3rd Palce Solution for ECCV 2022 Workshop SSLAD Track 1

Jun Yu, Shenshen Du, Renda Li, Gongpeng Zhao, Zhongpeng Cai, Bingyuan Zhang, and Ruiqiang Yang

University of Science and Technology of China
harryjun@ustc.edu.cn,
{dushens,rdli,zgp0531,czp_2402242823,sa22218182,xq9866}@mail.ustc.edu.cn

Abstract. This paper describes our solution for SSLAD Track 1 - 2D Object Detection at the ECCV 2022 workshop. The goal of semi-supervised object detection (SS-OD) is to train a powerful object detector by training from labeled data and massive unlabeled data with pseudo labels. In this work, we propose a simple and effective semi-supervised learning framework and a two-stages pseudo-label generation strategy for object detection. We use Cascade R-CNN with ConvNeXt backbone as our detector, utilize semi-supervised learning on a large number of unlabeled and a small number of labeled dataset, and choose a reasonable data enhancement strategy based on the characteristics of the SODA 10M dataset, as well as a post-processing strategy. Finally, our solution achieves 81.26 mAP in the public ranking of SSLAD Track 1 - 2D object detection, ranking the third place in this competition.

Keywords: SSLAD, semi-supervised, object detection

1 Introduction

This workshop SSLAD Track 1 - 2D aims to facilitate the development of future industrial-grade autonomous driving systems, where the required visual recognition models should be self-exploratory, self-training and self-adaptive, able to cope with a variety of emerging geographic environments, streets, cities, weather conditions, object labels, viewpoints or anomalous scenarios. To address this problem, many efforts have recently been made in self-supervised learning, large-scale pre-training, weakly supervised learning, and incremental/continuous learning to improve autonomous driving perception systems that deviate from the traditional path of supervised learning to enable autonomous driving solutions. Currently, various supervised learning methods have greatly improved many problems in autonomous driving due to the rise of large-scale annotated datasets and advances in computing hardware. However, these supervised learning methods are notoriously "data hungry", especially in the current autonomous driving domain. The performance of self-driving perception systems is highly dependent on the annotation scale of labeled bounding boxes, which makes them

impractical for many real-world industrial applications. Therefore, the direction of real-world autonomous driving development increasingly tends to transition from annotated to unannotated datasets. Thus, self-supervised learning and semi-supervised learning show a great need to drive the development of autonomous driving technologies.

In recent years, object detection[10, 14, 13] has made great progress in the field of computer vision, benefiting from the great success of deep learning[6]. By using large amounts of manually labeled data, the accuracy of object detection has been significantly improved. However, obtaining large amounts of manually labeled data, especially labeled data for object detection, requires accurate localization and classification, which is laborious. Semi-supervised learning has received increasing attention in recent years, as it provides methods to improve model performance using unlabeled data without large-scale annotated data. Most semi-supervised[8, 15] methods typically include input augmentation, perturbation, and consistency regularization. They tune the model to be invariant and robust to certain enhancements of the inputs, which requires that the outputs are consistent given the original and enhanced inputs. In this work, we propose a simple and effective Cascade r-cnn with ConvNeXt[12] backbone as a SSOD framework. Experimental results show that the framework yields good results in semi-supervised object detection.

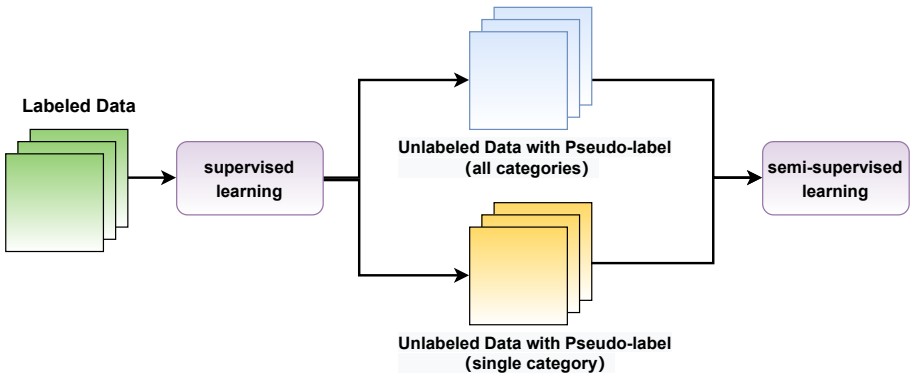

Fig. 1. shows a brief version of our pseudo-label based object detection method

In this work, Cascade r-cnn network, a very good method for object detection, does not overemphasize the backbone network, but uses a cascading approach to improve the IOU threshold to purify the samples, and achieves a 2-4 point improvement for different backbone networks and has good compatibility. ConvNeXt outperforms Swin Transformer[11] in COCO detection and ADE20K segmentation, while maintaining the simplicity and efficiency of the standard ConvNet module. As shown in Figure. 1, we performed supervised learning on the labeled data as a baseline. We use the labeled data to train the model and

then predict the unlabeled data. The unlabeled data is then used to obtain the pseudo-labels, which are then filtered and thresholded. For semi-supervised phase, we perform semi-supervised learning based on pseudo-label with the full categories and single category, respectively. And then retrained on both labeled and unlabeled images after data augmentation. After trying a series of data augmentation and training techniques, good results were achieved, especially in tricycle category.

2 Methods

We use two-stage object detection framework Cascade r-cnn[2] as our detection framework. Figure. 2 shows the overview of our solution. Our framework consists of three main parts, supervised learning phase, two-stages pseudo-label generation phase and semi-supervised learning phase.

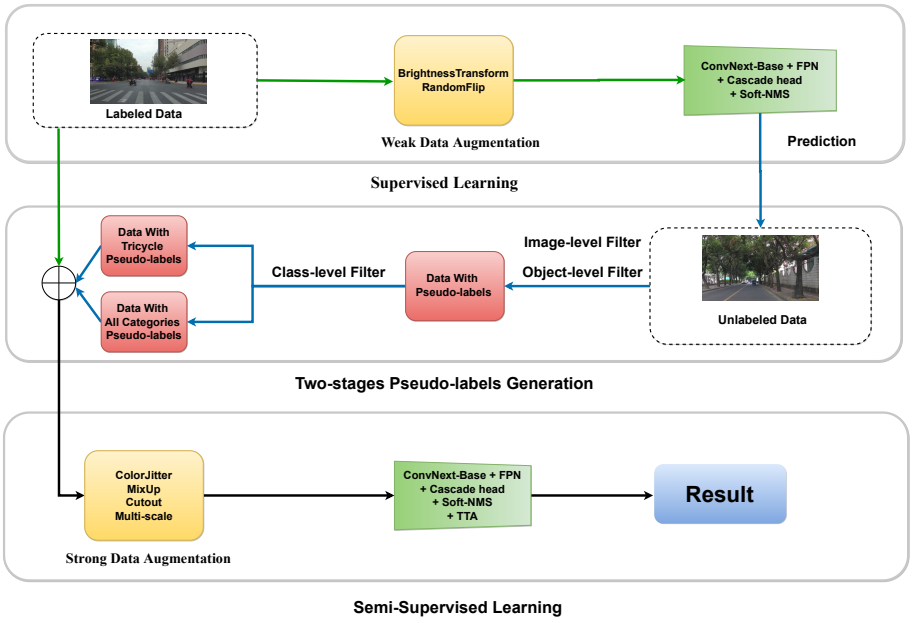

Fig. 2. An overview of our solution

2.1 Supervised Learning Phase

As used in soft-teacher[16], in the supervised learning phase, weak augmentation is used on both the training and validation sets to produce hard labels, mainly about *brightnesstransform* and *randomflip*. Our detector is Cascade r-cnn with ConvNeXt[12] as the backbone. ConvNeXt is a stronger model with

faster inference speed and higher accuracy compared to Swin Transformer, and ConvNeXt-XL achieves 87.8% accuracy on ImageNet 22K. The neck is the default FPN[9], Cascade head is used as the detection head. In the post-processing stage, softnms[1] is used for processing the redundant boxes.

2.2 Two-stages Pseudo-labels Generation

SODA10M[5] contains 10M unlabeled data and 20K labeled data, where 5K images are used as training set, 5K images as validation set, and 10K images as the test set. This dataset has 6 categories including *Pedestrian, Cyclist, Car, Truck, Tram, Tricycle*. We counted the number of each category and found that the number of tricycles is very small, which would cause the model to underlearn the category. Therefore we propose a two-stages pseudo-label generation approach to solve this problem. We generate pseudo-labels with the model trained in the first stage. We first infer about 25w unlabeled images and then go through image-level and object-level filers to generate pseudo-labels with good quality but only a part of pseudo-labels are used due to GPU limitation. If the quality of pseudo-label is inaccurate, it will affect the performance of the model. In order to solve the problem that the tricycle category has fewer training samples, we are able to reduce the imbalance in the number of this category by filtering at the class-level while ensuring the quality of the pseudo-labels with the rest pseudo-labels.

2.3 Semi-Supervised Learning

In this phase, the labeled data and the pseudo-labels generated in the second phase are used as the training set, which is then strongly enhanced with Color jitter, Mixup[7] and CutOut[4]. Color jitter can transform the brightness, contrast, saturationhe hue of the image, which makes the model more robust to the brightness variations of the images. Mixup increases the robustness of the model by remembering the wrong labels. CutOut enables the model to learn global information instead of local information. Multi-scale training increases the robustness of the model to scale transformations. The enhanced data are then trained with the same detector as in the first stage and post processing is used to obtain our results finally.

3 Experiments

In this section we will briefly describe the details of our implementation on SSLAD and the specific details of the experiment.

3.1 Dataset

The dataset used in SSLAD track1 is a Large-Scale 2D Self/Semi-Supervised Object Detection Dataset for Autonomous Driving, for short SODA10M, which contains 10M unlabeled images and 20k labeled images with 6 representative

object categories. SODA10M is designed for promoting significant progress of self-supervised learning and domain adaptation in autonomous driving. It is the largest 2D autonomous driving dataset until now and will serve as a more challenging benchmark for the community.

3.2 Metric

For this task, we use Mean Average Precision(mAP) in COCO API among all categories as our evaluation metric, that is, the mean over the APs of pedestrian, cyclist, car, truck, tram and tricycle. The IoU overlap threshold for pedestrian, cyclist, tricycle is set to 0.5, and for car, truck, tram is set to 0.7.

3.3 Implementation Details

The detector we use is the Cascade r-cnn based on ConvNeXt-Base, a model that has been better implemented in MMDetection[3]. AdamW is used as our optimizer with 0.0001 initial learning rate. We use 8*V100 for training, and we also perform multi-scale training with the batchsize 4*8.

3.4 Result

We first perform supervised training on the labeled data with strong data augmentation, and then migrate the augmentation method directly to the pseudo-label in the semi-supervised training phase. The following Table 1 shows our scores on local validation. We can find that mixup get a 5% increase on our local validation. And then we just perform weak data augmentation on our first phase supervised training to get teacher model.

Table 1. Performance of data augmentation of a single model with single scale testing on the validation set, baseline is cascade r-cnn with backbone ConvNext-tiny.

Method	mAP
Baseline	54.4
+ colorjitter	55.8
+ mixup	61.0
+ cutout	61.1
+ multi-scale	61.5

In the two-stages pseudo label generation phase, we infered 25w images with the teacher model. Then we filtered them through the two stages pseudo-labels generation, and then generated about 1w instances of tricycle, which greatly increased the number of tricycle while maintaining the quality of the pseudo label in that category. This allowed us to finally achieve 61.37AP on this class.

Table 2. Performance of different data on our local validation, split train data is a part of train and val set, the model is Cascade r-cnn with backbone ConvNeXt-base.

Data	tricycle AP	mAP
split train data	0.541	0.675
+ Pseudo-labels(all categories)	0.610	0.700
+ Pseudo-labels(only tricycle)	0.643	0.721

In the semi-supervised training phase, we directly migrated the data augmentation tested on supervised training to semi-supervised training. And firstly we split the all label data as a new train and val dataset with a ratio 9:1 as our baseline, and 25w pseudo-labels with all categories after image and object level filter are trained on our model. And we use Pseudo-labels(only tricycle) after class level filter to futher train our model with the Pseudo-labels and all of the labeled data, Table 2 shows our semi-supervised training results. And final test score achieved 81.26, which is the third place in this challenge. Tabel 3 shows the final result of the test set on the leaderboard.

Table 3. Final test leaderboard, our submission achives 81.26 and wins 3rd place.

User	Team name	mAP
gavin	MTCV	85.13
IPIU-XDU	IPIU-XDU	82.19
transformer	CMIC	81.26

4 Conclusions

This report details the key methods used in the ECCV 2022 SSLAD Challenge track 1-2D object detection. In this work, we propose a simple and effective semi-supervised object detection framework based on the Caseade r-cnn with ConvNeXt as backbone and a two-stages pseudo-label generation strategy that collaboratively exploits large-scale unlabeled data and few labeled data to learn a robust detector. By two-stages pseudo-label generation strategy, the AP of the tricycle category is greatly improved, Experiments show that our method effectively utilizes unlabeled data, and this approach progressively updates the original detector to improve accuracy and robustness. Finally, we obtained the third place in this competition with 81.26 mAP.

References

1. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms–improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision. pp. 5561–5569 (2017)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
3. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
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., Xu, C.: Soda10m: A large-scale 2d self/semi-supervised object detection dataset for autonomous driving (2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Hongyi Zhang, Moustapha Cisse, Y.N.D.D.L.P.: mixup: Beyond empirical risk minimization. International Conference on Learning Representations (2018), https://openreview.net/forum?id=r1Ddp1-Rb
8. Jeong, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. Advances in neural information processing systems **32** (2019)
9. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
10. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
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 (ICCV) (2021)
12. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
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. Wan, F., Wei, P., Jiao, J., Han, Z., Ye, Q.: Min-entropy latent model for weakly supervised object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1297–1306 (2018)
16. 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. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)