1st Place Solution for SSLAD Challenge 2022: 2D Object Detection

Jiawei Zhao, Xingyue Chen, Zhaolin Cui, Xuede Li, Junfeng Luo, Xiaolin Wei

Meituan Vision AI Department
{zhaojiawei12,chenxingyue02,cuizhaolin,lixuede,luojunfeng,weixiaolin02}@meituan.com

Abstract. In this report, we introduce the technical details of our solution for ECCV 2022 Workshop SSLAD Track 1 - 2D Object Detection. Large-scale object detection in autonomous driving is a challenging task due to time and GPU constraints. To tackle this problem, we first construct a strong baseline model with the two-stage detector Cascade RCNN. Then we propose a simple but effective pseudo-supervised learning method to iterative train better pretrain weights on more accurate pesudo-labels and more unlabeled images, which could effectively lead to better performance for supervised learning. At last, we adopt semi-supervised learning and ensemble. We achieve 85.13 mAP on the test set and win 1st place in this large-scale object detection challenge.

Keywords: Object Detection, Pseudo-supervised Learning, Large-scale

1 Introduction

In recent years, deep neural networks have achieved great success in object detection. In computer vision, CNNs are widely adapted to extract appearance information for classification, whose effectiveness has already been confirmed since the age of VGG [1]. After that, RPN [2] network and two-stage framework become a hit on object detection, which shows that both recognition and location could be easily solved by CNNs.

However, CNN still has many problems in object detection tasks. One of them is the expensive annotation cost of detection datasets. Especially for the daily-updated collected unlabeled images in autonomous driving scenes, employing various unlabeled images to improve model performance is an urgent demand.

To tackle these problems, we introduce an iterative pseudo-supervised learning method and win 1st place on the test leaderboard of the 2D object detection challenge. We will introduce our pipeline and detailed components in section 2. The experiment results and ablation studies are shown in section 3.

2 Approach

In this section, we introduce an effective pipeline with multiple training steps. As shown in Fig. 1, we first train a strong two-stage detector with ImageNet-1K

[3] pretrain as Fig. 1 (a), then we iterative predict the unlabeled images and retrain the detector with only pseudo labels as Fig. 1 (b) to conduct better pretrain weights, which plays a key role in our pipeline. At last, we apply full- and semi-supervised learning with bigger resolution pseudo-pretrain, to conduct the final submissions.

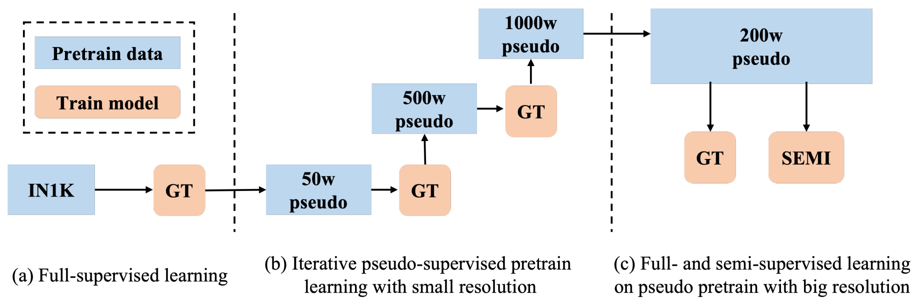

Fig. 1. Pipeline of our solution.

2.1 Base Detector

We adopt the two-stage Cascade RCNN [4] with FPN [5] as our baseline architecture. Furthermore, we apply swin-base [6] and convnext-base [7] with ImageNet-1K [3] pretrain as our backbone for a comprehensive representation.

2.2 Data Augmentation

To adapt to the diversity of weather and scenes in the test set, we apply multiple augmentation methods (*e.g.*Albu, MixUp, AutoAugmentV2 [8]) to enrich the samples during training. We first use color/bright augmentations (ColorJitter, RandomBrightnessContrast, RGBShift) and weather augmentations (RandomFog, Random Rain) implemented in Albu. Then two random samples are selected to fuse with a random weighted summation. At last, we apply the v2 policy implemented in AutoAugment.

2.3 Pseudo-supervised learning

To exploit the large-scale unlabeled images in SODA10M [9], we propose a simple but effective pseudo-supervised method to iterative train the pretrain weights with more accurate pseudo labels and more unlabeled images. In supervised learning, large-scale pseudo labels would introduce more noise and limit performance improvements. To solve this problem, we utilize pseudo labels to train a more robust representation as the pretrain weights for supervised learning. The detailed performance improvement of pseudo-pretrain is shown in Tab. 3.

To avoid the huge time consumption of training large-scale unlabeled images, we first resize images to a smaller resolution (360, 640) to train 50w/500w/1000w pseudo labels respectively. To align the feature distribution of small and big resolutions, we further resize images to a bigger resolution (1080, 1920) to train 200w pseudo labels. Notably, we filter the pseudo labels with confidence 0.8.

2.4 Semi-supervised learning

Semi-supervised object detection (SSOD) methods could effectively utilize unlabeled data to boost performance. We use the SOTA approach PseCo [10] for SSOD, which delves into two key components of semi-supervised learning (*i.e.*pseudo labeling and consistency training). We adapt PseCo [10] to the strong detector mentioned above with 50w unlabeled images and 1:1 sampling ratio.

2.5 SWA

To improve the robustness of our model, we further use Stochastic Weight Averaging (SWA [11]) to train extra 12 epochs with cyclical learning rates and select the average checkpoint as the final model.

3 Experiments

3.1 Datasets

In the ECCV2022 SSLAD Challenge Track1, SODA10M [9] is provided as the official dataset. SODA10M is a large-scale 2D objection detection dataset for autonomous driving, which contains 10M unlabeled images and 20K labeled images. All labeled images are annotated exhaustively with 6 categories(car, truck, pedestrian, tram, cyclist and tricycle). The labeled images are split into train(5K), validation(5K) and testing(10K) sets respectively. Only train set and validation set are used during training.

Unlabeled Data Sampling We find the test set is collected from Shanghai, Guangzhou, and Shenzhen, and the sunny images are about three times that of the rainy and overcast images. To keep the data distribution in pseudo- and semi-supervised learning, we random sample 50w/100w/200w/500w unlabeled images respectively by some dimensions (*e.g.*collection city, weather, period).

3.2 Implementation Details

In the initial supervised learning stage, we adopt multi-scale training and the image size ranges from (648, 1920) to (1080, 1920). We adopt AdamW optimizer with an initial learning rate 1e-4, betas (0.9, 0.999) and weight decay 0.05. The training epoch is set to 50 with the learning rate decayed by a factor of 10:1 at epochs 33 and 44. In the pseudo-supervised learning stage, the learning rate is set as 1e-4 and the training epoch is set as 24. In the inference stage, we adopt multi-scale augmentation and soft-NMS [12]. All experiments are conducted on 16 NVIDIA A100 GPUs. Our implementation is based on the open-source object detection toolbox MMDetection [13].

Table 1. Final performance on the test leaderboard.

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

3.3 Performance Analysis

The final submissions on test leaderboard are shown in Tab. 1, our approach achieved meanAP 85.13 and won the first prize by a big margin.

Table 2. Ablation study on the validation set. All results are conducted by a single swin-based model with ImageNet-1K pretrain.

epochs	img size	multi-scale	AutoAugV2	MixUp	Albu	SWA	val mAP
36	(1080, 1920)	✓					64.70
36	(1080, 1920)	✓	✓	✓			67.03
50	(1080, 1920)	✓	✓	✓			68.27
50	(1080, 1920)	✓	✓	✓	✓		70.86
50	(1080, 1920)	✓	✓	✓	✓	✓	71.05
50	(2048, 2666)	✓	✓	✓	✓	✓	71.61

In the initial supervised learning stage, we train a strong detector with ImageNet-1K pretrain on the train set. The detailed ablation studies on the validation set are shown in Tab. 2. Adopting more augmentations (*e.g.*Albu, MixUp, AutoAugmentV2) and more epochs could effectively improve the validation mAP from 64.70 to 70.86. SWA could further improve the performance slightly. Applying a bigger resolution from 1920 to 2666 is also beneficial.

In the pseudo-supervised learning stage, we first obtain reliable pseudo-labels on 50w unlabeled images by the best-performed model and train the network with a small resolution (360, 640). The model trained on pseudo labels is further used as the pretrain weights of supervised learning. We iterative generate more accurate pseudo-labels on more unlabeled images and train better pretrain weights for supervised learning. The detailed ablation studies on the validation set and test set are shown in Tab. 3. Notably, our proposed pseudo-supervised learning could effectively improve performance by almost 10%. Loading the 1000w pseudo-pretrain weights on the small resolution, we further retrain 200w pseudo labels with a bigger resolution to obtain better performance.

In the semi-supervised learning stage, we adopt PseCo [10] with 50w unlabeled images. At last, we replace the backbone from swin-base [6] to convnext-base [7] and retrain the above steps, Weighted Boxes Fusion (WBF [14]) is utilized to fuse the predicted results from these different models.

Table 3. Ablation study on the validation set and test set. All results are conducted with pseudo-pretrain.

pseudo-pretrain	pretrain size	ensemble	PseCo [10]	img size	val	test
50w	(360,640)			(1080,1920)	76.45	-
500w	(360,640)			(1080,1920)	79.28	-
1000w	(360,640)			(1080,1920)	81.44	83.75
200w	(1080,1920)			(1080,1920)	-	84.45
200w	(1080,1920)	WBF [14]	50w	(1080,1920)	-	85.13

3.4 Not Work Attempts

Self-supervised learning Inspired by the previous solution, we attempt to use self-supervised training (*e.g.*Simmim [15], BYOL [16], SoCo [17]) to obtain better weights as the initialization for subsequent training. However, the performance on downstream detection tasks is not as good as our proposed method mentioned above. with resource and time constraints, we abort these attempts early.

4 Conclusions

In this paper, we build a strong detector and introduce an iterative pseudo-supervised learning method for large-scale unlabeled images. With 1000w pseudo-pretrain, semi-supervised learning and model ensemble, we achieve 85.13 mAP and win 1st place on the test leaderboard.

References

1. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
2. 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)
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6) (2017) 84–90
4. Cai, Z., Vasconcelos, N.: Cascade r-cnn: high quality object detection and instance segmentation. IEEE transactions on pattern analysis and machine intelligence **43**(5) (2019) 1483–1498
5. 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. (2017) 2117–2125
6. 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. (2021) 10012–10022
7. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 11976–11986
8. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 113–123
9. 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)
10. 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)
11. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407 (2018)
12. 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. (2017) 5561–5569
13. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
14. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107** (2021) 104117
15. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 9653–9663
16. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33** (2020) 21271–21284

17. Wei, F., Gao, Y., Wu, Z., Hu, H., Lin, S.: Aligning pretraining for detection via object-level contrastive learning. Advances in Neural Information Processing Systems **34** (2021) 22682–22694