Point-Teaching: Weakly Semi-supervised Object Detection with Point Annotations

Yongtao Ge[1], Qiang Zhou[2], Xinlong Wang[3], Chunhua Shen[4], Zhibin Wang[2], Hao Li [2]
[1] The University of Adelaide
[2] Alibaba Group
[3] Beijing Academy of Artificial Intelligence
[4] Zhejiang University
yongtao.ge@adelaide.edu.au, {jianchong.zq, zhibin.waz, lihao.lh}@alibaba-inc.com, wangxinlong@baai.ac.cn, chunhuashen@zju.edu.cn

Abstract

Point annotations are considerably more time-efficient than bounding box annotations. However, how to use cheap point annotations to boost the performance of semi-supervised object detection remains largely unsolved. In this work, we present **Point-Teaching**, a weakly semi-supervised object detection framework to fully exploit the point annotations (WSSOD-P). Specifically, we propose a Hungarian-based point-matching method to generate pseudo labels for point-annotated images. We further propose multiple instance learning (MIL) approaches at the level of images and points to supervise the object detector with point annotations. Finally, we propose a simple-yet-effective data augmentation, termed point-guided copy-paste, to reduce the impact of the unmatched points. Experiments demonstrate the effectiveness of our method on a few datasets and various data regimes. In particular, Point-Teaching outperforms the previous best method Group R-CNN by 3.1 AP with 5% fully labeled data and 2.3 AP with 30% fully labeled data on MS COCO dataset. We believe that our proposed framework can largely lower the bar of learning accurate object detectors and pave the way for its broader applications. The code is available at https://github.com/YongtaoGe/Point-Teaching.

Introduction

Great progress has been achieved in object detection and segmentation in recent years (Ren et al. 2015; Redmon et al. 2016; Lin et al. 2017; Tian et al. 2019; He et al. 2017; Tian, Shen, and Chen 2020; Wang et al. 2021). Accurate object detectors can be trained using large fully-labeled datasets (Lin et al. 2014; Gupta, Dollár, and Girshick 2019). However, annotating large-scale object detection datasets are extremely expensive and time-consuming, as it requires the annotators to find all the objects of interest in the images and to draw a tight bounding box/segmentation mask for each of them.

How to train object detectors with fewer annotations has attracted increasing attention. Weakly supervised object detection (WSOD) methods (Song et al. 2014; Cinbis, Verbeek, and Schmid 2014; Bilen and Vedaldi 2016; Kantorov et al. 2016; Tang et al. 2017; Ren et al. 2020a) reduce the cost via replacing the box annotations with cheaper annotations, *e.g.*, image-level categories, point clicks and squiggles. Semi-supervised object detection (SSOD) methods (Jeong et al. 2019; Sohn et al. 2020; Liu et al. 2021; Zhou et al. 2021b; Yang et al. 2021) train object detectors with a small amount of fully-labeled images and large-scale unlabeled images. However, although both ways can reduce the annotation cost, the performance of the trained detectors is still far behind the fully-supervised counterpart.

In this paper, we aim to train object detectors with considerably fewer annotations while achieving comparable performance with the fully-supervised counterpart. To achieve this goal, there are two key problems: 1) what annotation formats to use and 2) how to train object detectors with such annotations. A cheap but effective annotation format for object detection should be 1) simple to annotate, 2) convenient to store and use 3) localization-aware. Among various weak formats, point click annotation stands out as it meets all the requirements. Point click provides a stronger prior of object location compared with image-level category annotation. Meanwhile, it does not require detailed and expensive location information such as object bounding box or segmentation masks, thus being considerably more time-efficient. According to (Papadopoulos et al. 2017a) and (Cheng, Parkhi, and Kirillov 2022), a box annotation takes 7 seconds while a point annotation takes 0.8-0.9 seconds. To achieve the best balance of detection performance and annotation cost, we adopt mixed annotation formats to construct the training dataset. In the following, we use *point annotated setting* to represent such a dataset which comprises a small number of fully annotated images and massive point annotated images. Under this setting, we are able to obtain abundant annotations in a relatively cheaper manner (Su, Deng, and Fei-Fei 2012; Russakovsky, Li, and Li 2015).

To fully utilize both the limited box annotations and abundant point annotations, we propose a novel weakly semi-supervised object detection framework, termed **Point-Teaching**. Inspired by Mean Teacher (Tarvainen and Valpola 2017) and Unbiased Teacher (Liu et al. 2021), we construct a Student model and a Teacher model with the same architecture. In each training iteration, weakly augmented point-labeled images are fed to the Teacher model to generate reliable pseudo-bounding boxes. The Student is then optimized on fully labeled and pseudo-labeled images with strong augmentation. The Teacher is updated via the Exponential Moving Average (EMA) of the Student. Within this basic framework, we propose three key components tailored for point annotations. First, we propose the hungarian-based point matching method to generate pseudo labels for point annotated images. A spatial cost and a classification cost are introduced to find the bipartite matching between point annotations and predicted box proposals.

We further propose multiple instance learning (MIL) approaches at the level of images and points to supervise the object detector with point annotations. Inspired by previous WSOD works (Bilen and Vedaldi 2016; Kantorov et al. 2016; Tang et al. 2017; Ren et al. 2020a), we perform image-wise MIL via treating the whole image as a bag of object proposals. These proposals are aggregated for predicting all presented classes in the image, supervised by image-level labels during the training. To leverage the location information of point annotations, we propose point-wise MIL, which selects the highest detection score proposal with the same class label as the only positive and suppresses the rest of proposals as negatives around the given point annotation. Finally, we propose the point-guided copy-paste augmentation strategy. The motivation is that there still exist some point annotations that have not been matched any proposals after the point matching. To further utilize those unmatched points, we maintain an online object bank and paste same-class objects to unmatched point annotations during the training. The point-guided copy-paste makes the distribution of generated pseudo labels closer to that of the ground truth.

Experiments demonstrate the effectiveness of our method on different datasets and various data regimes. Point-Teaching has the following advantages: 1) Our method can boost the performance of existing SSOD methods, *e.g.*, over the strong semi-supervised baseline method Unbiased Teacher (Liu et al. 2021), our detector achieves significant improvements of 9.1 AP with 0.5% fully labeled data on MS COCO. 2) Our method outperforms all existing WSSOD-P methods in all data regimes by a large margin. In particular, when using 30% fully labeled data from MS COCO, our method outperforms previous state-of-the-art WSSOD-P method Group R-CNN (Zhang et al. 2022) by 2.3 AP and Point DETR (Chen et al. 2021) by 3.4 AP.

Our main contributions are summarized as follows:

- We propose a simple and effective training framework for weakly semi-supervised object detection, termed **Point-Teaching**, which integrates point annotations into semi-supervised learning. The key components of Point-Teaching include Hungarian-based point-matching approach, image-wise and instance-wise MIL loss, and point-guided copy-paste augmentation.
- Extensive experiments are conducted on MS-COCO and VOC datasets to verify the effectiveness of our method. Point-Teaching significantly outperforms the existing methods (Chen et al. 2021; Zhang et al. 2022) and greatly narrows the gap between weakly semi-supervised and fully-supervised object detectors.
- We further extend Point-Teaching from WSSOD-P to weakly semi-supervised instance segmentation (WSSIS) and weakly-supervised instance segmentation (WSIS), setting a strong baseline for the two challenging tasks.

Related Work

Fully-supervised object detection. With the large-scale fully annotated detection datasets (Lin et al. 2014; Gupta, Dollár, and Girshick 2019), existing modern detectors have obtained great improvements in the object detection task. These detectors can be divided into three categories: two-stage detectors (Ren et al. 2015; Wu et al. 2020), one-stage detectors (Redmon et al. 2016; Liu et al. 2016; Tian et al. 2019) and the recent end-to-end detectors (Carion et al. 2020; Zhu et al. 2021; Zhou et al. 2021a). Faster RCNN is a popular two-stage detector that first generates region proposals and then refines these proposals in the second stage. Unlike two-stage detectors, one-stage detectors, such as YOLO (Redmon et al. 2016) and FCOS (Tian et al. 2019), directly output dense predictions of classification and regression without refinement. Recently, DETR (Carion et al. 2020) introduces the transformer encoder-decoder architecture to object detection and effectively removes the need for many hand-craft components, e.g. predefined anchors and non-maximum suppression (NMS) post-processing. Despite the great success, these detectors are trained with large amounts of expensive fully-labeled data. Therefore, a lot of work has been proposed to reduce the annotation cost.

Weakly-supervised object detection. There exist many WSOD works that focus on training object detector with weakly-labeled data. Most previous studies have two phases: proposal mining and proposal refinement. The proposal mining phase is formulated as the MIL problem to implicitly mine object locations with image-level labels. The proposal refinement phase aims at refine the object location with the predictions from the proposal mining phase. WSDDN (Bilen and Vedaldi 2016) proposes a two-stream network to simultaneously perform region selection and classification. The region-level scores from these two streams are then element-wise multiplied and transformed to image-level scores by summing over all regions. Following WSDDN (Bilen and Vedaldi 2016), ContextLocNe (Kantorov et al. 2016) introduces context information. OICR (Tang et al. 2017) presents a multi-stage refinement strategy to avoid the MIL detector being trapped in the local minimum. PCL (Tang et al. 2020) proposes to refine instance classifiers by clustering region proposals in an image to different clusters. MIST (Ren et al. 2020a) proposes a multiple-instance self-training framework. OIM (Lin et al. 2020) effectively mines all possible instances by introducing information propagation on spatial and appearance graphs. However, propagating image-level weak supervision to instance-level training data inevitably involves a large amount of noisy information and the performance of these methods are limited.

Semi-supervised object detection. Besides WSOD, SSOD addresses the problem by using a large amount of unlabeled data, together with a small set of labeled data. One popular SSOD technique is consistency regularization, which aims to regularize the detector's prediction with an image of different augmentations. CSD (Jeong et al. 2019) enforces the detector to make consistent predictions on an input image and its horizontally flipped counterpart. ISD (Jeong et al. 2021) proposes an interpolation-based method for SSOD. Another emerging SSOD approach is pseudo labeling, where a teacher model is trained on labeled data to generate pseudo labels on unlabeled data, and a student model is then trained on both labeled and pseudo-labeled data. STAC (Sohn et al. 2020) pre-trains a model on labeled data and fine-tunes it on both labeled and unlabeled data iteratively. Instance-Teaching (Zhou et al. 2021b) introduces a co-rectify scheme for alleviating confirmation bias of pseudo labels. Unbiased Teacher (Liu et al. 2021) proposes a class-balance loss to address the class imbalance issue in pseudo-labels and refine the teacher model via Exponential Moving Average (EMA). DenseTeacher (Zhou et al. 2022) replaces the sparse pseudo-boxes with the dense prediction as a united form of pseudo-label.

Weakly semi-supervised object detection. Image-level annotation is a kind of weak annotation compared to box annotation. However, it is not optimal for object detection since the lack of instance-level information. Recently, point supervision (Ren et al. 2020b; Chen et al. 2021; Papadopoulos et al. 2017b,a) has been employed in WSSOD. Papadopoulos et al. (Papadopoulos et al. 2017b,a) collect click annotation for the PASCAL VOC dataset and train an object detector through iterative multiple instance learning. UFO^2 (Ren et al. 2020b) proposes a unified object detection framework that can handle different forms of supervision simultaneously, including box annotation and point annotation. Point DETR (Chen et al. 2021) extends DETR (Carion et al. 2020) by adding a point encoder and thus can convert point annotations to pseudo box annotations. Group R-CNN (Zhang et al. 2022) proposes to use instance-level proposal grouping for each point annotation and thus can get a high recall rate. In this paper, we follow this setting and introduce several methods for improving the performance of WSSOD-P.

Method

Preliminaries

Problem definition. In this work, we study weakly semi-supervised object detection under the *point annotated setting*, in which the dataset consists of a small set of fully annotated images $D_F = \{(I_i, \{\hat{b}_{i,j}\})\}_{i=1}^{N_F}$ and a large set of point annotated images $D_P = \{(I_i, \{\hat{p}_{i,j}\})\}_{i=1}^{N_P}$. N_F and N_P are the numbers of fully labeled and point labeled images respectively. I denotes fully or point labeled images, i is the image index and j is the instance index of image I_i. For fully annotated images, the annotations $\{\hat{b}_{i,j}\}$ includes box coordinates $(\hat{b}_{i,j}^{x_1}, \hat{b}_{i,j}^{y_1}, \hat{b}_{i,j}^{x_2}, \hat{b}_{i,j}^{y_2})$ and class label $\hat{b}_{i,j}^{l}$. For point annotated images, the annotations $\{\hat{p}_{i,j}\}$ includes point location $(\hat{p}_{i,j}^{x}, \hat{p}_{i,j}^{y})$ and class label $\hat{p}_{i,j}^{l}$. For point-annotated images, we only need to randomly annotate one point for each object instance, thereby the annotation cost can be greatly reduced.

Overall Architecture

For a fair comparison, we take Faster RCNN with FPN (Ren et al. 2015) and ResNet-50 backbone (He et al. 2016) as our baseline object detector. Compared to the original Faster RCNN network (Ren et al. 2015), we add two additional parallel branches to the RCNN head, termed Objectness-I branch and the Objectness-P branch, respectively. The Objectness-I branch is used to suppress the likelihood of inconsistent classification predictions with image-level annotations and is optimized with image-wise MIL loss. The Objectness-P branch is developed to measure the quality of pseudo boxes at point level and is supervised with point-wise MIL loss. The key difference is that the Objectness-I branch selects the most probable region proposals for each class from the image-level bag with different classes. While the Objectness-P branch performs binary classification to select the most probable region proposal from the point-level bag that only contains region proposals of the same class.

The training pipeline of Point-Teaching is represented in Fig. 1. Inspired by Mean Teacher (Tarvainen and Valpola 2017) and Unbiased Teacher (Liu et al. 2021), there are two models with the same architecture, a student model and a teacher model. In each training iteration, weakly augmented point-labeled images from the dataset D_P are firstly fed to the Teacher for reliable pseudo labels; the Student is then optimized by labels from fully-labeled dataset D_F, and pseudo labels generated from the Teacher with strong augmentation; Finally, the Teacher is updated by EMA of the Student. Different from the original Unbiased Teacher (Liu et al. 2021), there are three key components within the proposed framework: *hungarian-based point matching strategy, point supervision with image-wise and instance-wise MIL loss, and point-guided copy-paste augmentation.*

Point Matching

In order to find the best matching between the annotated points and the predicted boxes, *i.e.*, to choose the best box prediction for each point annotation, we propose a simple point matching method, termed hungarian-based point matching. Specifically, we design two types of matching costs between annotated points and predicted boxes: a spatial cost and a classification cost. For the spatial cost, we consider two factors: 1) Predicted boxes that share the same class label with the given point annotation should have a low cost. 2) Predicted boxes with point annotations inside lead to a low cost. For the classification cost, higher confidence scores of the Classification branch and Objectness-P branch lead to a lower cost.

Formally, the cost matrix $\mathcal{L}_{\text{match}} \in \mathbb{R}^{N_p \times N_b}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\text{match}}(i,j) = &\underbrace{(1 - \mathbb{1}[\hat{p}_i \text{ in } b_j] \cdot \mathbb{1}[\hat{p}_i^l = b_j^l])}_{\text{spatial cost}} \\ &+ \underbrace{(1 - \sigma(\boldsymbol{s}_{j,\hat{p}_i^l}) \cdot \sigma^{\text{P}}(\boldsymbol{s}_{j,1}^{\text{P}}))}_{\text{classification cost}},\end{aligned} \tag{1}$$

where i is the index of the annotated points, and j is the index of the predicted boxes. $\mathcal{L}_{\text{match}}(i,j)$ denotes the matching cost between the annotated point $\hat{p}_i$ and the predicted box b_j.

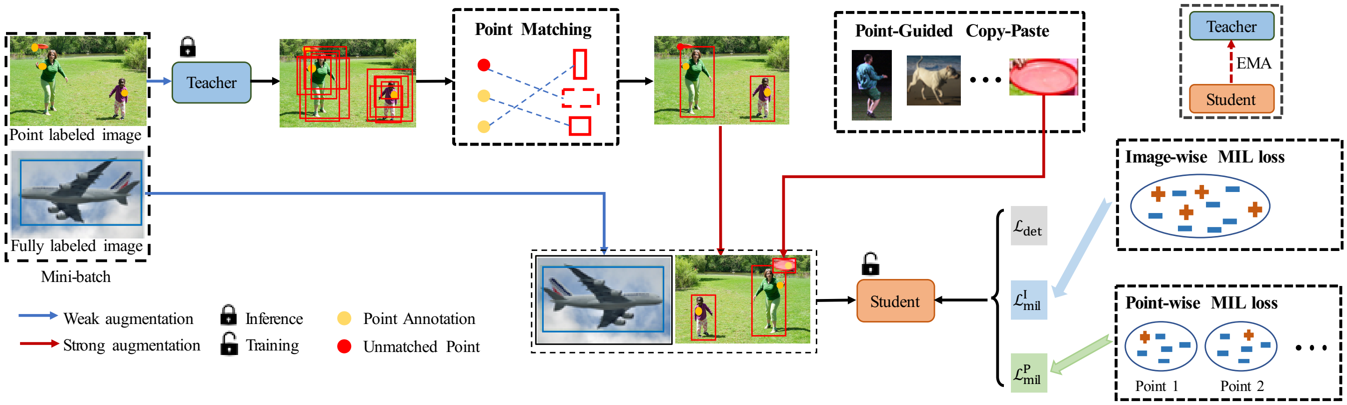

Figure 1: The training process of Point-Teaching. In each training iteration, the Teacher model first generates pseudo box annotations for the point-labeled images with weak augmentation. The Student model is then trained on fully-labeled images with weak augmentation and point-labeled images with strong augmentation. The Teacher model is gradually updated by the student model via EMA. Image-wise MIL loss constructs a bag containing all predicted boxes, and the number of positive boxes in the bag is uncertain. The point-wise MIL loss constructs a bag for each annotated point, and there is only one positive box in these bags.

N_p is the number of annotated points, and N_b is the number of predicted boxes. $\hat{p}_i^l$ indicates the class label of the annotated point $\hat{p}_i$, b_j^l indicates the class label of the predicted box b_j. We use $\boldsymbol{s} \in \mathbb{R}^{N_b \times (C+1)}$ and $\boldsymbol{s}^{\text{P}} \in \mathbb{R}^{N_b \times 2}$ to denote the outputs of Classification and Objectness-P branches, respectively, where C denotes the number of categories excluding the background. $\sigma(\cdot)$ represents the *softmax* operation on the Classification output along the second dimension. $\sigma^{\text{P}}(\cdot)$ represents the *softmax* operation on the Objectness-P output along the second dimension.

Once the cost matrix is defined, the point matching problem could be mathematically formulated as a bipartite matching problem as:

$$\hat{\pi} = \underset{\pi \in \mathfrak{S}_{N_b}}{\arg\min} \sum_{i}^{N_p} \mathcal{L}_{\text{match}}\left(i, \pi(i)\right), \tag{2}$$

where $\pi \in \mathfrak{S}_{N_b}$ indicates a permutation of N_b elements. This optimal assignment can be solved with the Hungarian algorithm (Kuhn 1955).

MIL Loss for Images and Points

In this section, we present the overall loss function $\mathcal{L}$ of Point-Teaching framework.

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_1 \mathcal{L}^{\text{I}}_{\text{mil}} + \lambda_2 \mathcal{L}^{\text{P}}_{\text{mil}}. \tag{3}$$

As shown in Eq. (3), the overall loss $\mathcal{L}$ consists of three parts: $\mathcal{L}_{\text{det}}$, $\mathcal{L}^{\text{I}}_{\text{mil}}$ and $\mathcal{L}^{\text{P}}_{\text{mil}}$, respectively. $\mathcal{L}_{\text{det}}$ represents the losses of the original object detector, *e.g.* classification loss and regression loss in RPN and ROI head of Faster RCNN. $\mathcal{L}^{\text{I}}_{\text{mil}}$ is image-wise MIL loss, which is proposed in WSDDN (Bilen and Vedaldi 2016). $\mathcal{L}^{\text{P}}_{\text{mil}}$ is our proposed point-wise MIL loss, which is defined below. λ_1 and λ_2 are hyper-parameters used to balance these three loss terms.

Image-wise MIL loss. Given the point annotations, we can easily obtain image-level labels $\{\hat{\phi}_c,\ c = 1, \cdots, C\}$. Image labels can help improve the performance of object detection in two ways. First, for categories that do not present in the image, the image-level supervision could help decrease the confidence score of the corresponding predicted boxes. Second, it helps detect the objects of the categories that present in the image.

Taking hundreds of predicted boxes as a bag, we only know the class labels of the entire bag and do not know the individual class label of each predicted box. Let us denote by $\boldsymbol{s}, \boldsymbol{s}^{\text{I}} \in \mathbb{R}^{N_b \times C}$ the output of Classification branch and Objectness-I branch, respectively; $\sigma^{\text{I}}(\cdot)$ the *softmax* operation on the first dimension. We share ROI features of box proposals with two fully-connected layers and then produce two score matrices $\sigma\left(\boldsymbol{s}\right), \sigma^{\text{I}}\left(\boldsymbol{s}^{\text{I}}\right) \in \mathbb{R}^{N_b \times C}$ by Classification branch and Objectness-I branch, respectively. Then the element-wise product of the two score matrix is a new score matrix $X^{\text{s}} \in \mathbb{R}^{N_b \times C}$, which can be formulated as: $X^{\text{s}} = \sigma\left(\boldsymbol{s}\right) \odot \sigma^{\text{I}}\left(\boldsymbol{s}^{\text{I}}\right)$. Finally, a sum pooling is applied to obtain image-level classification scores:

$$\phi_c = \sum_{i=1}^{N_b} X^{\text{s}}_{ic} = \sum_{i=1}^{N_b} \left[\sigma\left(\boldsymbol{s}_{i,c}\right) \odot \sigma^{\text{I}}\left(\boldsymbol{s}^{\text{I}}_{i,c}\right)\right]. \tag{4}$$

Based on the obtained image-level labels and image-level classification scores, the introduced image-wise MIL loss is defined as the sum of binary cross-entropy loss across all categories:

$$\mathcal{L}^{\text{I}}_{\text{mil}} = -\sum_{c=1}^{C} \left(\hat{\phi}_c \log(\phi_c) + (1 - \hat{\phi}_c) \log(1 - \phi_c)\right), \tag{5}$$

where C is the number of categories, $\hat{\phi}_c \in \{0, 1\}^C$ is the image-level one-hot labels, and ϕ_c denotes the predicted image-wise classification scores.

Point-wise MIL loss. To perform multiple instance learning at the point level, we construct a bag with part of the predicted boxes for each annotated point, as shown in Fig. 1. For example, the constructed bag Ψ_i for the annotated point $\hat{p}_i$ consists of those predicted boxes that enclose point $\hat{p}_i$ and have the same class label as $\hat{p}_i$. In other words, $\Psi_i = \{b_j \mid \mathbb{1}[\hat{p}_i \text{ in } b_j] \cdot \mathbb{1}[\hat{p}_i^l = b_j^l]\}$, in which $\hat{p}_i^l$ denotes the class label of annotated point $\hat{p}_i$, and b_j^l denotes the class label of the predicted box b_j. Unlike the bag of image-wise MIL loss, there is only one positive box proposal inside Ψ_i, defined as the best-predicted box corresponding to the annotated point $\hat{p}_i$. Assuming we know how to calculate the bag-level confidence score φ_i for bag Ψ_i, we can define the proposed point-wise MIL loss as:

$$\mathcal{L}_{\text{mil}}^{\text{P}} = -\sum_{i=1}^{N_p} \log(\varphi_i), \tag{6}$$

where N_p denotes the number of annotated points, and $\mathcal{L}_{\text{mil}}^{\text{P}}$ is the sum of the binary cross-entropy loss for all annotated points.

Next, we explain how to compute the bag-level confidence score φ_i corresponding to bag Ψ_i. To help find out the best box proposal inside Ψ_i, we add the Objectness-P branch. This branch performs binary classification to predict whether the box is the best prediction inside bag Ψ_i, and its output is denoted as $\text{s}^{\text{P}} \in \mathbb{R}^{N \times 2}$. Since there should be only one positive box inside the bag Ψ_i, we use a slightly different way to compute φ_i. As shown in Eq. (7):

$$\varphi_i = \sum_{k=1}^{|\Psi_i|} \left[\sigma(\boldsymbol{s}_{k,\hat{p}_i^l}) \odot \sigma^{\text{P}}(\boldsymbol{s}_{k,1}^{\text{P}}) \odot \prod_{m!=k} \sigma^{\text{P}}(\boldsymbol{s}_{m,0}^{\text{P}}) \right], \tag{7}$$

in which $\sigma(\cdot)$ and $\sigma^{\text{P}}(\cdot)$ denote *softmax* operation as described earlier, $|\Psi_i|$ indicates the number of predicted boxes in bag Ψ_i. Comparing Eq. (4) and Eq. (7), we can find that the element-wise multiplication before accumulation is different. Taking the k^{th} predicted box in bag Ψ_i as an example. In addition to multiplying the positive confidence score of the two branches (i.e., $\sigma(\boldsymbol{s}_{k,\hat{p}_i^l}) \cdot \sigma^{\text{P}}(\boldsymbol{s}_{k,1}^{\text{P}})$), we also multiply the negative confidence scores of the Objectness-P branch of the remaining boxes in bag Ψ_i (i.e., $\prod_{m!=k} \sigma^{\text{P}}(\boldsymbol{s}_{m,0}^{\text{P}})$). With the help of negative confidence scores, the proposed point-wise MIL loss can encourage that each bags have and only have one positive box with the highest positive confidence score, while the positive confidence score of remaining boxes is suppressed. The pseudo-code of point-wise MIL loss based on PyTorch is provided in the supplementary.

Point-Guided Copy-Paste

During the point matching, we observe that some of the annotated points are not matched with any predicted boxes, and these unmatched points usually correspond to difficult instances to be detected (e.g. instances from minority classes). Ignoring these unmatched points may cause the class imbalance of the generated pseudo boxes. The confirmation bias in pseudo boxes further reinforces the imbalance issue. To alleviate the impact of these unmatched points, we propose a simple data augmentation strategy termed *point-guided copy-paste*. Different from naively copying ground-truth boxes from one labeled image to another unlabeled image like Simple Copy-Paste (Ghiasi et al. 2021), we maintain a dynamic object bank as depicted in Fig. 1, which will be updated with ground truth object patches (cropped based on box annotation) from fully labeled images and pseudo object patches from point labeled images during each training iteration. For each unmatched point after the point matching stage, we randomly select an object patch with the same class label from the object bank, and paste the selected patch near the point on the original image. The effectiveness of point-guided copy-paste augmentation is verified in Table 7.

Experiment

Datasets

We mainly benchmark our proposed method on the large-scale dataset MS-COCO (Lin et al. 2014). Following (Chen et al. 2021), we synthesize the point annotations by randomly sampling a point inside the annotated box. Then we discard the box annotations of point-labeled images. Specifically, We randomly selected 0.5%, 1%, 2%, 5% , 10% and 30% from the 118k labeled images as the fully-labeled set, and the remainder is used as the point-labeled set. Model performance is evaluated on the COCO2017 val set. We also conduct experiments on PASCAL VOC (Everingham et al. 2010). The VOC results are in Appendix.

Implementation Details

We implement our proposed Point-Teaching framework based on the Detectron2 toolbox (Wu et al. 2019). For fair comparison with existing works (Sohn et al. 2020; Zhou et al. 2021b; Liu et al. 2021), we take Faster RCNN with FPN (Ren et al. 2015) as our object detector and ResNet-50 (He et al. 2016) as backbone. The feature weights are initialized by the ImageNet pretrained model. Our method mainly contains three hyperparameters: τ, λ_1 and λ_2, which indicates the score threshold of the pseudo boxes, the loss weight of image-wise MIL loss and the loss weight of point-wise MIL loss, respectively. We set $\tau = 0.05$, $\lambda_1 = 1.0$ and $\lambda_2 = 0.05$ unless otherwise specified.

We use $AP_{50:95}$ (denoted as AP) as evaluation metric. On Pascal VOC, the models are trained for 40k iterations on 8 GPUs and with batch size 32, which contains 16 box-labeled images and 16 point-labeled images respectively. Other training and testing details are the same as the original Unbiased-Teacher (Liu et al. 2021).

Ablation Study

When conducting ablation experiments, we choose 1% MS-COCO protocol and take a quick learning schedule of 90k iterations and a smaller batch size of 32, containing 16 box-labeled images and 16 point-labeled images, respectively.

Effects of point location. We verify the effectiveness of point annotation location to Point-Teaching between two point location schemes: a center point and an arbitrary point on objects. As shown in Table 2, when using the center point on objects as our annotation, Point-Teaching achieves 25.2 AP. While we randomly sample points inside the box annotation, the performance only slightly drops 0.01% AP, showing that Point-Teaching is insensitive to the location of point annotation.

Method	Type	0.5%	1%	5%	10%	30%
Supervised	FSOD	6.83 ± 0.15	9.05 ± 0.16	18.47 ± 0.22	23.86 ± 0.81	31.99 ± 0.82
CSD (Jeong et al. 2019)	SSOD	7.41 ± 0.21	10.51 ± 0.06	18.63 ± 0.07	24.46 ± 0.08	-
STAC (Sohn et al. 2020)	SSOD	9.78 ± 0.53	13.97 ± 0.35	24.38 ± 0.12	28.64 ± 0.21	-
Instant-Teaching (Zhou et al. 2021b)	SSOD	-	18.05 ± 0.15	26.75 ± 0.05	30.40 ± 0.05	-
Unbiased Teacher (Liu et al. 2021)	SSOD	16.94 ± 0.23	20.16 ± 0.12	27.84 ± 0.11	31.39 ± 0.10	-
DenseTeacher (Zhou et al. 2022)	SSOD	-	22.38 ± 0.31	27.20 ± 0.20	**37.13 ± 0.12**	-
Point DETR (Chen et al. 2021)	WSSOD-P	-	-	26.2	30.4	34.8
Group R-CNN (Zhang et al. 2022)	WSSOD-P	-	-	30.1	32.6	35.4
Point-Teaching	WSSOD-P	**26.02 ± 0.09**	**28.34 ± 0.02**	**33.15 ± 0.07**	35.18 ± 0.09	**38.20 ± 0.10**

Table 1: Comparison of our proposed Point-Teaching with other SSOD (without point-level labels) and WSSOD-P (with point-level labels) methods on COCO val. set. All these models use R50-FPN as the backbone network. Point-Teaching are trained with a batch size of 64 (32 fully-labeled images and 32 point-labeled images) and 180k iterations. Note that the upper bound of 100% fully supervised model is 40.2 AP (Wu et al. 2019).

Point Location	$\mathrm{AP}_{50:95}$	AP_{50}
random	25.18	48.26
center	25.19	48.28

Table 2: Comparison of the effectiveness of the point location on the COCO validation set. 'random' and 'center' indicate the annotation location on objects.

Point Matching	$\mathrm{AP}_{50:95}$	AP_{50}
None	20.2	36.5
Hungarian	**25.2**	**48.3**

Table 3: Comparison of detection accuracy on the COCO val. set by varying the point matching methods when selecting pseudo box annotations.

Effects of point matching. We explore the impact of our proposed Hungarian-based point matching method on the model performance. In this experiment, we set the loss weights of λ_1 and λ_2 to 0. As shown in Table 3, when point matching is not used, the model reaches 20.2 AP, as reported in Unbiased-Teacher (Liu et al. 2021). Taking point annotations into consideration and using our proposed Hungarian matching, the model reaches 25.2 AP, which improves the AP with 5.0 absolute points.

λ_1	λ_2	$\mathrm{AP}_{50:95}$	AP_{50}
0.5	0	25.00	47.88
1.0		**25.66**	**49.04**
1.5		25.01	48.46

Table 4: Comparison of detection accuracy on COCO val. set when varying the loss weight λ_1 of image-wise MIL loss.

Loss weight λ_1 of image-wise MIL loss. We conduct experiments to explore the effect of loss weight λ_1 of image-wise MIL loss. In these experiments, we use Hungarian-based point matching and set the loss weight λ_2 of point-wise MIL loss to 0, *i.e.* the Objectness-P branch is not optimized during training and the Objectness-P score is removed in Eq. (1) when computing the cost matrix. As shown in Table 4, when loss weight λ_1 reaches 1.0, the model achieves the highest AP. If not specified, in other experiments, we will set λ_1 to 1.0 by default.

λ_1	λ_2	$\mathrm{AP}_{50:95}$	AP_{50}
0	0.025	25.85	49.63
	0.05	**25.97**	**49.90**
	0.1	25.74	49.77
	0.15	25.40	48.82

Table 5: Comparison of detection accuracy on COCO val. set when varying the loss weight λ_2 of point-wise MIL loss.

Loss weight λ_2 of point-wise MIL loss. We conduct experiments to explore the effect of loss weight λ_2 of point-wise MIL loss. In these experiments, we use Hungarian-based point matching and set the loss weight λ_1 of image-wise MIL loss to 0. As shown in Table 5, when loss weight λ_2 reaches 0.05, the model achieves the highest AP. If not specified, in other experiments, we will set λ_2 to 0.05 by default.

τ	λ_1	λ_2	$\mathrm{AP}_{50:95}$	AP_{50}
0.01	1.0	0.05	26.24	**50.71**
0.05			**26.28**	50.44
0.1			26.17	50.01
0.15			26.19	49.94

Table 6: Comparison of detection accuracy on the COCO val. set when varying the score threshold τ

Score threshold τ. The score threshold τ is used to filter out low-quality pseudo boxes. We conduct experiments to explore the effect of score threshold τ. When conducting these experiments, we use Hungarian-based point matching and set the loss weights of λ_1 and λ_2 to 1.0 and 0.05 respectively. As shown in Table 6, when τ reaches 0.05, the model achieves the highest AP. If not specified, in other experiments, we set τ to 0.05 by default.

H. PM	I. MIL	P. MIL	P. CP	$AP_{50:95}$
				20.2
✓				25.2
✓	✓			25.7
✓		✓		26.0
✓	✓	✓		26.3
✓	✓	✓	✓	**27.3**

Table 7: The effect of each element proposed in this work. H. PM indicates hungarian-based point matching, I. MIL denotes image-wise MIL loss and P. MIL indicates point-wise MIL loss, P. CP indicates point-guided copy-paste augmentation.

Factor-by-factor experiment. We conduct a factor-by-factor experiment on our proposed Hungarian-based point matching, image-wise MIL loss, point-wise MIL loss, and point-guided copy-paste. As shown in Table 7, each element of our proposed Point-Teaching has a positive impact on the performance of the model. When all these elements are combined, the model reaches the highest performance, *i.e.*, 27.3 AP.

Comparison with State-of-the-art Methods

We verify our method with previous studies on COCO-standard dataset. As shown in Table 1, our method consistently surpasses all previous SSOD models (CSD, Instance Teaching, Unbiased Teacher) and WSSOD-P models (Group R-CNN and Point DETR) in all data regimes that 0.5% to 30% data are fully-labeled. The results also indicate that Point-Teaching is robust to fewer fully-label data compared to previous methods, *e.g.* Point-Teaching outperforms Point DETR (Chen et al. 2021) by 6.95 AP and Group RCNN (Zhang et al. 2022) by 3.1 AP under 5% COCO labeled data.

Extensions: Point-Teaching for Instance Segmentation

In order to demonstrate the generality of Point-Teaching, we extend our framework to weakly semi-supervised instance segmentation. In this experiment, Mask RCNN with ResNet-50 backbone is used as our detector and only fully-labeled data has box and mask annotations. As shown in Table 8, Point-Teaching significantly improves the performance in all data regimes. This result indicates that Point-Teaching can benefit from only a small amount of mask annotations. Thus, it is a promising approach to reduce the annotation cost in weakly semi-supervised instance segmentation tasks.

Method	Backbone	COCO Labeled Setting			
		1%	2%	5%	10%
Supervised	R50-FPN	10.8	14.5	18.7	22.6
Point-Teaching	R50-FPN	23.5	25.9	30.7	33.3

Table 8: Point-Teaching for weakly semi-supervised instance segmentation on COCO val. set. Results are reported with mask $AP_{50:95}$. All models are trained with a batch size of 32 (16 fully-labeled images and 16 point-labeled images) and 180k iterations.

Method	Backbone	COCO Labeled Setting
		30%
Supervised	R50-FPN	22.1
Point-Teaching	R50-FPN	28.0 (↑5.9)

Table 9: Point-Teaching for weakly-supervised instance segmentation on COCO validation set. Results are reported with mask $AP_{50:95}$.

We further extend our framework to weakly supervised instance segmentation. In this scenario, we supervise the instance segmentation training with **only box and point annotations**. Specifically, we train Mask RCNN with ResNet-50 under 30% COCO labeled setting. The whole training pipeline contains two stages. In the first stage, we use the proposed Point-Teaching framework to get a well-trained teacher model. In the second stage, we fix the teacher model with zero EMA update rate and use the proposed hungarian-based point matching method to generate pseudo-bounding boxes, and the student model is supervised with both annotated and pseudo-annotated boxes with three additional loss terms, *e.g.* point loss, project loss (Hsu et al. 2019; Tian et al. 2021) and pairwise loss (Tian et al. 2021). More details about loss functions can be found in the supplementary materials. As shown in Table 9, Point-Teaching achieves 28.0 mask AP without mask annotation, outperforming the supervised baseline by 5.9 AP.

Conclusion

In this work, we presented Point-Teaching, a novel weakly semi-supervised framework for object detection and instance segmentation. It can effectively leverage point annotation with the proposed hungarian-based point matching strategy, image-wise MIL loss, point-wise MIL loss, and point-guided copy-paste augmentation. These contributions enable our framework significantly outperforms all previous works by a large margin in all data regime settings. We hope that our work can inspire the community to design more practical object detectors with limited human annotations.

Acknowledgments

This work was supported by National Key R&D Program of China (No. 2022ZD0118700).

References

Bilen, H.; and Vedaldi, A. 2016. Weakly Supervised Deep Detection Networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Proc. Eur. Conf. Comp. Vis.*

Chen, L.; Yang, T.; Zhang, X.; Zhang, W.; and Sun, J. 2021. Points as Queries: Weakly Semi-supervised Object Detection by Points. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 8823–8832.

Cheng, B.; Parkhi, O.; and Kirillov, A. 2022. Pointly-Supervised Instance Segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Cinbis, R. G.; Verbeek, J. J.; and Schmid, C. 2014. Multi-fold MIL Training for Weakly Supervised Object Localization. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2): 303–338.

Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.-Y.; Cubuk, E. D.; Le, Q. V.; and Zoph, B. 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2918–2928.

Gupta, A.; Dollár, P.; and Girshick, R. B. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *Proc. IEEE Int. Conf. Comp. Vis.*

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Hsu, C.-C.; Hsu, K.-J.; Tsai, C.-C.; Lin, Y.-Y.; and Chuang, Y.-Y. 2019. Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior. In *Proc. Advances in Neural Inf. Process. Syst.*

Jeong, J.; Lee, S.; Kim, J.; and Kwak, N. 2019. Consistency-based Semi-supervised Learning for Object detection. In *Proc. Advances in Neural Inf. Process. Syst.*

Jeong, J.; Verma, V.; Hyun, M.; Kannala, J.; and Kwak, N. 2021. Interpolation-based semi-supervised learning for object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 11602–11611.

Kantorov, V.; Oquab, M.; Cho, M.; and Laptev, I. 2016. ContextLocNet: Context-Aware Deep Network Models for Weakly Supervised Localization. In *Proc. Eur. Conf. Comp. Vis.*

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Lin, C.; Wang, S.; Xu, D.; Lu, Y.; and Zhang, W. 2020. Object Instance Mining for Weakly Supervised Object Detection. In *Proc. AAAI Conf. Artificial Intell.*

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proc. IEEE Int. Conf. Comp. Vis.*

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Proc. Eur. Conf. Comp. Vis.*

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. E.; Fu, C.; and Berg, A. C. 2016. SSD: Single Shot MultiBox Detector. In *Proc. Eur. Conf. Comp. Vis.*

Liu, Y.; Ma, C.; He, Z.; Kuo, C.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021. Unbiased Teacher for Semi-Supervised Object Detection. In *Proc. Int. Conf. Learn. Representations*.

Papadopoulos, D. P.; Uijlings, J. R. R.; Keller, F.; and Ferrari, V. 2017a. Extreme clicking for efficient object annotation. In *Proc. IEEE Int. Conf. Comp. Vis.*

Papadopoulos, D. P.; Uijlings, J. R. R.; Keller, F.; and Ferrari, V. 2017b. Training object class detectors with click supervision. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Redmon, J.; Divvala, S. K.; Girshick, R. B.; and Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proc. Advances in Neural Inf. Process. Syst.*

Ren, Z.; Yu, Z.; Yang, X.; Liu, M.; Lee, Y. J.; Schwing, A. G.; and Kautz, J. 2020a. Instance-Aware, Context-Focused, and Memory-Efficient Weakly Supervised Object Detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Ren, Z.; Yu, Z.; Yang, X.; Liu, M.; Schwing, A. G.; and Kautz, J. 2020b. UFO^2: A Unified Framework Towards Omni-supervised Object Detection. In *Proc. Eur. Conf. Comp. Vis.*

Russakovsky, O.; Li, L.; and Li, F. 2015. Best of both worlds: Human-machine collaboration for object annotation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2121–2131. IEEE Computer Society.

Sohn, K.; Zhang, Z.; Li, C.; Zhang, H.; Lee, C.; and Pfister, T. 2020. A Simple Semi-Supervised Learning Framework for Object Detection. *arXiv: Comp. Res. Repository*, abs/2005.04757.

Song, H. O.; Lee, Y. J.; Jegelka, S.; and Darrell, T. 2014. Weakly-supervised Discovery of Visual Pattern Configurations. In *Proc. Advances in Neural Inf. Process. Syst.*

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing Annotations for Visual Object Detection. In Chen, Y.; Ipeirotis, P. G.; Law, E.; von Ahn, L.; and Zhang, H., eds., *Proc. Workshop of AAAI Conf. Artificial Intell.* AAAI Press.

Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; and Yuille, A. L. 2020. PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(1): 176–191.

Tang, P.; Wang, X.; Bai, X.; and Liu, W. 2017. Multiple Instance Detection Network with Online Instance Classifier Refinement. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. Advances in Neural Inf. Process. Syst.*

Tian, Z.; Shen, C.; and Chen, H. 2020. Conditional Convolutions for Instance Segmentation. In *Proc. Eur. Conf. Comp. Vis.*

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *Proc. IEEE Int. Conf. Comp. Vis.*

Tian, Z.; Shen, C.; Wang, X.; and Chen, H. 2021. BoxInst: High-Performance Instance Segmentation with Box Annotations. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Wang, X.; Zhang, R.; Shen, C.; Kong, T.; and Li, L. 2021. SOLO: A Simple Framework for Instance Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*

Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; and Fu, Y. 2020. Rethinking Classification and Localization for Object Detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2. Accessed: 2019-09-05.

Yang, Q.; Wei, X.; Wang, B.; Hua, X.-S.; and Zhang, L. 2021. Interactive self-training with mean teachers for semi-supervised object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 5941–5950.

Zhang, S.; Yu, Z.; Liu, L.; Wang, X.; Zhou, A.; and Chen, K. 2022. Group R-CNN for Weakly Semi-supervised Object Detection with Points. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 9417–9426.

Zhou, H.; Ge, Z.; Liu, S.; Mao, W.; Li, Z.; Yu, H.; and Sun, J. 2022. Dense teacher: Dense pseudo-labels for semi-supervised object detection. In *Proc. Eur. Conf. Comp. Vis.*, 35–50. Springer.

Zhou, Q.; Yu, C.; Shen, C.; Wang, Z.; and Li, H. 2021a. Object Detection Made Simpler by Eliminating Heuristic NMS. *arXiv: Comp. Res. Repository*, abs/2101.11782.

Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and Li, H. 2021b. Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proc. Int. Conf. Learn. Representations*.