{cls}}^{\text{u}}) + L_{\text{loc}}^{\text{roihead}}(\hat{\mathcal{Y}}_{\text{loc}}^{\text{u}}, \mathcal{Y}_{\text{loc}}^{\text{u}}) \tag{7}$$

where $L_{\text{sup}}$ is the supervised loss computed based on the labeled data $\mathcal{X}^{\mathcal{L}}$ and the corresponding label $\mathcal{Y}^{\mathcal{L}}$ (including $\mathcal{Y}_{\text{cls}}^{\text{l}}$ and $\mathcal{Y}_{\text{loc}}^{\text{l}}$). $L_{\text{sup}}$ is the summation of the classification loss $L_{\text{cls}}^{\text{rpn}}$ and localization loss $L_{\text{loc}}^{\text{rpn}}$ generated by the RPN, and the classification loss $L_{\text{cls}}^{\text{roihead}}$ and localization loss $L_{\text{loc}}^{\text{roihead}}$ generated by the RoI head of the student model. $L_{\text{unsup}}$ is the unsupervised loss computed based on the unlabeled data $\mathcal{X}^{\mathcal{U}}$ and its corresponding reliably classified pseudo-labels $\mathcal{Y}_{\text{cls}}^{\text{u}}$ and reliably located pseudo-labels $\mathcal{Y}_{\text{loc}}^{\text{u}}$. $L_{\text{unsup}}$ has a similar computation procedure as $L_{\text{sup}}$. Note that $\mathcal{Y}_{\overline{\text{cls}}}^{\text{l}}$ and $\mathcal{Y}_{\overline{\text{cls}}}^{\text{u}}$ include 0–1 labels for the background or foreground of the image. $L_{\text{scale}}$ represents the multi-scale regularization loss, as defined in Equation (4). $\alpha$ and $\beta$ are the weighted coefficients of the loss function; here, $\alpha = 2$ and $\beta = 1$. The classification loss $L_{cls}$ ($L_{\text{cls}}^{\text{rpn}}$ or $L_{\text{cls}}^{\text{roihead}}$) is computed with Equation (8). The localization loss $L_{loc}$ ($L_{\text{loc}}^{\text{rpn}}$ or $L_{\text{loc}}^{\text{roihead}}$) is defined with Equation (10).

$$L_{cls}(\hat{\mathcal{Y}}, \mathcal{Y}) = \frac{1}{N_{\text{cls}}} \sum_{i=1}^{N_{\text{cls}}} l_{cls}(\hat{y}_{\text{i}}, y_{\text{i}}) \tag{8}$$

$$l_{cls}(p, y) = -y(1-p)^{\gamma} \log(p) - (1-y)p^{\gamma} \log(1-p) \tag{9}$$

$$L_{loc}(\hat{\mathcal{R}}, \mathcal{R}) = \frac{1}{N_{\text{loc}}} \sum_{i=1}^{N_{\text{loc}}} l_{loc}(\hat{r}_{\text{i}}, r_{\text{i}}) \tag{10}$$

$$l_{loc}(\hat{r}, r) = \begin{cases} 0.5(\hat{r} - r)^2, & \text{if } |\hat{r} - r| < 1 \\ |\hat{r} - r| - 0.5, & \text{otherwise} \end{cases} \tag{11}$$

where $N_{\text{cls}}$ indicates the number of ground-truth labels for the supervised loss or reliably classified pseudo-labels for the unsupervised loss. $N_{\text{loc}}$ indicates the number of ground-truth bounding boxes for the supervised loss or reliably located pseudo-labels for the unsupervised loss. $\hat{y}_{\text{i}}$ is the highest classification probability predicted by the student model, and $y_{\text{i}}$ is the corresponding ground truth. $\hat{r}_{\text{i}}$ is the region proposal predicted by the student model, and $r_{\text{i}}$ is the corresponding ground truth. $\gamma = 2.0$ in this paper.

### 2.5. Datasets

In our experiment, we validate our proposed method using the MS COCO [36] and Pascal VOC [37] datasets. Table 1 shows the number of images and categories in the datasets. The reasons for selecting these two datasets are as follows: (1) The MS COCO and Pascal VOC datasets consist of images depicting common objects of complex scenes. The COCO dataset is a large-scale dataset with over 100,000 images and 80 common categories, while the VOC dataset has a lot of images and 20 common categories. These datasets include a large number of samples and various categories of objects, as well as annotations (bounding boxes and categories), for each image. This is a convenient method for us to conduct effective experiments and analyses in different situations. (2) The MS COCO and Pascal VOC datasets are two baseline datasets that have been widely used in other semi-supervised object detection studies [17,22–26,30], and, thus, they will facilitate a fair comparison between our method and other state-of-the-art methods. For comparison, we follow the same settings used in previous works.

(1) In the COCO2017-Train dataset, 1%, 2%, 5%, and 10% images are randomly sampled as labeled data, and the rest of the images are used as unlabeled data. These labeled data and unlabeled data are used as the training data. The COCO2017-Val dataset is used as the test data.

(2) The VOC2007-Train and -Val datasets are used as labeled data and the VOC2012-Train and-Val datasets are used as unlabeled data. These labeled data and unlabeled data are used as the training data. The VOC2007-Test dataset is used as the test data.

**Table 1.** Detailed description of MS COCO and Pascal VOC datasets. The number of images and categories of each dataset are shown.

| Dataset | Train | Val | Test | Total | Categories |
|---|---|---|---|---|---|
| VOC2007 | 2501 | 2510 | 4952 | 9963 | 20 |
| VOC2012 | 5717 | 5832 | - | 11,549 | 20 |
| COCO2017 | 118,287 | 5000 | - | 123,287 | 80 |

## 3. Experiments

We conducted quantitative and qualitative experiments to evaluate the performance of our proposed semi-supervised object detection method.

### *3.1. Settings and Details*

#### 3.1.1. Evaluation Metrics

To quantitatively measure the performance of our semi-supervised object detection method, we adopted mAP50, mAP75, and mAP50:95 as evaluation metrics. The formula to determine the mean average precision (mAP) is as follows:

$$mAP = \int_0^1 P(R)dR \tag{12}$$

$$P = \frac{TP}{TP + FP} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

where $P$ is the precision, $R$ is the recall, $TP$ represents the number of positive samples predicted as positive, $FP$ represents the number of negative samples predicted as positive, and $FN$ represents the number of positive samples predicted as negative. Given an Intersection over Union (IoU = Intersection/Union) of two bounding boxes, we can draw a precision–recall (P-R) curve, and the area under the P-R curve is the mAP. So, mAP50 represents the mAP when the IoU is 0.5; mAP75 represents the mAP when the IoU is 0.75; and mAP50:95 represents the mean of the mAP values when the IoU is 0.5, 0.55, 0.6,..., 0.95.

#### 3.1.2. Implementation Details

Our experiments were performed on Ubuntu 16.04 with Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz and NVIDIA Tesla T4 GPU (×4). The proposed method was built on the Detectron2 [38] deep learning framework.

As for data augmentation, Horizontal Flipping is used as a weak augmentation method, while Color Jittering, Gaussian Blur, and Cutout [39] are applied as strong augmentation methods. The batch sizes of both the labeled data and unlabeled data are set as 4. The learning rate is 0.005, and the momentum is 0.9. The number of training iterations for the following ablation experiment is 120K. The thresholds are $\tau_1$ = 0.7, $\tau_2$ = 0.3, and $\tau_3$ = 0.95. The weighted coefficients $\alpha$ and $\beta$ of the loss function are set as 2 and 1, respectively. The student model uses the SGD optimizer and Gradient Descent method to update

the parameters. The teacher model depends on the EMA to update its parameters based on the parameters of the student model, as shown in Equation (15).

$$\theta_t^{\mathrm{T}} = \psi\theta_{t-1}^{\mathrm{T}} + (1-\psi)\theta_t^{\mathrm{S}} \tag{15}$$

where $\theta^{\mathrm{T}}$ is the teacher model's parameter, $\theta^{\mathrm{S}}$ is the student model's parameter, $t$ is the current iteration, and $\psi$ is set as 0.9996.

We also conduct additional experiments to determine other hyper-parameters of our method. $\tau_1$ is set as 0.7, which is consistent with the settings of the Unbiased Teacher method [26]. $\tau_2$ and $\tau_3$ are the important hyper-parameters in the BBRP strategy. Table 2 shows the performance of our method with different values of $\tau_2$ and $\tau_3$. We can observe from Table 2 that the model achieves the highest mAP when $\tau_2 = 0.3$ and $\tau_3 = 0.95$. Table 3 shows the performance of our method with different values of $\alpha$ and $\beta$ for the loss function. We can observe that when $\alpha = 3.0$, the training of the model cannot converge. Our method achieves the highest mAP when $\alpha = 2.0$ and $\beta = 1.0$.

**Table 2.** The performance of our method with different $\tau_2$ and $\tau_3$ values for BBRP. It can be seen that our method achieves the highest mAP when $\tau_2 = 0.3$ and $\tau_3 = 0.95$.

| $\tau_2$ | $\tau_3$ | mAP50:95 |
|---|---|---|
| 0.2 | 0.98 | 21.90 |
| | 0.95 | 21.95 |
| | 0.9 | 21.82 |
| | 0.8 | 21.34 |
| | 0.7 | 20.67 |
| 0.3 | 0.98 | 22.21 |
| | 0.95 | 22.46 |
| | 0.9 | 22.13 |
| | 0.8 | 21.65 |
| | 0.7 | 20.89 |
| 0.4 | 0.98 | 21.70 |
| | 0.95 | 21.65 |
| | 0.9 | 21.51 |
| | 0.8 | 21.19 |
| | 0.7 | 20.38 |

**Table 3.** The performances of our method with different $\alpha$ and $\beta$ for the loss function. It can be seen that our method achieves the highest mAP when $\alpha = 2.0$ and $\beta = 1.0$.

| $\alpha$ | $\beta$ | mAP50:95 |
|---|---|---|
| 1.0 | 1.0 | 21.12 |
| | 2.0 | 21.02 |
| | 3.0 | 20.78 |
| 2.0 | 1.0 | 22.46 |
| | 2.0 | 22.32 |
| | 3.0 | 21.95 |
| 3.0 | 1.0 | Cannot Converge |
| | 2.0 | Cannot Converge |
| | 3.0 | Cannot Converge |

During training, the mAP curve of our method evolves, as shown in Figure 4, using the COCO 1% labeled data. As a comparison, we also draw the mAP curve of the baseline semi-supervised method, Unbiased Teacher. We can see that both methods converge as the iterations increase.

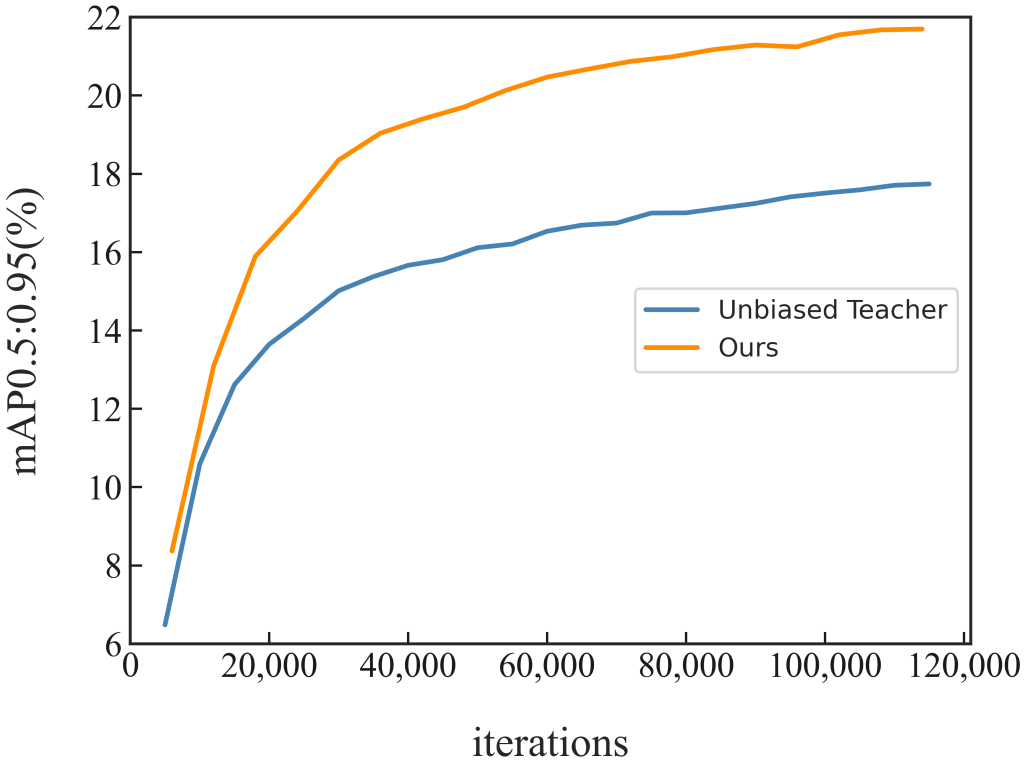


**Figure 4.** mAP curves of Unbiased Teacher method and our method during training with COCO 1% labeled data. Both methods converge as the iterations increase.

### *3.2. Comparison with State-of-the-Art Methods*

To verify the effectiveness of our method, we compared our method with other state-of-the-art semi-supervised object detection methods, as shown in Table 4: consistency-based semi-supervised learning (CSD) [17], STAC [22], Instant-Teaching [23], Humble Teacher [24], Robust Teacher [25], Unbiased Teacher [26], and MUM [30]. The supervised method is a baseline that only uses labeled data for fully supervised training. The CSD method introduces consistency loss for an image and its horizontally flipped counterpart into semi-supervised learning. The STAC method adopts a self-training and augmentation-driven consistency regularization framework. The Instant-Teaching method co-rectifies the erroneous predictions of two structurally identical but independently trained models on the same image. The Humble Teacher method uses the EMA and soft-label mechanism to improve semi-supervised models. The Robust Teacher method introduces a new self-correcting pseudo-label module and pseudo-label-guided copy–paste technology to generate more reliable predictions for semi-supervised detection. The Unbiased Teacher method employs a real-time pseudo-label generation method. The MUM method presents an effective data augmentation method for semi-supervised models.

For a fair comparison, all methods use the Faster R-CNN with a Resnet-50 backbone [40] as the object detector. We can see from Table 4 that, compared with fully supervised training, when using COCO 1%, 2%, 5%, and 10% labeled data, our presented method improves mAP50:95 by 13.41%, 13.65%, 11.26%, and 9.03%, respectively. Compared with the Unbiased Teacher method, with the same four different proportions of labeled data, our proposed method improves mAP50:95 by 4.32%, 4.12%, 3.08%, and 3.76%. Our presented method outperforms the other semi-supervised methods under comparison in most scenarios. We can also see from Table 5 that, on the VOC dataset, our method outperforms other methods in terms of mAP50 and mAP50:95.

**Table 4.** Comparison with other state-of-the-art methods using the COCO dataset (to determine the metric mAP50:95 under different ratios of labeled data). Our method achieves the best results in all four situations.

| | mAP50:95 | | | |
|---|---|---|---|---|
| **Method** | **1%** | **2%** | **5%** | **10%** |
| Supervised | 9.05 | 12.70 | 18.47 | 23.86 |
| CSD [17] | 10.51 | 13.93 | 18.63 | 22.46 |
| STAC [22] | 13.97 | 18.25 | 24.38 | 28.64 |
| Instant-Teaching [23] | 18.05 | 22.45 | 26.75 | 30.40 |
| Humble Teacher [24] | 16.96 | 21.72 | 27.70 | 31.61 |
| Robust Teacher [25] | 17.91 | 21.88 | 25.81 | 28.81 |
| Unbiased Teacher [26] | 18.14 | 22.23 | 26.65 | 29.13 |
| MUM [30] | 21.88 | 24.84 | 28.52 | 31.87 |
| Our Method | 22.46 | 26.35 | 29.73 | 32.89 |

**Table 5.** Comparison with other state-of-the-art methods using the VOC dataset. Our method achieves the best results for the metrics mAP50 and mAP50:95.

| **Method** | **mAP50** | **mAP50:95** |
|---|---|---|
| Supervised | 76.70 | 43.60 |
| CSD [17] | 74.70 | - |
| STAC [22] | 77.45 | 44.64 |
| Instant-Teaching [23] | 79.20 | 50.00 |
| Humble Teacher [24] | 80.94 | 53.04 |
| Robust Teacher [25] | 80.24 | 53.47 |
| Unbiased Teacher [26] | 79.30 | 53.50 |
| MUM [30] | 80.04 | 52.31 |
| Our Method | 81.29 | 55.26 |

Due to the additional computational cost introduced by our innovative method compared to that of the traditional Unbiased Teacher method, we performed a comparative experiment to test the training speed of the model to evaluate its training efficiency. We found that the Unbiased Teacher model takes 0.347 s per iteration for training, while our model takes 0.471 s per iteration. Our model requires slightly more training time than the Unbiased Teacher model. The inference speed of a single image using our model or the Unbiased Teacher model is 0.116 s on a Tesla T4 GPU.

### *3.3. Ablation Study*

To investigate the effectiveness of our proposed method's components, we conducted the ablation experiments on partially labeled data, as shown in Table 6. As an example, the experiment was performed using COCO 1% labeled data. Aug indicates the Scale Jittering strategy, MSR (RoI head) indicates the multi-scale regularization strategy for the RoI head, MSR (RPN) indicates the multi-scale regularization strategy for the RPN, and BBRP indicates the bounding box re-prediction strategy.

**Table 6.** Ablation experiments to test multi-scale regularization strategy and bounding box re-prediction strategy. ✓indicates that the corresponding component is used in our proposed method for the ablation experiment. Each component of our method is effective.

| **Aug** | **MSR (RPN)** | **MSR (RoI Head)** | **BBRP** | **mAP50** | **mAP75** | **mAP50:95** |
|---|---|---|---|---|---|---|
| | | | | 33.50 | 16.65 | 17.71 |
| ✓ | | | | 35.62 | 18.44 | 19.18 |
| ✓ | ✓ | | | 36.18 | 19.24 | 19.86 |
| ✓ | ✓ | ✓ | | 37.02 | 20.02 | 20.74 |
| ✓ | ✓ | ✓ | ✓ | 37.21 | 21.44 | 22.46 |

From Table 6, we can see that the Scale Jittering strategy can improve the detection accuracy of our method by 1.47% in mAP50:95. With the MSR strategy for the RPN, our method achieves a 0.68% improvement in mAP50:95. With the MSR strategy for the RoI head, our method achieves a 0.88% improvement in mAP50:95. In summary, with the MSR strategy, our method increases mAP50 by 1.40%, mAP75 by 1.58%, and mAP50:95 by 1.56%. With the BBRP strategy, our method obtains a 0.19% improvement in mAP50, 1.42% improvement in mAP75, and 1.72% improvement in mAP50:95. Finally, when combining the MSR strategy and BBRP strategy, our method increases by 3.71%, 4.79%, and 4.75% in mAP50, mAP75, and mAP50:95, respectively.

Up to now, the ablation experiment has proved the effectiveness of each component of our method. The reasons for this are as follows: (1) The multi-scale regularization strategy exploits the multi-scale information from the unlabeled data, thereby boosting the model's detection performance of multi-scale objects and also improving the model's robustness. (2) The bounding box re-prediction strategy selects reliably located pseudo-labels during training to further constrain the model for better localization accuracy when using unlabeled data.

### *3.4. Visualization*

Figure 5 visualizes the prediction results of our method and those of the representative semi-supervised Unbiased Teacher method. As an example, the experiment is carried out with 1% labeled data. It can be seen that the Unbiased Teacher model suffers from false detection (the sports balls indicated by the red arrows in Figure 5a), missed detection (the elephant and persons indicated by the red arrows in Figure 5b,c), and duplicate detection (the flying person indicated by the red arrow in Figure 5c). These issues lead to unreliable pseudo-labels in the teacher model and affect the student model's learning. In contrast, our method extracts the multi-scale features from the unlabeled data using the MSR strategy and robustly detects multi-scale objects in the images. Therefore, MSR can effectively solve these incorrect detection issues and improve the generalization capability of the model. On the other hand, as shown in Figure 5d, the Unbiased Teacher method only roughly locates the objects under some circumstances and ignores part of the object. However, our proposed method employs BBRP to select the reliably located pseudo-labels to improve the localization ability of our model. Therefore, our semi-supervised method can accurately detect the target objects.

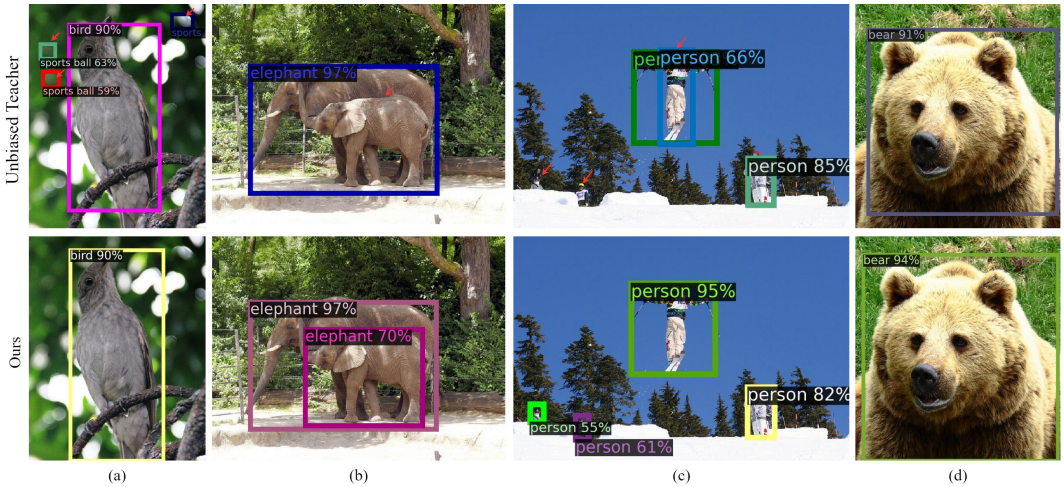


**Figure 5.** Detection visualization of the results from the Unbiased Teacher method and our proposed method. (**a**) the sports balls pointed by three red arrows indicate the false detection. (**b**) the elephant pointed by the red arrow indicates the missed detection. (**c**) the flying person pointed by the red arrow indicates the duplicate detection, and the other two persons pointed by the red arrows indicate the missed detection. (**d**) the bear indicates the inaccurate localization. The visualization shows that our approach accurately detects the target objects.

## 4. Discussion

Semi-supervised object detection methods often encounter false detection and duplicate detection due to the limited amount of labeled data. During semi-supervised training, it is challenging for a poor-performance teacher model to provide high-quality pseudo-labels. Therefore, we introduce a multi-scale regularization strategy to boost the semi-supervised object detector's performance. On the other hand, semi-supervised object detection methods typically use a thresholding method to filter pseudo-labels of unlabeled data and utilize them for the computation of unsupervised loss. However, these reliably classified pseudo-labels only produce accurate classification results and uncertain localization results. Hence, the localization information of unlabeled data is often ignored.

Our method is a semi-supervised object detection method based on multi-scale regularization and bounding box re-prediction. The multi-scale regularization strategy effectively alleviates the issue of false detection and duplicate detection by constraining the model to produce consistent predictions for images at different scales. This improvement notably enhances the robustness of the model. The bounding box re-prediction strategy improves the localization capability of the detector by using the RoI head of the Faster R-CNN to perform re-prediction on the detection results. It considers the bounding boxes with similar re-prediction results as reliably located pseudo-labels, which are then used for computing the unsupervised localization loss. This approach effectively enhances the detector's localization ability. Our semi-supervised object detection framework can be applied in many scenarios, such as autonomous driving, industrial inspection, and agriculture automation.

Although the two proposed strategies effectively enhance the accuracy of semi-supervised object detectors, there are still some issues to be addressed:

(1) Training efficiency. Our method introduces additional computational costs, which decreases the training efficiency to some extent.
(2) Domain adaptation. When the model begins semi-supervised learning, the training of the model becomes unstable due to the distribution gap between the labeled data and unlabeled data. We will try domain adaptation to alleviate this issue.
(3) Label assignment. Our method obtains reliably classified pseudo-labels using the thresholding method, which is a simple label assignment strategy. In the future, we will try other strategies to improve the performance of our method.

We will further improve our semi-supervised object detection method in our future work based on these perspectives.

## 5. Conclusions

Based on the teacher–student framework, this paper proposes a semi-supervised object detection method with multi-scale regularization and bounding box re-prediction. With a small amount of labeled data and a large amount of unlabeled data, our method trains a semi-supervised detection model to detect various objects in the images. Furthermore, to alleviate the detection issues suffered by the traditional semi-supervised object detection methods, our method proposes a multi-scale regularization strategy to effectively improve the generalization of the model. Additionally, our method presents a novel bounding box re-prediction strategy to select the reliably located pseudo-labels for the accurate location of objects in the unlabeled data. The experiments show that our semi-supervised method has a promising detection performance. Moreover, as a regularization strategy, multi-scale regularization can be applied to emerging technologies such as Transformer and its variants in contrastive language-image learning, human action recognition, semantic segmentation, and so on. While bounding box re-prediction can be combined with other novel semi-supervised object detection methods, such as semi-supervised vision Transformers and so on. In the future, we will mainly focus on the training efficiency of the method, domain adaptation between labeled data and unlabeled data, and label assignment to further improve our method.

**Author Contributions:** Conceptualization, Y.S.; Methodology, Y.S., C.L. and G.Y.; Software, C.L. and H.Y.; Validation, C.L., M.C. and Q.J.; Formal Analysis, M.C. and Y.S.; Investigation, C.L. and Y.S.; Resources, Y.S. and Q.J.; Data Curation, C.L.; Writing—Original Draft Preparation, C.L. and Q.J.; Writing—Review and Editing, C.L., Y.S., M.C. and Q.J.; Visualization, R.Z. and M.C.; Supervision, G.Y., Q.J. and Y.S.; Project Administration, Q.J. and Y.S.; Funding Acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported, in part, by the National Natural Science Foundation of China, grant 61671255.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data is contained within the article.

**Acknowledgments:** This paper and research would not have been possible without the support of the Transportation Safety and Intelligence Analysis Team from Nan Tong University.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Bochkovskiy, A.; Wang, C.; Liao, H.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
2. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
3. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Repbulic of Korea, 27 October–2 November, 2019; pp. 9626–9635. [CrossRef]
4. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787. [CrossRef]
5. Wang, C.; Bochkovskiy, A.; Liao, H.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475. [CrossRef]
6. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
7. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
8. Laine, S.; Aila, T. Temporal Ensembling for Semi-Supervised Learning. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
9. Athiwaratkun, B.; Finzi, M.; Izmailov, P.; Wilson, A.G. There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
10. Xie, Q.; Luong, M.; Hovy, E.H.; Le, Q.V. Self-Training With Noisy Student Improves ImageNet Classification. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 10684–10695. [CrossRef]
11. Qiao, S.; Shen, W.; Zhang, Z.; Wang, B.; Yuille, A.L. Deep Co-Training for Semi-Supervised Image Recognition. In Proceedings of the Computer Vision-ECCV 2018-15th European Conference, Munich, Germany, 8–14 September 2018; pp. 142–159. [CrossRef]
12. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
13. Luo, Y.; Zhu, J.; Li, M.; Ren, Y.; Zhang, B. Smooth Neighbors on Teacher Graphs for Semi-Supervised Learning. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8896–8905. [CrossRef]
14. Maaløe, L.; Sønderby, C.K.; Sønderby, S.K.; Winther, O. Auxiliary Deep Generative Models. In Proceedings of the Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, 19–24 June 2016; pp. 1445–1453.
15. Springenberg, J.T. Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016.
16. Zhang, J.; Wang, X.; Zhang, D.; Lee, D.J. Semi-Supervised Group Emotion Recognition Based on Contrastive Learning. *Electronics* **2022**, *11*, 3990. [CrossRef]
17. Jeong, J.; Lee, S.; Kim, J.; Kwak, N. Consistency-based Semi-supervised Learning for Object detection. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019; pp. 10758–10767.

18. Zhou, H.; Ge, Z.; Liu, S.; Mao, W.; Li, Z.; Yu, H.; Sun, J. Dense Teacher: Dense Pseudo-Labels for Semi-supervised Object Detection. In Proceedings of the Computer Vision-ECCV 2022-17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 35–50. [CrossRef]
19. Guo, Q.; Mu, Y.; Chen, J.; Wang, T.; Yu, Y.; Luo, P. Scale-Equivalent Distillation for Semi-Supervised Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 14502–14511. [CrossRef]
20. Li, G.; Li, X.; Wang, Y.; Wu, Y.; Liang, D.; Zhang, S. PseCo: Pseudo Labeling and Consistency Training for Semi-Supervised Object Detection. In Proceedings of the Computer Vision-ECCV 2022-17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 457–472. [CrossRef]
21. Miyato, T.; Maeda, S.; Koyama, M.; Ishii, S. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1979–1993. [CrossRef] [PubMed]
22. Sohn, K.; Zhang, Z.; Li, C.; Zhang, H.; Lee, C.; Pfister, T. A Simple Semi-Supervised Learning Framework for Object Detection. *arXiv* **2020**, arXiv:2005.04757.
23. Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; Li, H. Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; pp. 4081–4090. [CrossRef]
24. Tang, Y.; Chen, W.; Luo, Y.; Zhang, Y. Humble Teachers Teach Better Students for Semi-Supervised Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; pp. 3132–3141. [CrossRef]
25. Li, S.; Liu, J.; Shen, W.; Sun, J.; Tan, C. Robust Teacher: Self-correcting pseudo-label-guided semi-supervised learning for object detection. *Comput. Vis. Image Underst.* **2023**, *235*, 103788. [CrossRef]
26. Liu, Y.; Ma, C.; He, Z.; Kuo, C.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; Vajda, P. Unbiased Teacher for Semi-Supervised Object Detection. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual, 3–7 May 2021.
27. Lin, T.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
28. Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; Liu, Z. End-to-End Semi-Supervised Object Detection with Soft Teacher. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021; pp. 3040–3049. [CrossRef]
29. Feng, Z.; Wang, F. Semi-Supervised Object Detection Algorithm Based on Localization Confidence Weighting. *Comput. Eng. Appl.* **2023**, *accepted*.
30. Kim, J.; Jang, J.; Seo, S.; Jeong, J.; Na, J.; Kwak, N. MUM: Mix Image Tiles and UnMix Feature Tiles for Semi-Supervised Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 14492–14501. [CrossRef]
31. Cai, X.; Luo, F.; Qi, W.; Liu, H. A Semi-Supervised Object Detection Algorithm Based on teacher–student Models with Strong-Weak Heads. *Electronics* **2022**, *11*, 3849. [CrossRef]
32. Liu, Y.; Ma, C.; Kira, Z. Unbiased Teacher v2: Semi-supervised Object Detection for Anchor-free and Anchor-based Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 9809–9818. [CrossRef]
33. Chen, B.; Li, P.; Chen, X.; Wang, B.; Zhang, L.; Hua, X. Dense Learning based Semi-Supervised Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 4805–4814. [CrossRef]
34. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
35. Lin, T.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
36. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision-ECCV 2014-13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755. [CrossRef]
37. Everingham, M.; Gool, L.V.; Williams, C.K.I.; Winn, J.M.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
38. Pham, V.; Pham, C.; Dang, T. Road Damage Detection and Classification with Detectron2 and Faster R-CNN. In Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA, 10–13 December 2020; pp. 5592–5601. [CrossRef]

39. Devries, T.; Taylor, G.W. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv* **2017**, arXiv:1708.04552.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]