Learn to Detect Corner Case with Semi-supervised Learning

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 corner case detection in autonomous driving technology, using semi-supervised learning to help improve the detection performance of novel object instances. Based on CODA, a real-world road corner case dataset for object detection in autonomous driving, 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 3 with the score of 3.06, proving the potential of semi-supervised learning for open-world object detection.

Keywords: Semi-supervised Learning, Corner Case Detection, Autonomous Driving

1 Introduction

Object detection based on deep learning [9, 15, 16, 19, 20, 25] has achieved impressive results in recent years, but these results are based on the supposition that all of the classes to be detected are available during the training phase. Inevitably, in the real world, there are novel objects that are unseen or rare, i.e., the out-of-distribution samples, which mainly consist of two categories, namely 1) instance of novel class and 2) novel instance of common class, as shown in Fig . 1. These novel cases of object detection in autonomous driving may result in severe consequences, putting lives at risk. Thanks to CODA [4], a novel real-world road corner case dataset for object detection in autonomous driving, consisting of approximately 10k carefully selected road driving scenes with image domain tags, well-aligned lidar data and high-quality bounding box annotation for 43 representative object categories. This benchmark was used to host ECCV 2022 SSLAD challenge track 3, which intends to discover novel methods for detecting corner cases among common traffic participants in the real world.

Fig. 1. Examples of corner cases.

Recently, more and more work is focusing on open world object detection. ORE [6] proposes a novel computer vision problem called: 'Open World Object Detection' and introduce a strong evaluation protocol and provide a novel solution based on contrastive clustering and energy based unknown identification. MViT [14] proposes to train with aligned image-text pairs based on multi-scale deformable attention and late vision-language fusion, which can effectively bridge the lack of a top-down supervision signal governed by human-understandable semantics. OW-DETR [3] introduces a novel end-to-end transformer-based framework for open-world object detection, consisting of attention-driven pseudo-labeling, novelty classification, and objectness scoring. Benefits from multi-scale contextual information and less inductive bias, OW-DETR enables knowledge transfer from known classes to the unknown class and can better discriminate between unknown objects and backgrounds.

In contrast to the incremental learning paradigm used in open-world object detection, we focus on using a semi-supervised learning approach to explore the identification of unknown categories and the generalization of out-of-distribution objects. 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 [12, 23]. Inspired by the great success of SSL in image classification, a growing body of work has attempted to apply SSL to object detection [7, 10, 13, 17, 22], 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.

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 [13], 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

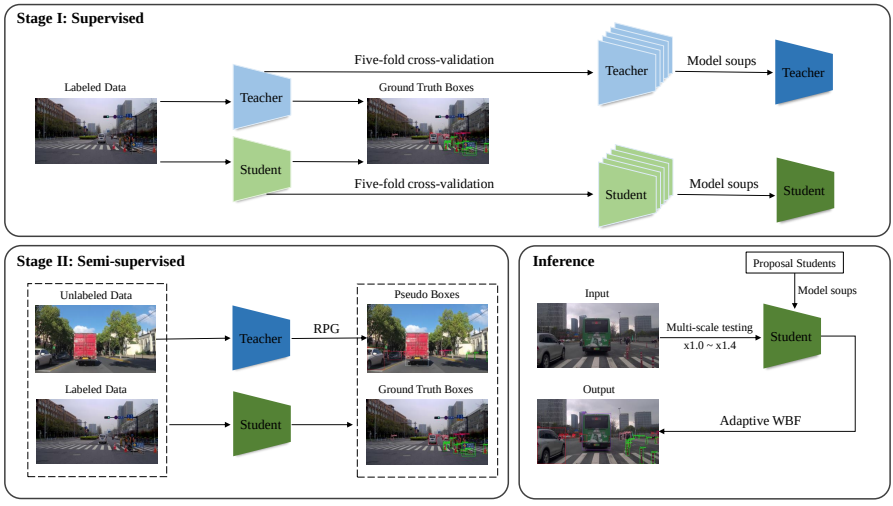

Fig. 2. The overview of our proposed method.

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 based on CODA [8] validation dataset. After that, 10K images from SODA10M [4] labeled with common categories are fed to the teacher model to obtain the pseudo-labels for corner cases. Finally, the pseudo-labels and their corresponding unlabeled images are combined with the whole validation dataset to retrain the student model in the semi-supervised training stage. Furthermore, we design a powerful post-processing pipeline containing model soups [21], multi-scale testing, and Adaptive WBF (AWBF). We present AWBF, which is an upgraded version of WBF [18]. Instead of the constantly fixed weights used in WBF [18], the prediction uncertainty of different models for a test image is evaluated to adaptively provide dynamic weights for candidate fusion results in AWBF.

2 Method

The overall framework of our proposed method is shown in Fig. 2. We will introduce the basic framework in the following parts.

2.1 Supervised training

we use Cascade Mask R-CNN [1,5] as detector and CBNetv2 [9] with Swin-B [11] as the backbone to form the teacher model. The fed masks are created by sim-

ply filling pixels within bounding boxes that have been annotated. The student model adopts single detector Scaled-YOLOv4P5 [19]. 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.

Following the current popular semi-supervised object detectors [10, 17], we also use a predefined confidence threshold δ to filter out prediction boxes with low confidence.

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 [24], and Cutout [2], 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 [18] 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 [18]. 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

We simply unify all corner categories into a single category, the category Corner. In the supervised training stage, we use SODA10M [4] pre-trained weight to initial the teacher model and the student model. 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 100 epochs. In the semi-supervised training stage, we use the weight trained in the supervised training stage to initial the student model. The confidence threshold δ used by the teacher to filter pseudo-labels is set to 0.4. 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 59, 69, 79, 89, and 99 to integrate a robust student.

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

Student	Teacher	Five-fold	Model soups [21]	Self-training	TTA+AWBF	Sum
✓						3.04
✓		✓	✓			3.11
✓		✓	✓		✓	3.27
	✓					3.13
	✓	✓	✓			3.19
	✓	✓	✓		✓	3.29
✓	✓		✓	✓		3.26
✓	✓		✓	✓	✓	3.35

Table 2. State-of-the-art performance on the test set, our method achieves 3.06 and wins 2nd place.

User	Team	AR-agnostic-corner	AR-agnostic	AP-agnostic	AP-common	Sum
gavin	MTCV	0.80	0.85	0.78	0.66	3.09
IPIU-XDU	**IPIU-XDU**	0.79	0.85	0.77	0.64	3.06
haooooooo	edl	0.76	0.81	0.70	0.55	2.83

3.2 Results

The effectiveness of different parts of our proposed method is shown in Tab . 1. We randomly sample 0.8k images from the CODA validation dataset as the local validation dataset. Tab . 2 shows the final results of ECCV 2022 Workshop SSLAD Track 3-Corner Case 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 for open-world object detection. Specifically, 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 3-Corner Case Detection Challenge, which demonstrates the potential of semi-supervised learning for open-world object detection.

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. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
3. Gupta, A., Narayan, S., Joseph, K., Khan, S., Khan, F.S., Shah, M.: Ow-detr: Open-world detection transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9235–9244 (2022)
4. 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)
5. 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)
6. Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5830–5840 (2021)
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. Li, K., Chen, K., Wang, H., Hong, L., Ye, C., Han, J., Chen, Y., Zhang, W., Xu, C., Yeung, D.Y., et al.: Coda: A real-world road corner case dataset for object detection in autonomous driving. arXiv preprint arXiv:2203.07724 (2022)
9. 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)
10. 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)
11. 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)
12. 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)
13. 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)
14. Maaz, M., Rasheed, H., Khan, S., Khan, F.S., Anwer, R.M., Yang, M.H.: Class-agnostic object detection with multi-modal transformer. arXiv preprint arXiv:2111.11430 (2021)
15. 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)
16. 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)

17. 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)
18. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107**, 104117 (2021)
19. 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)
20. 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)
21. 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)
22. 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)
23. 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)
24. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
25. 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)