Two-stage Detector Provides Better Knowledge for YOLO in Semi-supervised Object Detection

Xiaoqiang Lu*, Yuting Yang, Lingling Li,
Xu Liu, Fang Liu, and Licheng Jiao

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University
* xqlu@stu.xidian.edu.cn

Abstract. In the paper, we focus on researching 2D object detection in autonomous driving technology, using semi-supervised learning to help improve detection algorithm performance with limited labeled data. Based on SODA10M, the first large-scale object detection benchmark dataset, we present a simple yet efficient teacher-student scheme. Specifically, we employ a well-trained two-stage detector as a teacher model to provide YOLO with high-quality and diverse pseudo-labels for self-training. Besides, we design a robust inference pipeline consisting of model soups, multi-scale testing, and adaptive WBF, which can improve detection performance while incurring no additional training costs. The experimental results show that our proposed method achieves 2nd place in ECCV 2022 SSLAD challenge track 1 with mAP of 82.19.

Keywords: Semi-supervised Learning, Object Detection, Autonomous Driving

1 Introduction

Object detection based on deep learning [8,13,14,17,18,24] has achieved impressive results in recent years, but these results are supported by a large amount of labeled data which requires intensive annotation at the box-level. In the real world, we have easy access to vast amounts of raw, unlabeled data, but hand-labeled data is difficult to access due to staff expertise and time costs. Thanks to SODA10M [5], the first and largest-scale object detection benchmark for autonomous driving, contains 20K fully-annotated and 10M unlabeled road images. This benchmark was used to host ECCV 2022 SSLAD challenge track 1, which intends to evaluate existing methodologies for building the next generation of industrial-level autonomous driving systems using self-supervised and semi-supervised learning.

Currently, semi-supervised learning (SSL) methods are divided into two main types. One is based on the theory of consistency regularization, while the other is a self-training method based on pseudo-labeling [11, 21]. Consistency regularization requires the model output to remain consistent in the face of various perturbations, such as sample, feature, and network perturbations [2, 22]. The

pseudo-labeling method creates artificial labels for unlabeled data by maintaining the categories most likely to be predicted by the teacher model to retrain the student model and minimize entropy.

Inspired by the great success of SSL in image classification, a growing body of work has attempted to apply SSL to object detection [7, 9, 12, 15, 20], which has proven to be an effective way to decrease the number of manually annotated samples and increase the performance of detector by utilizing unlabeled data. STAC [15] proposes a semi-supervised object detection framework that combines self-training and consistency regularization based on strong data augmentations. Based on STAC, Unbiased teacher [9] introduces an EMA approach to update the teacher model to obtain progressively higher quality pseudolabels to alleviate the severe foreground background imbalance and foreground instance imbalance in pseudolabels, and uses Focal Loss to train the student model. Soft teacher [20] proposes a semi-supervised object detection framework based on end-to-end pseudo-labeling, in which two sets of pseudo-labels are generated for unlabelled images and used to train the classification branch as well as the regression branch of the student model. PseCo [7] presents noisy pseudo box learning and multi-view scale-invariant learning to integrate object detection properties into semi-supervised object detection. STAC-YOLO [12] proposes a simple yet efficient framework based on YOLO with ensemble learning and a reliable pseudo-labels generation strategy.

In the self-training based SSL methods, the quality and diversity of pseudo-labels directly determine the training effectiveness of the model. High-quality pseudo-labels can alleviate the common confirmation bias problem in SSL, while the diversity of pseudo-labels can enable the prediction decision boundary in low-density regions and prevent model training from falling into local optimization. Following the Teacher-Student framework based on YOLO [12], we observe the pseudo-labels generated by the teacher model, which has a similar structure to the student model, are highly coupled to the output of the student model. Although this problem can be alleviated by injecting strong data augmentations, there is still much to explore. To address it, we propose a two-stage self-training paradigm using a two-stage detector as a teacher model to provide a diversity of pseudo-supervised signals for a single-stage detector student model. To be specific, we firstly utilize a five-fold cross-validation ensemble learning method to obtain a robust teacher model in the supervised training stage. After that, 160K unlabeled images are fed to the teacher model to obtain the initial pseudo-labels, which are then processed through the reliable pseudo-labels generation strategy [12] to generate the final 12K high-quality pseudo-labels. Finally, the pseudo-labels and their corresponding unlabeled images are combined with the whole training dataset to retrain the student model in the semi-supervised training stage. Furthermore, we design a powerful post-processing pipeline containing model soups [19], multi-scale testing, and Adaptive WBF (AWBF). We present AWBF, which is an upgraded version of WBF [16]. Instead of the constantly fixed weights used in WBF [16], the prediction uncertainty of different models

for a test image is evaluated to adaptively provide dynamic weights for candidate fusion results in AWBF.

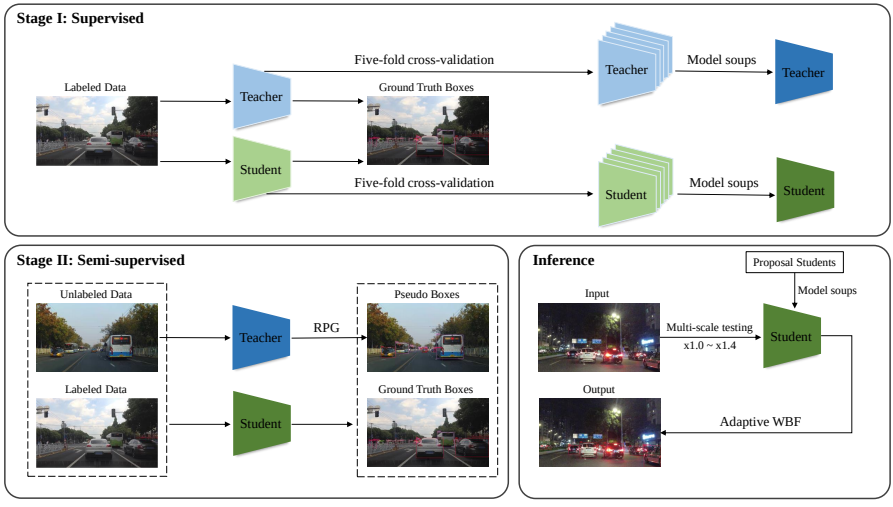

Fig. 1. The overview of our proposed method. In stage I, labeled data are used to train the teacher model and the student model in a supervised manner. The five models obtained from five-fold cross-validation are then passed through Model soups [19] to obtain the final teacher and the final student. In stage II, the teacher model trained in stage I is firstly employed to predict unlabeled data. The predictions are then sampled by RPG [12], and finally, the sampled predictions and their corresponding images after strong data augmentation are combined with labeled data to retrain the student model. During inference, we first utilize Model soups to generate a robust student from proposal students. Then all inputs are fed into the student model at five scales, and finally the predictions are fused by Adaptive WBF to obtain the final output.

2 Method

The overall framework of our proposed method is shown in Fig. 1. We will introduce the basic framework in the following parts.

2.1 Supervised training

we use Cascade Mask R-CNN [1, 6] as detector and CBNetv2 [8] with Swin-B [10] as the backbone to form the teacher model. The fed masks are created by simply filling pixels within bounding boxes that have been annotated. The student model adopts single detector Scale-YOLOv4P5 [17]. The classical ensemble learning method five-fold cross-validation is used to provide multiple models for

model integration. On the labeled data, the teacher model and student model are trained in a regular manner, supervised by the ground-truth boxes:

$$L^l_{T/S} = L^l_{T/S_{cls}} + L^l_{T/S_{reg}}, \tag{1}$$

where $L^l_{T/S}$ represents the supervised loss of the teacher (student) model.

Overfitting is unavoidable in this task due to the limited labeled images and complex driving scenes. To alleviate it, we employ week data augmentations including random resize, random flip, and colorjitter in this stage.

2.2 Semi-supervised training

Semi-supervised object detection aims to train a model in a semi-supervised setting to perform box-level classification and localization, where a small set of labeled dataset $D^l = \{(x^l_i, y^l_i)\}^{N^l}_{i=1}$ and a large unlabeled dataset $D^u = \{x^u_i\}^{N^u}_{i=1}$ are available. x^l_i, y^l_i, and N^l represent the i-th labeled input image, ground truth, and the number of labeled dataset, respectively. x^u_i and N^u represent the i-th unlabeled input image and the number of unlabeled dataset, respectively. In most works, the overall loss can be formulated as:

$$L = L^l + \lambda L^u, \tag{2}$$

where λ can be utilized to balance the supervised loss L^l and the unsupervised loss L^u. Since λ is a hyper-parameter that needs to be carefully chosen, this goes against the principle of minimalism in the proposed framework. Therefore, we replace λ by resampling the labeled dataset D^l until N^l approximates N^u. Then the overall loss is defined as:

$$L = \alpha L^l + L^u, \tag{3}$$

where α is fixed value, representing the ratio of N^u to N^l.

The current popular semi-supervised object detectors, such as STAC [15] and Unbiased teacher [9], all use a predefined confidence threshold to filter out prediction boxes with low confidence. However, there are issues with class imbalance in the data and substantial variances in object size, which result in differences in the model's capacity to detect objects in distinct categories. Therefore, we adopt the two-stage pseudo-labels sampling strategy RPG proposed in STAC-YOLO [12].

To alleviate accumulation of error gradients caused by some low accuracy pseudo-labels, we adopt strong data augmentation consisting of random resize, random flip, colorjitter, Mosaic, Mixup [23], and Cutout [4], which provides powerful perturbations to optimize the student and forces its predictions to be consistent as well as alleviates overfitting to false positive predictions.

2.3 Adaptive WBF

WBF [16] first sorts all of the bounding boxes in decreasing confidence score order. It then generates a new list of possible box fusions by determining whether

the IoU is greater than a predefined threshold and attempting to match the fused boxes to the original boxes. The coordinates of the new bounding boxes and confidence scores are then created using the formula in [16]. In this process, a fixed predefined weight is applied to the predictions from the different models.

However, we find that for some images, the predictions from different models and different scales can vary. As a result, using fixed weights limits the performance of fusion. For this reason, we present the adaptive WBF method, which applies adaptive weights to each candidate sample to be fused to better complement each other's strengths and weaknesses.

Assuming that we have K predictions $P_s = \{(\mathbf{X}^k, \mathbf{Y}^k, \mathbf{W}^k, \mathbf{H}^k, \sigma^k)\}_{k=1}^{K}$ for the test image s, where $\mathbf{X}^k$ represents a column vector consisting of the upper left corner coordinates of all prediction boxes from the k-th prediction, with the dimension equal to the total number of prediction boxes, and other similar. The following equation is used to calculate the certainty of the image s in the k-th prediction:

$$C_s^k = \frac{\sum_{j=1}^{J} \sigma_j^k}{N_J^k}, \ if \ \sigma_j^k > 0.01, \tag{4}$$

where N_J^k represents the total number of prediction bounding boxes that satisfy the condition $\sigma_j^k > 0.01$. A higher value of C_s^k means that this prediction should be given a larger weight. For vector $\mathbf{C}_s^K$, which consists of C_s^k, we use the reciprocal of the weighted median as the final weight.

3 Experiments

3.1 Implementation Details

In the supervised training stage, we use ImageNet [3] pre-trained weight to initial the teacher model, and the student model is trained from scratch. All experiments are conducted on 8 NVIDIA V100 GPUs with a batch size of 8 for the teacher and a batch size of 64 for the student. The SGD optimizer with an initial learning rate 0.01, momentum 0.937, and weight decay 0.0005 is used for training the student, and the adamW with an initial learning rate 0.00005 and weight decay 0.05 is used for training the teacher. The box head of the teacher uses GIoU loss for regression and cross entropy loss for classification. The training epoch of the teacher is set to 12 with the learning rate decayed by a factor of 0.1 at epoch 8 and 11, and the student is trained 150 epochs. In the semi-supervised training stage, we use the weight trained in the supervised training stage to initial the student model. During inference, the sizes of multi-scale testing are 1280, 1408, 1536, 1664, and 1792. For Model soups, we utilize epoch 8-epoch 12 to integrate a robust teacher and epoch 79, 89, 99, 109, 119 to integrate a robust student.

3.2 Results

The effectiveness of different parts of our proposed method is shown in Tab . 1. For the baseline training, the student model and the teacher model achieved

Table 1. The performance of our proposed method on the validation set. TTA and AWBF represent test time augmentation and adaptive WBF respectively.

Student	Teacher	Five-fold	Model soups [19]	RPG [12]	TTA+AWBF	mAP
✓						68.15
✓		✓	✓			71.24
✓		✓	✓		✓	74.38
	✓					68.77
	✓	✓	✓			71.14
	✓	✓	✓		✓	73.96
✓	✓		✓	✓		77.18
✓	✓		✓	✓	✓	79.73

Table 2. State-of-the-art performance on the test set, our method achieves 82.19 and wins 2nd place.

User	Team	Pedestrian	Cyclist	Car	Truck	Tram	Tricycle	mAP
gavin	MTCV	83.61	89.55	94.35	90.70	88.68	63.91	85.13
IPIU-XDU	**IPIU-XDU**	82.06	87.23	92.92	88.09	84.63	58.21	82.19
transformer	CMIC	79.83	85.54	92.02	85.91	82.56	61.70	81.26

68.15 and 68.77 respectively. When adding five-fold cross-validation, Model soups, and TTA with AWBF to the student and the teacher in the supervised training stage, they improve the mAP by 6.23% and 5.19%. In the semi-supervised training stage, our teacher-student scheme increases the mAP of the student from 68.15 to 79.73, which provides a significant gain of 11.58%..

Tab . 2 shows the final results of ECCV 2022 Workshop SSLAD Track 1-2D object detection challenge.

4 Conclusion

In the work, we extend the classical self-training paradigm and propose a simple yet efficient self-training framework based on teacher-student scheme. In contrast to previous approaches that used a similar teacher-student network structure, we introduce a two-stage detector as a teacher to impart richer knowledge to the single-stage student detector. Through enriching the diversity and improving the quality of pseudo-labels, our method assists the model in alleviating the confirmation bias issue that is inherent in semi-supervised learning and avoiding overfitting noisy data to prevent falling into local optimization. Besides, we construct a strong inference pipeline, resulting in higher performance without additional training costs. Finally, our proposed method achieves 2nd place in ECCV 2022 SSLAD Track 1-2D Object Detection Challenge.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: high quality object detection and instance segmentation. IEEE transactions on pattern analysis and machine intelligence **43**(5), 1483–1498 (2019)
2. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622 (2021)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
5. Han, J., Liang, X., Xu, H., Chen, K., Hong, L., Mao, J., Ye, C., Zhang, W., Li, Z., Liang, X., et al.: Soda10m: A large-scale 2d self/semi-supervised object detection dataset for autonomous driving. arXiv preprint arXiv:2106.11118 (2021)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
7. Li, G., Li, X., Wang, Y., Zhang, S., Wu, Y., Liang, D.: Pseco: Pseudo labeling and consistency training for semi-supervised object detection. arXiv preprint arXiv:2203.16317 (2022)
8. Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J., Ling, H.: Cbnetv2: A composite backbone network architecture for object detection. arXiv preprint arXiv:2107.00420 (2021)
9. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480 (2021)
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
11. Lu, X., Cao, G., Gou, T.: Semi-supervised landcover classification with adaptive pixel-rebalancing self-training. In: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. pp. 4611–4614. IEEE (2022)
12. Lu, X., Cao, G., Zhang, Z., Yang, Y., Jiao, L., Liu, F.: A simple semi-supervised learning framework based on yolo for object detection. arXiv preprint arXiv:2005.04757 (2020)
13. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
15. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020)
16. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107**, 104117 (2021)

17. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Scaled-yolov4: Scaling cross stage partial network. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 13029–13038 (2021)
18. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696 (2022)
19. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: International Conference on Machine Learning. pp. 23965–23998. PMLR (2022)
20. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3060–3069 (2021)
21. Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y.: St++: Make self-training work better for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4268–4277 (2022)
22. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)
23. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
24. Zhang, Z., Lu, X., Cao, G., Yang, Y., Jiao, L., Liu, F.: Vit-yolo: Transformer-based yolo for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2799–2808 (2021)