

A APPENDIX

In this supplementary material, we provide additional experimental quantitative results, model size comparison, as well as bounding boxes visualization to further support the effectiveness of our proposed `Consistent-Teacher`. In addition, we delineate more experimental details, implementation information, and hyper-parameter settings of our method. Our code is also attached for your reference.

A MORE DETAILS IN `CONSISTENT-TEACHER`

A.1 INCONSISTENCY MEASUREMENT.

Inconsistency refers to the fact that the pseudo boxes may be highly inaccurate and vary greatly at different stages of training. Therefore, we measure the pseudo-bboxes variation across different training steps. Specifically, we store the checkpoints every 4000 training steps. We then run inference using these checkpoints on a subset with 5000 images from the unlabeled set. The prediction output from the previous checkpoint is then set as GT and we evaluate the mAP of the current checkpoint with the previous predictions. Therefore, a higher mAP implies a more consistent pseudo targets. Then the inconsistency is measured by accumulating $1 - mAP$ for these checkpoints to reflect the accumulated effect of noisy targets.

B VERIFY THE INCONSISTENCY IN SSOD

Assignment Inconsistency under Noisy Pseudo Labels. To illustrate that the conventional IOU-based or heuristic label assignment is problematic in SSOD, we intentionally inject random noise to the ground-truth bounding boxes and testify the assignment consistency by quantifying the assignment IOU (A-IOU) of clean and noisy assignments. Suppose a bounding box $b = (x_1, y_1, x_2, y_2)$ is assigned to a set of k anchors $A = \{a_1, \dots, a_k\}$. We add Gaussian noise to its coordinate with a noise ratio ρ , so that $b' = (x_1 + \epsilon_{x_1} \times w, y_1 + \epsilon_{y_1} \times h, x_2 + \epsilon_{x_2} \times w, y_2 + \epsilon_{y_2} \times h)$, in which w and h are width and height of the box. $\epsilon_{x_1}, \epsilon_{y_1}, \epsilon_{x_2}, \epsilon_{y_2}$ are sampled from a normal distribution $\mathcal{N}(0, \rho)$. The perturbed box b' is matched to a new set of l anchors $A' = \{a'_1, \dots, a'_l\}$. The A-IOU is computed as the intersection-of-union between A and A' . The higher A-IOU score suggests the assignment is more robust to label noise.

We testify the assignment consistency under two scenario. First, we calculate the assignment IOU with different degrees of noise ratio $\rho \in \{0.1, 0.2, \dots, 0.5\}$ using the final model. Second, we would like to investigate how the assignment consistency change through training. We report the A-IOU at different time of training with a constant $\rho = 0.1$. We compare our ASA with IOU-based assigner Ren et al. (2015); Lin et al. (2017b); Liu et al. (2016) and ATSS assigner Zhang et al. (2020) with Mean Teacher RetinaNet baseline on COCO 10%. All modules except for the assignment are kept the same to provide a fair comparison. For both evaluations, we randomly select 1000 images from `val2017` to compute the A-IOU. Figure 9 visualize the $\text{mean} \pm \text{std}$ A-IOU between clean and noisy label at different training time and different noise ratio ρ . In Figure 9(a), both ATSS and our ASA provides higher A-IOU compared with the broadly applied IOU-based assignment. However, ATSS is still based on heuristic matching rule between label and anchor boxes. ASA, instead, steadily improves itself as the detector becomes more accurate. In Figure 9(b), we see that IOU-based assignment fails to maintain the initial assignment when the large magnitude of noise is introduced in the labels. Given the noisy nature of pseudo label in SSOD, our experiment suggests that IOU-based assignment is incapable of maintaining the assignment consistency in SSOD. In contrast, our ASA strategy still performs well under server noise scenario. This experiment supports our argument that the proposed consistent assignment strategy is robust to label noise in SSOD.

Classification and Regression Inconsistency. We unveil the regression and classification mismatch problem in SSOD by identifying the mismatch between the high-score and high-IOU predictions. We obtain the confidence-IOU pairs on `val2017` using `Consistent-Teacher` and Mean Teacher RetinaNet when trained on COCO 10% data, and analyze the correlation between the two variables. We apply linear regression and measures the standard error to reflect the correlation between confidences and IOUs. Smaller error indicates higher correlation.

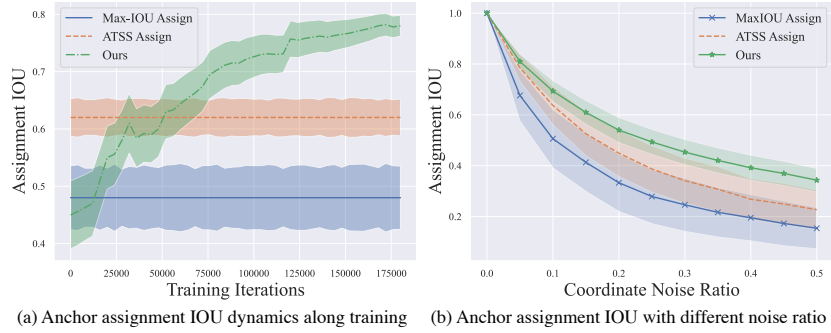


Figure 9: Assignment IOU score between ground-truth and the noisy bounding boxes (a) at different time of training and (b) using different noise ratio.

Table 6: Classification and Regression inconsistency analysis using IOU-Confidence linear regression (LR) error. We also provide the Mean Teacher IOU-Confidence plot on the right.

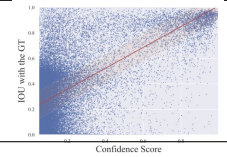
	LR Standard Error	
Mean Teacher	0.109	
Consistent-Teacher	0.080	

Table 6 provides the LR standard error for Consistent-Teacher and Mean Teacher RetinaNet. The right scatter figure displays the confidence-IOU of Mean Teacher. We observe clear cls-reg misalignment on semi-supervised detectors: numerous low-confident predictions possess high IOU score. It indicates that classification confidence does not provides a strong enough clue for an accurate regression result, which give rise to erroneous pseudo-label noise during training. The high LR error of 0.109 with Mean teacher also demonstrates this point. On the contrary, our Consistent-Teacher largely eliminates the mismatch between the two tasks with a lower LR error of 0.080. It supports our arguments that Consistent-Teacher can align the classification and regression sub-tasks and reduce the mismatch in SSOD.

C SEMI-SUPERVISED DETECTION RESULTS VISUALIZATION

C.0.1 QUALITATIVE COMPARISON WITH BASELINE.

We further compare the baseline Mean Teacher RetinaNet with our Consistent-Teacher by visualizing the predicted bounding boxes on val2017 under the COCO 10% protocol. In Figure 10, we plot the predicted and ground-truth bounding boxes in Violet and Orange respectively, alongside with the false positive bboxes highlighted in Red.

There are 3 general properties that we could observed in our demonstration.

1. First, Consistent-Teacher fits the situation of crowded object localization better, whereas Mean Teacher often mistakes the intersection of two overlapped objects as a new instance. For example, in the scenes of zebras or sheep, Mean Teacher often gives a false positive output in the overlapping area of the two objects, while Consistent-Teacher largely resolves the inaccurate positioning problem through the adaptive anchor selection mechanism.
2. Secondly, we see that under the semi-supervised setting, the Mean Teacher RetinaNet would either predict the wrong class for the correct location or regress an inaccurate bounding box despite its high classification confidence. For example, birds are sometimes misidentified as airplanes even when the localization is accurate. It is mainly attributed to the inconsistency of classification and regression tasks, i.e. the features required for

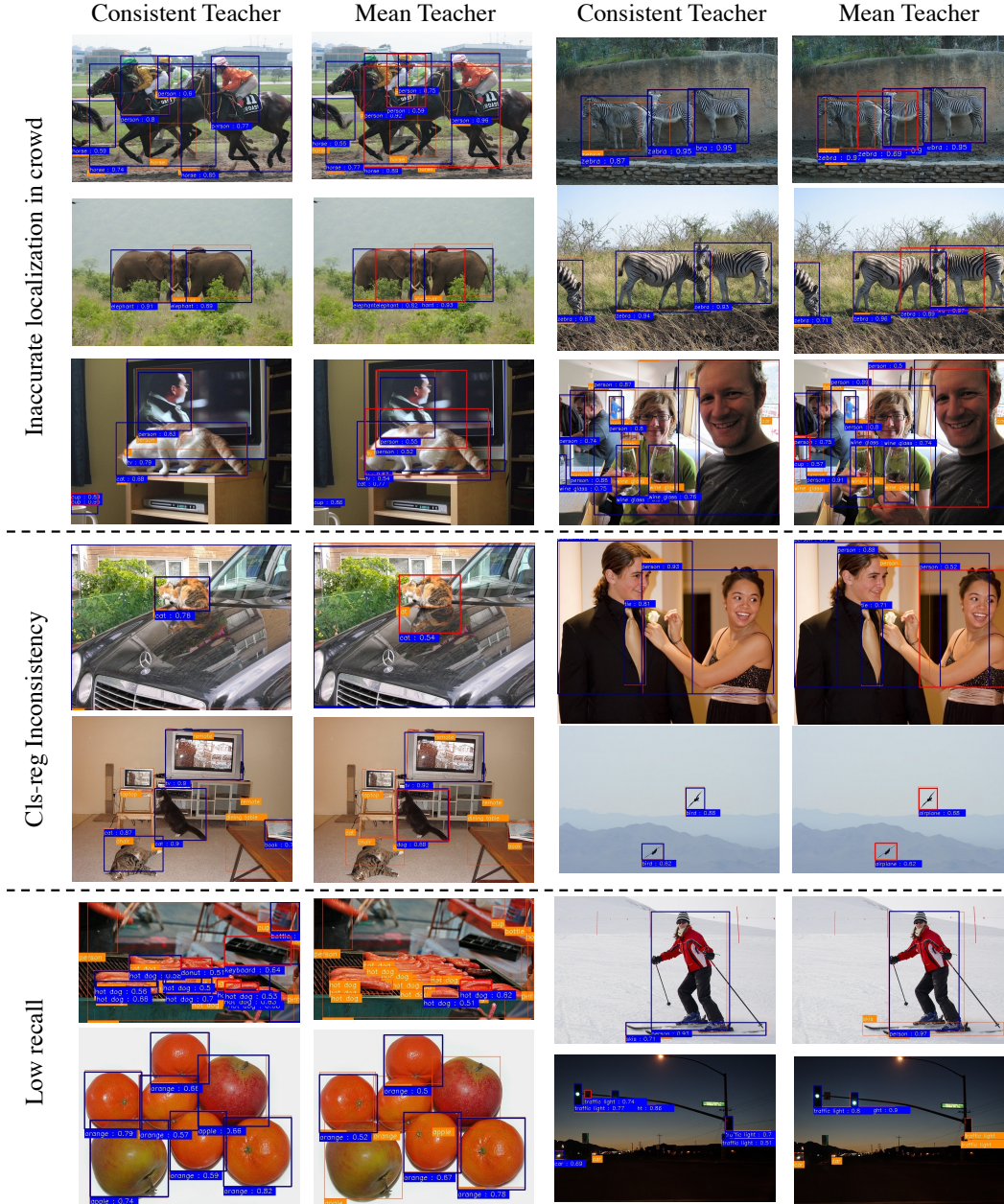


Figure 10: Qualitative comparison on the COCO%10 evaluation. The bounding boxes in Orange is the ground-truth, and Violet refers to the prediction. Red highlights the false positive predictions.

regression may not be optimal for classification. Consistent-Teacher effectively discriminates similar categories using the FAM-3D to select the features dynamically.

3. Third, Consistent-Teacher embraces higher recall since it is capable of detecting small or crowded instances which Mean Teacher fails to point out. For example, Consistent-Teacher discovers most of the hot dogs on the grill while Mean Teacher neglects most of them.

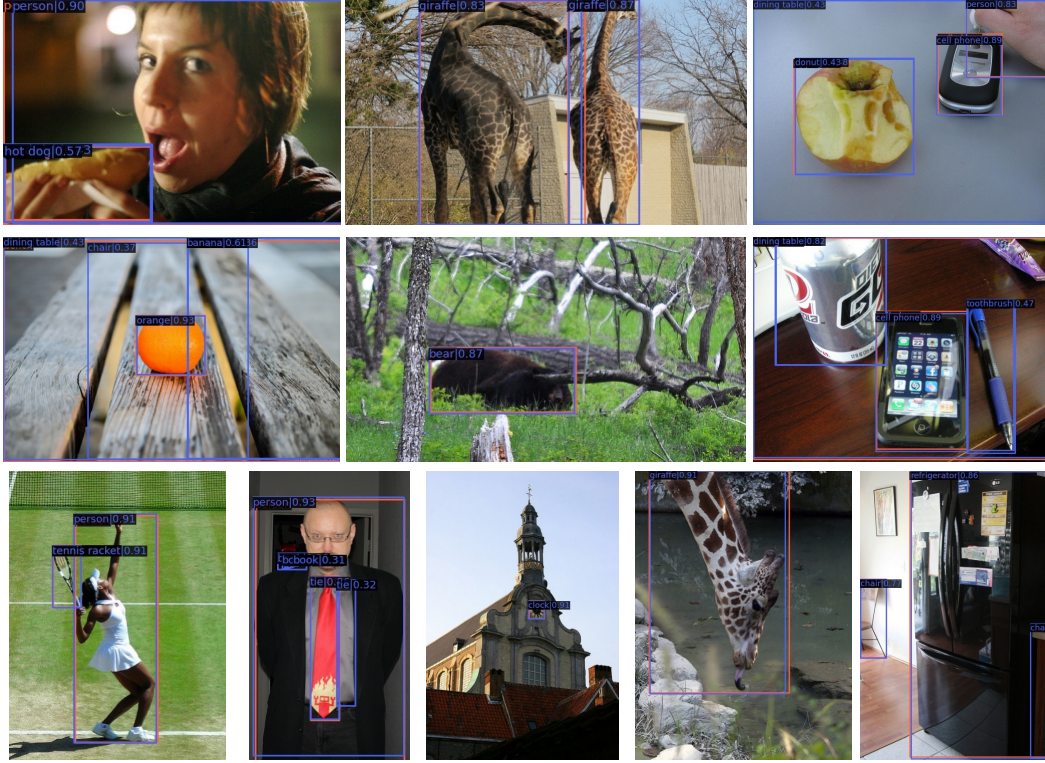


Figure 11: Good detection results for the COCO%10 evaluation. The bounding boxes in Orange is the ground-truth, and Violet refers to the prediction.

C.0.2 GOOD CASES AND FAILURE CASES.

We provide more examples to showcase the good and failure examples produced by Consistent-Teacher on COCO val2017 in Figure 11 and Figure 12. Although our proposed method achieved gratifying performance on a series of SSOD benchmarks, we can still point out its deficiencies in Figure 12. First, the trained detector lacks robustness to some out-of-distribution samples, for example, cartoon characters on street signs are recognized as real people, and reflections in mirrors are recognized as objects. Second, our detection performance is poor for some classes with small sizes, such as toothbrushes, hair dryers, etc. Third, Consistent-Teacher also tends to treat parts of the object as a whole, such as the head of the giant panda as a separate animal (in the lower left corner), and the dial of a clock as the entire clock (on the right of the panda).

D EXPERIMENT AND HYPER-PARAMETER SETTINGS

D.1 DATASETS AND DATA PREPROCESSING.

D.1.1 MS-COCO 2017.

The Microsoft Common Objects in Context (MS-COCO) is a large-scale object detection, segmentation, key-point detection, and captioning dataset. We use COCO2017 in our experiments for SSOD, which includes 118K training and 5K validation images along with bounding boxes of 80 object categories.

D.1.2 PASCAL VOC 2007-2012.

The PASCAL Visual Object Classes (VOC) dataset contains 20 object categories alongside with pixel-level segmentation annotations, bounding box annotations, and object class annotations. The

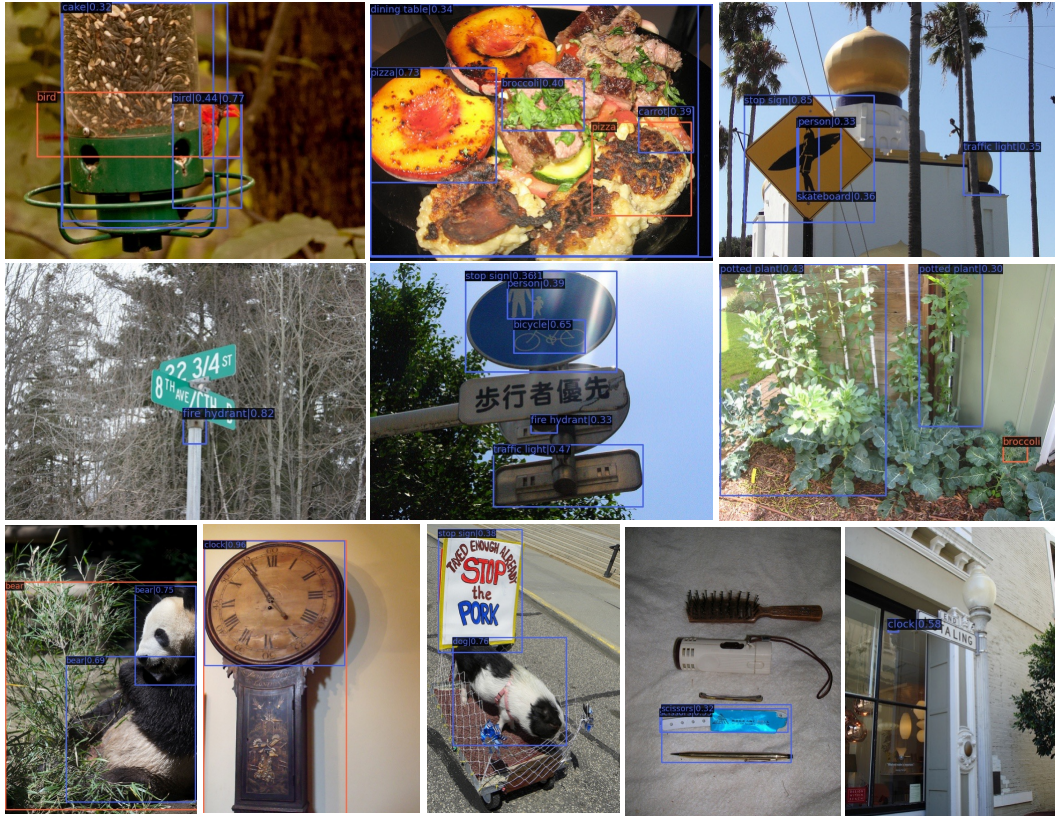


Figure 12: Failure detection results for the COCO%10 evaluation. The bounding boxes in Orange is the ground-truth, and Violet refers to the prediction.

official VOC 2007 `trainval` set is adopted as the labeled set with 5011 images and the 11540 images from VOC 2012 `trainval` set is used as unlabeled data in this study. We evaluate on the VOC 2007 test set.

D.1.3 DATA AUGMENTATIONS.

We use the same data augmentations as described in Soft Teacher Xu et al. (2021), including a labeled data augmentation in Table 7, a weak unlabeled augmentation in Table 8 and a strong unlabeled augmentation in Table 9.

D.2 IMPLEMENTATION DETAILS

We implement our `Consistent-Teacher` based on `MMDetection`⁴ framework with the data preprocessing code from the open-sourced `Soft-Teacher`⁵ and `google ssl-detection`⁶. We train our detectors on 8 NVIDIA Tesla V100 GPUs. It takes approximately 3 days for an 180K training. Each GPU contains 1 labeled image and 4 unlabeled images. The source code is attached in a separate zip file.

⁴<https://github.com/open-mmlab/mmdetection>

⁵<https://github.com/microsoft/SoftTeacher>

⁶https://github.com/google-research/ssl_detection/

Table 7: Data augmentation for labeled image training.

Transformation	Description	Parameter Setting
RandomResize	Resize the image to a the height of h randomly sampled from $h \sim U(h_{min}, h_{max})$, while keeping the height-width ratio unchanged.	$h_{min} = 400, h_{max} = 1200$ in MS-COCO $h_{min} = 480, h_{max} = 800$ in PASCAL-VOC
RandomFlip	Randomly horizontally flip a image with probability of p .	$p = 0.5$
OneOf	Select one of the transformation in a transformation set T .	$T = \text{TransAppearance}$

Table 8: Weak data augmentation for unlabeled image.

Transformation	Description	Parameter Setting
RandomResize	Resize the image to a the height of h randomly sampled from $h \sim U(h_{min}, h_{max})$, while keeping the height-width ratio unchanged.	$h_{min} = 400, h_{max} = 1200$ in MS-COCO $h_{min} = 480, h_{max} = 800$ in PASCAL-VOC
RandomFlip	Randomly horizontally flip a image with probability of p .	$p = 0.5$

Table 9: Strong data augmentation for unlabeled image.

Transformation	Description	Parameter Setting
RandomResize	Resize the image to a the height of h randomly sampled from $h \sim U(h_{min}, h_{max})$, while keeping the height-width ratio unchanged.	$h_{min} = 400, h_{max} = 1200$ in MS-COCO $h_{min} = 480, h_{max} = 800$ in PASCAL-VOC
RandomFlip	Randomly horizontally flip a image with probability of p .	$p = 0.5$
OneOf	Select one of the transformation in a transformation set T .	$T = \text{TransAppearance}$
OneOf	Select one of the transformation in a transformation set T .	$T = \text{TransGeo}$
RandErase	Randomly selects K rectangle region of size $\lambda h \times \lambda w$ in an image and erases its pixels with random values, where (h, w) are height and width of the original image.	$K \in U(1, 5)$ $\lambda \in U(0, 0.2)$

Table 10: Appearance transformations, called TransAppearance.

Transformation	Description	Parameter Setting
Identity	Returns the original image.	
Autocontrast	Maximizes the image contrast by setting the darkest (lightest) pixel to black (white).	
Equalize	Equalizes the image histogram.	
RandSolarize	Invert all pixels above a threshold value T .	$T \in U(0, 1)$
RandColor	Adjust the color balance of image. $C = 0$ returns a black&white image, $C = 1$ returns the original image.	$C \in U(0.05, 0.95)$
RandContrast	Adjust the contrast of image. $C = 0$ returns a solid grey image, $C = 1$ returns the original image.	$C \in U(0.05, 0.95)$
RandBrightness	Adjust the brightness of image. $C = 0$ returns a black image, $C = 1$ returns the original image.	$C \in U(0.05, 0.95)$
RandSharpness	Adjust the sharpness of image. $C = 0$ returns a blurred image, $C = 1$ returns the original image.	$C \in U(0.05, 0.95)$
RandPolarize	Reduce each pixel to C bits.	$C \in U(4, 8)$

Table 11: Geometric transformations, called TransGeo.

Transformation	Description	Parameter Setting
RandTranslate X	Translate the image horizontally by $\lambda \times \text{image width}$.	$\lambda \in U(-0.1, 0.1)$
RandTranslate Y	Translate the image vertically by $\lambda \times \text{image height}$.	$\lambda \in U(-0.1, 0.1)$
RandRotate Y	Rotates the image by θ degrees.	$\theta \in U(-30^\circ, 30^\circ)$
RanShear X	Shears the image along the horizontal axis with rate R .	$R \in U(-0.480, 0.480)$
RanShear Y	Shears the image along the vertically axis with rate R .	$R \in U(-0.480, 0.480)$