BoxTeacher: Exploring High-Quality Pseudo Labels for Weakly Supervised Instance Segmentation

Anonymous authors
Paper under double-blind review

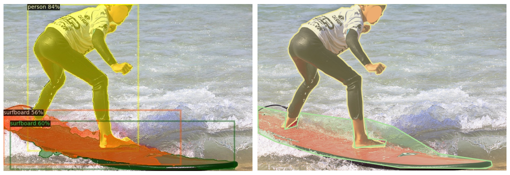
Abstract

Labeling objects with pixel-wise segmentation requires a huge amount of human labor compared to bounding boxes. Most existing methods for weakly supervised instance segmentation focus on designing heuristic losses with priors from bounding boxes. While, we find that box-supervised methods can produce some fine segmentation masks and we wonder whether the detectors could learn from these fine masks while ignoring low-quality masks. To answer this question, we present BoxTeacher, an efficient and end-to-end training framework for high-performance weakly supervised instance segmentation, which leverages a sophisticated teacher to generate high-quality masks as pseudo labels. Considering the massive noisy masks hurt the training, we present a mask-aware confidence score to estimate the quality of pseudo masks, and propose the noise-aware pixel loss and noise-reduced affinity loss to adaptively optimize the student with pseudo masks. Extensive experiments can demonstrate effectiveness of the proposed BoxTeacher. Without bells and whistles, BoxTeacher remarkably achieves 34.4 mask AP and 35.4 mask AP with ResNet-50 and ResNet-101 respectively on the challenging MS-COCO dataset, which outperforms the previous state-of-the-art methods by a significant margin. The code and models will be available later.

1 Introduction

Instance segmentation, aiming at recognizing and segmenting objects in images, is a fairly challenging task in computer vision. Fortunately, the rapid development of object detection methods (Ren et al., 2017; Tian et al., 2019; Carion et al., 2020) has greatly advanced the emergence of numbers of successful methods (He et al., 2017; Cai & Vasconcelos, 2021; Wang et al., 2020a;b; Tian et al., 2020; Bolya et al., 2019) for effective and efficient instance segmentation. With the fine-grained human annotations, recent instance segmentation methods can achieve impressive results on challenging the MS-COCO dataset (Lin et al., 2014). Nevertheless, labeling instance-level segmentation is much complicated and time-consuming, *e.g.*, labeling an object with polygon-based masks requires 10.3× more time than that with a 4-point bounding box (Cheng et al., 2022a).

Recently, a few works (Hsu et al., 2019; Lee et al., 2021; Tian et al., 2021; Wang et al., 2021; Li et al., 2022; Lan et al., 2021) explore weakly-supervised instance segmentation with bounding box annotations. These methods tend to adopt the multiple instance learning (MIL) to transform box annotations to segmentation annotations. In addition, (Tian et al., 2021; Li et al., 2022; Lan et al., 2021) further explore the affinity relations among pixels from low-level colors or features to design relation-based losses. These weakly supervised methods can effectively train instance segmentation methods (He et al., 2017; Tian et al., 2020; Wang et al., 2020b) without pixel-wise or polygon-based annotations and obtain fine segmentation masks. As shown in Fig. 1, BoxInst (Tian et al., 2021) can output a few high-quality segmentation masks and segment well on the object boundary, *e.g.*, the person, even performs better than the ground-truth mask in details though other objects may be badly segmented. Naturally, we wonder if the generated masks of box-supervised methods, especially the high-quality masks, could be qualified as pseudo segmentation labels to further improve the performance of weakly supervised instance segmentation.



BoxInst, 30.7 AP

Ground Truth

Figure 1: **Segmentation Masks from BoxInst**. BoxInst (ResNet-50 (He et al., 2016)) can produce some fine segmentation masks with weak supervisions from bounding boxes and images.

Method	AP	AP_{50}	AP_{75}
BoxInst (Tian et al., 2021)	30.7	52.5	31.2
Self-Training, 1×	31.0	53.1	31.6
Self-Training, 3×	31.3	53.8	31.7
BoxTeacher, 3×	34.2	56.0	35.4

Table 1: **Self-Training with Pseudo Labels on MS-COCO `val`.** We explore multi-stage self-training to train a CondInst (Tian et al., 2020) with the pseudo labels generated by BoxInst. However, the performance improvements are limited.

To answer this question, we first employ the naive self-training to evaluate the performance of using box-supervised pseudo masks. Given the generated instance masks from BoxInst, we propose a simple yet effective box-based mask assignment to assign pseudo masks to ground-truth boxes. And then we train the CondInst (Tian et al., 2020) with the pseudo masks and ground-truth boxes, which has the same architecture with BoxInst and consists of a fully convolutional detector (Tian et al., 2019) and a dynamic mask head. Tab. 1 shows that using self-training brings minor improvements and fails to unleash the power of high-quality pseudo masks, which can be attributed to two obstacles, *i.e*., (1) the naive self-training fails to filter low-quality masks, and (2) the noisy pseudo masks hurt the training using fully-supervised pixel-wise loss.

To address these problems, we present BoxTeacher, an end-to-end training framework for weakly supervised instance segmentation, which takes advantage of high-quality pseudo masks produced by box supervision. BoxTeacher is composed of a sophisticated *Teacher* and a *Student*, in which the teacher generates high-quality pseudo instance masks along with the mask-aware confidence scores to estimate the quality of masks. Then the proposed box-based mask assignment will assign the pseudo masks to the ground-truth boxes. The student is normally optimized with the ground-truth boxes and pseudo masks through box-based loss and noise-aware pseudo mask loss, and then progressively updates the teacher via Exponential Moving Average (EMA). In contrast to the naive multi-stage self-training, BoxTeacher is more simple and efficient. The proposed mask-aware confidence score effectively reduces the impact of low-quality masks. More importantly, pseudo labeling can mutually improve the student and further enforce the teacher to generate higher-quality masks, therefore pushing the limits of the box supervision. BoxTeacher can serve as a general training paradigm and is agnostic to the methods for instance segmentation.

To benchmark the proposed BoxTeacher, we adopt CondInst (Tian et al., 2020) as the basic segmentation method, which is a single-stage instance segmentation method and yields high-resolution masks. On the challenging MS-COCO dataset (Lin et al., 2014), BoxTeacher surprisingly achieves 34.4 mask AP and 35.4 mask AP based on ResNet-50 (He et al., 2016) and ResNet-101 respectively, which remarkably outperforms the counterparts. Furthermore, BoxTeacher with Swin Transformer (Liu et al., 2021b) obtains 40.0 mask AP as a weakly approach for instance segmentation.

2 Box-supervised Instance Segmentation via Naive Self-Training

Revisiting Box-supervised Methods. Note that *box-only* annotations is sufficient to train an object detector, which can accurately localize and recognize objects. Box-supervised methods (Tian et al., 2021; Li et al., 2022; Lan et al., 2021) based on object detectors mainly exploit two exquisite losses to supervise mask predictions, *i.e*., the multiple instance learning (MIL) loss and the pairwise relation loss. Concretely, according to the bounding boxes, the MIL loss can determine the positive and negative bags of pixels of the predicted masks. Pairwise relation loss concentrates on the local relations of pixels from low-level colors or features, in which neighboring pixels have the similar color will be regarded as a positive pair and should output similar probabilities. The MIL loss and pairwise relation loss enables the box-supervised methods to produce the complete segmentation masks, and even some high-quality masks with fine details.

Naive Self-Training with Box-supervised Methods. Recently, self-training (Fralick, 1967) has been widely used in semi-supervised (Yuan et al., 2021; Sohn et al., 2020a;b; Xu et al., 2021; Liu et al., 2021a; Chen et al., 2021; Tarvainen & Valpola, 2017), which aims to train new models on large-scale unlabeled datasets via multi-stage pseudo labeling and training. In the pseudo labeling stage, an existed model f, trained on the labeled dataset $\mathbb{X}^l$, can be applied to generate predictions on the unlabeled dataset $\mathbb{X}^u$ as the pseudo labels. And then a new model g can be trained with the pseudo-labeled $\mathbb{X}^u$ and pre-labeled $\mathbb{X}^l$.

Considering that the box-supervised methods can produce some high-quality segmentation masks without mask annotations, we propose to adopt self-training to utilize the high-quality masks as pseudo annotations to train an instance segmentation method with full supervision. Specifically, we adopt the successful BoxInst (Tian et al., 2021) to generate pseudo instance masks on the given dataset $\mathbb{X} = \{\mathcal{X}, \mathcal{B}^g\}$, which only contains the box annotations. For each input image $\mathcal{X}$, let $\{\mathcal{B}^p, \mathcal{C}^p, \mathcal{M}^p\}$ denote the predicted bounding boxes, confidence scores, and predicted instance masks, respectively. We propose a simple yet effective box-based assignment algorithm in Alg. 1 to assign the predicted instance masks to the box annotations via the confidence scores and intersection-over-union (IoU) between ground-truth boxes $\mathcal{B}^g$ and predicted boxes $\mathcal{B}^p$. The hyper-parameters τ_{iou} and τ_c are set to 0.5 and 0.05, respectively. The assigned instance masks will be rectified by removing the parts beyond the bounding boxes. Then, we adopt the dataset $\hat{\mathbb{X}} = \{\mathcal{X}, \mathcal{B}^g, \mathcal{M}^g\}$ with pseudo instance masks to train an approach, *e.g.*, CondInst (Tian et al., 2020).

Naive Self-Training is Limited. Tab. 1 and Tab. 6 provide the experimental results of using naive self-training pseudo masks. Compared to the pseudo labeler, using self-training brings minor improvements and even fails to surpass the pseudo labeler. We attribute the limited performance to two issues, *i.e.*, the naive self-training fails to exclude low-quality masks and the fully-supervised loss is sensitive to the noisy pseudo masks.

Algorithm 1: Labeling pseudo instance masks with ground-truth bounding boxes

```
Input: ground-truth boxes $\mathcal{B}^g \in \mathbb{R}^{K\times 4}$, predicted boxes $\mathcal{B}^p \in \mathbb{R}^{N\times 4}$, predicted instance masks
        $\mathcal{M}^p \in \mathbb{R}^{N\times H\times W}$, confidence score $\mathcal{C}^p \in \mathbb{R}^{N\times 1}$.
Parameter: IoU threshold $\tau_{\text{iou}}$, confidence threshold $\tau_c$.
Output: assigned instance masks $\mathcal{M}^g \in \mathbb{R}^{K\times H\times W}$.
Initialize output masks $\mathcal{M}^g$ with empty (0), assignment index $A \in \mathbb{R}^K$ with $-1$;
Filter the predictions by the confidence threshold $\tau_c$;
Sort the confidence score $\mathcal{C}^p$ in descending order with output indices $S \in \mathbb{N}^N$;
foreach prediction $i$ in $S$ do
    Initialize matched IoU: $u \leftarrow -1$, matched index: $v \leftarrow -1$;
    for $j = 1$ to $K$ do
        $iou_{ij} = \texttt{ComputeIoU}(\mathcal{B}^p_i, \mathcal{B}^g_j)$;
        if $A_j > 0$ then
            continue;
        end
        if $iou_{ij} \geq \tau_{iou}$ and $iou_{ij} \geq u$ then
            $u \leftarrow \text{iou}_{ij}$, $v \leftarrow i$;
        end
        if $v > 0$ then
            Assign predicted mask $\mathcal{M}^p_i$ to output mask $\mathcal{M}^g_v$ ;
            $A_j \leftarrow i$;
        end
    end
end
```

3 BoxTeacher

In this section, we present BoxTeacher, an end-to-end training framework, which aims to unleash the power of high-quality of pseudo masks. In contrast to multi-stage self-training, BoxTeacher, consisting of a teacher and a student, simultaneously facilitates the training of the student and pseudo labeling of the teacher. The mutual optimization is beneficial to both the teacher and the student, thus leading to higher performance for box-supervised instance segmentation.

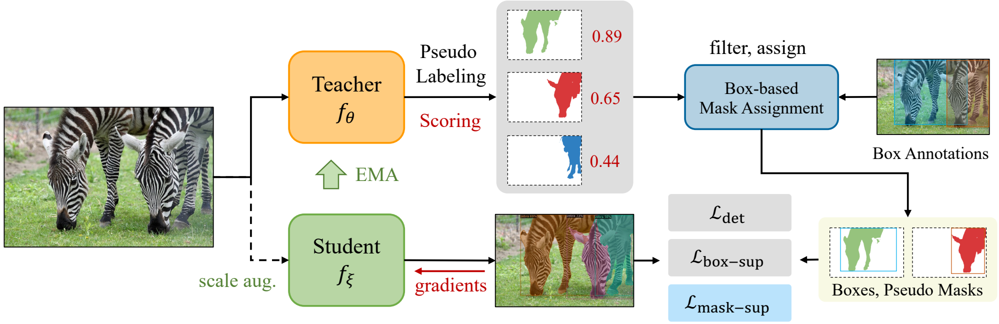


Figure 2: **The Architecture of BoxTeacher.** Images are firstly fed into the *Teacher* to obtain the pseudo masks and estimate the quality of masks. Then the box-based mask assignment filters and assigns pseudo masks to box annotations. The *Student* adopt the scale-augmented images (*i.e.*, multi-scale training) and pseudo masks to update the parameters by gradient descent and then update the Teacher with exponential moving average (EMA).

3.1 Architecture

The overall architecture of BoxTeacher is depicted in Fig. 2. BoxTeacher is composed of a teacher and a student, which shares the same model. Given the input image, the teacher f_θ straightforwardly generate the predictions, including the bounding boxes, segmentation masks, and *mask-aware confidence scores*. Similarly, we apply the Alg. 1 to assign the predicted masks to the ground-truth annotations and the confidence score in Alg. 1 is substituted with the *mask-aware confidence score*. The augmented images (*i.e.*, scale augmentation) are fed into the student f_ξ and the student is optimized under the box supervision and the mask supervision. To acquire high-quality pseudo masks, we adopt the exponential moving average to gradually update the teacher from student (Tarvainen & Valpola, 2017), *i.e.*, $f_\theta \leftarrow \alpha \cdot f_\theta + (1-\alpha) \cdot f_\xi$ (α is empirically set 0.999).

Mask-aware Confidence Score. Considering that the generated pseudo masks are noisy and unreliable, which may hurt the performance, we define the mask-aware confidence score to estimate the quality of the pseudo masks. Inspired by (Wang et al., 2020a), we denote $m_i^b \in \mathbb{R}^{H\times W}$ and $m_i \in \mathbb{R}^{H\times W}$ as the box-based binary masks and sigmoid probabilities of the i-th pseudo mask with the detection confidence c_i, the mask-aware confidence score s_i is defined as follows:

$$s_i = \sqrt{c_i \cdot \frac{\sum_{x,y}^{H,W} \mathbb{1}(m_{i,x,y} > \tau_m) \cdot m_{i,x,y} \cdot m_{i,x,y}^b}{\sum_{x,y}^{H,W} \mathbb{1}(m_{i,x,y} > \tau_m) \cdot m_{i,x,y}^b}}, \tag{1}$$

where $\mathbb{1}(\cdot)$ is the indicator function, τ_m is the threshold for binary masks and set to 0.5. The mask-aware score calculates the average probability score of the positive masks inside the ground-truth boxes, and the higher average probability means more confident pixels in the mask. In addition, we explore several kinds of quality scores and compare with the mask-aware score in experiments.

Training Loss. BoxTeacher can be end-to-end optimized with box annotations and the generated pseudo masks. The overall loss is defined as: $\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{box-sup}} + \mathcal{L}_{\text{mask-sup}}$, which consists of the standard detection loss $\mathcal{L}_{\text{det}}$, the box-supervised loss $\mathcal{L}_{\text{box-sup}}$, and the mask-supervised loss $\mathcal{L}_{\text{mask-sup}}$. We inherit the detection loss defined in FCOS (Tian et al., 2020). For box-supervised mask loss $\mathcal{L}_{\text{box-sup}}$, we follow previous works (Tian et al., 2021; Hsu et al., 2019; Lan et al., 2021) and adopt the max-projection loss and the color-based pairwise relation loss (Tian et al., 2021).

3.2 Noise-aware Pseudo Mask Loss

The goal of BoxTeacher is to take advantage of high-quality pseudo masks in a fully supervised manner while reduce the impact of the noisy or low-quality instance masks. To this end, we present the noise-aware pseudo mask loss in Eq. 2. Ideally, BoxTeacher can leverage the pseudo masks to calculate the fully-supervised pixel-wise segmentation loss, *e.g.*, dice loss (Milletari et al., 2016). Besides, we also propose a novel *noise-reduced mask affinity loss* $\mathcal{L}_{\text{affinity}}$ to enhance the pixel-wise segmentation with neighboring pixels. Further, we employ the proposed mask-aware confidence scores $\{s_i\}$ as weights for the pseudo mask loss, which adaptively scales the weights for pseudo masks of different qualities. The total pseudo mask loss is defined as follows:

$$\mathcal{L}_{\text{mask-sup}} = \frac{1}{N_p} \sum_{i=1}^{N_p} s_i \cdot (\lambda_{\text{pixel}} \mathcal{L}_{\text{pixel}}(m_i^p, m_i^g) + \lambda_{\text{affinity}} \mathcal{L}_{\text{affinity}}(m_i^p, m_i^g)), \tag{2}$$

where m_i^p and m_i^g denotes the i-th predicted masks and pseudo masks, N_p denotes the number of valid pseudo masks, λ_{pixel} and $\lambda_{\text{affinity}}$ are set to 0.5 and 0.1 respectively.

Noise-reduced Mask Affinity Loss. Considering that pixels tend to have similar labels with neighboring pixels, we exploit the label affinities among pixels within local regions to lower the impact of noisy pixels. Given the i-th pixel sigmoid probability g_i of the pseudo segmentation, we first calculate the refined pixel probability $\tilde{g_i}$ with its neighboring pixels, which is defined as follows:

$$\tilde{g_i} = \frac{1}{2}(g_i + \frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} g_j), \tag{3}$$

where $\mathcal{P}$ denotes the set of neighboring pixels, *e.g.*, a 3×3 region. This simple refinement can reduce the outliers and enhance the pixels with local context. Inspired by recent works (Ahn & Kwak, 2018; Ru et al., 2022) which explore pixel-wise affinity for weakly semantic segmentation for noisy labels, we present a simple noise-reduced mask affinity loss and define the affinity μ_{ij} between i-th and j-th pixels as follows:

$$\mu_{ij} = \tilde{g_i} \cdot \tilde{g_j} + (1 - \tilde{g_i}) \cdot (1 - \tilde{g_j}), \tag{4}$$

where $\tilde{g_i}$ and $\tilde{g_j}$ are refined pixels which encode the local context. Then the *noise-reduced mask affinity loss* for i-th pixel is defined as follows:

$$\mathcal{L}_{\text{affinity}} = -\frac{\sum_{j \in \mathcal{P}} \mathbb{1}(\mu_{ij} > \tau_a)(\log(p_i \cdot p_j) + \log((1 - p_i) \cdot (1 - p_j)))}{\sum_{j \in \mathcal{P}} \mathbb{1}(\mu_{ij} > \tau_a)}, \tag{5}$$

where $j \in \mathcal{P}$ are the neighboring pixels of the i-th pixel and τ_a is set to 0.5 as default.

4 Experiments

In this section, we mainly evaluate the proposed BoxTeacher on the challenging MS-COCO dataset (Lin et al., 2014) and the Cityscapes dataset (Cordts et al., 2016), and provide extensive ablations to analyze the proposed BoxTeacher. We also refer the readers to the Appendix for additional ablations and visualizations.

Datasets. The COCO dataset contains 80 categories and $110k$ images for training, $5k$ images for validation, and $20k$ images for testing. The Cityscapes dataset, aiming for perception in driving scenes, consists of 5000 street-view high-resolution images, in which 2975, 500, and 1525 images are used for training, validation, and testing, respectively. Foreground objects in Cityscapes are categorized into 8 classes and fine-annotated with pixel-wise segmentation labels instead of polygons adopted in COCO, thus making the labeling process much costly. For weakly supervised instance segmentation, we only keep the bounding boxes and ignore the segmentation masks during training.

Implementation Details The proposed BoxTeacher is implemented based on PyTorch (Paszke et al., 2019) and the Detectron2 toolbox (Wu et al., 2019). We mainly adopt CondInst (Tian et al., 2020) as the meta method for instance segmentation. All backbone networks are initialized with the ImageNet-pretrained weights and the BatchNorm layers are frozen.

4.1 Experiments on MS-COCO Instance Segmentation

Experimental Setup. Following the training recipes (Tian et al., 2019; 2020; 2021), BoxTeacher is trained over 8 GPUs with 16 images per batch. Unless specified, we adopt the standard $1\times$ schedule ($90k$ iterations) (He et al., 2017; Wu et al., 2019) with the SGD and the initial learning rate 0.01. For comparisons with the state-of-art methods, we scale up the learning schedule to $3\times$ ($270k$ iterations). For images input to the student, we adopt the multi-scale augmentation which randomly resizes images from 640 to 800. While the images fed into the teacher are fixed to 800×1333.

Table 2: **COCO Instance Segmentation.** Comparisons with state-of-the-art methods on COCO `test-dev`. With the same backbone or learning schedule, BoxTeacher surprisingly surpasses the counterparts by a large margin (more than 2.0 mask AP).

Method	Backbone	Schedule	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Mask-supervised methods.								
Mask R-CNN (He et al., 2017)	R-50-FPN	1×	35.5	57.0	37.8	19.5	37.6	46.0
CondInst (Tian et al., 2020)	R-50-FPN	1×	35.9	57.0	38.2	19.0	38.6	46.7
CondInst (Tian et al., 2020)	R-101-FPN	3×	39.1	60.9	42.0	21.5	41.7	50.9
SOLO (Wang et al., 2020a)	R-101-FPN	6×	37.8	59.5	40.4	16.4	40.6	54.2
SOLOv2 (Wang et al., 2020a)	R-101-FPN	6×	39.7	60.7	42.9	17.3	42.9	57.4
Box-supervised methods.								
BBTP (Hsu et al., 2019)	R-101-FPN	1×	21.1	45.5	17.2	11.2	22.0	29.8
BBAM (Lee et al., 2021)	R-101-FPN	1×	25.7	50.0	23.3	-	-	-
BoxCaseg Wang et al. (2021)	R-101-FPN	1×	30.9	54.3	30.8	12.1	32.8	46.3
BoxInst (Tian et al., 2021)	R-50-FPN	3×	32.1	55.1	32.4	15.6	34.3	43.5
BoxInst (Tian et al., 2021)	R-101-FPN	3×	33.2	56.5	33.6	16.2	35.3	45.1
BoxLevelSet (Li et al., 2022)	R-101-FPN	3×	33.4	56.8	34.1	15.2	36.8	46.8
BoxLevelSet (Li et al., 2022)	R-101-DCN-FPN	3×	35.4	59.1	36.7	16.8	38.5	51.3
DiscoBox (Lan et al., 2021)	R-50-FPN	3×	32.0	53.6	32.6	11.7	33.7	48.4
DiscoBox (Lan et al., 2021)	R-101-DCN-FPN	3×	35.8	59.8	36.4	16.9	38.7	52.1
DiscoBox (Lan et al., 2021)	X-101-DCN-FPN	3×	37.9	61.4	40.0	18.0	41.1	53.9
BoxTeacher	R-50-FPN	1×	32.9	54.1	34.2	17.4	36.3	43.7
BoxTeacher	R-50-FPN	3×	34.4	56.5	35.9	18.8	37.5	45.0
BoxTeacher	R-101-FPN	3×	35.4	57.8	37.2	19.5	39.1	46.3
BoxTeacher	Swin-Base-FPN	3×	40.0	64.3	41.9	23.1	43.8	53.4

Main Results. Tab. 2 shows the main results on MS-COCO `test-dev`. In comparison with other state-of-the art methods, we evaluate the proposed BoxTeacher with different backbone networks, *i.e.*, ResNet (He et al., 2016) and Swin Transformer (Liu et al., 2021b), and under different training schedules, *i.e.*, $1\times$ and $3\times$. It's clear that BoxTeacher with ResNet-50 achieves 32.9 mask AP, which outperforms other box-supervised methods even with longer schedules. Compared to recent box-supervised methods (Tian et al., 2021; Li et al., 2022; Lan et al., 2021), BoxTeacher brings about significant 2.0 mask AP improvements on different backbones under the same setting. With the stronger backbones, *e.g.*, Swin Transformer (Liu et al., 2021b), BoxTeacher can surprisingly obtain 40.0 mask AP on the challenging MS-COCO, which is highly competitive as a weakly supervised method for instance segmentation.

4.2 Ablation Experiments

Effects of Pseudo Mask Loss. In Tab. 3, we explore the different pseudo mask loss for BoxTeacher. Firstly, we apply the box-supervised loss proposed in (Tian et al., 2021) achieves 30.7 mask AP (the gray row). As shown in Tab. 3, directly applying binary cross entropy (bce) loss with pseudo masks leads to severe performance degradation, which can be attributed to the foreground/background imbalance and noise in pseudo masks. Using dice loss to supervise the training with pseudo masks can bring significant improvements in comparison to the baseline. In addition, we adopt the weakly average projection loss proposed in (Wang et al., 2022), which aims for coarse pseudo instance masks. Tab. 3 shows that average projection loss $\mathcal{L}_{\text{avg}}$ is inferior to the fully-supervised dice loss. Adding mask affinity loss $\mathcal{L}_{\text{affinity}}$ provides 0.4 mask AP gain based on the dice loss. Moreover, we ablate the loss weights in pseudo mask loss in Tab. 4.

Effects of Mask-aware Confidence Score. Tab. 5 explores several different scores to estimate the quality of pseudo masks in an unsupervised manner, *i.e.*, (1) classification scores (cls), (2) matched IoU between predicted boxes and ground-truth boxes (iou), (3) mean entropy of the pixel probabilities of pseudo masks (mean-entropy: $s = 1 + \frac{1}{HW}\sum_{i,j}^{H,W}(p_{ij}\log p_{ij} + (1-p_{ij})\log(1-p_{i,j}))$), (4) the proposed mask-aware score (mask-aware). As Tab. 5 shows, using the proposed mask-aware confidence score leads to better performance for BoxTeacher. Notably, measuring the quality of predicted masks is critical but challenging for leverage pseudo masks. Fig. 3 shows the mask scores

Table 3: **Pseudo Mask Loss.** We evaluate the effects of different loss for BoxTeacher. $\mathcal{L}_{\text{avg}}$ is average projection loss proposed in (Wang et al., 2022).

$\mathcal{L}_{\text{pixel}}$	$\mathcal{L}_{\text{affinity}}$	$\mathcal{L}_{\text{avg}}$	AP	AP_{50}	AP_{75}
✗	-	-	30.7	52.5	31.2
bce	-	-	28.9	49.2	29.5
dice	-	-	31.8	53.1	32.8
✗	-	✓	31.4	52.9	32.2
dice	✓	-	32.2	53.5	33.2

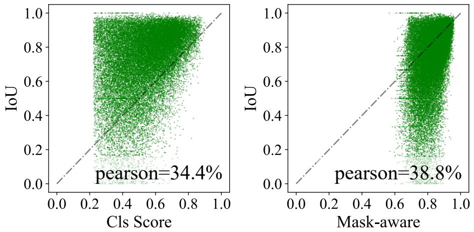
Table 4: **Effect of the Weights of Pseudo Mask Loss.** We adopt $\lambda_{\text{pixel}} = 0.5$ and $\lambda_{\text{affinity}} = 0.1$ as the default setting.

λ_{pixel}	$\lambda_{\text{affinity}}$	AP	AP_{50}	AP_{75}
0.1	-	31.4	53.0	32.4
0.5	-	31.8	53.1	32.8
1.0	-	31.5	52.8	32.3
0.5	0.1	32.2	53.5	33.2
0.5	0.5	31.7	52.8	32.8

Table 5: **The Effects of Mask-aware Confidence Score.** We evaluate different mask scores for BoxTeacher, and it shows that the proposed mask-aware confidence performs better.

Mask Score	AP	AP_{50}	AP_{75}
✗	32.2	53.5	33.2
cls	32.0	53.5	33.1
iou	32.2	53.5	33.4
mean-entropy	31.8	53.3	32.6
mask-aware	32.6	53.5	33.8

Figure 3: **Visualizations of the Mask Scores *v.s.* Mask IoU.** We adopt the COCO `val` images to compare the mask score with the IoU between pseudo masks and GT masks.



compared to the IoU between pseudo masks and ground-truth masks, mask-aware confidence score has a higher correlation with the practical mask quality. Accurate quality estimation can effectively reduce the impact of noisy masks and stabilize the training.

Comparisons with Self-Training Paradigm. We adopt the box-supervised approach, *i.e.*, BoxInst (Tian et al., 2021), to generate pseudo masks, which is pre-trained on COCO with *box-only* annotations. And then we assign the pseudo masks to the ground-truth boxes through the assignment Alg. 1. As shown in Tab. 6, the improvements provided by self-training are much limited and the naive self-training even performs worse than the training with *box-only* annotations, *e.g.*, CondInst with R-50 and 3× schedule obtains 31.3 AP with pseudo masks, but inferior to the box-supervised version (31.8 AP). Though the self-training scheme enables the supervised training with pseudo masks and achieves comparable performance, we believe the high-quality pseudo masks are not well exploited. Significantly, BoxTeacher achieves higher mask AP compared to both self-training, in an end-to-end manner without complicated steps or procedures for label generation.

Table 6: **Comparison with Naive Self-Training.** As discussed in Sec. 2, we leverage the pre-trained BoxInst to generate pseudo mask labels and assign the pseudo masks to the ground-truth boxes. Then we adopt the pseudo masks and train the CondInst with different schedules and backbones. †: the mask AP achieved by the pseudo labeler, *i.e.*, BoxInst, with *box-only* annotations. ‡: the ideal mask AP could be achieved by CondInst if trained with box annotations following BoxInst.

Method	Backbone	Schedule	Pseudo Label	$\text{AP}^{\dagger}$	$\text{AP}^{\ddagger}$	AP	AP_{50}	AP_{75}
CondInst	R-50	1×	BoxInst, R-50	30.7	30.7	31.0	53.1	31.6
CondInst	R-50	3×	BoxInst, R-50	30.7	31.8	31.3	53.8	31.7
CondInst	R-50	3×	BoxInst, R-101	33.0	31.8	32.5	54.9	33.2
CondInst	R-101	3×	BoxInst, R-101	33.0	33.0	32.9	55.4	33.7
BoxTeacher	R-50	1×	End-to-End	-	-	32.6	53.5	33.8
BoxTeacher	R-50	3×	End-to-End	-	-	34.2	56.0	35.4
BoxTeacher	R-101	3×	End-to-End	-	-	35.2	57.1	36.8

Effects of Exponential Moving Average. To see whether EMA could partially bring some performance improvements, we re-train BoxInst with EMA to obtain the averaged model to evaluate the

Table 7: **Ablations on Exponential Moving Average.** We evaluate the performance of BoxInst *w/* or *w/o* EMA to make it clear whether the improvements are brought by EMA in BoxTeacher.

Method	w/ EMA	AP^{bbox}	AP	AP_{50}	AP_{75}
BoxInst	✗	39.3	30.6	52.2	31.0
BoxInst	✓	39.4	30.7	52.5	31.2

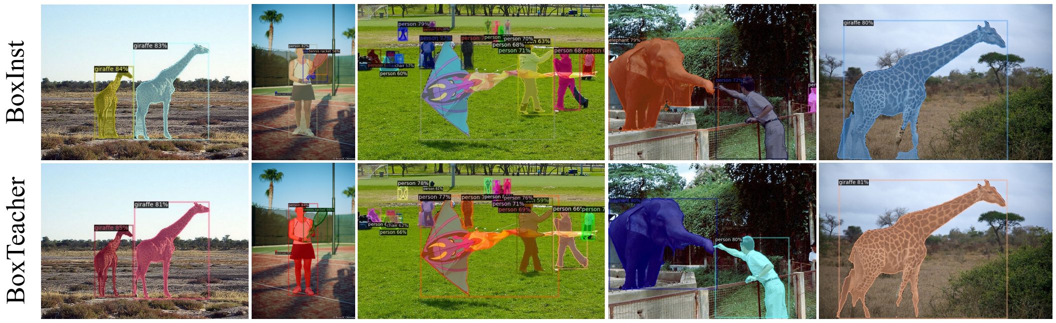


Figure 4: **Qualitative Comparisons.** We compare the qualitative results with BoxInst on the COCO `val`. Compared with BoxInst, BoxTeacher outputs segmentation masks with better boundaries and less false-positive background regions.

performance. Tab. 7 shows that applying EMA has little impact to the final performance, proving that the improvements of BoxTeacher are mainly brought by the effects of pseudo masks.

Qualitative Comparisons. We compare the qualitative results between BoxTeacher and BoxInst in Fig. 4. BoxInst relies on the color-based pairwise relation loss to separate the objects apart from other objects or the background. However, it might lead to some mistakes, *e.g.*, the foreground object has similar color with the background and hollow areas are neglected due to the large stride when pairwise relations are built. Fig. 4 shows that BoxTeacher can alleviate those problems based on the refined pixel affinity.

4.3 Experiments on Cityscapes Instance Segmentation

Experimental Setup. Following previous methods (He et al., 2017; Tian et al., 2020), we train all models for $24k$ iterations with 8 images per batch. The initial learning rate is 0.005. Cityscapes contains high-resolution images (2048×1024), and we randomly resize images from 800 to 1024 for the student and keep the original size for the teacher during training. In addition, we also adopt the COCO pre-trained models ($1\times$ schedule) to initialize the weights for higher performance.

Main Results. Tab. 8 shows the evaluation results on Cityscapes `val`. The proposed BoxTeacher outperforms the box-supervised methods significantly, especially with the COCO pre-trained weights. Though performance gap between fully supervised methods and weakly supervised methods become larger than that in MS-COCO, the human labour of labeling pixel-wise segmentation for a high-resolution Cityscapes image is much costly (90 minutes per image). And we hope future research can bridge the gap between box-supervised methods and mask-supervised methods for high-resolution images.

5 Related Work

Instance Segmentation. Methods for fully supervised instance segmentation can be roughly divided into two groups, *i.e.*, single-stage methods and two-stage methods. Single-stage methods (Bolya et al., 2019; Tian et al., 2020; Xie et al., 2020; Zhang et al., 2020) tend to adopt single-stage object detectors, *e.g.*, FCOS (Tian et al., 2019), to localize and recognize objects, and then generate segmentation masks through object enmbeddings or dynamic convolution (Chen et al., 2020). Wang *et al.* present box-free SOLO (Wang et al., 2020a) and SOLOv2 (Wang et al., 2020b),

Table 8: **Cityscapes Instance Segmentation.** Comparisons with state-of-the-art methods for mask AP on Cityscapes `val`. †: our re-produced results on Cityscapes.'`fine`' denotes the Cityscapes `train` with fine annotations while '`fine`+COCO' denotes using COCO pre-trained weights. For box-supervised methods, we remove the fine-grained mask annotations in Cityscapes.

Method	Backbone	Data	AP	AP_{50}
Mask-supervised methods.				
Mask R-CNN (He et al., 2017)	R-50-FPN	`fine`	31.5	-
CondInst (Tian et al., 2020)	R-50-FPN	`fine`	33.0	59.3
CondInst (Tian et al., 2020)	R-50-FPN	`fine` + COCO	37.8	63.4
Box-supervised methods.				
BoxInst† (Tian et al., 2021)	R-50-FPN	`fine`	19.0	41.8
BoxInst† (Tian et al., 2021)	R-50-FPN	`fine` + COCO	24.0	51.0
BoxLevelSet† (Li et al., 2022)	R-50-FPN	`fine`	20.7	43.3
BoxLevelSet† (Li et al., 2022)	R-50-FPN	`fine` + COCO	22.7	46.6
BoxTeacher	R-50-FPN	`fine`	21.7	47.5
BoxTeacher	R-50-FPN	`fine` + COCO	26.8	54.2

which are independent of object detectors. SparseInst (Cheng et al., 2022b) and YOLACT (Bolya et al., 2019), aiming for real-time inference, achieve great trade-off between speed and accuracy. Two-stage methods (He et al., 2017; Huang et al., 2019; Kirillov et al., 2020; Cheng et al., 2020) adopt bounding boxes from object detectors and RoIAlign (He et al., 2017) to extract the RoI (region-of-interest) features for object segmentation, *e.g*., Mask R-CNN (He et al., 2017). Several methods (Huang et al., 2019; Cheng et al., 2020; Kirillov et al., 2020) based on Mask R-CNN are proposed to refine the segmentation masks for high-quality instance segmentation. Recently, many approaches (Carion et al., 2020; Fang et al., 2021; Cheng et al., 2021b;a; Dong et al., 2021; Zhang et al., 2021) based on transformers (Vaswani et al., 2017; Dosovitskiy et al., 2021) or the Hungarian algorithm (Stewart et al., 2016) have made great progress in instance segmentation.

Weakly Supervised Instance Segmentation. Considering the huge cost of labeling instance segmentation, weakly supervised instance segmentation using image-level labels or bounding boxes gets lots of attention. Several methods (Zhou et al., 2018; Zhu et al., 2019; Ahn et al., 2019; Arun et al., 2020) exploit image-level labels to generate pseudo masks from activation maps. Khoreva et.al. (Khoreva et al., 2017) propose to generate pseudo masks with GrabCut (Rother et al., 2004) from given bounding boxes. BoxCaseg (Wang et al., 2021) leverages a saliency model to generate pseudo object masks for training Mask R-CNN along with the multiple instance learning (MIL) loss. Recently, many box-supervised methods (Hsu et al., 2019; Tian et al., 2021; Lan et al., 2021; Li et al., 2022) combines the MIL loss or pairwise relation loss from low-level features obtain impressing results with box annotations.

6 Conclusions

In this paper, we explore the naive self-training with pseudo labeling for box-supervised instance segmentation, which is much limited by the noisy pseudo masks. To address this issue, we present an effective training framework, namely BoxTeacher, which contains a collaborative teacher and student for mutually generating high-quality masks and training with pseudo masks. We adopt mask-aware confidence scores to measure the quality of pseudo masks and noise-aware mask loss to train the student with pseudo masks. In the experiments, BoxTeacher achieves promising improvements on both COCO and Cityscapes datasets, indicating that the proposed training framework is effective and can achieve higher level of weakly supervised instance segmentation.

References

Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018.

Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019.

Aditya Arun, C. V. Jawahar, and M. Pawan Kumar. Weakly supervised instance segmentation by learning annotation consistent instances. In *ECCV*, 2020.

Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: real-time instance segmentation. In *ICCV*, 2019.

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1483–1498, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021.

Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *CVPR*, 2020.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2021a.

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021b.

Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *CVPR*, 2022a.

Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving Mask R-CNN. In *ECCV*, 2020.

Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, and Wenyu Liu. Sparse instance activation for real-time instance segmentation. In *CVPR*, 2022b.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. SOLQ: segmenting objects by learning queries. In *NeurIPS*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, 2021.

Stanley C. Fralick. Learning to recognize patterns without a teacher. *IEEE Trans. Inf. Theory*, 1967.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, 2017.

Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In *NeurIPS*, pp. 6582–6593, 2019.

Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring R-CNN. In *CVPR*, 2019.

Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020.

Shiyi Lan, Zhiding Yu, Christopher B. Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S. Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *ICCV*, 2021.

Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. BBAM: bounding box attribution map for weakly supervised semantic and instance segmentation. In *CVPR*, 2021.

Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xiansheng Hua, and Lei Zhang. Box-supervised instance segmentation with level set evolution. In *ECCV*, 2022.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

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 *ICLR*, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021b.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

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.*, 2017.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. ”grabcut”: interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 2004.

Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *CVPR*, 2022.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020a.

Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *CoRR*, abs/2005.04757, 2020b. URL https://arxiv.org/abs/2005.04757.

Russell Stewart, Mykhaylo Andriluka, and Andrew Y. Ng. End-to-end people detection in crowded scenes. In *CVPR*, 2016.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *ICCV*, 2019.

Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020.

Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *CVPR*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Xinggang Wang, Jiapei Feng, Bin Hu, Qi Ding, Longjin Ran, Xiaoxin Chen, and Wenyu Liu. Weakly-supervised instance segmentation via class-agnostic learning with salient images. In *CVPR*, 2021.

Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: segmenting objects by locations. In *ECCV*, 2020a.

Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. In *NeurIPS*, 2020b.

Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M. Alvarez. Freesolo: Learning to segment objects without annotations. In *CVPR*, 2022.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *CVPR*, 2020.

Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *ICCV*, 2021.

Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic segmentation with strong data augmentation*. In *ICCV*, 2021.

Rufeng Zhang, Zhi Tian, Chunhua Shen, Mingyu You, and Youliang Yan. Mask encoding for single shot instance segmentation. In *CVPR*, 2020.

Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In *NeurIPS*, 2021.

Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *CVPR*, 2018.

Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David S. Doermann, and Jianbin Jiao. Learning instance activation maps for weakly supervised instance segmentation. In *CVPR*, 2019.

A Appendix

A.1 Ablation Studies

Does data augmentation help? Consistency regularization has successfully boosted semi-supervised methods (Xu et al., 2021; Sohn et al., 2020b; Chen et al., 2021; Yuan et al., 2021) and act as an indispensable component in recent semi-supervised tasks. In consistency-based self-training, the teacher generates pseudo labels with weak/no perturbation while the student adopts strong perturbation. And the consistency regularization enforces the outputs of the student to be consistent with the those of the teacher, thus facilitating the student to be more robust and invariant to the augmentations and enhance the feature representation learning.

In this study, we explore the effect of data augmentation on the proposed BoxTeacher, and apply augmentation to the input images of the student. Specifically, we defined two levels of simple data augmentation in Tab. 9. Tab. 10 provides the evaluation results of using strong/weak data augmentation. As Tab. 10 shows, both strong and weak augmentation hurt the performance of CondInst and BoxTeacher under the $1\times$ schedule. Differently, BoxTeacher is more robust to the augmentations as the AP drops 0.4 compared to CondInst. However, BoxTeacher remarkably benefits more from the strong data augmentation when increasing the schedule from $1\times$ to $3\times$. In comparison to CondInst, BoxTeacher with strong augmentation will enforce the consistency between the student and teacher. Interestingly, Tab. 10 indicates that using strong augmentation is merely beneficial to the weakly supervised instance segmentation ($+0.6$ AP), but has no effect to the fully supervised object detection ($+0.1$AP), suggesting that consistency regularization might facilitate the learning from noisy pseudo masks.

Table 9: **Specifications of Data Augmentation.**

Type	Probability	Parameters
	Weak Augmentation	
Color Jittering	0.2	(brightness, contrast, saturation, hue) = (0.2, 0.2, 0.2, 0.1)
	Strong Augmentation	
Color Jittering	0.8	(brightness, contrast, saturation, hue) = (0.4, 0.4, 0.4, 0.1)
Grayscale	0.2	-

Table 10: **The Effects of Data Augmentation.** We explore whether strong data augmentation will be beneficial to BoxTeacher, which has been widely exploited in semi-supervised methods. We apply weak and strong augmentation to both CondInst and the proposed BoxTeacher.

Method	Schedule	weak aug.	strong aug.	AP^{bbox}	AP	AP_{50}	AP_{75}
CondInst	1×			39.6	36.2	56.0	38.8
CondInst	1×	✓		39.6	35.6	56.3	38.0
CondInst	1×		✓	39.4	35.1	55.8	37.3
CondInst	3×			41.9	37.5	58.5	40.1
CondInst	3×		✓	42.0	**37.6**	58.7	40.0
BoxTeacher	1×			39.4	32.6	53.5	33.8
BoxTeacher	1×	✓		39.1	32.4	53.0	33.7
BoxTeacher	1×		✓	38.6	32.2	52.7	33.7
BoxTeacher	3×			41.7	34.2	56.0	35.4
BoxTeacher	3×		✓	41.8	**34.8**	56.2	36.3

A.2 Qualitative Results

Fig. 5 provides visualization results of the proposed BoxTeacher on the COCO *test-dev*. Even with *box-only* annotations, BoxTeacher can output high-quality segmentation masks with fine boundaries.

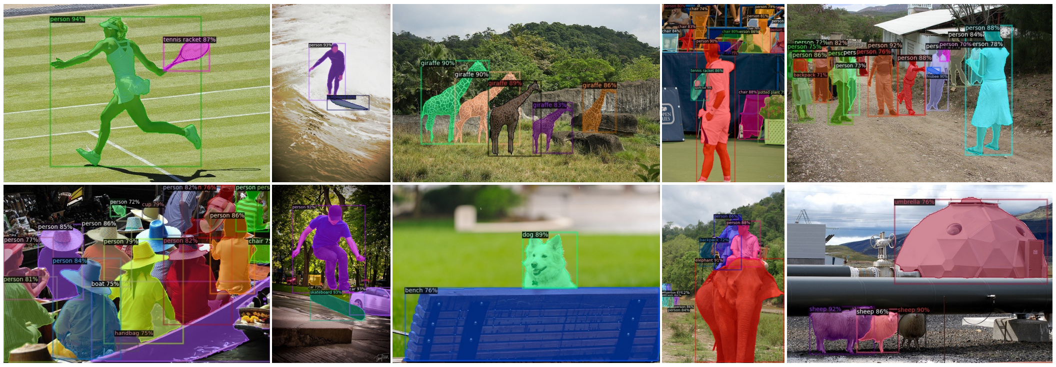


Figure 5: **Visualization Results.** We provide the visualization results of BoxTeacher with ResNet-101 on the COCO `test-dev`. The proposed BoxTeacher can produce the high-quality segmentation results, even in some complicated scenes.