Ambiguity-Resistant Semi-Supervised Learning for Dense Object Detection

Chang Liu[1,*], Weiming Zhang[2,*], Xiangru Lin[2], Wei Zhang[2,†], Xiao Tan[2], Junyu Han[2],
Xiaomao Li[1,3], Errui Ding[2], Jingdong Wang[2]
[1] Shanghai University, [2] Baidu Inc, [3] Shanghai Artificial Intelligence Laboratory
{liuchang123,lixiaomao}@shu.edu.cn
{zhangweiming,linxiangru,zhangwei99,tanxiao01,hanjunyu,dingerrui,wangjingdong}@baidu.com

Table 1. Performance of supervised baselines on COCO-Standard.

Methods	COCO-Standard			
	1%	2%	5%	10%
FCOS	9.05 ±0.31	14.40 ±0.28	20.69 ±0.22	26.01 ±0.15
RetinaNet	9.40 ±0.35	14.13 ±0.33	20.12 ±0.16	26.14 ±0.08

1. More Implementation Details

1.1. Supervised baselines

In this section, we detail both anchor-free (FCOS [8]) and anchor-based (RetinaNet [4]) supervised baselines of the proposed ARSL. Both baselines employ ResNet-50 [2] as the backbone and use FPN [3] as the neck. In inference, a score threshold (0.05) is used to filter backgrounds and retain top-1000 detection results per feature pyramid. The Non-Maximum Suppression (NMS) is then employed with the IoU thresh 0.6 per class to obtain final results. Their supervised performance on the COCO-Standard setting are given in Tab. 1.

FCOS Implementations. For FCOS, following other SSOD methods [1, 6], we adopt the original implementation [8] as the baseline.

RetinaNet Implementations. RetinaNet is not employed in most SSOD methods, since its original performance is slightly lower than FCOS due to the lack of various tricks. To obtain comparable performance, we simply add tricks of FCOS to RetinaNet, e.g., adding GN [9] in heads, using GIoU [7] loss for the localization task, and setting one anchor per location. Moreover, the IoU-based assignment is replaced with the ATSS [10] assignment. With these modifications, RetinaNet achieves competitive performance with FCOS, as shown in Tab. 1.

*Co-first author (Equal Contribution).
†Corresponding author.
This work was done when Chang Liu was an intern at Baidu Inc.

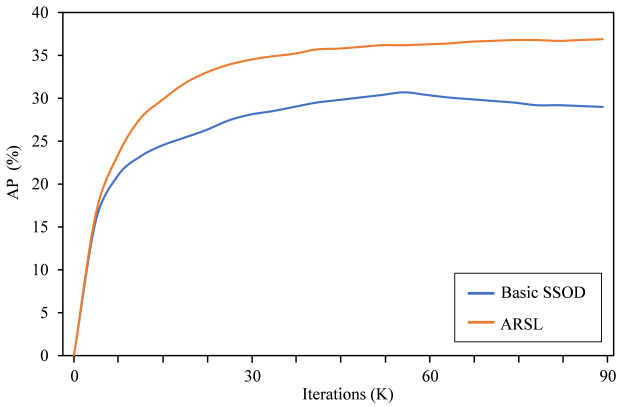

Figure 1. The AP curves of the basic SSOD framework and proposed ARSL on 10% split of the COCO-Standard setting.

1.2. Data Augmentation

For data augmentation, we follow the basic augmentation setting in Unbiased Teacher [5] without whistles and bells. The weak augmentations contain multi-scale training and random horizontal flip. The strong augmentations additionally include color jittering, grayscale, Gaussian blur, and cutout. The details are given in Tab. 2.

1.3. Training Hyper-parameters

The hyper-parameters of the training process are summarized in Tab. 3..

2. More Experiments on COCO-Standard

In this section, we present more experiments on the 10% split of the COCO-standard setting. The results are obtained on the COCO val set.

Positive Threshold of TSA. The influence of the positive threshold τ_{pos} in TSA is analyzed in Tab. 4. Substantially, a higher τ_{pos} guarantees the quality of positives and ignores more candidate samples. Under the fixed-value scheme, the best accuracy of 35.4% AP is obtained when τ_{pos} is set to

Table 2. Details of weak and strong data augmentations. 'Prob represents the probability to apply the corresponding augmentation.

Augmentation	Weak Augmentation	Strong Augmentation	Descriptions
Random Resize	short edge $\in (640, 800)$	short edge $\in (640, 800)$	The short edge of image is random resized from 640 to 800.
Horizontal Flip	Prob = 0.5	Prob = 0.5	Flip an image horizontally.
Color Jittering	-	Prob = 0.8	Including brightness, contrast, saturation, and hue. Brightness factor, contrast factor, and saturation factor is uniformly chosen from (0.6, 1.4). Hue factor is uniformly chosen from (0.9, 1.1).
Grayscale	-	Prob = 0.2	Convert an RGB image to grayscale.
Gaussian Blur	-	Prob = 0.5	Gaussian filter with $\sigma_x = 0.1$ and $\sigma_y = 2.0$ is applied.
Cutout (first)	-	Prob = 0.7	Randomly erase the rectangle region with the scale of (0.05, 0.2) and ratio of (0.3, 3.3).
Cutout (second)	-	Prob = 0.5	Randomly erase the rectangle region with the scale of (0.02, 0.2) and ratio of (0.1, 6.0).
Cutout (third)	-	Prob = 0.3	Randomly erase the rectangle region with the scale of (0.02, 0.2) and ratio of (0.05, 8.0).

Table 3. Training hyper-parameters on MS COCO and PASCAL VOC.

Hyper-parameter	COCO Standard & VOC	COCO Full
EMA rate	0.9996	0.9996
Unsupervised loss weight	2.0	2.0
Batch size for labeled data	32	32
Batch size for unlabeled data	32	32
Learning rate	0.02	0.02
Training iterations	90K	360K

0.4 and 0.5. With the adaptive τ_{pos}, the accuracy is improved to 35.6% AP. This substantiates that dynamically calculating τ_{pos} based on the statistics of samples, achieves a better trade-off between the quality and quantity of positives, and therefore bolsters the performance.

Unsupervised Loss Weight. We also investigate the effect of unsupervised loss weight β. As shown in Tab. 5, the SSOD performance is insensitive to β in a relatively large range (from 1.5 to 2.5). The best performance is obtained when β is set to 2.0.

AP Curves. We compare the detailed AP curves on the whole training process between the proposed ARSL and the basic SSOD framework. As shown in Fig. 1, under the basic SSOD framework, the performance tends to decline after about 60K iterations. We conjecture that the model is suppressed by the selection and assignment ambiguities. On the contrary, the proposed ARSL gets continuous improvements from the whole training process.

3. Qualitative Results

Fig. 2 exhibits the qualitative comparison of detection results in common and dense-object scenes between the basic SSOD framework and the proposed ARSL. Both detectors are trained on the 10% split of COCO-Standard. Compared with the basic SSOD framework, ARSL generates a large amount of detection results with higher quality. Consequently, it establishes a better foundation for pseudo labels and bolsters the SSOD performance.

Table 4. Investigation on the positive threshold τ_{pos} of task-separation assignment without classification and localization mining. 'Adaptive' represents the thresh is dynamically calculated based on the statistics.

Value of τ_{pos}	0.2	0.3	0.4	0.5	Adaptive
AP	34.8	35.2	35.4	35.4	35.6

Table 5. Investigation on unsupervised loss weight β.

Value of β	1.0	1.5	2.0	2.5	3.0
AP	36.1	36.8	36.9	36.5	35.9

4. Limitations and Future Works

In this paper, we propose ARSL to tackle the ambiguity of pseudo labels. Concretely, JCE effectively mitigates the selection ambiguity by jointly quantifying the quality of classification and localization. TSA alleviates the assignment ambiguity by separately exploiting positives for the two tasks. Although ARSL has shown remarkable improvements on both anchor-based and anchor-free one-stage detectors, it remains challenging to further improve the quality of pseudo labels. There also exist other problems in SSOD, such as class imbalance between labeled and unlabeled data, domain shift among datasets. These topics are the core problems to improve SSOD performance and generality, and worth exploring in future research.

(a)

(b)

Figure 2. Qualitative comparison on 10% split of COCO-Standard. (a) ARSL and (b) Basic SSOD Framework.

References

[1] Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4815–4824, 2022. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

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

[6] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022. 1

[7] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 1

[8] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1922–1933, 2020. 1

[9] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 1

[10] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. 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*, pages 9759–9768, 2020. 1