CISO: Co-iteration Semi-Supervised Learning for Visual Object Detection

Jianchun Qi*, Minh Nguyen, Wei Qi Yan
Auckland University of Technology, Auckland 1010 New Zealand
yhy5508@autuni.ac.nz*, minh.nguyen@aut.ac.nz, weiqi.yan@aut.ac.nz

Abstract. Semi-supervised learning offers a solution to the high cost and limited availability of manually labeled samples in supervised learning. In semi-supervised visual object detection, the use of unlabeled data can significantly enhance the performance of deep learning models. In this study, we introduce an end-to-end framework, named CISO (Co-iteration Semi-Supervised Learning for Object Detection), that integrates a knowledge distillation approach and a collaborative, iterative semi-supervised learning strategy. To maximize the utilization of pseudo-label data and address the scarcity of pseudo-label data due to high threshold settings, we propose a mean iteration approach where all unlabeled data is applied in each training iteration. Pseudo-label data with high confidence is extracted based on an ever-changing threshold (average intersection over union of all pseudo-labeled data). This strategy not only ensures the accuracy of the pseudo-label but also optimizes the use of unlabeled data. Subsequently, we apply a weak-strong data augmentation strategy to update the model. Lastly, we evaluate CISO using the Swin Transformer model and conduct comprehensive experiments on MS-COCO. Our framework delivers impressive results, outperforming state-of-the-art methods by 2.16 mAP and 1.54 mAP with 10% and 5% labeled data, respectively.

Keywords: Semi-supervised · Data Augmentation · Transformer · CISO

1 Introduction

Deep learning [2, 20, 55, 58] has achieved remarkable results in computer vision, natural language processing, and speech recognition [46, 57]. Visual object detection, a fundamental task in the field of computer vision, has seen the emergence of deep neural network-based algorithms. To enhance algorithmic performance accuracy and prevent overfitting in large models, training on a large-scale dataset is crucial. However, manually annotating such data poses a significant challenge. Therefore, semi-supervised learning [4, 6, 31], which labels only a small fraction of large-scale data and effectively utilizes a large amount of unlabeled data to improve model performance, has received increasing attention.

Currently, popular semi-supervised learning strategies include consistent regularization [4, 21, 22, 27, 35, 36, 43, 49, 51, 53]. This approach's basic idea is to separate different data points in low-density regions, and ensure similar data points

yield similar outputs. This consistency means the network's prediction remains the same as the original if the unlabeled input data is perturbed. Consistency regularization compares the model outputs in terms of their spatial distribution, independent of the labels, making it suitable for semi-supervised learning. Additionally, advances in semi-supervised learning have been associated with effective data augmentation development [17, 39, 48]. Data augmentation not only increases the data amount for training, improving the model's generalization but also adds noisy data to enhance the network's robustness [19, 38]. Presently, several data augmentation strategies have been effectively employed to improve semi-supervised learning models' training performance [4, 41, 49].

In recent years, most object detection research has primarily focused on developing robust detectors [9, 26, 47]. Significant progress has also been made in semi-supervised object detection [3, 15, 22, 25, 40, 42, 52]. The recently proposed STAC [40] has paved the way for semi-supervised learning applications for visual object detection. The instant-teaching method [60] further improves on STAC, achieving significant results in the field of SSOD and providing valuable insights for subsequent SSOD research. The instant-teaching improvement has two aspects; one is the use of an instant pseudo-label generation model, the other is the proposed co-rectify scheme to address confirmation bias due to pseudo-label. However, pseudo-label ineffectiveness stems from two main issues: (1) An increase in incorrect pseudo-labels leads to excessive noise and misdirects model learning. (2) Overconfident pseudo-labels are not updated and tend to cause model overfitting.

Therefore, in this paper, we propose a new SSOD framework, CISO, to address these problems. We maintain all the unlabeled data during each training iteration, that is, the pseudo-label data obtained from the first training is not discarded but reintroduced into the unlabeled data. This allows all the unlabeled data to be fully utilized in several iterations to correct each other and reduce the number of incorrect pseudo-labels. Considering that such a setup may lead to the repeated acquisition of high confidence pseudo-labels and the need to alleviate overfitting, we propose Mean Iteration. This approach involves training the models using pseudo-labels with IoU values greater than the average value and labeled data.

Since the pseudo-label is generated differently each time, the average value of the IoU after each iteration also changes, achieving the purpose of updating the pseudo-label. The advantage of CISO is that it maximizes pseudo-label usage and continuously improves the quality of the pseudo-label. Moreover, we inherit the end-to-end concept from instant-teaching and the weak-strong data augmentation approach from STAC. However, we integrate knowledge distillation with semi-supervised learning to achieve an end-to-end framework. For weak-strong data augmentation, we also adopt cropping, rotating, flipping, translating, and the new Cutmix.

We choose the MS-COCO dataset [23] to test our CISO framework. The performance is evaluated using the same experimental protocol as the STAC [40] and instant-teaching methods [60], that is, we select 1%, 5%, and 10% of the amount of labeled data for performance evaluation. It is worth noting that our proposed CISO framework outperforms most SSOD methods, achieving superior performance. The contributions of this paper are as follows:

(1) We propose CISO, a collaborative, iterative SSOD framework that extensively leverages unlabeled data. Besides, knowledge distillation and weak-strong data augmentation are also applied to our framework for the purpose of improving model accuracy and efficiency.

(2) To reduce the number of incorrect pseudo-label and avoid the overfitting problem caused by using the inability to update pseudo-label, we propose Mean Iteration method, a scheme for pseudo-label selection based on the IoU average value.

(3) We test CISO using the MS-COCO dataset and conduct extensive experiments. The sresults show that our proposed method achieved advanced performance. We also performed ablation experiments to conduct the analytics of our method.

In the rest of the paper, we present related work in Section 2. Our methodology is discussed in Section 3. Section 4 presents the analysis of the experimental results. Finally, our conclusions are drawn in Section 5.

2 Related Works

2.1 Visual Object Detection

Visual object detection is a popular research direction in computer vision, and it is widely employed in various industries, which can reduce the consumption of labor costs and has important social significance [14, 16, 28, 37, 47]. At present, visual object detection algorithms can be grouped into two categories, one is an end-to-end and one-stage network [24, 32, 44] which dominates in training efficiency, such as YOLO family [32, 45], the other is a two-stage network [9, 10, 33] which requires the use of region proposal CNN for feature extraction and classification, such as ResNet and Faster R-CNN [33].

Until recently, Transformers with a self-attention mechanism has also been employed in various tasks, including visual object detection, image classification, image segmentation, and video detection. Transformer models have not only received increasing attention but also have achieved good results [26], such as DETR for visual object detection [5]. However, the very majority of these methods require training based on large amounts of labeled data, which is very labor-intensive and time-consuming. Therefore improving the performance of object detection models through semi-

supervised learning has gradually been required and needs us to pay attention to it. We adopt Swin Transformer [26] in this article to develop the framework.

2.2 Semi-supervised Learning

Semi-supervised learning [59] aims to generate pseudo-label for unlabeled data samples by training a small number of labeled data samples, typically with much larger amount of unlabeled data than labelled data. The methods [1, 2, 13] apply semi-supervised learning to visual object detection. The core idea of Semi-Supervised Object Detection (SSOD) is to make full use of unlabeled data to improve the performance of the model. Currently, consistency-based learning and pseudo-label-based learning are the two main research directions of SSOD. The former can be referred to as a soft pseudo label, while the latter is a hard pseudo label. Early SSOD methods include CSD [15], which is based on consistent learning and proposes background elimination.

While STAC [40] proposes a SSOD method based on the hard pseudo label and also used consistency learning. After that, instant-teaching [60] improved on STAC by implementing instant pseudo-label training. The unbiased teacher [25] approach addressed the class imbalance problem. Moreover, data augmentation is effective in improving SSOD [22, 25, 60], such as Mixup [60] and Cutout [22]. Based on these approaches, we focus on the efficient use of unlabeled data as a means to improve model performance.

2.3 Knowledge Distillation

Knowledge distillation, which is essentially model compression [12, 54], is proposed to be applied to classification tasks in a simple way. Unlike quantization and pruning methods, knowledge distillation proposes a teacher-student network, where the output of teacher network is knowledge, and the student network learns to transfer knowledge for distillation. The performance and accuracy of the teacher network are higher, and the network structure is more complex than that of student network. There are two methods of knowledge acquisition in knowledge distillation; one is to use one-stage features [29, 30, 34], the other is to transfer knowledge through multi-stage information [11, 18, 51]. Knowledge distillation can lead to better model performance, reduce model latency, and compress network parameters [12]. Therefore, in this article, we consider adding a knowledge distillation method to our framework to improve the model performance.

3 Our Method

3.1 The Structure of Our Framework

Fig. 1 illustrates our CISO framework. We split the whole training process into three stages. In the first stage, small batches of randomly selected labeled data are trained in the Student model, while pseudo-label is generated for the unlabeled data by using the Teacher model, reliable data and unreliable data were selected according to the threshold $\tau \geq$ Mean (IoU). In the second stage, the labeled data and the reliable data are fed into the student learning model for training at the same time. At this point, the unreliable data generated in the first stage is released back into the unlabeled data, the pseudo-label is generated in the full unlabeled data. Finally, the reliable data selection process is repeated. Note that our Mean Iteration iterates four times and performs weak-strong data augmentation based on the data in each iteration. In the third stage, all the reliable data, unreliable data, and labeled data are fed into the model for training, the final detection model is obtained.

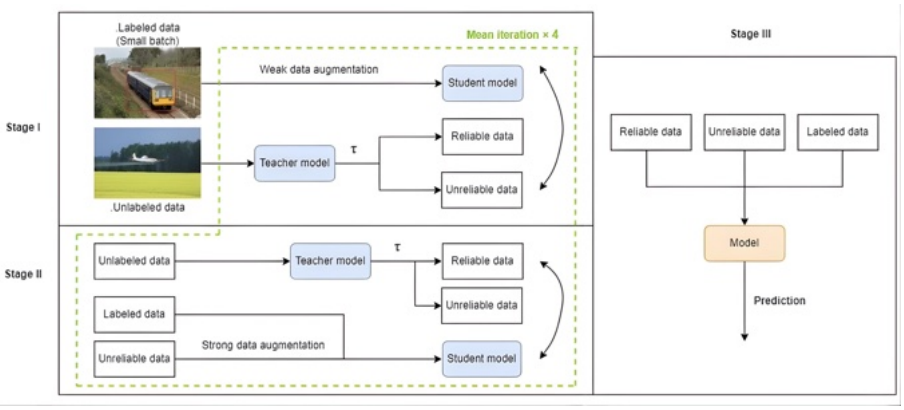

Fig. 1 The proposed semi-supervised object detection framework CISO. We are use of the Teacher model in knowledge distillation to generate pseudo-label for the unlabeled data and train iterations with the Student model. We only select pseudo-label with τ greater than or equal to the mean of τ. During the training period, the number of Mean Iteration was 4. We also conducted weak-strong data augmentation based on the given data

3.2 CISO: Co-iteration SSL for Object Detection

Pseudo labeling. A plethora of experiments have demonstrated that the efficient use of pseudo-label data can improve the accuracy of algorithms [3, 22, 25], leading to considerations of leveraging pseudo-label data to enhance model performance by proposing co-iteration semi-supervised learning based on knowledge distillation [61]. This differs from both classical STAC [40] and Instant-teaching [57]. STAC pioneered the application of SSL to object detection tasks by conducting self-training with pseudo-label and augmenting the data with consistent regularization. This method requires training the teacher model in advance and then training the student model. In

contrast, our CISO achieves end-to-end transfer of parameter data between models by using knowledge distillation to complete semi-supervised learning. Moreover, while Instant-Teaching is also end-to-end and our CISO inherits its self-training method, CISO retains all the unlabeled data instead of removing the unlabeled data (i.e., pseudo-label data with high confidence). Furthermore, we propose Mean Iteration, in which the threshold τ is continuously updated with our proposed method to enhance pseudo-label utilization and model performance.

To describe CISO in detail, we initially train each iteration by simultaneously generating a pseudo-label for the unlabeled data, using both pseudo-label data and a small amount of labeled data. Specifically, in each data batch, the labeled and unlabeled data are randomly sampled according to a set ratio, usually 1:10. Following that, we employ two models during the training process, namely, the teacher model and the student model for knowledge distillation. The teacher model is responsible for generating a pseudo-label for the unlabeled data, while the student model is responsible for conducting the training. Notably, the teacher model is the student model updated with the Exponential Moving Average (EMA). This end-to-end approach eliminates the need for complex multi-stage training schemes.

CISO also implements Mean Iteration, which facilitates mutual reinforcement between the pseudo-label and detection training process, rendering the training results increasingly effective. The details of Mean Iteration will be described later. Finally, all data, both labeled and unlabeled, are combined in the network to train the model and obtain the final detection model. Furthermore, for comparison purposes with STAC and Instant-Teaching, we perform weak-strong data augmentation based on the unlabeled data. In this approach, the weakly augmented data are inferred in the initial model to obtain the corresponding prediction scores. The pseudo-label of the corresponding data is obtained according to a threshold τ, while the strongly augmented data is then passed through the model to obtain the prediction scores and calculate the loss with the pseudo-label.

Overall, we train the model with the same loss function used in STAC [40] and Instant-teaching [57], which are the consistency regularization loss and the cross-entropy loss. The supervised loss consists of a classification loss function L_ce and a bounding box regression loss function L_1, as shown in Eq. 1.

$$L_s = \sum_s [\frac{1}{n} \sum_i L_{ce}(P(c_i \mid \alpha(Xs)), G(c_i)) + \frac{\lambda}{n} \sum_i G(c_i) L_1(P(r_i \mid \alpha(Xs)), G(r_i))] \quad (1)$$

where *s* is the index of the labeled image, *i* is the index of the anchor in the image, *n* is the total number of generated bounding boxes, $P(c_i)$ is the predicted probability of anchor *i* becoming an object in image *X*, and $G(c_i)$ is the label of anchor *i*. Then, $P(r_i)$ is the predicted generated bounding boxes coordinates, and $G(r_i)$ is the actual labeled coordinates.

As for the unsupervised loss part, the predicted probability distribution and frame coordinates of the model obtained by a small batch of weakly augmented unlabeled data are firstly calculated by using Eq. (2), and the pseudo-label is converted into hard labels as the finally obtained labels by Eq. (3).

$$G(c_i^u), G(r_i^u) = P(c_i, t_i \mid \alpha(Xu)) \tag{2}$$

$$\hat{G}(c_i^u) = argmax(c_i^u) \tag{3}$$

Thus, the unsupervised loss function is written as Eq. 4, which is shown as

$$L_u = \sum_u [\frac{1}{n} \sum_i L_{ce}\left(P(c_i \mid A(Xu)), \hat{G}(c_i^u)\right) + \frac{\lambda}{n} \sum_i (M(c_i^u) \geq \tau) L_1(P(r_i \mid A(Xs)), G(r_i^u))] \tag{4}$$

where *u* is the index of the unlabeled image, $\hat{G}(c_i^u)$ and $G(r_i^u)$ are the pseudo-label generated by the model itself, $\mathrm{M}(c_i^u)$ denotes the maximum prediction value, and τ is the confidence level.

Combined Eq. 1 and Eq. 4, the final loss function can be written as Eq. 5, where λ_u is the unsupervised loss weight.

$$L_{total} = \lambda_u L_u + L_s \tag{5}$$

Mean Iteration. CISO makes use of a portion of the labeled data to train the student model, while the teacher model generates a pseudo-label for the unlabeled data. In this step, we calculate the Intersection over Union (IoU) of all the pseudo-labeled data, and then determine the average of these IoU values to set the threshold for generating pseudo-labels. Furthermore, taking the mean value of IoU as the threshold τ, two types of pseudo-label data are generated, i.e., pseudo-labels with high confidence and pseudo-labels with low confidence. We consider the pseudo-labels with τ greater than the mean τ to be reliable labels, and the remaining pseudo-labels to be unreliable labels. Afterwards, the student model is trained a second time using both the labeled data and the reliable label data. After training, the teacher model is applied to predict the unlabeled data and generate both reliable and unreliable label data again. It is worth

noting that the pseudo-labeled data are generated randomly each time, so the reliable and unreliable labeled data are different with each iteration. To achieve iterative training, we retain all the unlabeled data in each training cycle of the Student model, without removing any of the classified unlabeled data from the pseudo-label data.

The proposed approach allows the threshold τ to be continuously updated from one iteration to the next. Since previous semi-supervised learning methods are prone to adopting pseudo-label data with a high threshold τ (e.g., τ = 0.9), this leads to data imbalance. Therefore, our CISO makes the best use of the pseudo-label data and ensures the accuracy of the pseudo-label data due to collaborative iterations. We conducted only four iterations of the experiment. Upon conducting a fifth iteration, there were no additional variations in what the model learned, which we will describe in detail in the ablation study. The results show that our method leads to improved model performance.

Weak-strong data augmentation. The SSL method using consistent regularization is closely related to data augmentation, which enables the model to gain much information in pseudo-label data playing a positive impact. Regarding soft augmentation, we conducted cropping, rotating, flipping, and translation to improve the quality of the labeled data in the pre-training period if the quality of the pseudo-labeled data was low. While for substantial augmentation, we harnessed Cutmix [56] for consistent learning on unlabeled data. Cutmix was chosen because it can apply both hard and soft fusion to two images, allowing the information from the entire image to be utilized without the dataset changing after image mixing. Furthermore, Cutmix does not loose the region information as Cutout does, which affects the training efficiency, nor does it introduce some of the pseudo-pixel information as Mixup does. By utilizing both weak and strong data augmentation, we increase the amount of data and noises, improve the robustness and generalization ability of the model and avoid overfitting. Fig. 2 illustrates the strategies for different classes of strong and weak data augmentation strategies.

Specifically, as shown in the Cutmix image section in Fig. 2, two images were randomly selected for the combination to generate a new training sample; given unlabeled data U_i, two images $U_1 = (X_{U_1}, Y_{U_1})$ and $U_2 = (X_{U_2}, Y_{U_2})$, the new sample is $N = (X_n, Y_n)$. We completed a regional dropout from the U_1sample by combining the corresponding regions in the U_2 sample where the U_1 sample is dropped as:

$$X = \boldsymbol{M} \odot X_{U1} + (\mathbf{1} - \boldsymbol{M}) \odot X_{U2} \tag{6}$$

$$Y = \lambda Y_{U1} + (\mathbf{1} - \lambda) Y_{U2} \tag{7}$$

where X is the image sample and Y is the image label. The λ is employed as the ratio of the combined regions of image U_1 and U_2, and as with Cutmix, we set the λ to be in the range (0, 1), where M is the binary mask indicating where images U_1 and U_2

were extracted. Besides, **1** indicates that the value of the mask matrix element is set to **1**. Finally, element-wise multiplication $\odot$ is used in Eq. (7).

$$r_X \sim \text{Unif}\,(0, W), \qquad r_W = W\sqrt{1-\lambda}, \tag{8}$$
$$r_Y \sim \text{Unif}\,(0, H), \qquad r_H = H\sqrt{1-\lambda}$$

Afterwards, Eq. (8) show how the extracted mask region is calculated. We use the same random method as Cutmix and define the coordinates of the mask region $C = (r_{X,} r_Y, r_W, r_H)$, where W is the width of the image U_i, H is the length of the image U_i, and r_{X} and r_{Y} are selected in the ranges (0, W) and (0, H), respectively.

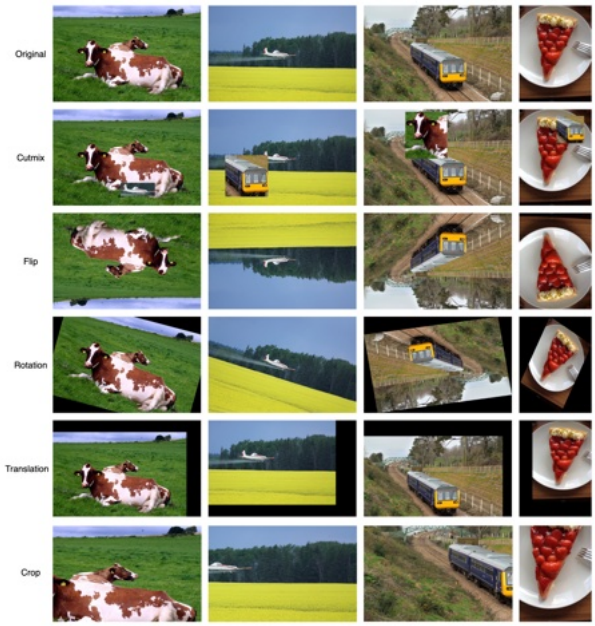

Fig. 2 Visualization of weak data augmentation and strong data augmentation strategies together. The first two are the original image and the strong data augmentation Cutmix. The remaining ones are the weak data augmentation, from top to bottom: Flipping, rotating, translating/shifting, and cropping

4 Experiments

4.1 Datasets

We propose the semi-supervised visual object detection framework CISO and conduct performance evaluation on the large-scale dataset MS-COCO [23] and PASCAL VOC

[8]. MS-COCO is a dataset for visual object detection, segmentation, and other scenarios. It has a total of 330K images, of which over 200K images were labeled, and it also has 80 object classes and 91 stuff categories. We adopt the same experimental protocol as STAC [40] and Instant-Teaching [60], that is, we randomly select 1%, 5%, and 10% of the labeled data for testing, and the rest of image samples are employed as unlabeled data. Our mAP is presented based on 80 object classes. Then, we applied *VOC07* and *VOC12* from the PASCAL VOC dataset as labeled and unlabeled sets, respectively. In the PASCAL VOC dataset.

4.2 Implementation Details

We applied the CISO framework to Swin Transformer. In this article, we are use of τ, λ_u, and λ. The three hyperparameters λ_u and λ are set to 1.0 and 1.0 respectively, while τ is dynamic, i.e., $\tau \geq$ Mean (IoU). The initialization of our network weights is all performed by the ImageNet pre-training model. We selected 1%, 5%, and 10% MS-COCO protocols, the experiments were performed by using a quick learning schedule.

Furthermore, our training parameters were kept consistent with STAC and Instant-Teaching, as detailed in Table 1.

Table 1 Training parameters of our framework

Classes	Parameters
Initial learning rate	0.01
Momentum	0.90
Weight decay	1e-4
Training step	180K
Learning rate decays (120K, 165K)	10

Although we adopted Swin Transformer as the feature extractor, we took use of Faster R-CNN as the detector to make a fair comparison with the experimental results of other models. Besides, we also conducted an experiment using the same backbone network ResNet-50 as the other model to verify the validity of our model.

4.3 Results

In the last two years, semi-supervised visual object detection methods have gradually gained attention. We compare our method with other state-of-the-art semi-supervised object detection methods and report the mAP and AP values for each protocol, the results of the comparison are shown in Table 2 and Table 3. Based on the experimental protocols, we find out that our proposed CISO outperformed all other SSOD methods to achieve the state-of-the-art outcome, which is evident that collaborative iteration and

mean thresholding strategy significantly improved the performance of semi-supervised visual object detection.

Specifically, in the Table 2, under the 1% protocol, our CISO's mAP value reached 22.00, an improvement up to 1.54 mAP; under the 5% protocol, our CISO increased the mAP value from Soft Teacher [50] method from 30.74 to 30.90, resulting in an improvement of 0.16 mAP values; under 10% protocol, our CISO improved the mAP value from Soft Teacher's result from 34.04 to 36.20, which improves the mAP value by 2.16. Finally, compared with the new semi-supervised learning baseline, LabelMatch [7], our mAPs is 0.71 higher under 10% of the protocol. Even for our experiments using ResNet-50 as the backbone network, CISO still outperforms other models, with mAPs of 21.04, 29.50, and 34.20 for 1%, 5%, and 10% protocols, respectively. The adoption of Swin Transformer indicates that our method is also applicable to the Transformer model with a self-attention mechanism. As depicted in Table 3, when we used *VOC07* and *VOC12* dataset as labeled and unlabeled data respectively, our CISO* increased the $AP_{50:95}$ from 50.00 to 51.77 compared to the Instant Teaching. Afterwards, we added 20 categories of MS-COCO dataset to the unlabeled data. When there is more unlabeled data, we also found that the $AP_{50:95}$ of CISO* is 3.03 higher than that of the Instant Teaching. In addition, for the application of Swin Transformer, our method's $AP_{50:95}$ is also higher than other methods, verifying the effectiveness of our model.

Table 2 Comparisons of mAP results of different semi-supervised object detection methods using MS-COCO dataset. Ours (CISO)* indicate that we are use of ResNet-50 as the backbone network for the implementation, Ours (CISO) shows Swin Transformer was selected as the backbone network for the implementation

Method		1%	5%	10%
Anchor based	Supervised	9.05±0.16	18.47±0.22	23.86±0.81
	CSD [15]	10.20±0.15	18.90±0.10	24.50±0.15
	STAC [40]	13.97±0.35	24.38±0.12	28.64±0.21
	DETReg [5]	14.58±0.30	24.80±0.20	29.12±0.20
	Instant Teaching [60]	18.05±0.15	26.75±0.05	30.40±0.05
	ISMT [51]	18.88±0.38	26.37±0.24	30.53±0.52
	Unbiased Teacher [25]	20.75±0.12	28.27±0.11	31.50±0.10
	Soft Teacher [50]	20.46±0.39	30.74±0.08	34.04±0.14
	LabelMatch [7]	25.81±0.28	32.70±0.18	35.49±0.17
Anchor free	HT [43]	16.96±0.36	27.70±0.15	31.61±0.28
	Ours (CISO)*	21.04±0.18	29.50±0.21	34.20±0.12
	Ours (CISO)	22.00±0.17	30.90±0.15	36.20±0.26

We observed that the improvement in mAP value became more prominent as the amount of labeled data increased, from an improvement 1.54 mAP in the 1% protocol to an improvement 2.16 mAP in the 10% protocol. We find that this problem is related to the fact that we released the pseudo-labeled data back into the unlabeled data. This might be due to the release of the pseudo-labeled data, which leads to a higher probability

of extracting duplicate pseudo-labeled data again in the next iteration. We leave this consideration for later investigation. Moreover, Fig. 3 shows the prediction results.

Table 3 Comparisons of AP results of different semi-supervised object detection methods using PASCAL VOC dataset.

labeled	Unlabeled	Methods	AP_{50}	$AP_{50:95}$
VOC07	None	Supervised	72.75	42.04
VOC07	VOC12	CSD[15]	74.70	-
		STAC[40]	77.45	44.64
		Instant Teaching [60]	79.20	50.00
		Ours (CISO)*	80.39	51.77
		Ours (CISO)	81.44	52.98
VOC07	VOC12 + COCO (20 classes)	CSD[15]	75.10	-
		STAC[40]	79.08	46.01
		Instant Teaching [60]	79.90	50.80
		Ours (CISO)*	83.03	53.83
		Ours (CISO)	84.48	55.30

Fig. 3 The prediction results of our framework

5 Ablation Study

5.1 Implementation Details Analysis of the number of Mean Iterations

In Fig. 1, we detailed that the mean iteration part in the green dashed box is required to iterate for a number of 4 iterations, so we analyze the impact of the number of Mean Iteration in this section. We tested the model under the protocol of 10% MS-COCO, with the remaining 90% being unlabeled data. The experimental results are shown in Table 4, where we see that six experiments were conducted with the number of iterations set to 1, 2, 3, 4, 5, and 6, respectively. As the number of iterations varies from 1 to 6, we conclude that the performance of the model is getting progressively better.

However, starting from iteration number 5, the performance of the model tends to level off. By 6-th iteration, the mAP has been improved only 0.06. Therefore, the performance and efficiency of the model will remain optimal when the number of iterations reaches number 4.

Table 4 Comparisons of mAP with different Mean iterations

The number of Mean iterations	mAP
1	27.40
2	29.80
3	33.60
4	36.20
5	36.40
6	36.46

5.2 Strong Data Augmentation

Since data augmentation strategies affect model performance in semi-supervised visual object detection models, we are use of weak-strong data augmentation strategies in CISO. However, the impact of solid data augmentation on model performance is more significant. For a fair comparison, we took advantage of the Cutmix strategy while retaining the Color+Cutout strategy.

In Table 5, we summarize the mAP values using the different robust data augmentation strategies. If we use only the Color+Cutout and Geometric strategies, the mAP value of our method does not improve much, only 1.26. Furthermore, the model performance is improved by using the Cutmix strategy, with an mAP value improvement of 0.50 compared to using Mixup and Mosaic. This validates our conjecture that the Cutmix strategy improves pseudo-label quality by not adding pseudo-pixel information to the data. CISO obtained the highest mAP value of 29.70 using the Cutmix data augmentation method. The analysis suggests that we are able to improve the performance of SSOD using Cutmix. The tests in this section are based on a 5% MS-COCO protocol.

Table 5 Comparisons of mAP values of CISO with different strong data augmentation. For fair comparison, we keep the Color+Cutout strategy

Methods	Strong data augmentations					mAP
	Color+Cutout	Geometric	Mixup	Mosaic	Cutmix	
STAC	√	√				23.14
Instant Teaching	√		√	√		25.60
CISO	√	√				24.40
	√		√	√		29.20
	√				√	29.70

5.3 Analysis of τ

The confidence threshold τ is a significant coefficient in semi-supervised target detection, and its setting directly affects the performance of the model. As other SSOD methods have taken a constant τ, we set τ to be dynamically changing, and obtain pseudo-label according to the criterion that τ is greater than or equal to the mean value. Since the reliable data and unreliable data generated after each iteration is different, the average value τ taken each time is dynamic τ (by using 10% MS-COCO protocol).

We see from Table 6 that the highest model performance is achieved if τ is averaged, with a mAP 36.20. Moreover, the mAP of the model continues decreasing as τ decreases. This confirms our hypothesis that the quality of the pseudo-label improves when τ is dynamic. Finally, whether there is a more suitable dynamic τ other than the mean value that can be applied to SSOD is the subject of our future research work.

Table 6 Comparisons of mAP values with various values of confidence threshold τ

τ	mAP
0.30	29.4
0.50	31.60
0.70	33.60
0.90	34.80
Mean (IoU)	36.20

5.4 Analysis of λ_u

Our study investigates the impact of the balance coefficient λ_u on the model's performance by incorporating it into the loss function. In this section, we conduct testing using the 10% MS-COCO protocol. We set the values of τ to the dynamic mean and test the model with different values of λ_u, specifically 0.25, 0.50, 1.00, 2.00, 3.00, and 4.00. Our results, presented in Table 7, demonstrate that the model performs the best when λ_u is set to 1.0. However, if λ_u=2.0, though the performance of the model decreases, the mAP is 35.80, which is only 0.40 lower than 36.20. Furthermore, although the model performance decreases with the change of other values of λ_u, the mAP value decreases most at λ_u =0.25 by 5. We observe that our proposed framework is relatively robust to λ_u.

Table 7 Comparison of mAP values with various vlues of balance coefficient λ_u

λ_u	mAP
0.25	30.20
0.50	32.50
1.00	36.20
2.00	35.60
3.00	32.90
4.00	31.40

5.5 Analysis of Mean Iteration

In addition to the mean τ, we also propose mean iterations to improve the quality of the pseudo-label by using the unlabeled data as much as possible. This is performed based on the dynamic mean τ and focuses on releasing the pseudo-label extracted in each iteration into the unlabeled data. As shown in Table 8, the mAP value without Mean Iteration is 33.10, which is lower than the value 3.10. Furthermore, Fig. 4 shows the visualization of pseudo-labels of the unlabeled data. This result is generated based on whether or not the Mean Iteration strategy is used. We see that using the Mean Iteration strategy is effective in generating more accurate pseudo-label, which in turn improves model performance. In this section, we still test it with the 10% MS-COCO protocol.

Table 8 Comparison of mAP values with various vlues of balance coefficient λ_u

Mean Iteration	mAP
	33.10
√	36.20

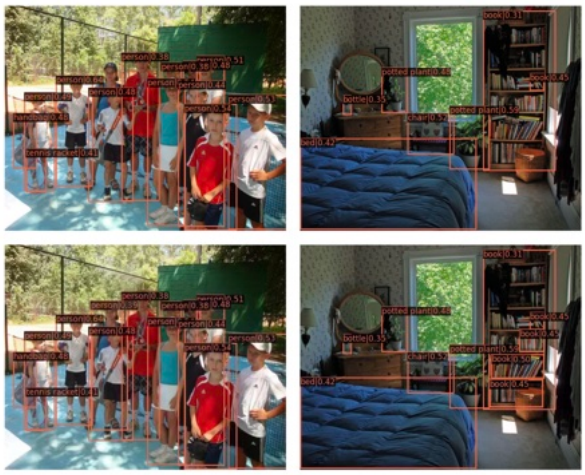

Fig. 4 The predicted pseudo-label. The top two images and the bottom two images were obtained from the non-Mean Iteration training and Mean Iteration training, respectively

5.6 Analysis of the Size of Unlabeled Data

Finally, analysis of the size of unlabeled data is also essential. Therefore, we evaluated under the 5% and 10% protocols of MS-COCO dataset. The dimensions of unlabeled data were set according to 1, 2, 4, and 8 times the labeled data. Table 9 shows the comparison results of mAP values with variable scales of unlabeled data. We can see that our method outperforms STAC and Instant Teaching, indicating that CISO can efficiently utilize pseudo label data.

Table 9 Comparison of mAP values with various scales of unlabeled data

Methods	Labeled size	Unlabeled size				
		1×	2×	4×	8×	Full
STAC[40]	5% COCO	19.81	20.79	22.09	23.14	24.38
Instant Teaching [60]		23.60	24.30	25.30	25.60	25.60
Ours (CISO) *		26.71	27.63	28.28	28.60	28.65
STAC[40]	10% COCO	25.38	26.52	27.33	27.95	28.64
Instant Teaching [60]		28.80	29.00	29.20	29.50	29.53
Ours (CISO) *		32.10	32.42	32.67	32.87	32.91

6 Conclusion

Our research presents a novel semi-supervised object detection (SSOD) learning strategy, CISO, which employs knowledge distillation and weak-strong data augmentation techniques on unlabeled data. In addition, it makes full use of unlabeled data for iterative training. To tackle the problem of model overfitting, caused by the inability to update pseudo-labels, we introduce a Mean Iteration scheme. Our work effectively leverages unlabeled data to enhance model performance. While we evaluate CISO on the Swin Transformer with a self-attentive mechanism, our approach can be applied to other detectors as well. We conduct extensive experiments on the MS-COCO and PASCAL VOC datasets, and our proposed method demonstrates impressive performance, surpassing other state-of-the-art methods with higher mAP values. Currently, our research does not address the selection of training samples and merely selects training data randomly from the dataset. However, in practical applications, labeled and unlabeled data may not adhere to the assumption of independent and identically distributed data, since unlabeled data may originate from scenarios different from those of the labeled data. Therefore, our future work will focus on improving the performance of the SSOD model by exploring methods for selecting training samples that account for such distribution differences.

References

1. Arazo E, Ortego D, Albert P, O'Connor NE, McGuinness K (2020) Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: Proceedings of International Joint Conference on neural networks, pp 1-8
2. Bachman P, Alsharif O, Precup D (2014) Learning with pseudo-ensembles, NIPS
3. Bar A, Wang X, Kantorov V, Reed CJ, Herzig R, Chechik G, Rohrbach A, Darrell T, Globerson A (2022) DETReg: Unsupervised pretraining with region priors for object detection. In: Proceedings of the IEEE International Conference on computer vision, pp 14605-14615
4. Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel CA (2019) Mixmatch: A holistic approach to semi-supervised learning, NIPS
5. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. ECCV: 213-229

6. Chapelle O, Schölkopf B, Zien A (2010) Semi-supervised learning. Adaptive Computation and Machine Learning. MIT Press 21(1): 2
7. Chen B, Chen W, Yang S, Xuan Y, Song J, Xie D, Pu S, Song M, Zhuang Y (2022) Label matching semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14381-14390
8. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vision 88: 303-338
9. Girshick R (2015) Fast R-CNN. In: Proceedings of the IEEE International Conference on computer vision, pp 1440-1448
10. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: Proceedings of the IEEE International Conference on computer vision, pp 2961-2969
11. Heo B, Kim J, Yun S, Park H, Kwak N, Choi JY (2019) A comprehensive overhaul of feature distillation. In: Proceedings of the IEEE International Conference on computer vision, pp 1921-1930
12. Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv:1503.02531
13. Iscen A, Tolias G, Avrithis Y, Chum O (2019) Label propagation for deep semi-supervised learning. In: Proceedings of the IEEE International Conference on computer vision, pp 5070-5079
14. Janiesch C, Zschech P, Heinrich K (2021) Machine learning and deep learning. Electronic Markets 31(3): 685-695
15. Jeong J, Lee S, Kim J, Kwak N (2019) Consistency-based semi-supervised learning for object detection, NIPS
16. Joseph KJ, Khan S, Khan FS, Balasubramanian VN (2021) Towards open world object detection. In: Proceedings of the IEEE International Conference on computer vision, pp 5830-5840
17. Kim J, Hur Y, Park S, Yang E, Hwang SJ, Shin J (2020) Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning, NIPS
18. Komodakis N, Zagoruyko S (2017) Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. ICLR
19. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Communications of the ACM 60(6): 84-90
20. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553): 436-444
21. Lee DH (2013) Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. ICML: pp 896
22. Li K, Liu C, Zhao H, Zhang Y, Fu Y (2021) Ecacl: A holistic framework for semi-supervised domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 8578-8587
23. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: Common objects in context. ECCV: 740-755
24. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. ECCV: 21-37
25. Liu YC, Ma CY, He Z, Kuo CW, Chen K, Zhang P, Wu B, Kira Z, Vajda P (2021) Unbiased teacher for semi-supervised object detection. arXiv:2102.09480
26. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on computer vision, pp 10012-10022
27. Miyato T, Maeda SI, Koyama M, Ishii S (2018) Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans on Pattern Anal Mach Intell 8: 1979-1993
28. Papageorgiou C, Poggio T (2000) A trainable system for object detection. International Journal of computer vision 38: 15-33
29. Park W, Kim D, Lu Y, Cho M (2019) Relational knowledge distillation. In: Proceedings of the IEEE International Conference on computer vision, pp 3967-3976

30. Passalis N, Tefas A (2018) Probabilistic knowledge transfer for deep representation learning. CoRR 1(2): 5
31. Rasmus A, Berglund M, Honkala M, Valpola H, Raiko T (2015) Semi-supervised learning with ladder networks, NIPS
32. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE International Conference on computer vision, pp 779-788
33. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks, NIPS
34. Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y (2014) FitNets: Hints for thin deep nets. ICLR
35. Sajjadi M, Javanmardi M, Tasdizen T (2016) Mutual exclusivity loss for semi-supervised deep learning. IEEE Int Conf Image Process (ICIP), 1908-1912
36. Sajjadi M, Javanmardi M, Tasdizen T (2016) Regularization with stochastic transformations and perturbations for deep semi-supervised learning, NIPS
37. Shrivastava A, Gupta A, Girshick R (2016) Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE International Conference on computer vision, pp 761-769
38. Simard PY, Steinkraus D, Platt JC (2003) Best practices for convolutional neural networks applied to visual document analysis. ICDAR
39. Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel CA, Cubuk ED, Kurakin A, Li CL (2020) Fixmatch: Simplifying semi-supervised learning with consistency and confidence, NIPS
40. Sohn K, Zhang Z, Li CL, Zhang H, Lee CY, Pfister T (2021) A simple semi-supervised learning framework for object detection, CVPR
41. Suzuki T (2022) Teachaugment: Data augmentation optimization using teacher knowledge. In: Proceedings of the IEEE International Conference on computer vision, pp 10904-10914
42. Tang Y, Chen W, Luo Y, Zhang Y (2021) Humble teachers teach better students for semi-supervised object detection. In: Proceedings of the IEEE International Conference on computer vision, pp 3132-3141
43. Tarvainen A, Valpola H (2017) Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, NIPS
44. Tian Z, Shen C, Chen H, He T (2019) Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE International Conference on computer vision, pp 9627-9636
45. Wang CY, Bochkovskiy A, Liao HYM (2022) YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696
46. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification. In: Proceedings of the IEEE International Conference on computer vision, pp 3156-3164
47. Wu X, Sahoo D, Hoi SC (2020) Recent advances in deep learning for object detection. Neurocomputing 396: 39-64
48. Xie Q, Dai Z, Hovy E, Luong T, Le Q (2020) Unsupervised data augmentation for consistency training, NIPS
49. Xie Q, Luong MT, Hovy E, Le QV (2020) Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE International Conference on computer vision, pp 10687-10698
50. Xu M, Zhang Z, Hu H, Wang J, Wang L, Wei F, Bai X, Liu Z (2021) End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 3060-3069
51. Yang F, Wu K, Zhang S, Jiang G, Liu Y, Zheng F, Zhang W, Wang C, Zeng L (2022) Class-aware contrastive semi-supervised learning. In: Proceedings of the IEEE International Conference on computer vision, pp 14421-14430
52. Yang Q, Wei X, Wang B, Hua XS, Zhang L (2021) Interactive self-training with mean teachers for semi-supervised object detection. In: Proceedings of the IEEE International Conference on computer vision, pp 5941-5950

53. Yang X, Song Z, King I, Xu Z (2022) A survey on deep semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering, 1-20
54. Yim J, Joo D, Bae J, Kim J (2017) A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of IEEE International Conference on computer vision, pp 4133-4141
55. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) Bisenet: Bilateral segmentation network for real-time semantic segmentation. ECCV: 325-341
56. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y (2019) Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE International Conference on computer vision, pp 6023-6032
57. Yu J, Jiang Y, Wang Z, Cao Z, Huang T (2016) Unitbox: An advanced object detection network. In: Proceedings of the ACM International Conference on multimedia, pp 516-520
58. Zhao ZQ, Zheng P, Xu ST, Wu X (2019) Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems 30: 3212-3232
59. Zhai X, Oliver A, Kolesnikov A, Beyer L (2019) S4l: Self-supervised semi-supervised learning. In: Proceedings of the IEEE International Conference on computer vision, pp 1476-1485
60. Zhou Q, Yu C, Wang Z, Qian Q, Li H (2021) Instant-teaching: An end-to-end semi-supervised object detection framework. In: Proceedings of the IEEE International Conference on computer vision, pp 4081-4090
61. Yan, W (2023) Computational Methods for Deep Learning – Theory, Algorithms, and Implementations (2nd Edition), Springer.
62. Yan, W (2019) Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics (3rd Edition) Springer
63. Qi J, Nguyen M, Yan W (2022) Waste classification from digital images using ConvNeXt. PSIVT, pp.1-13.
64. Qi J, Nguyen M, Yan W (2022) Small visual object detection in smart waste classification using Transformers with deep learning, IVCNZ, pp.301-314.