A Simple Semi-Supervised Learning Framework based on YOLO for Object Detection

Xiaoqiang Lu, Guojin Cao, Zixiao Zhang, Yuting Yang, Licheng Jiao, Fang Liu
School of Artificial Intelligence, Xidian University
Xi'an, Shaanxi Province,710071, China
{xqlu,20181214133,zhangzx1999,Ytyang_1}@stu.xidian.edu.cn lchjiao@mail.xidian.edu.cn

Abstract

SODA10M is the first large-scale object detection benchmark for autonomous driving, which aims at facilitating a safe, ever-evolving and robust autonomous driving system. Despite the large amount of data acquired, rare annotations are available. Semi-supervised learning methods can help to get use of extensive existing data resources and reduce the cost of manual labeling, which is of great significance for visual scene understanding. In this work, we propose a simple yet efficient semi-supervised learning framework and a reliable pseudo-labels generation strategy for object detection. To begin with, a well-trained powerful teacher model has been obtained by the ensemble learning strategy, which unlabeled images could apply to produce excellent pseudo-labels for self-training. The experimental results show that our method achieves the third place on ICCV2021 SSLAD challenge track 1 (mAP is 81.27).

1. Introduction

Autonomous driving [9] means that the vehicle perceives the surrounding environment through sensors, and changes the driving behavior in real time to complete the driving task without human intervention. Autonomous driving can reduce the occurrence of traffic accidents, increase the utilization rate of road traffic resources, and save residents' travel costs. Therefore, the research on autonomous driving technology is of great significance. Autonomous driving technology based on computer vision uses observation images of visual sensors as input and driving actions as output. As an essential module in the visual perception system, object detection in road images plays one of the most critical roles for autonomous driving. Performances of current object detection approaches, however, may be limited by the currently available datasets, due to the drawbacks of existing benchmarks. These limitations include the lacking diversity of data sources and limited labeled datasets. To boost the development of real-world autonomous driving systems, the first and largest-scale object detection benchmark for autonomous driving (SODA10M) [6] was developed, which contains 10 million road images. The SODA10M data set can be distinguished from the existing data set in terms of scale, diversity and generalization. This benchmark was used to hold the ICCV2021 SSLAD Challenge, which aims to investigate the current methods for constructing the next-generation of industrial-level autonomous driving systems by self-supervised and semi-supervised learning.

With the development of deep learning, various object detection algorithms have been proposed. Existing object detectors are mostly categorized by whether they have a region-of-interest proposal step (two-stage) [17, 5, 20, 3] or not (one-stage) [14, 16, 2, 11, 10]. Recently, following the one-stage detector design, YOLO series [14, 16, 2, 15, 1] have attracted substantial attention due to their efficiency and simplicity. They extract the most advanced detection technologies available at the time (e.g., the SPP module [7] for YOLOv3, Mish [13] activation for YOLOv4) and optimize the implementation for best practice. Hence, we selected Scaled-YOLOv4 as our baseline model.

Semi-supervised learning (SSL) [21] has received growing attention in recent years as it provides means of using unlabeled data to improve model performance when large-scale annotated data is not available. The majority of the recent SSL methods typically consist of input augmentations , perturbations, and consistency regularization. They regularize the model to be invariant and robust to certain augmentations on the input, which requires the outputs given the original and augmented inputs to be consistent. In this work, we propose a simple yet efficient SSL framework based on YOLO for object detection. As shown in Figure 1, we first obtain 5 models through ensemble learning, and then use them to predict 160k unlabeled data. The Reliability Pseudo-labels Generation (RPG) strategy is used in the prediction process. After obtaining approximately 12k high-quality pseudo-labels, we add them to the abeled data for self-training until the model converges. Then we use the

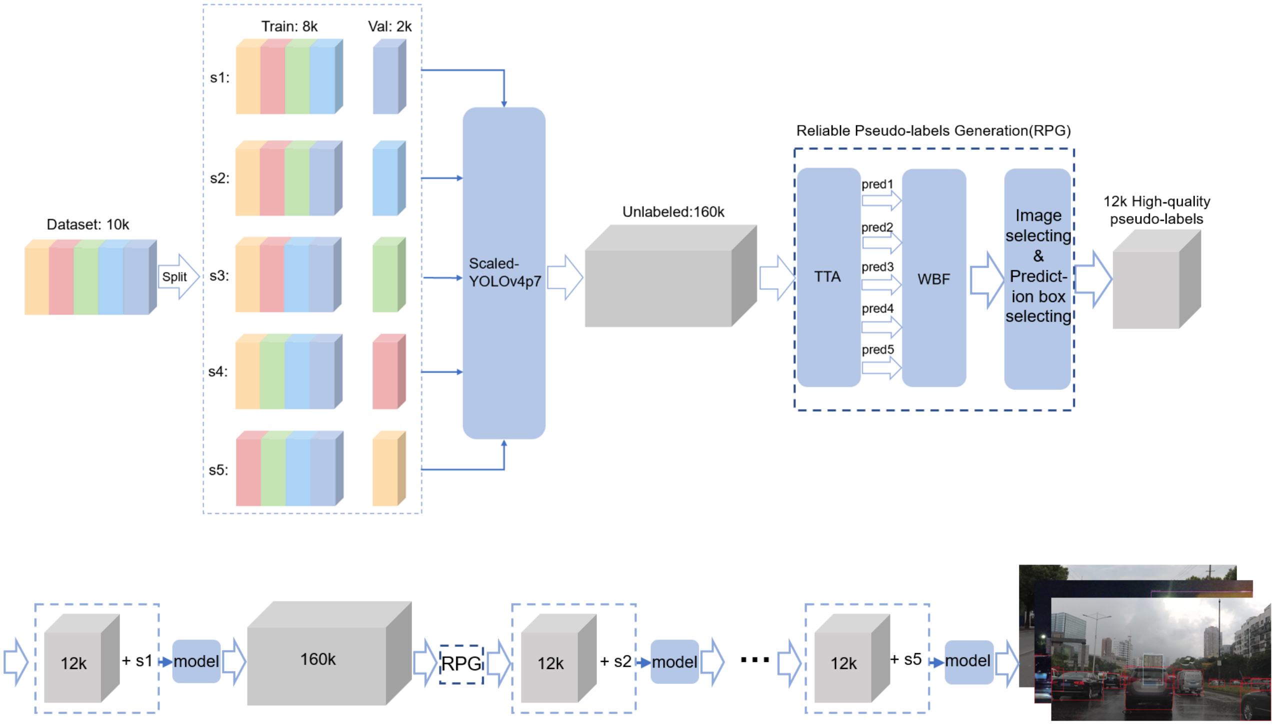

Figure 1. The whole Semi-Supervised Learning Framework

convergent model to re-predict the unlabeled data, and also use the RPG strategy to obtain higher-quality pseudo-labels for self-training. Repeat the above process until the labeled data is fully utilized.

2. Methods

2.1. Ensemble Learning for Supervised Stage

In the ICCV2021 SSLAD Challenge Track1 - 2D object detection, only 10k fully-annotated training sets and validation sets are provided. Due to limited annotated samples, how to train an excellent teacher model which unlabeled data can apply to generate high-qualified pseudo labels has become the basis of semi-supervised training methods. Based on this consideration, we adopt the idea of ensemble learning [4]. That is, we first randomly shuffle the 10k labeled data, and then split them into 5 data sets. Whenever one of them is selected as the validation set, the remaining four data are integrated into the training set. After this data division method, the samples obtained (s1, s2, s3, s4, s5) are respectively subjected to supervised training, and the model selects Scaled-YOLOv4p7.

2.2. Reliable Pseudo-labels Generation

The current popular semi-supervised object detectors, such as STAC [18] and Unbiased teacher [12], all use a fixed confidence threshold to filter out prediction boxes with low confidence. But they do not consider that the sensitivity of each category to the confidence threshold may vary greatly. For example, category 3 (car) in the SODA10M data is not sensitive to the confidence threshold, that is, this category can be filtered with a larger threshold; while category 6 (tricycle) is very sensitive to the confidence threshold, so it needs to use a smaller threshold to filter.

In the RPG process, the five models obtained in the previous stage are firstly multi-scale predicted separately, and then the predictions of multi-scales of each model are fused by Weighted Boxes Fusion (WBF) [19], after which the fused results are then WBF processing is performed to obtain preliminary reliable pseudo-labels.

The computational cost of using these unlabeled images directly is obviously too high, so we perform image-level selecting and prediction box-level selecting successively. In image-level selecting, we set the image-level confidence threshold for each category of the 160k prediction results, that is, take the maximum value of the confidence of each category of each image, and retain a certain number of images by a certain confidence threshold. Similar to the operation in image-level selecting, in prediction box-level selecting, we set prediction box-level confidence threshold for each category of the prediction boxes in the retained images, and then retain a certain number of prediction boxes by a certain confidence threshold.

	pedestrian	cyclist	car	truck	tram	tricycle
Threshold on image level	0.938	0.955	0.98	0.968	0.968	0.7
Threshold on bounding-box level	0.4	0.4	0.2	0.2	0.1	0.1

Table 1. The Selecting Thresholds on image level and bounding-box level.

	first	+s1	+s2	+s3	+s4	+s5	GPU days
YOLOR-P6	0.559	0.601	0.625	0.638	0.644	0.649	0.62×4
Scaled YOLOv4p7	0.595	0.622	0.634	0.641	0.650	0.655	3.21×4

Table 2. The local evaluation of some models after training through our simple yet efficient semi-supervised learning framework.

2.3. Self-training for Semi-supervised Stage

In STAC, the pseudo-labels used by the student is fixed. When the student is gradually improving, the immutable pseudo-labels will obviously drag down the student's progress. In order to solve this problem, Unbiased teacher gradually transfers the weight of the student to update the teacher via Exponential Moving Average (EMA) [8]. With the improvement of detection accuracy, the teacher can generate more accurate and stable pseudo-labels to optimize the student. The two models can continuously evolve together to improve the accuracy of detection. Obviously, an immutable teacher will hinder the progress of Student, but a constantly changing teacher may also mislead student. Therefore, we used a compromise solution. We first combine the 12k high-quality pseudo-labels obtained in Section 2.2 with the sample s1 in Section 2.1 to train until the model converges, and then re-process the converged model through RPG. After that, the processed higher-quality pseudo-labels and the sample s2 are jointly trained until the model converges, and this process is repeated until the labeled data is fully utilized. In this process, our teacher model has been updated 4 times, which is very helpful for guiding the student model correctly and stably.

3. Experiments

3.1. Implement Details

In the supervised training phase, We do not use ImageNet pre-trained models, and all scaled-YOLOv4 models are trained from scratch and the adopted tool is SGD optimizer. For the first-generation teacher model, we choose Scaled-YOLOv4p7. We first use weak data augmentation methods (rotation, translation, flip, or color jittering) to train 200 epochs, and then followed by using stronger data augmentation methods (Mosaic, Mixup) to train 100 epochs.

In the process of Reliable Pseudo-labels Generation, the size of multi-scale prediction is [1536, 1664, 1792, 1920, 2048], the weight of multi-scale fusion is 1, 2, 3, 2, 1, and the weight of multi-model fusion is 1, 1, 1, 1, 1. The image-level selecting threshold and the prediction box-level selecting threshold will vary with the update of the teacher model and the number of unlabeled images that need to be retained. Table 1 is the threshold table under the condition of the initial teacher model and 12k unlabeled images.

In the self-training phase, we first perform 3 epochs warm-up training. During the warm-up process, the momentum of the optimizer SGD is set to 0.8, and one-dimensional linear interpolation is used to update the learning rate of each iteration. After warm-up training, the cosine annealing function is used to attenuate the learning rate, where the initial learning rate is 0.02, and the minimum learning rate is 0.2*0.01. The time for self-training is 200 epochs. During training, strong and weak data augmentation strategies are used, with a probability of 0.5. The weight decays are set to 5104. All experiments were performed with 4 Tesla V100.

3.2. Evaluation Metrics

For this task, SODA10M use Mean Average Precision(mAP) in COCO API among all categories as 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. Experimental Results

Model	AP	AP50
YOLOR-P6	0.559	0.763
YOLOR-W6	0.556	0.756
YOLOR-E6	0.571	0.783
YOLOR-D6	0.576	0.792
Scale-YOLOv4p5	0.577	0.791
Scale-YOLOv4p6	0.583	0.796
Scale-YOLOv4p7	**0.595**	**0.801**

Table 3. The local evaluations based on the 7 first-generation teacher models under the s1 data set.

Table 3 shows the local evaluations based on the 7 first-generation teacher models under the s1 data set. The results show that Scaled-YOLOv4p7 has achieved the best performance. Table 2 shows the local evaluation of some models after training through our simple yet efficient semi-supervised learning framework. The results show that Scaled-YOLOv4p7 achieves the best performance, but

YOLOR-P6 demonstrates significant ability to learn information from unlabeled data and ultimately achieves similar performance to Scaled-YOLOv4p7. If the cost of training time is considered, YOLOR-P6 would be the most cost-effective model.

4. Conclusion

This report details the key technologies used in ICCV2021 SSLAD challenge track1-2D object detection. Inspired by STACUnbiased teacher, we propose a simple yet efficient semi-supervised object detection framework based on YOLO and a reliable pseudo-label generation strategy. Experiments show that our approach makes efficient use of unlabeled data while making full use of labelled data, and this approach progressively updates the original detector to improve accuracy and robustness. In the end, after multi-model fusion, we achieved the third place with mAP of 81.27.

References

[1] ultralytics/yolov5, Aug. 2021. original-date: 2020-05-18T03:45:11Z.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv:2004.10934 [cs, eess]*, Apr. 2020. arXiv: 2004.10934.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving Into High Quality Object Detection. pages 6154–6162, 2018.

[4] Thomas G. Dietterich. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, Lecture Notes in Computer Science, pages 1–15, Berlin, Heidelberg, 2000. Springer.

[5] Ross Girshick. Fast R-CNN. pages 1440–1448, 2015.

[6] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. SODA10M: Towards Large-Scale Object Detection Benchmark for Autonomous Driving. *arXiv:2106.11118 [cs]*, June 2021. arXiv: 2106.11118.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, Sept. 2015. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[8] A. J. Lawrance and P. a. W. Lewis. An exponential moving-average sequence and point process (EMA1). *Journal of Applied Probability*, 14(1):98–113, Mar. 1977. Publisher: Cambridge University Press.

[9] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J. Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, Michael Sokolsky, Ganymed Stanek, David Stavens, Alex Teichman, Moritz Werling, and Sebastian Thrun. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 163–168, June 2011. ISSN: 1931-0587.

[10] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. pages 2980–2988, 2017.

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision ECCV 2016*, Lecture Notes in Computer Science, pages 21–37, Cham, 2016. Springer International Publishing.

[12] 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. *arXiv:2102.09480 [cs]*, Feb. 2021. arXiv: 2102.09480.

[13] Diganta Misra. Mish: A Self Regularized Non-Monotonic Neural Activation Function. page 13.

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *arXiv:1506.02640 [cs]*, May 2016. arXiv: 1506.02640.

[15] Joseph Redmon and Ali Farhadi. YOLO9000: Better, Faster, Stronger. pages 7263–7271, 2017.

[16] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*, Apr. 2018. arXiv: 1804.02767.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[18] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A Simple Semi-Supervised Learning Framework for Object Detection. *arXiv:2005.04757 [cs]*, Dec. 2020. arXiv: 2005.04757.

[19] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, Mar. 2021. arXiv: 1910.13302.

[20] Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. *arXiv:1911.09070 [cs, eess]*, July 2020. arXiv: 1911.09070.

[21] Xiaojin Zhu and Andrew B. Goldberg. Introduction to Semi-Supervised Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, Jan. 2009. Publisher: Morgan & Claypool Publishers.