Dual Decoupling Training for Semi-Supervised Object Detection with Noise-Bypass Head

Shida Zheng*, **Chenshu Chen***, **Xiaowei Cai, Tingqun Ye, Wenming Tan**†

Hikvision Research Institute
{zhengshida, chenchenshu, caixiaowei6, yetingqun, tanwenming}@hikvision.com

Abstract

Pseudo bounding boxes from the self-training paradigm are inevitably noisy for semi-supervised object detection. To cope with that, a dual decoupling training framework is proposed in the present study, *i.e.* clean and noisy data decoupling, and classification and localization task decoupling. In the first decoupling, two-level thresholds are used to categorize pseudo boxes into three groups, *i.e.* clean backgrounds, noisy foregrounds and clean foregrounds. With a specially designed noise-bypass head focusing on noisy data, backbone networks can extract coarse but diverse information; and meanwhile, an original head learns from clean samples for more precise predictions. In the second decoupling, we take advantage of the two-head structure for better evaluation of localization quality, thus the category label and location of a pseudo box can remain independent of each other during training. The approach of two-level thresholds is also applied to group pseudo boxes into three sections of different location accuracy. We outperform existing works by a large margin on VOC datasets, reaching 54.8 mAP (+1.8), and even up to 55.9 mAP (+1.5) by leveraging MS-COCO *train2017* as extra unlabeled data. On MS-COCO benchmark, our method also achieves about 1.0 mAP improvements averaging across protocols compared with the prior state-of-the-art.

1 Introduction

Great progress has been made in object detection owing to the development of deep convolutional neural networks. Yet, training an accurate detector still demands a large well-annotated dataset, which is labor-consuming. To lessen the demand for expensive labeled data, semi-supervised learning (SSL) in image classification has been widely researched, while semi-supervised object detection (SSOD) is still an open issue with little literature about it.

Recent SSOD works following the self-training paradigm show better performance than the consistency-based method (Jeong et al. 2019). For self-training, a teacher model generates a set of predicted boxes on unlabeled images, with a fixed confidence threshold filtering out possible negative predictions. Remaining boxes, named pseudo boxes, are then used as targets for training a student model. The lower the threshold, the higher the recall for pseudo boxes, yet the more wrongly classified the boxes, *i.e.* false positives, interfering with the convergence of semi-supervised training. The higher the threshold, the higher the precision for pseudo boxes, but positive recalls with low confidence may be filtered, undermining the ability of trained models to detect tiny or hard samples. It seems fair to conclude that there is always a trade-off between recall and precision when there is only one confidence threshold, and thus noise in pseudo boxes is unavoidable.

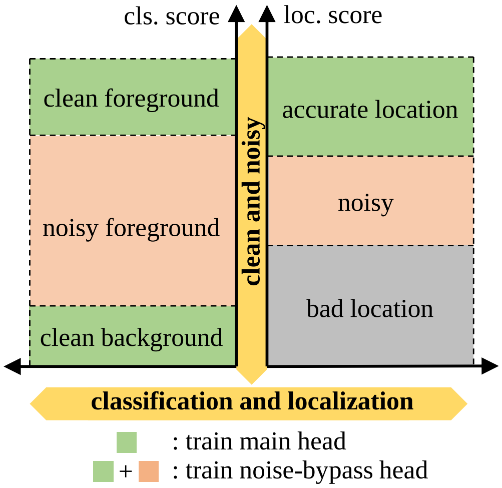


Figure 1: Illustration of 2-dimension decoupling. X-axis is about tasks. Localization is decoupled from classification with the help of the localization score. Pseudo-labels with scores from 0 to 1 are split into three parts for each task, and noisy ones teach the noise-bypass head as ground truths.

Since confidence is based on outputs of classifiers, it can be argued that confidence alone may be able to qualify classification, but not localization. Considering that, Liu et al. remove the localization training for pseudo boxes, which is not a best solution in our perspective.

In the present study, a dual decoupling training (DDT) framework is proposed to overcome the above-mentioned problems in SSOD. DDT introduces two-dimensional decoupling for self-training, specifically, clean and noisy data

*Equal contributions

†Corresponding author.

decoupling as well as classification and localization task decoupling, as shown in Figure 1.

(1) *decoupling clean and noisy data*. Most existing studies decide on the threshold without much deliberation. For example, STAC (Sohn et al. 2020b) uses 0.9, Unbiased Teacher (Liu et al. 2021a) uses 0.7, and ISMT (Yang et al. 2021) uses 0.9. By contrast, we use two thresholds to get three types of pseudo boxes. Boxes with their classification score above the higher threshold, which implies that their category label is likely to be correct, are called "clean foregrounds", while those with their classification score below the lower threshold "clean backgrounds". All the other boxes are treated as noisy foregrounds, as they are a mix of false positives and true positives. Inspired by Luo and Yang, the noise-bypass ROI head (bypass head for short) is designed to learn noisy pseudo boxes, parallel with the original main ROI head (main head for short). The bypass head is responsible for digging out general categorizing and locating information for shared backbone networks from noisy pseudo boxes. And the main head handles clean data to reach better convergence and performance.

(2) *decoupling classification and localization tasks*. Object detection handles classification and localization tasks simultaneously, but classification confidence is the only score to evaluate predictions, which is not a proper way to estimate localization quality. In this paper, an approach is provided to calculate the localization score via the bypass head. As a result, each pseudo box has two scores attached, including the classification score and the localization score. Like what we do to decouple noise from clean pseudo category labels, another two thresholds are used to discriminate pseudo boxes with different localization quality. During semi-supervised training, classification and localization are treated separately, which implies that there may be a pseudo box with clean classification but noisy localization. We will elaborate on it further in Section 3.

Our DDT framework follows the teacher-student scheme with asymmetrical augmentations. The teacher generates pseudo boxes on weakly augmented unlabeled images, and the same images after strong augmentation will then feed the student for training. The student is learnable, and the teacher updates weights via exponential moving average (EMA) from the student to generate stable pseudo annotations.

Strong augmentations in this work not only include intra-data augmentations like photometric distortion and Cutout (DeVries and Taylor 2017), but also inter-data augmentations like Mixup (Zhang et al. 2018; Guo, Mao, and Zhang 2019) and Mosaic (Bochkovskiy, Wang, and Liao 2020; Zhou et al. 2021). Inter-data augmentations alleviate the overfitting problem of two-head structure and gain extra improvements in our framework. Attributed to the proposed dual decoupling training, the teacher-student scheme, and the inter-data augmentations, we make significant progress from the current SSOD best performance on VOC datasets (Everingham et al. 2010). On experimental protocols of MS-COCO (Lin et al. 2014), we achieve state-of-the-art as well.

The main contributions of this paper are as follows:

- Data decoupling: two thresholds are adopted to differentiate clean and noisy data and devise the noise-bypass head to learn noisy data while the main head focusing on clean data. Data decoupling enhances the feature representation of shared networks and eliminates negative impacts of noise on final predictors.
- Task decoupling: the localization score is calculated with the help of the noise-bypass head, independent of confidence. As we fully utilize localization guidance, better localization performance is achieved.
- With sufficient experiments on VOC and MS-COCO datasets, the method proposed has been verified and our work is a new state-of-the-art in SSOD.

2 Related Works

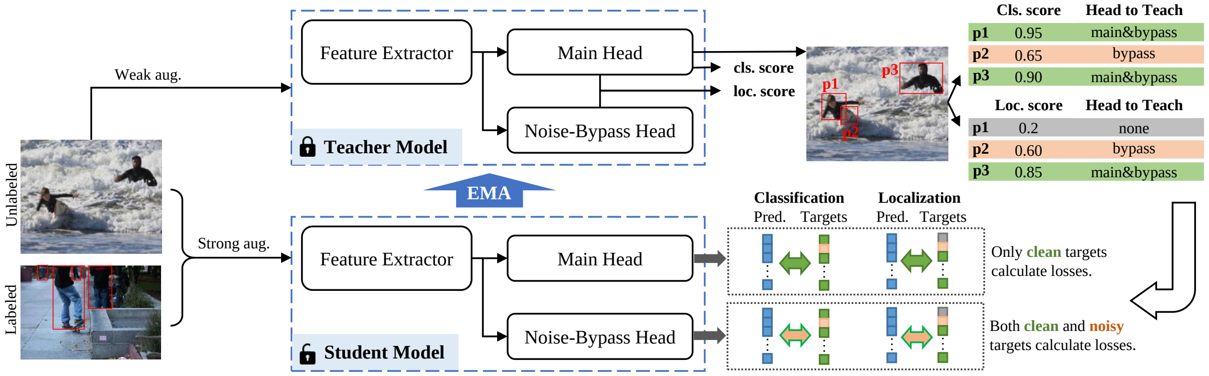
Semi-Supervised Learning in Image Classification

SSL in image classification aims at training models on labeled data along with a large amount of unlabeled data. Self-training (Lee et al. 2013; Bachman, Alsharif, and Precup 2014), also called pseudo-labeling, is a popular approach for SSL, which adopts a pretrained model to annotate pseudo labels on unlabeled images. Another widely-used approach is consistency regularization, which encourages consistency among outputs from different views of one and the same image. Different views usually result from perturbations, including model jittering (Bachman, Alsharif, and Precup 2014), data augmentations (Berthelot et al. 2019, 2020; Sohn et al. 2020a; Xie et al. 2020), feature augmentations (Kuo et al. 2020), and adversarial disturbance (Miyato et al. 2018; Yu et al. 2019). Mean teacher (Tarvainen and Valpola 2017) introduces a teacher-student dual-model framework with the teacher updating its weights from the EMA weights of the student and inspires successors (Verma et al. 2019; Liu et al. 2021a; Tang et al. 2021b). To further improve performance, uncertainty-aware methods (Mukherjee and Awadallah 2020; Rizve et al. 2021) are studied, which leverage uncertainty estimators to select better calibrated pseudo labels. This paper follows some thoughts of SSL but steps further to address the localization problem for detection tasks.

Semi-Supervised Learning in Object Detection

Thanks to deep convolutional neural networks, a variety of approaches have emerged, and facilitated the development of object detection (Ren et al. 2015; Zhang et al. 2020; Li et al. 2020; Zhang et al. 2021; Feng et al. 2021) based on costly man-annotated datasets (Everingham et al. 2010; Lin et al. 2014). In this paper, we limit the scale of labeled datasets and combine labeled datasets with unlabeled ones to train a detector in the SSOD paradigm.

As for SSOD, CSD Jeong et al. follows consistency-based algorithm and encourages original and flipped views of images to predict similar results. ISD (Jeong et al. 2021) adds extra interpolation-regularization on the basis of CSD and performs better. Another consistency-based method (Tang et al. 2021a) perturbates features and forces outputs to be consistent. STAC (Sohn et al. 2020b) is a self-training method. It uses a pretrained model to generate pseudo boxes on unlabeled images and train a student model with unlabeled images strongly augmented. Instant-teaching (Zhou



	Cls. score	Head to Teach
p1	0.95	main&bypass
p2	0.65	bypass
p3	0.90	main&bypass

	Loc. score	Head to Teach
p1	0.2	none
p2	0.60	bypass
p3	0.85	main&bypass

Figure 2: The overview of the proposed DDT framework. Firstly, the pretrained teacher generates pseudo boxes on weakly augmented unlabeled images. Then classification and localization of pseudo boxes are separated into clean and noisy targets with their scores (clean in green and noisy in orange). When training the student, the main head learns from clean targets, while the bypass head learns from both clean and noisy targets. The teacher updates weights from the student via EMA after every iteration. Labeled images with annotations only teach the main head. Best viewed in color.

et al. 2021) generates pseudo boxes instantly in a mini-batch and designs a dual-model mutual learning algorithm. ISMT (Yang et al. 2021) takes historical pseudo boxes into account but the performance is not competitive. All of the self-training methods mentioned above use one fixed threshold to generate hard pseudo boxes and train localization simply, but here we refine the pseudo category label and pseudo box location with two-level thresholds by implementing a more elaborate algorithm. Although Humble Teacher (Tang et al. 2021b) does not use hard pseudo boxes, its usage of class-dependent regression is not adequately proved. In this paper, localization quality is evaluated in a more reasonable way, and the localization score helps our decoupling method achieve better performance.

3 Methodology

Notations

In SSOD, there is a labeled dataset $\mathcal{X} = \{x_i^l, \mathcal{Y}_i^l\}_{i=1}^{N_l}$ and an unlabeled dataset with pseudo boxes $\mathcal{U} = \{x_i^u, \mathcal{Y}_i^u\}_{i=1}^{N_u}$. x_i^l and x_i^u are the i-th labeled and unlabeled image respectively. N_l and N_u are numbers of images in each dataset. $\mathcal{Y}_i = \{(c_j, s_j^c, b_j, s_j^b)\}_{j=1}^{J}$ is the annotation set of the i-th image, where c_j, s_j^c, b_j and s_j^b represent the category label, the classification score, the box location and the localization score respectively for the j-th annotation, with $s_j^c, s_j^b \in [0, 1]$. For ground truths, we define $s_j^c = s_j^b = 1$. J is the number of annotations of the i-th image.

DDT Framework Overview

Figure 2 illustrates the whole pipeline of our DDT framework. DDT consists of a student model θ_s and a teacher model θ_t sharing the same network structure. First of all, the student is pretrained on the labeled dataset and then copies the weights to the teacher. As a result, the teacher is capable of producing adequate pseudo boxes and the student has better initialized weights. The total loss for supervised pretraining can be formulated as follows,

$$\mathcal{L}_{pre} = \mathcal{L}_{pre}^{rpn} + \mathcal{L}_{pre}^{M} + \mathcal{L}_{pre}^{B} \tag{1}$$

where $\mathcal{L}_{pre}^{rpn}$ is the RPN loss, while $\mathcal{L}_{pre}^{M}$ and $\mathcal{L}_{pre}^{B}$ are ROI losses of the main and bypass head during pretraining. We pretrain both heads with different random initialization.

After warm-up, semi-supervised training starts. In each iteration, both labeled images and unlabeled images are sampled as a batch of inputs with a ratio of 1:1. Weakly augmented unlabeled images are fed into the teacher. Then the resulting boxes from the main head with confidence above the lower classification threshold τ_l^c are kept as pseudo boxes. Additionally, each pseudo box will be attached with a localization score. After generating pseudo boxes, all image-target pairs in a batch will be strongly augmented and fed to the student model.

For two-stage detectors, there is an RPN producing K_i proposals for the i-th image. We regard all pseudo boxes as foreground targets when assigning proposals. Then ROI features z_k are extracted by ROIAlign (He et al. 2017) and sent to the subsequent heads, where k is the index of proposals. There are two sets of output predictions. One is from the main head and the other is from the bypass head. Each set has classification logits and localization offsets. The ROI losses for two branches are as follows,

$$\mathcal{L}^{roi} = \mathcal{L}^{M} + \mathcal{L}^{B} \tag{2}$$

where,

$$\mathcal{L}^{M} = \sum_i \sum_k \left(w_j^{Mc} \mathcal{L}^{Mc}(\hat{c}_{jk}^{M}, c_j) + w_j^{Mb} \mathcal{L}^{Mb}(\hat{b}_{jk}^{M}, b_j) \right) \tag{3}$$

$$\mathcal{L}^{B} = \sum_i \sum_k \left(w_j^{Bc} \mathcal{L}^{Bc}(\hat{c}_{jk}^{B}, c_j) + w_j^{Bb} \mathcal{L}^{Bb}(\hat{b}_{jk}^{B}, b_j) \right) \tag{4}$$

where the uppercase letters M and B represent items about the main and bypass head respectively. The lowercase letters c and b refer to the classification and localization task. $\hat{c}_{jk}$ or $\hat{b}_{jk}$ means the prediction from the k-th proposal which is

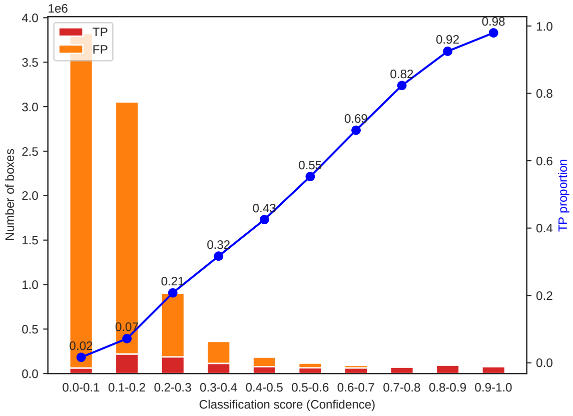


Figure 3: The histogram for pseudo box candidates. The height of bars indicates the number of boxes with different confidence. Red parts are for TPs and orange parts are for FPs. Every point on the line graph is the TP proportion. Data are from the unlabeled part of 10% COCO.

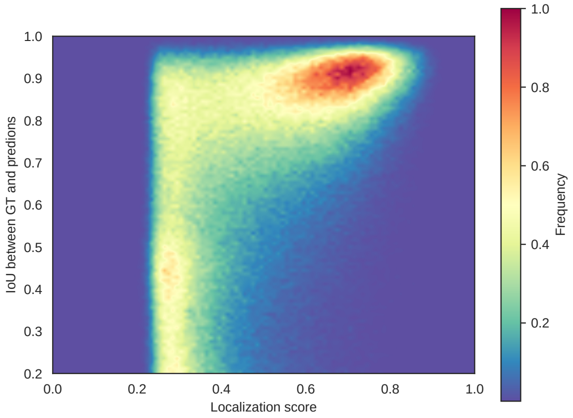


Figure 4: The density heatmap between the localization score and the actual localization quality. Red means more pseudo boxes than blue. Data are from the unlabeled part of 10% COCO.

assigned to the j-th target. i is the index of images in a mini-batch. There are four weights in equations above, which play an important role in our decoupling training. These weights can be formulated as a matrix like,

$$\begin{bmatrix} w_j^{Mc} & w_j^{Mb} \\ w_j^{Bc} & w_j^{Bb} \end{bmatrix} = \begin{bmatrix} \frac{s_j^c \mathbb{I}[s_j^c > \tau_h^c]}{\sum \mathbb{I}[s_j^c > \tau_h^c]} & \frac{s_j^b \mathbb{I}[s_j^b > \tau_h^b]}{\sum \mathbb{I}[s_j^b > \tau_h^b]} \\ \frac{s_j^c \mathbb{I}[s_j^c > \tau_l^c]}{\sum \mathbb{I}[s_j^c > \tau_l^c]} & \frac{s_j^b \mathbb{I}[s_j^b > \tau_l^b]}{\sum \mathbb{I}[s_j^b > \tau_l^b]} \end{bmatrix} \tag{5}$$

where $\mathbb{I}$ is the indicator function, τ_h^c is the higher classification threshold, τ_h^b is the higher localization threshold, τ_l^c is the lower classification threshold, and τ_l^b is the lower localization threshold. These thresholds are used to discriminate clean and noisy parts of classification and localization targets. The clean-noisy boundaries of Figure 1 illustrate the four thresholds intuitively.

At the end of an iteration, the teacher updates weights from the student in the EMA mechanism with a given EMA ratio α,

$$\theta_t \leftarrow (1-\alpha)\theta_t + \alpha\theta_s, \alpha \in [0,1] \tag{6}$$

Dual Decoupling Training

To decouple clean and noisy classification targets, we need a score s_j^c to estimate the classification quality of pseudo boxes, for which confidence is a good choice, as it is from the prediction distribution of the classifier and naturally has the capacity to qualify classification. What's more, Figure 3 shows that with the increasing confidence, the proportion of TPs is increasing too, indicating that there is a strong positive correlation between classification accuracy and confidence. If a single threshold is utilized to get pseudo boxes, a low threshold can recall extra TPs, but there are numerous FPs in pseudo boxes. A high threshold assures that the pseudo boxes are clean, but some positive instances are missing, which may be hard samples and worth learning. Therefore, we adopt two-level threshold strategy to get clean pseudo boxes with a high enough threshold and noisy yet valuable pseudo boxes with a low threshold. The rest boxes are likely to be a real background and filtered. As a result, Figure 3 illustrates that the proposed two-level threshold strategy is reasonable.

Besides, in order to decouple clean and noisy localization targets like what we do to classification above, we need another score s_j^b to estimate the localization quality. Jiang et al. explicate that the correlation between confidence and localization quality is low, thus confidence is not a valid score. In the present paper, the bypass head is capable of categorizing and locating boxes (see Figure 6), though noisy data are involved while training. Moreover, one box from the main head has a counterpart box from the bypass head, since they share the same ROI features. Consequently, our estimation method makes use of the bypass head, and calculates the localization score as follows,

$$s_j^b = s_j^c \cdot IoU(b_j^M, b_j^B) \tag{7}$$

where b_j^M and b_j^B are the output boxes from the main and bypass head respectively. They are both from the teacher model and b_j^M is the localization guidance b_j in ROI losses. The confidence term indicates *objectness*, as localization for a background box makes no sense. The IoU term indicates *certainty of localization*. Combining these two terms, we get the phenomenon in Figure 4, which shows our localization score can reflect actual localization quality. The independent localization score decouples localization from classification, which improves our performance by promoting the localization precision. Also, the two-level threshold strategy is also applied to localization, resulting in three parts of pseudo boxes with different localization quality. So far, both classification and localization targets have been split into clean and noisy parts, which will be exploited by the bypass head.

Noise-Bypass Head

The name, *noise-bypass head*, is inspired by the bypass capacitor in Electronics, which shorts high-frequency noise to ground and keeps the voltage constant.

Method	Labeled	Unlabeled	AP	AP50	AP75
Supervised	VOC07	-	42.6	72.6	47.5
CSD†	VOC07	VOC12	42.7	76.7	-
STAC (Sohn et al. 2020b)			44.6	77.5	-
ISMT (Yang et al. 2021)			46.2	77.2	-
Instant-teaching (Zhou et al. 2021)			48.7	78.3	52.0
Unbiased Teacher (Liu et al. 2021a)			48.7	77.4	51.1
Humble Teacher (Tang et al. 2021b)			53.0	80.9	-
DDT (ours)			**54.7**	**82.4**	**59.8**
CSD†	V0C07	VOC12 + MS-COCO20	43.6	77.1	-
STAC (Sohn et al. 2020b)			46.0	79.1	-
ISMT (Yang et al. 2021)			49.6	77.8	-
Instant-teaching (Zhou et al. 2021)			50.8	79.9	55.7
Unbiased Teacher (Liu et al. 2021a)			50.3	78.8	54.9
Humble Teacher (Tang et al. 2021b)			54.4	81.3	-
DDT (ours)			**55.9**	**82.5**	**61.1**

Table 1: Results on VOC datasets compared with existing works. AP50 and AP75 are reported for a thorough comparison. MS-COCO20 represents the subset of MS-COCO *train2017* with the same classes as VOC. †: the re-implementation version of Tang et al.

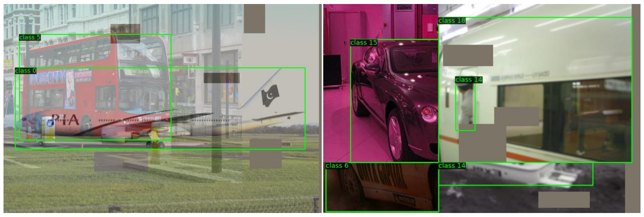
In this paper, more noise is deliberately absorbed into pseudo boxes because we want to extract more knowledge from it. A naive way is to train a single-head model with a mix of clean and noisy data. It may be good for backbone networks to extract more diverse information, but predictors will suffer from the noisy targets, leading to inaccurate results. We specially design a bypass head, which is parallel with the original main head in the network structure. The main head only learns clean targets, while the bypass head focuses on noise, so uncertain pseudo annotations will not interfere with the main head training. The extra noisy annotations provide approximate location of objects, and training them on a separate branch introduces regularization into the underlying backbone to some extent, improving the network's generalization capability.

The bypass head is a training assistant and can be removed when training is done. Therefore, the resulting model is the teacher model with the main head only (exactly the same as regular Faster RCNN in this paper). Figure 6 shows the performance between different models with different heads.

Augmentations

Weak augmentation here is the random horizontal flip. Strong augmentation includes intra- and inter-data augmentations. Intra-data augmentations only deal with one image, such as random horizontal flip, photometric distortion, random gaussian blur and Cutout. Inter-data augmentations fuse two or more images to augment one image, *e.g.* Mixup and Mosaic.

Our implementation follows Zhou et al. for Mixup and Yolov4 (Bochkovskiy, Wang, and Liao 2020) for Mosaic. Figure 5 visualizes these two augmentations. Inter-data augmentations in our method only mix labeled images with unlabeled ones. We believe they can enrich the context of input images and alleviate the overfitting problem.



(a) Mixup

(b) Mosaic

Figure 5: Visualization for inter-data augmentations. Intra-data augmentations like photometric distortion and Cutout can also be found in the visualized pictures.

4 Experiments and Analysis

Datasets

Our method DDT is benchmarked on two popular detection datasets. One is PASCAL VOC, which includes VOC07 and VOC12 datasets. In VOC07, we treat the *trainval* set as labeled data (5,011 images) and evaluate performance on the *test* set. Data from VOC12 *trainval*(11,540 images) and the subset of MS-COCO with the same classes as VOC (about 95k images) are used as extra unlabeled data. For MS-COCO, we randomly sample 1%/2%/5%/10% data from MS-COCO *train2017* as the labeled data with the rest data as the unlabeled data. Additionally, the entire MS-COCO *train2017* (about 118k images) and MS-COCO *unlabeled* (about 123k images) are used for the SSOD experiment, which validates our approach on large-scale datasets. The mean average precision (AP) at IoU=0.5:0.95 is the metrics. For VOC datasets, AP50 and AP75 are also reported for better comparison.

Method	1% COCO	2% COCO	5% COCO	10% COCO	COCO-full
Supervised	9.05±0.16	12.70±0.15	18.47±0.22	23.86±0.81	37.63
CSD†	11.12±0.15	14.15±0.13	18.79±0.13	22.76±0.09	38.52
STAC (Sohn et al. 2020b)	13.97±0.35	18.25±0.25	24.38±0.12	28.64±0.21	39.21
ISMT (Yang et al. 2021)	18.88±0.74	22.43±0.56	26.37±0.24	30.53±0.52	39.64
Instant-teaching (Zhou et al. 2021)	18.05±0.20	22.45±0.15	26.75±0.05	30.40±0.05	40.20
Unbiased Teacher (Liu et al. 2021a)	20.75±0.12	24.30±0.07	28.27±0.11	31.50±0.10	41.30
Humble Teacher (Tang et al. 2021b)	16.96±0.38	21.72±0.24	27.70±0.15	31.61±0.28	42.37
DDT (ours)	**18.62±0.42**	**24.52±0.20**	**29.24±0.16**	**32.80±0.22**	**41.90**
DDT (ours)‡	**19.44±0.32**	**25.20±0.16**	**29.92±0.12**	**33.46±0.18**	**42.40**

Table 2: The mAP at IoU=0.5:0.95 on MS-COCO val2017. Different percentages of labeled MS-COCO *train2017* are used to train models. We report margins of error for metrics, which come from multiple experiments with different random seeds. ‡: the final confidence threshold is 0.001 as Zhou et al. do.

Implementation Details

Our DDT framework is based on the two-stage detector Faster RCNN (Ren et al. 2015) with feature pyramid networks (Lin et al. 2017). The backbone network is ResNet-50 (He et al. 2016) initialized by the ImageNet-pretrained model. The classification loss is cross entropy loss for the RPN and focal loss for the ROI heads. DDT introduces four hyper-parameters to decouple the clean and noisy data. We set $\tau_l^c = 0.4$, $\tau_h^c = 0.6$, $\tau_l^b = 0.6$ and $\tau_h^b = 0.8$ unless otherwise specified. The optimizer we use is SGD with a momentum of 0.9. The size of a mini-batch is 32 with 16 labeled and 16 unlabeled images. We do not stop training until the model converges and the learning rate keeps constant during semi-supervised training, with 0.04 for VOC and 0.02 for MS-COCO. EMA ratio is set as $\alpha = 1e-4$. We adopt 8 NVIDIA Tesla V100 GPUs for all experiments.

Results on PASCAL VOC

On PASCAL VOC datasets, we compare the proposed DDT framework with the state-of-the-art methods in SSOD. As summarized in Table 1, our DDT outperforms state-of-the-art approaches by a large margin under both experimental settings. Compared with the latest proposed Humble Teacher, we improve AP from 53.0 to 54.7 when VOC12 is the only unlabeled dataset. Given additional unlabeled MS-COCO data (images of VOC classes), we improve AP from 54.4 to 55.9. Instant-teaching calculates regression loss for all pseudo boxes, while Unbiased Teacher does not learn location of pseudo boxes at all. Hence, we also report the AP75 metric and compare DDT with these two typical algorithms. The improvement of AP75 is more than that of AP, which demonstrates that our performance benefits from the higher localization precision.

Results on MS-COCO

Experiments on different settings of MS-COCO are shown in Table 2. Compared with the supervised baseline, we improve about 10 mAP after semi-supervised training, even when the labeled data are rare (1% COCO has 1100 labeled data for 80 classes). For the 2%, 5%, 10% protocol, we outperform the best of existing works, improving 1.85 absolute AP when 10% labeled data are used. If a lower final threshold is set (from 0.05 to 0.001), we achieve a higher AP for each protocol. To sum up, the state-of-the-art performance validates that our method is effective and outstanding.

Main		Bypass		Loc		
Cls.	Loc.	Cls.	Loc.	Score	AP	AP75
○	○	○	○	Conf.	31.7	32.2
●	●	●	●	Conf.	29.5	31.0
○	○	●	○	Conf.	32.2	32.6
○	○	●	○	Eq. 7	32.6	34.4
●	○	●	●	Eq. 7	30.2	32.6
○	○	○	●	Eq. 7	32.2	34.1
○	○	●	●	Eq. 7	32.7	35.0

Table 3: Ablation study for different clean and noisy data training combination on 10% COCO. ○: train with clean data. ●: train with both clean and noisy data.

Ablation Study

For every ablation study below, we use the 10% COCO protocol with the single data fold for fast experiments.

Dual-Decoupling with the bypass head Firstly, we validate effectiveness of our dual decoupling training. Thanks to the flexible framework, we can switch between different combinations of decoupling for ablation experiments, as shown in Table 3.

The first two rows are clean and noisy baseline experiments, respectively. If we add noisy data to the bypass head training (row 3), we can find the improvement compared with the clean baseline. Row 4 experiments our task decoupling and uses proposed localization score instead of confidence. Although AP improvement is marginal, AP75 gains 1.8, which shows our better localization. Row 5 introduces clean localization targets to the noisy baseline, and improves 0.7 AP. Row 6 introduces noisy localization targets to the clean baseline, and improves 0.5 AP. If dual decoupling training is fully utilized, as shown in row 7, we get the best performance among all experiments in Table 3, thus

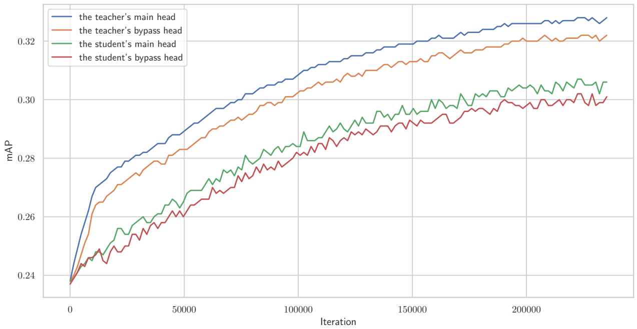


Figure 6: Performance of the four heads in DDT framework during semi-supervised training (after warm-up).

demonstrates the effectiveness of our proposed dual decoupling training.

The Performance of Four Heads Our DDT framework has four heads to produce predictions: the main head of the teacher, the bypass head of the teacher, the main head of the student and the bypass head of the student. The performance trends of these heads are displayed in Figure 6. It shows that the performance of the teacher is consistently higher than the student, which echoes the conclusion of Liu et al.. As what we have expected, the performance of the main head is also consistently higher than the bypass head, which supports our practice of treating the teacher model with the main head as the final model.

Intra-Data			Inter-Data		
Color	Blur	Cutout	Mixup	Mosaic	AP
√					30.1
√	√				30.5
√	√	√			31.4
√	√	√	√		32.4
√	√	√	√	√	32.7

Table 4: Ablation study on different augmentations.

Augmentations Next, we progressively study different strong augmentations from intra- to inter-data augmentations. As shown in Table 4, our method benefits from stronger augmentations. Among all augmentations, Mixup improves the performance from 31.4 to 32.4 and Mosaic gains 0.3 AP further.

Potential and extensibility Soft Teacher (Xu et al. 2021) achieves high performance in SSOD recently. In its training config, 0.5∼1.5 multi-scale training and 0.5∼1.5 multi-scale jittering for pseudo generation are used. For fair comparison, we conduct experiments following the same multi-scale configs and show higher AP in 5%/10%/Full COCO than Soft Teacher in Table 5. Besides, we experiment DDT on stronger base detectors like Cascade RCNN (Cai and Vasconcelos 2018) and stronger backbones like Swin-Large (Liu et al. 2021b). Results in Table 6 demonstrate our effectiveness. Above all, our method has potential for higher performance and extensibility for different networks.

Method	Protocol	AP
Soft Teacher	5% COCO	30.7
Soft Teacher	10% COCO	34.0
Soft Teacher	Full COCO	44.5
Ours	5% COCO	31.5 (+0.8)
Ours	10% COCO	35.1 (+1.1)
Ours	Full COCO	45.1 (+0.6)

Table 5: Experiments under 0.5-1.5 multi-scale training and pseudo generating following Soft Teacher (Xu et al. 2021).

Base Detector	Protocol	Supervised	Semi.
R50-CRCNN	10% COCO	26.7	35.1
SwinL-CRCNN	10% COCO	42.7	47.6
SwinL-CRCNN	Full COCO	50.8	53.3

Table 6: Extensibility experiments for our proposed method.

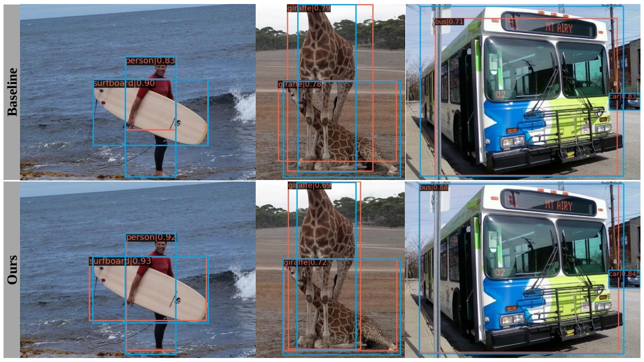


Figure 7: Visualization of localization quality. Red boxes are predictions and blue boxes are ground-truths. Top and bottom predictions are from the baseline (the first row in Table 4) and our proposed DDT.

Qualitative visualization Additionally, some qualitative results are visualized in Figure 7. The man in the first image, the giraffes in the second image and the bus in the third image illustrate that our method has more precise location than the baseline (the first row in Table 3).

5 Conclusion

In this paper, we dive into two essential parts of the self-training SSOD problem, which are noisy pseudo boxes and localization quality estimation. In order to address these two problems, we propose the DDT framework with two-dimensional decoupling: clean and noisy data decoupling as well as classification and localization decoupling. The introduced noise-bypass head successfully extracts extra knowledge out of noise while keeping the main predictor precise. Our method is validated by abundant experiments above and outperforms existing works on PASCAL VOC by a large margin and obtains SOTA on the MS-COCO benchmark.

References

Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with Pseudo-Ensembles. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2020. ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection.

Cai, Z.; and Vasconcelos, N. 2018. Cascade R-CNN: Delving into High Quality Object Detection. In *CVPR*.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout.

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

Feng, C.; Zhong, Y.; Gao, Y.; Scott, M. R.; and Huang, W. 2021. TOOD: Task-aligned One-stage Object Detection.

Guo, H.; Mao, Y.; and Zhang, R. 2019. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3714–3722.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jeong, J.; Lee, S.; Kim, J.; and Kwak, N. 2019. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32: 10759–10768.

Jeong, J.; Verma, V.; Hyun, M.; Kannala, J.; and Kwak, N. 2021. Interpolation-Based Semi-Supervised Learning for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11602–11611.

Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; and Jiang, Y. 2018. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, 784–799.

Kuo, C.-W.; Ma, C.-Y.; Huang, J.-B.; and Kira, Z. 2020. Featmatch: Feature-based augmentation for semi-supervised learning. In *European Conference on Computer Vision*, 479–495. Springer.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896.

Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 21002–21012. Curran Associates, Inc.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021a. Unbiased Teacher for Semi-Supervised Object Detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *International Conference on Computer Vision (ICCV)*.

Luo, W.; and Yang, M. 2020. Semi-supervised Semantic Segmentation via Strong-Weak Dual-Branch Network. In *Computer Vision – ECCV 2020*, 784–800. Springer International Publishing.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1979–1993.

Mukherjee, S.; and Awadallah, A. 2020. Uncertainty-aware Self-training for Few-shot Text Classification. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 21199–21212. Curran Associates, Inc.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2021. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020a. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 596–608. Curran Associates, Inc.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020b. A Simple Semi-Supervised Learning Framework for Object Detection.

Tang, P.; Ramaiah, C.; Wang, Y.; Xu, R.; and Xiong, C. 2021a. Proposal learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2291–2301.

Tang, Y.; Chen, W.; Luo, Y.; and Zhang, Y. 2021b. Humble Teachers Teach Better Students for Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3132–3141.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.

Verma, V.; Lamb, A.; Kannala, J.; Bengio, Y.; and Lopez-Paz, D. 2019. Interpolation Consistency Training for Semi-supervised Learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 3635–3641. International Joint Conferences on Artificial Intelligence Organization.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10687–10698.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-End Semi-Supervised Object Detection with Soft Teacher.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5941–5950.

Yu, B.; Wu, J.; Ma, J.; and Zhu, Z. 2019. Tangent-normal adversarial regularization for semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10676–10684.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond empirical risk minimization. In *mixup: Beyond empirical risk minimization*.

Zhang, H.; Wang, Y.; Dayoub, F.; and Sunderhauf, N. 2021. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8514–8523.

Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; and Li, S. Z. 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9759–9768.

Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and Li, H. 2021. Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4081–4090.