Semi-Supervised Object Detection via Multi-instance Alignment with Global Class Prototypes

Aoxue Li Peng Yuan Zhenguo Li
Huawei Noah's Ark Lab, China
lax@pku.edu.cn, yuanpeng126@huawei.com, Li.Zhenguo@huawei.com

Abstract

Semi-Supervised object detection (SSOD) aims to improve the generalization ability of object detectors with large-scale unlabeled images. Current pseudo-labeling-based SSOD methods individually learn from labeled data and unlabeled data, without considering the relation between them. To make full use of labeled data, we propose a Multi-instance Alignment model which enhances the prediction consistency based on Global Class Prototypes (MA-GCP). Specifically, we impose the consistency between pseudo ground-truths and their high-IoU candidates by minimizing the cross-entropy loss of their class distributions computed based on global class prototypes. These global class prototypes are estimated with the whole labeled dataset via the exponential moving average algorithm. To evaluate the proposed MA-GCP model, we integrate it into the state-of-the-art SSOD framework and experiments on two benchmark datasets demonstrate the effectiveness of our MA-GCP approach.

1. Introduction

With a large amount of labeled data available, deep learning has shown superior performance when solving object detection task. However, it is very costly to collect sufficient labeled data for each object category. Fortunately, there are large amount of unlabeled data available which can be collected from social media and websites. Semi-supervised object detection (SSOD) is proposed to improve the generalization ability of object detectors with large-scale unlabeled images [8, 18, 23, 28, 31]. In SSOD, we are given a labeled dataset and an unlabeled dataset. SSOD aims to learn a object detector with good generalization ability by using these labeled and unlabeled images.

To achieve this goal, existing SSOD approaches usually adopt two strategies: consistency-based SSOD [12, 13, 30] and pseudo-labeling-based SSOD [18, 22, 26–29, 31]. Consistency-based approaches train the target detector by minimizing inconsistency between prediction results of unlabeled data with different perturbations. Their performance highly depends on the design of perturbations and the measurement of consistency. Recently, pseudo-labeling-based approaches become popular. As shown in the yellow box of Figure 1, they adopt a teacher-student learning framework. Specifically, the pseudo labels of unlabeled images are first estimated by using a teacher detector followed by a label refinement module. Then, they jointly train a student detector with both labeled and unlabeled images. The detection losses of labeled and unlabeled images are used to optimize the parameters of student detector. Teacher detector's parameters are updated with the parameters of student detector via Exponential Moving Average (EMA) algorithm or pretrained with all labeled images. However, these models individually train the student detector with labeled data and unlabeled data. That is, their detection losses of labeled and unlabeled data update the student detector individually, without considering the relation between them.

This work fully leverages the labeled images to improve SSOD by developing a Multi-instance Alignment model with Global Class Prototypes (MA-GCP). Our key insight is to better estimate prediction consistency of unlabeled images by using the reliable information learned from all labeled images. By enforcing the consistency regularization in the pseudo-labeling-based framework, our approach can improve its detection performance.

Specifically, we assume each class is represented by a prototype in the feature space. As shown in the green box of Figure 1, each prototype is progressed with region features of the corresponding category via EMA algorithm. Since each prototype is updated with all labeled instances during the whole training process, we call it global class prototype. Then, we compute the class distributions of unlabeled images' proposals based on its visual similarity with each global class prototype. After that, we impose consistency between each pseudo ground-truth proposal and its candidate proposals with high Intersection-of-Union (IoU) scores by minimizing the cross-entropy loss of their class

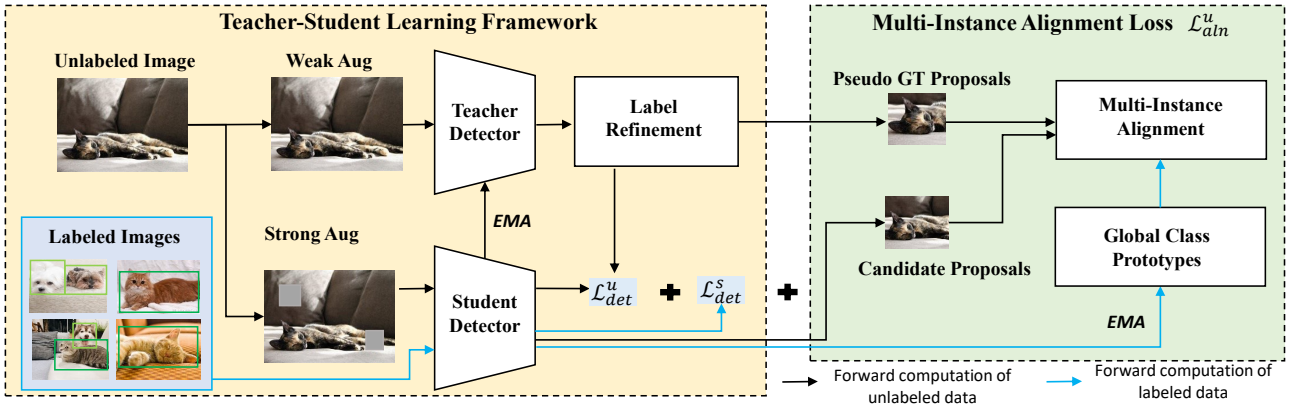

Figure 1. Illustration of the proposed approach. Our approach is implemented based on a pseudo-labeling-based SSOD framework, as shown in the yellow box. It estimates the pseudo labels for unlabeled images using a teacher detector followed by a label refinement module and then jointly trains a student detector with both labeled and unlabeled images. As shown in the green box, to make full use of labeled images, we propose a multi-instance alignment model based on global class prototypes which are learned with all labeled images via EMA algorithm. The detection loss of labeled images $\mathcal{L}^s_{det}$ and unlabeled images $\mathcal{L}^u_{det}$ and multi-instance alignment loss $\mathcal{L}^u_{aln}$ are used to train the student detector, and teacher detector's parameters are updated with student detector's parameters via EMA algorithm.

distributions. Different from previous consistency-based SSOD [2, 12, 13] which use batch-wise prototypes as references or directly compute prediction results without references, our model leverages these global class prototypes to produce more reliable consistency regularization and thus improves the detection accuracy of SSOD.

To evaluate the performance of our MA-GCP approach, we integrate it into the state-of-the-art pseudo-labeling-based framework. Experimental results on the PASCAL VOC and MSCOCO datasets demonstrate that our approach outperforms the competing models, and thus obtains the state-of-the-art results.

In summary, our main contributions are as follows:

- We propose a multi-instance alignment model with global class prototypes to enhance prediction consistency of pseudo-labeling-based SSOD approaches.
- We propose to learn global class prototypes with all labeled images via the EMA algorithm and then employ them to estimate prediction consistency of unlabeled images. These global class prototypes help to estimate more reliable prediction consistency and thus benefit SSOD.
- Extensive experiments demonstrate that our MA-GCP approach has achieved a consistent improvement on two benchmark datasets and obtains the state-of-the-art results.

2. Related Work

Semi-supervised image classification (SSIC) aims to exploit a large amount of unlabeled data to improve classification accuracy [4, 11, 14, 19, 24]. Most of them adopt consistency regularization which penalizes the inconsistency of prediction results of an unlabeled image of different augmentation views [1, 9, 19, 24]. Recently, some data-augmentation methods are designed to tackle semi-supervised image classification and have shown superior performance [3, 4, 25]. This work aims to tackle a more challenging task – SSOD, where not only classification task but location task needs to be dealt with.

Existing SSOD approaches are grouped into two categories: consistency-based approaches [12, 13, 30] and pseudo-labeling-based approaches [18, 22, 28, 29, 31]. Consistency-based approaches follow semi-supervised classification and learn from unlabeled data by enforcing the prediction consistency between unlabeled images in different augmentation views. For example, Jeong et al. imposed a prediction consistency between each unlabeled image and its horizontally flipped variant [12]. They further developed an interpolation-based SSOD (ISD) approach which produces reliable mix-up patches from two input images and maximizes the consistency between mix-up patches and original patches [13]. They combine this ISD method with previous consistency-based approach to promote the detection performance. These aforementioned approaches focus on the design of the augmentation method to improve sample diversity. Our approach not only improve spatial diversity by introducing high-IoU candidates of pseudo ground-truth proposals, but also propose to strengthen the consistency based on reliable global class prototypes.

Recently, pseudo labeling (or self-training) strategy becomes popular in SSOD. Sohn et al. pre-trained a teacher

detector using labeled images and generated pseudo-labels of unlabeled images to fine-tune the target(student) detector [21]. Their pseudo-labels are generated only once and are fixed through out the rest of training. Many follow-up work propose to simultaneously update teacher detector and student detector in an end-to-end manner. Liu et al. employed EMA strategy to train a student detector and a gradually progressing teacher in a mutually-beneficial manner [18]. Tang et al. also utilized EMA to update teacher detector and proposed a data ensemble method to produce reliable pseudo labels for unlabeled images [22]. Xu et al. developed an end-to-end soft teacher mechanism where the classification loss of each unlabeled bounding box is weighed by the classification score produced by the teacher network [27]. However, these approaches individually learn from labeled and unlabeled data. To make full use of labeled data, we propose a multi-instance alignment model based on global class prototypes. By inserting the proposed alignment model into a teacher-student training framework, our approach can improve its detection accuracy.

3. Method

In this section, we first provide the formulation of SSOD and a popular teacher-student framework. Then, an overview briefly introduces the proposed model. After that, we give technical details of two key components, i.e., global class prototypes and multi-instance alignment. The overall training objective is finally provided.

3.1. Preliminary: Teacher-Student Framework for Semi-Supervised Object Detection

Before introducing SSOD approaches, we first provide the definition of SSOD. In SSOD, we are given a set of labeled images $\boldsymbol{D}_s = \{(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s)\}_{i=1}^{N_s}$ and a set of unlabeled images $\boldsymbol{D}_u = \{\boldsymbol{x}_j^u\}_{j=1}^{N_u}$, where N_s and N_u are the total number of labeled and unlabeled images, respectively. Here, $\boldsymbol{y}_i^s$ denotes the image annotation of image $\boldsymbol{x}_i^s$. The goal of SSOD is to learn a good detector with $\boldsymbol{D}_s$ and $\boldsymbol{D}_u$.

Recent attempts [18, 22, 26, 27, 29, 31] exploit teacher-student framework to address SSOD. Specifically, as shown in the yellow box of Figure 1, given an unlabeled image $\boldsymbol{x}_i^u$, we first feed it into a weak augmentation model $\mathbf{T}_w$ and a strong augmentation module $\mathbf{T}_s$ to obtain the inputs of teacher detector $\mathbf{M}_t$ and student detector $\mathbf{M}_s$, respectively. Then, we feed the weak augmented image $\mathbf{T}_w(\boldsymbol{x}_i^u)$ into the teacher detector $\mathbf{M}_t$ followed by a label refinement post-processing module $\mathbf{H}$ to produce pseudo labels for each unlabeled image. After that, the strong augmented image $\mathbf{T}_s(\boldsymbol{x}_i^u)$ is fed into student model $\mathbf{M}_s$ to predict detection results. A detection loss of unlabeled image $\mathcal{L}_{det}^u$ is used to minimize the differences between prediction results $\mathbf{M}_s(\mathbf{T}_s(\boldsymbol{x}_i^u))$ and pseudo labels. Simultaneously, each labeled image $\boldsymbol{x}_j^s$ is fed into student detector $\mathbf{M}_s$ to predict detection result and a detection loss $\mathcal{L}_{det}^s$ is used to minimize the difference between its prediction result and its ground-truth label. By combining the two losses together, given an unlabeled image set $\boldsymbol{D}_u$ and a labeled image set $\boldsymbol{D}_s$, we compute the total loss of the teacher-student framework $\mathcal{L}_{det}^{st}$, which is formulated in Equation (1).

$$\mathcal{L}_{det}^{st}(\boldsymbol{D}_s, \boldsymbol{D}_u) = \mathcal{L}_{det}^{s}(\boldsymbol{D}_s) + \lambda_u \mathcal{L}_{det}^{u}(\boldsymbol{D}_u). \quad (1)$$

where λ_u is a hyper-parameter to balance the detection losses of labeled and unlabeled images. During the training stage, the total loss is used to optimize the parameters of student detector and label refinement module. The teacher detector and student detector share the same network architecture. The parameters of teacher detector are pretrained by all labeled images and then fixed or slowly progressed with student detector parameters via EMA algorithm. During the test stage, the student detector is used to predict detection results of test images.

3.2. Overview

Although the teacher-student framework has achieved promising results in SSOD, it didn't consider the relations between labeled and unlabeled images. To make full use of labeled images, we propose a multi-instance alignment model which enforces the proposal-level consistency based on global class prototypes learned from all labeled images. Our MA-GCP approach benefits from the reliable global class prototypes and can learn more robust visual features for object detection. We insert our MA-GCP model into the SOTA SSOD model [27] and employ it as an additional consistency regularization for performance improvement.

3.3. Global Class Prototype Learning

In our model, we represent each class with a global prototype. These global class representations are initialized by Gaussian noise and then updated with features of ground-truth proposals via EMA algorithm. Specifically, during each training iteration, given a labeled image $\boldsymbol{x}_i^s$, we first compute its RoI feature set of ground-truth proposals by using the RoI head of student detector. The set of RoI features is denoted by $\boldsymbol{F}_i^{gt} = \{(\boldsymbol{f}_{i,j}^{gt}, \boldsymbol{y}_{i,j}^{gt})\}$, where $\boldsymbol{f}_{i,j}^{gt}$ denotes the RoI feature of j-th ground-truth proposal, $\boldsymbol{y}_{i,j}^{gt} \in \boldsymbol{C}$ denotes its class label and $\boldsymbol{C}$ denotes the set of all training object categories. After that, we average these RoI features by class and obtain a local prototype for each class:

$$\boldsymbol{v}_k = \begin{cases} \dfrac{\sum_{i,j} \boldsymbol{f}_{ij}^{gt} \mathbb{1}(\boldsymbol{y}_{i,j}^{gt} = k)}{\sum_{i,j} \mathbb{1}(\boldsymbol{y}_{i,j}^{gt} = k)} & ,\sum_i \mathbb{1}(\boldsymbol{y}_{i,j}^{gt} = k) > 0, \\ \mathbf{0} & ,\sum_i \mathbb{1}(\boldsymbol{y}_{i,j}^{gt} = k) = 0. \end{cases} \quad (2)$$

where $\boldsymbol{v}_k$ denotes the local prototype of k-th class in $\boldsymbol{C}$, $\mathbf{0}$ denotes a zero vector and $\mathbb{1}$ denotes an indicator opera-

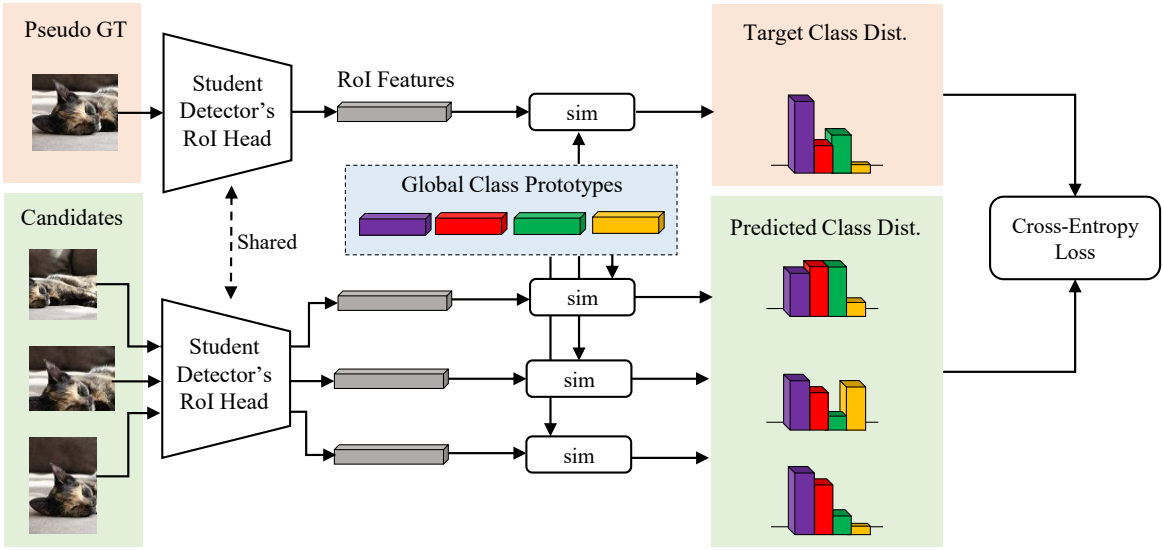

Figure 2. Illustration of the proposed multi-instance alignment model. The multi-instance alignment model takes as inputs the pseudo ground-truth proposals obtained by teacher detector and candidate proposals obtained by the student detector, and enforces their consistency by minimizing the cross-entropy loss between their class distributions based on global class prototypes.

tor. The global class prototypes are updated with local prototypes via the EMA algorithm. The process of updating global class prototypes is formulated in Equation (3).

$$\boldsymbol{g}_k = \alpha \cdot \boldsymbol{g}_k + (1-\alpha) \cdot \boldsymbol{v}_k. \tag{3}$$

where $\boldsymbol{g}_k$ denotes the global prototype of the k-th class in $\boldsymbol{C}$. α is momentum parameter and empirically set to be 0.99. By doing so, we obtain a set of global class prototypes and will employ it as reference for multi-instance alignment.

3.4. Multi-instance Alignment

With these global class prototypes, we propose a multi-instance alignment model which minimizes the difference between the class distributions of pseudo ground-truth proposals and their high-IoU candidate proposals, as illustrated in Figure 2. These class distributions are computed with global class prototypes as references. Specifically, the alignment model takes as inputs the pseudo ground-truth proposals obtained by the label refinement module, RoI proposals obtained by the student detector and global class prototypes. For each pseudo ground-truth proposal $\boldsymbol{r}_{i,j}^{pgt}$ of a given unlabeled image $\boldsymbol{x}_i^u$, RoI proposals with a high intersection-of-union(IoU) score are selected to construct its candidate proposal set, which is denoted by $\boldsymbol{R}_{i,j}^c = \{\boldsymbol{r}_{i,j,z}^c\}$. Then, we feed these pseudo ground-truth proposals and their candidate proposal sets into the RoI head of student detector, and obtain their RoI features, which are denoted by $\{\boldsymbol{f}_{i,j}^{pgt}\}$ and $\boldsymbol{F}_{i,j}^c = \{\boldsymbol{f}_{i,j,z}^c\}$, respectively. Here, j is the index of pseudo ground-truth proposals for unlabeled image $\boldsymbol{x}_i^u$ and z is the index of candidate proposals of the j-th pseudo ground-truth proposals. After that, we produce the class distribution of $\boldsymbol{f}_i^{pgt}$ based on a softmax over its similarities with the global class prototypes in the feature space:

$$p(y=k|\boldsymbol{f}_{i,j}^{pgt}) = \frac{\exp(sim(\boldsymbol{f}_{i,j}^{pgt}, \boldsymbol{g}_k))}{\sum_k \exp(sim(\boldsymbol{f}_{i,j}^{pgt}, \boldsymbol{g}_k))}. \tag{4}$$

where sim denotes the similarity metric between RoI features and global class prototypes. Here, we empirically use cosine similarity and the diagnosis study of its form is provided in Section 4.4. In the same way, we can obtain the class distribution of candidate proposals. For each pseudo ground-truth proposal, we enforce its prediction consistency with its candidate proposals by minimizing the cross-entropy loss $\mathcal{L}_{aln}^u$ between their class distributions. The formulation of $\mathcal{L}_{aln}^u$ and the outline of computing $\mathcal{L}_{aln}^u$ is given in Algorithm 1.

$$\mathcal{L}_{aln}^u = \sum_j \sum_z \sum_{k \in \boldsymbol{C}} -p(y=k|\boldsymbol{f}_{i,j}^{pgt}) \log p(y=k|\boldsymbol{f}_{i,j,z}^c). \tag{5}$$

Note that, the proposed multi-instance alignment model is different from previous consistency-based SSOD approaches [2, 12, 13, 30] from two aspects: (1) We enforce the consistency between pseudo ground-truth proposals and their candidates with high IoU scores, rather the same candidate proposals cropped from images and their horizontal-

Algorithm 1 Multi-instance Alignment Loss Computation

Require: labeled dataset $\boldsymbol{D}_s = \{(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s)\}$, unlabeled dataset $\boldsymbol{D}_u = \{\boldsymbol{x}_i^u\}$, object category set $\boldsymbol{C}$, student detector $\mathbf{M}_s$, teacher detector $\mathbf{M}_t$, label refinement module $\mathbf{H}$.

Ensure: Multi-instance alignment loss $\mathcal{L}_{aln}$.

for each $(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s) \in \boldsymbol{D}_s$ **do**
  Extract $\boldsymbol{f}_{i,j}^{gt}$ of ground-truth proposals by using $\mathbf{M}_s$;
  for $k \in \boldsymbol{C}$ **do**
    Compute local prototype $\boldsymbol{v}_k$ according to Eq. (2);
    Update global prototype $\boldsymbol{g}_k$ according to Eq. (3);
  end for
end for
for each $\boldsymbol{x}_i^u \in \boldsymbol{D}_u$ **do**
  Obtain pseudo ground-truth proposals $\boldsymbol{r}_{i,j}^{pgt}$ by using $\mathbf{M}_t$ followed by $\mathbf{H}$;
  Obtain candidate proposal set $\boldsymbol{R}_{i,j}^c = \{\boldsymbol{r}_{i,j,z}^c\}$ with high IoU scores by using $\mathbf{M}_s$;
  Compute RoI features $\boldsymbol{f}_{i,j}^{pgt}$ of $\boldsymbol{r}_{i,j}^{pgt}$ by using $\mathbf{M}_s$;
  Compute RoI features $\boldsymbol{f}_{i,j,z}^{c}$ of $\boldsymbol{r}_{i,j,z}^{c}$ by using $\mathbf{M}_s$;
  for $k \in \boldsymbol{C}$ **do**
    Compute class distributions $p(y = k|\boldsymbol{f}_{i,j}^{pgt})$ and $p(y = k|\boldsymbol{f}_{i,j,z}^{c})$ according to Eq. (4);
  end for
  Compute $\mathcal{L}_{aln}^u$ according to Eq. (5).
end for

flipped variants [12]. In this way, we increase the diversity of input proposal pairs and thus benefits SSOD. (2) The consistency is computed based on global class prototypes learned from labeled images, while previous approaches either use batch-wise prototypes [2] or directly compute consistency without any references of labeled image [12, 13, 30], which are less reliable and suffer from class imbalance issue.

3.5. The Overall Training Objective Function

The proposed multi-instance alignment model can be flexibly incorporated with the teacher-student framework to enhance its detection performance. By combining the detection loss $\mathcal{L}_{det}^{st}$ of original detector with $\mathcal{L}_{aln}$, the overall objective function is formulated in Equation (6).

$$\mathcal{L}_{overal} = \mathcal{L}_{det}^{st} + \lambda_a \mathcal{L}_{aln}^u. \tag{6}$$

where λ_a denotes the weight of multi-instance alignment loss $\mathcal{L}_{aln}$ and a diagnosis study of λ_a is provided in Section 4.4.

4. Experimental Results and Discussion

In this section, to evaluate the effectiveness of our approach, we conduct three groups of experiments: 1) Comparison with state-of-the-art SSOD approaches; 2) Ablation study to test the effectiveness of key components and select hyper-parameters; 3) Qualitative analysis.

4.1. Experimental Setup

We evaluate our approach on two benchmark datasets, i.e., PASCAL VOC [7] and MSCOCO [17] datasets. PASCAL VOC consists of 20 object categories. In SSOD, 5,011 labeled images from trainval set of VOC2007 and 11,540 unlabeled images from trainval set of VOC2012 dataset are used for training, and 4,952 images from VOC2007 test set for testing. As in [12, 13], the mean average precision (mAP) with IoU threshold 0.5 is used as the evaluation metric. MSCOCO is more challenging dataset which consists of 80 object categories. There are 118k, 5k, and 123k images in its training, validation and unlabeled sets, respectively. Following [21,27], there are two data splits for SSOD: partially-labeled split and fully-labeled split. The partially-labeled split randomly selects 1%, 5% or 10% labeled images from training set of MSCOCO as labeled dataset and the remaining images in the training set are used to construct the unlabeled dataset. The fully-labeled split is more practical, which uses the whole training set of MSCOCO as the labeled dataset and the whole unlabeled set of MSCOCO as the unlabeled dataset. In both data splits, the test set for evaluation is the validation set of MSCOCO. The mean average precision with IoU threshold ranging from 0.5 to 0.95 is used for evaluation metric as in [27]. To avoid sampling randomness in the partially-labeled split, we report averaged results over 5 data folds as in recent SSOD approaches [18, 27].

4.2. Implementation Details

Our approach is implemented based on a recent pseudo-labeling-based SSOD [27] which achieves the state-of-the-art results on benchmark datasets. The baseline detection framework is Faster RCNN [20] equipped with Pyramid Feature Network [16]. The backbone is ResNet50 [20]. For PASCAL VOC dataset, the model is trained for 60k iterations on 8 GPUs with 5 image per GPU. For each training iteration, the ratio of the number of labeled and unlabeled samples is 0.25. We train the whole model by using SGD [15] with momentum. The learning rate is initialized to 0.01 and is divided by 10 at 40k iteration and 50k iteration. The weight decay and the momentum are set to 0.0001 and 0.9, respectively. For MSCOCO dataset, We train our full model with the same learning scheme as in [27].

4.3. Comparison with State-of-the-Arts

PASCAL VOC. Table 1 provides comparative results on the PASCAL VOC dataset. From this table, we can observe that our approach outperforms the SoftTeacher baseline and achieves state-of-the-art results. Specifically, our approach exceeds the SoftTeacher baseline over 1.40% ab-

Settings	Models	Labeled dataset	Unlabeled dataset	mAP(%)
Fully-Supervised	Faster RCNN+FPN	VOC2007	None	76.30
	Faster RCNN+FPN	VOC2007+VOC2012	None	82.17
Semi-Supervised	CSD [12]	VOC2007	VOC2012	74.70
	STAC [21]	VOC2007	VOC2012	77.45
	ISD [13]	VOC2007	VOC2012	74.40
	ISMT [29]	VOC2007	VOC2012	77.23
	UGMP [26]	VOC2007	VOC2012	78.60
	UnbiasedTeacher [18]	VOC2007	VOC2012	77.37
	HumbleTeacher [22]	VOC2007	VOC2012	80.94
	Instant-Teaching [31]	VOC2007	VOC2012	79.90
	SoftTeacher* [27]	VOC2007	VOC2012	80.32
	MA-GCP(Ours)	VOC2007	VOC2012	**81.72**

Table 1. Results on Pascal VOC, evaluated on the VOC07 test set. Soft Teacher* denotes the results obtained by the official implementation of [27] with the same training scheme as ours. The mAP with IoU threshold 0.5 is used as the evaluation metric. Our approach not only outperforms competing SSOD models and achieves comparable results with fully-supervised baseline trained with all samples from both VOC2007 and VOC2012 datasets.

Model	1% labeled samples	5% labeled samples	10% labeled samples
STAC [21]	13.97± 0.35	24.38±0.12	28.64± 0.21
ISMT [29]	18.88±0.74	26.37±0.24	30.53±0.52
Unbiased Teacher [18]	20.75 ±0.12	28.27±0.11	31.50±0.10
HumbleTeacher [22]	16.96±0.38	27.70±0.15	31.61±0.28
Instant-Teaching [31]	18.05±0.15	26.75±0.05	30.40±0.05
SoftTeacher [27]	20.46±0.39	30.74±0.08	34.04±0.14
MA-GCP(Ours)	**21.30**±0.28	**31.67**±0.16	**35.02**±0.26

Table 2. Results on MSCOCO under the partially-labeled setting, evaluated on the MSCOCO's validation set. The mAP(%) with IoU threshold ranging from 0.5 to 0.95 is used as the evaluation metric. We report averaged results over 5 data folds as in most recent SSOD approaches [21, 27]. Our approach consistently outperforms competing SSOD models under different ratios of labeled images.

Model	$AP_{0.5:0.95}(\%)$
STAC [21]	39.20
ISMT [29]	39.64
UnbiasedTeacher [18]	41.30
HumbleTeacher [22]	42.37
Instant-Teaching [31]	40.20
SoftTeacher [27]	44.50
MA-GCP(Ours)	**45.92**

Table 3. Results on MSCOCO under the fully-labeled setting, where the whole training set of MSCOCO is used as labeled dataset and the whole unlabeled set of MSCOCO is used unlabeled dataset. The evaluation metric is the same as that of Table 2. Our approaches yields better results than competing SSOD approaches.

solute points in terms of mAP and outperforms the supervised baseline over 5.43% absolute points. Furthermore, our approach achieves comparable results with the supervised baseline trained with all labeled samples from both VOC2007 and VOC2012 dataset. This demonstrates the effectiveness of our MA-GCP approach in SSOD.

MSCOCO. Tables 2&3 provide comparative results under partially-labeled and fully-labeled settings on the MSCOCO dataset, respectively. For more challenging MSCOCO dataset, our approach still consistently outperforms competing SSOD models under both two settings. Compared with the state-of-the-art SoftTeacher, our approach yields 0.84, 0.93, and 0.98 points improvement under 1%, 5% and 10% labeled settings, respectively. Our approach is shown to be more effective when the ratio of labeled images are larger. It can be expected – more labeled samples help to learn more reliable global class prototypes and thus lead to more benefit to SSOD. For more practical fully-labeled setting, our MA-GCP model achieves 45.92% mAP, which outperforms SoftTeacher by 1.42 absolute points, which is bigger than that of partial-labeled setting. This indicates the benefit leaded by our MA-GCP

Model	mAP(%)
Baseline	80.32
Baseline+MA(Contrast)	80.53
Baseline+MA(Siamese)	80.79
Baseline+MA(GCP)	**81.72**

Table 4. Ablation study on contribution of key components of our model on the PASCAL VOC dataset. The evaluation metric is the same as that of Table 1. Notations: 'Baseline' –the state-of-the-art SSOD model [27]; 'MA(Costrastive)' – the multi-instance alignment model which aligns the RoI features of pseudo ground-truth proposals and their high-IoU candidates by contrastive learning; 'MA(Siamese)' – the multi-instance alignment model which aligns the RoI features of pseudo ground-truth proposals and their high-IoU candidates by Siamese learning; 'GCP' – the global class prototypes proposed in Section 3.3. The consistently improvement over Baseline shows the effectiveness of the key components proposed in our MA-GCP model.

model will not diminish when more labeled images are available. That is, our approach is suitable for practical scenarios where both large-scale labeled and unlabeled dataset are available.

4.4. Ablation Studies

Effect of Key Components. We first conduct ablation studies to validate the effectiveness of key components in our MA-GCP approach. Here, we compare our full model with three stripe-down versions: The simplest version is 'Baseline', i.e., the state-of-the-art SSOD [27]. The other two strip-down versions align the RoI features of pseudo ground-truth proposals and their high-IoU candidates by visual similarity, rather than class distributions over global class prototypes. Specifically, we follow the recent self-supervised learning approaches and develop two ways to measure their visual similarities. The first way is based on contrastive learning [5, 10], which is denoted by 'MA(Costrastive)': we feed these RoI features into a projection head $\mathbf{B}$ and then add a contrastive loss on these projected RoI features. The alignment loss formulated in Equation (5) is reformulated as follows.

$$\mathcal{L}^u_{CtAln} = -\sum_j \sum_z \log \frac{\exp(\mathbf{B}(\boldsymbol{f}^c_{i,j,z}), \mathbf{B}(\boldsymbol{f}^{pgt}_{i,j}))}{\sum_{k \neq j} \exp(\mathbf{B}(\boldsymbol{f}^c_{i,k,z}), \mathbf{B}(\boldsymbol{f}^{pgt}_{i,j}))} \tag{7}$$

The second way is based on Siamese learning [6], which is denoted by 'MA(Siamese)': we first feed these RoI features into a projection head $\mathbf{B}$ and then feed the projected RoI features of candidate proposals into a predictor head $\mathbf{S}$. We optimize the model with a new alignment loss which maximizes the cosine distance between projected RoI features of pseudo ground-truth proposals and predicted RoI features of their candidate proposals. The new alignment loss is defined in Equation (8).

$$\mathcal{L}^u_{SeAln} = -\sum_j \sum_z cos(\mathbf{S}(\mathbf{B}(\boldsymbol{f}^c_{i,j,z})), \mathbf{B}(\boldsymbol{f}^{pgt}_{i,j})) \tag{8}$$

In Equation (8), $cos(\cdot,\cdot)$ denotes the cosine distance between two inputs. As in [6], the projected RoI features of pseudo ground-truth proposals $\mathbf{B}(\boldsymbol{f}^{pgt}_{i,j})$ doesn't back-propagate the gradients.

Table 4 provides the comparative results of these models on the PASCAL VOC dataset. From this table, we can observe that: 1) Multi-instance alignment implemented by three different ways can improve the detection accuracy of Baseline model. This indicates the proposed alignment model is effective for SSOD. 2) Our alignment based on global class prototypes achieves much better results then those based on contrastive learning or Siamese learning. These results demonstrate the superior of the proposed global class prototypes, which is one of key contributions of this work. This can be expected – the global class prototypes learned from labeled images provide reliable guidance to consistency regularization and thus lead to better detection performance.

Model	mAP(%)
Negative L_2-Distance	81.04
Cosine Similarity (ours)	**81.72**

Table 5. The diagnosis study of the different form of similarity functions. The evaluation metric is exactly same as in Table 4. Notations: 'Negative L_2-Distance' – negative L_2 norm of differences between two inputs, which is formulated in Equation (10); 'Cosine(ours)' – cosine similarity formulated in Equation (9).

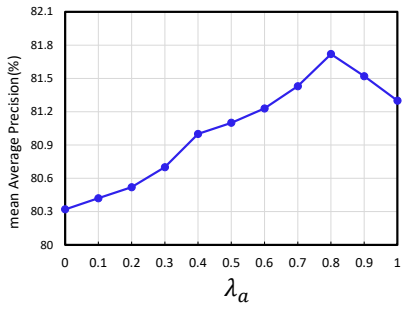

Figure 3. Diagnosis experiment over the hyper-parameter λ_a in Equation (6). The evaluation metric is the same as in Table 4.

The Form of Similarity Function. In our experiments, we implement our similarity function as a cosine similarity between two input feature vectors, which is formulated in Equation (9).

$$sim(\boldsymbol{f}^{pgt}_{i,j}, \boldsymbol{g}^k) = \frac{\boldsymbol{f}^{pgt}_{i,j}(\boldsymbol{g}^k)^T}{\|\boldsymbol{g}^k\| \cdot \|\boldsymbol{f}^{pgt}_{i,j}\|}. \tag{9}$$

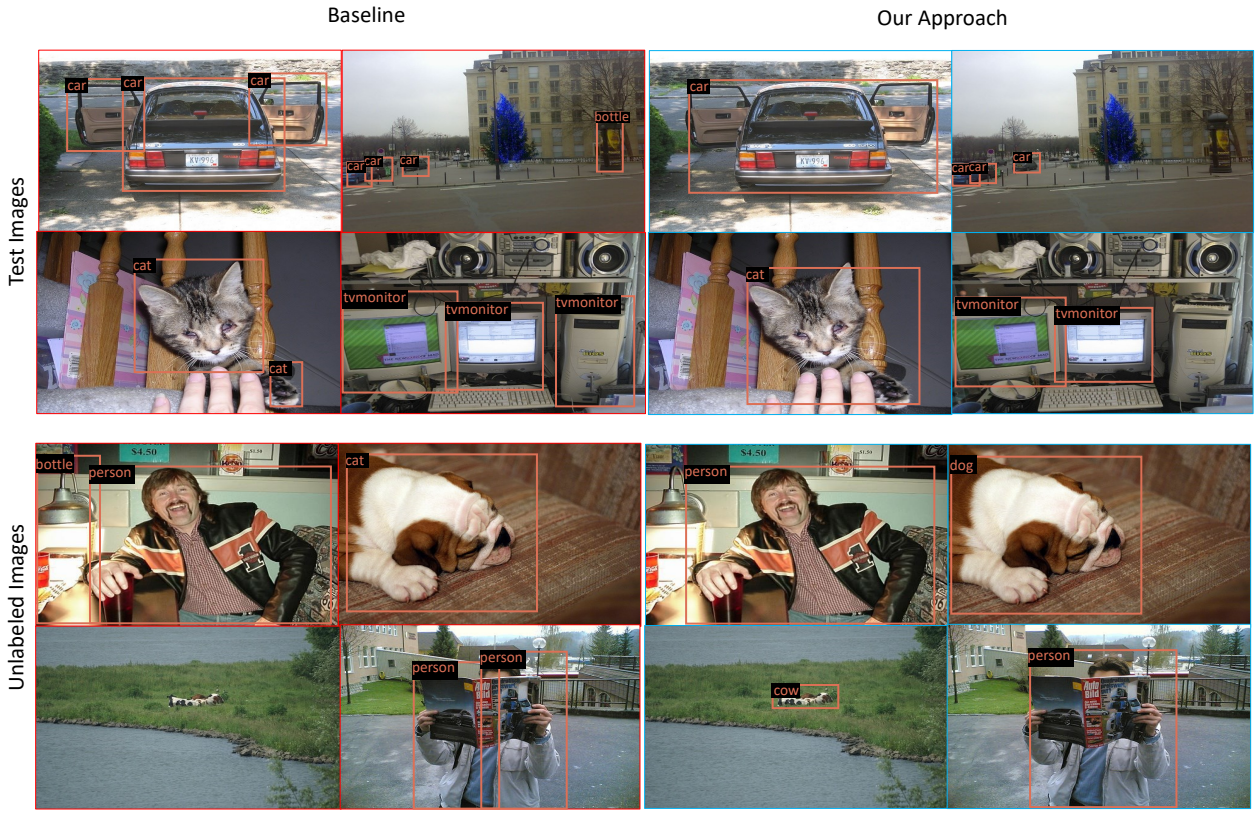

Figure 4. Qualitative visualizations of the detected novel objects obtained by baseline models and our approach on the PASCAL VOC dataset. We show our model achieves much better detection result than baseline model.

An alternative way is using the negative L_2 norm of differences between two inputs as their similarity, which is formulated as in Equation (10).

$$sim(\boldsymbol{f}_{i,j}^{pgt}, \boldsymbol{g}^k) = -\|\boldsymbol{f}_{i,j}^{pgt} - \boldsymbol{g}^k\|. \tag{10}$$

Table. 5 provides the comparative results of the alternative methods and our solution on PASCAL VOC dataset. We can observe that our method consistently outperforms the negative L_2 distance strategy. This demonstrate that our cosine similarity solution is more suitable than the negative L_2-distance.

The Weight of Multi-instance Alignment Loss. We conduct a diagnosis experiment over the important hyperparameter λ_a in Equation (6). The results of different λ_a values are illustrated in Figure 3. We can observe that we obtained the best performance when $\lambda_a = 0.8$. Thus, we set λ_a to be 0.8 in our experiments.

4.5. Qualitative Results

We provide qualitative visualizations of the detected results on unlabeled set and test set of PASCAL VOC in Figure 4. We show our model achieves much better detection result than baseline model, thanks to the multi-instance alignment model based on reliable global class prototypes. Moreover, we also illustrate some failure cases of our approach and discuss these failure cases in the supplementary material. We can observe that our approach struggles in detecting objects in complex backgrounds or rare views. This might be a future direction that needs to be investigated. Please be careful to apply this model in the situation where failures lead to serious adverse consequences.

5. Conclusion

In this paper, we propose to make full use of labeled images in SSOD by developing a multi-instance alignment model based on global class prototypes. The global class prototypes learned by using all labeled training images are shown to be the reliable guidance for improving SSOD. With the reliable guidance, we can enhance the consistency between prediction results of teacher detector and student detector. By inserting our MA-GCP approach into the state-of-the-art SSOD model, we obtain a strong solution for object detection.

References

[1] Abulikemu Abuduweili, Xingjian Li, Humphrey Shi, Cheng-Zhong Xu, and Dejing Dou. Adaptive consistency regularization for semi-supervised transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6923–6932, 2021. 2

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael G. Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *IEEE Conference on Computer Vision, ICCV*, pages 3060–3069, 2021. 2, 4, 5

[3] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations, ICLR*, 2020. 2

[4] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems, NeurIPS*, pages 5050–5060, 2019. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, volume 119, pages 1597–1607, 2020. 7

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 15750–15758, 2021. 7

[7] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision, IJCV*, 88(2):303–338, 2010. 5

[8] Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, and Ram Nevatia. NOTE-RCNN: noise tolerant ensemble RCNN for semi-supervised object detection. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 9507–9516, 2019. 1

[9] Chengyue Gong, Dilin Wang, and Qiang Liu. Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 13683–13692, 2021. 2

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 9726–9735, 2020. 7

[11] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: Similar pseudo label exploitation for semi-supervised classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 15099–15108, 2021. 2

[12] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Advances in Neural Information Processing Systems, NeurIPS*, pages 10758–10767, 2019. 1, 2, 4, 5, 6

[13] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 11602–11611, 2021. 1, 2, 4, 5, 6

[14] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations, ICLR*, 2017. 2

[15] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, and Lawrence D. Hubbard, Wayne E.and Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 5

[16] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 936–944, 2017. 5

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanann, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision, ECCV*, pages 740–755, 2014. 5

[18] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *International Conference on Learning Representations, ICLR*, 2021. 1, 2, 3, 5, 6

[19] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1979–1993, 2019. 2

[20] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 5

[21] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *Arxiv abs/2005.04757*, 2020. 3, 5, 6

[22] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3132–3141, 2021. 1, 2, 3, 6

[23] Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandréa, Robert J. Gaizauskas, and Liming Chen. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2119–2128, 2016. 1

[24] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural*

Information Processing Systems, NIPS, pages 1195–1204, 2017. 2

[25] Dong Wang, Yuan Zhang, Kexin Zhang, and Liwei Wang. Focalmix: Semi-supervised learning for 3d medical image detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3950–3959, 2020. 2

[26] Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4568–4577, 2021. 1, 3, 6

[27] Mengde Xu, Zheng Zhang, Jianfeng Wang Han Hu, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *IEEE Conference on Computer Vision, ICCV*, pages 3060–3069, 2021. 1, 3, 5, 6, 7

[28] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 7283–7292, 2019. 1, 2

[29] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5941–5950, 2021. 1, 2, 3, 6

[30] Chao Ye, Huaidong Zhang, Xuemiao Xu, Weiwei Cai, Jing Qin, and Kup-Sze Choi. Object detection in densely packed scenes via semi-supervised learning with dual consistency. In *International Joint Conference on Artificial Intelligence, IJCAI*, pages 1245–1251, 2021. 1, 2, 4, 5

[31] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4081–4090, 2021. 1, 2, 3, 6