1st Place Solution for ICCV 2021 Workshop SSLAD Track 1-2D object detection challenge

Yingjia Bu, Chengfeng Xu, Shida Zheng, Mengyue Zhang, Jiexiang Wang,
Chenhao Li, Tao Hu, Chenshu Chen, Jin Wang, Dahu Shi, Wenming Tan
Hikvision Research Institute
{buyingjia, xuchengfeng, zhengshida, zhangmengyue, wangjiexiang,
lichenhao, hutao, chenchenshu, wangjin, shidahu, tanwenming}@hikvision.com

Abstract

Autonomous driving technology has been significantly accelerated in recent years because of its great potential in reducing accidents, saving human lives and improving efficiency but turns out to be harder than expected, probably due to the difficulty of labeled data collection for model training. Motivated by recent powerful advances of self-supervised and semi-supervised learning, a promising direction is to learn a robust detection model by collaboratively exploiting large-scale unlabeled data and a few labeled data. Specifically, we design a Multi-instance Detection Contrastive Representation Learning (MDCRL) method for self-supervised learning and an efficient semi-supervised learning solution. We obtain 85.24 mAP, which win the 1st place in ICCV 2021 Workshop SSLAD Track 1-2D object detection challenge.

1. Introduction

Pretraining and finetuning has been the dominant paradigm of training deep neural networks in computer vision. Downstream tasks usually leverage pretrained weights learned on large labeled datasets such as ImageNet [1] for initialization. As a result, supervised ImageNet pretraining has been prevalent throughout the field. Recently, self-supervised pretraining [2, 3, 4, 5, 6, 7, 8, 9] has achieved considerable progress and alleviated the dependency on labeled data.

One of the main restriction of autonomous driving is that the annotation cost of self-driving datasets is much more expensive. Considering that vehicles keep collecting unlabeled data when driving, self-supervised and semi-supervised learning might be a promising method to ease the desire for labeled data and improve model performances.

An obvious representation gap exists between image-level pretraining and the object-level bounding boxes of object detection. Image-level representations are sub-optimal for dense prediction tasks such as object detection. A potential reason is that image-level pretraining may overfit to holistic representations and fail to learn properties that are important outside of image classification. To alleviate this presentation gap, we have two difficult problems must to solve: (1) how to define contrastive learning samples when using non-object centric images, especially for high resolution images with multi-instance. (2) how to bridge the gap in network architecture between pretraining and finetuning. We propose an efficient self-supervised representation learning method (MDCRL) which contains a multi-instance generator module (MIG) and an all Detection network Contrastive Representation Learning module (DetCRL).

MIG training with a few labeled data can get task-specific contrastive learning samples, which are very import for contrastive learning. Object detection often involves dedicated modules, e.g., feature pyramid network (FPN), and special-purpose sub-networks, e.g., R-CNN[10] head. In contrast to common contrastive learning methods where only feature backbones are pretrained, DetCRL performs contrastive representation learning over all the network modules used in detectors, including backbone, neck and head. As a result, all layers of the detectors can be well-initialized.

2. Method

In this section, we first describe the overview pipeline of our solution. And then, we introduce our self-supervised learning method MDCRL explicitly, which play an import role in our solution.

2.1. Overview

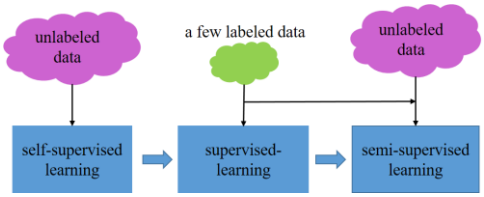

Figure 1 pipeline of our solution

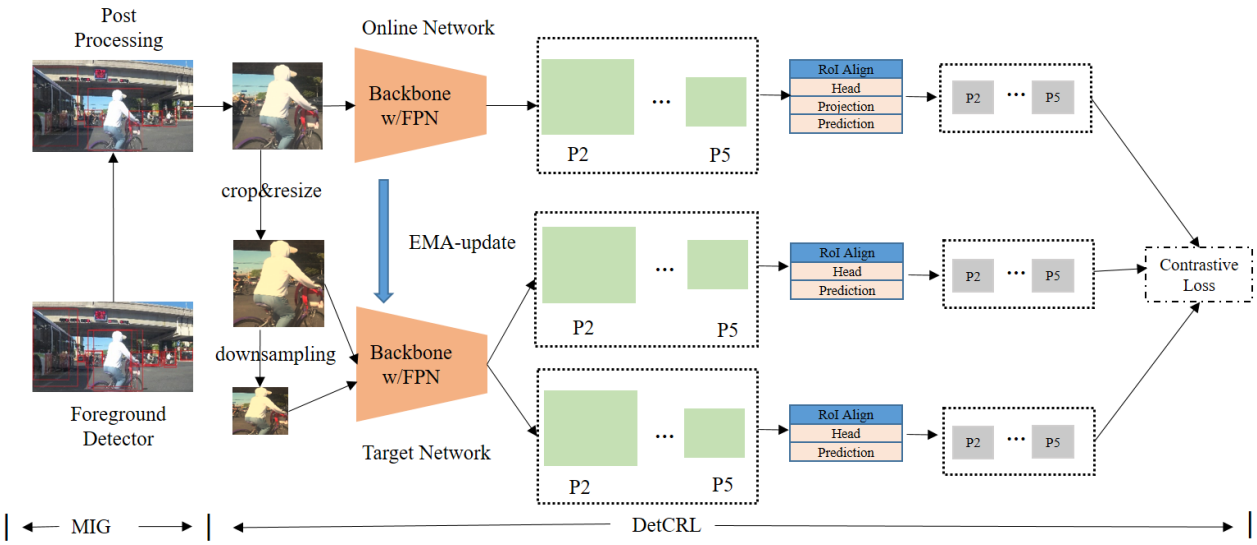

Figure 2 pipeline of MDCRL

As illustrated in Figure 1, the pipeline of our solution contains three parts: (1) self-supervised learning module with large number of unlabeled data; (2) supervised learning module with a few labeled data; (3) semi-supervised learning module with a few labeled data and large number of unlabeled data.

Self-supervised Learning We mainly apply MDCRL method based CNN network (e.g. ResNext, EfficientNet-v2) on SODA10M datasets[15].

Supervised Learning We apply Cascade-RCNN[14] detector on training dataset and validation dataset. All networks are well-initialized with MDCRL pretraining network except RPN modules used in detectors. Additionally, multi-scale training, flipping and mirrors are basic configurations in our experiments.

Semi-supervised Learning In order to using unlabeled datasets more efficiently, we also apply Unbiased Teacher[17] and Soft Teacher[18] two semi-supervised learning methods with some practical tricks(e.g. Mixup, Moasic).

2.2. MDCRL

Self-supervised learning method (MDCRL) contains two key modules: MIG and DetCRL.

As illustrated in Figure 2, MIG, a multi-instance generator, contains a Foreground Detector and a Post-Processing module. As the image is not non-object centric, which background accounts for more, but not importantly. First, we train the foreground detector with a few labeled data (5k images for training) to get many task-specific instances. Second, we merge some instances with an IOU@0.3, getting the final task-specific contrastive learning samples. Through the above two steps, we get object centric image crops which are very important for contrastive learning.

As illustrated in Figure 2, DetCRL, an all Detection network Contrastive Representation Learning module, learns robust representation abilities from large unlabeled images.

MDCRL is very similar with SoCo[16] except data preprocessing. We replace selective search[13] module with our MIG module that gets task-specific contrastive learning samples and has better performance. Due to the limited space, for more details about DetCRL, you can get from SoCo.

3. Experiments

In this section, we first introduce the dataset and implementation details. Finally, we report the results on the challenge test server.

3.1. Dataset and Evaluation Metrics

SODA10M is a large-scale 2D object detection dataset for Autonomous Driving, which contains 10 million unlabeled images and 20k images fully-annotated with 6 representative categories (pedestrian, cyclist, car, truck, tram, tricycle). For labeled set there are 5K images for training, 5K images for validation and 10K images for testing. To improve diversity, the images are collected every ten seconds per frame within 32 different cities under different weather conditions, periods and location scenes.

For this task, we use Mean Average Precision(mAP) in COCO[12] 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.

3.2. Implementation details

The Pytorch framework is employed to implement our method. In our experiments, we use two CNN-based networks: ResNext101-32-16d and EfficientNet-v2. We select the Cascade-RCNN as our object detector, which gets better performance in many common detection datasets, e.g. COCO detection datasets.

In self-supervised training phrase, we only use 10M unlabeled images; LARS optimizer; SyncBN[11] in backbone, neck, and head; 64GPUs with 32 samples per GPU; training 50epochs with one image selecting 4 samples preprocessing by MIG module as default.

In semi-supervised training phrase, we only use 100K unlabeled images for time and GPUS constraint and we use two semi-supervised learning methods (Unbiased Teacher and Soft Teacher).

Validation To verify the MDCRL method effectiveness, we first training ResNext101-32-16d in 5K images for training and test in 5K images for validation. The experimental results with different pretrained method are shown in Table 1. MDCRL can get robust representation abilities, the model improve the mAP from 65.2 to 80.1, demonstrating the effectiveness of our method.

		Cascade RCNN 1x
Pre-trained Dataset	Method	mAP
IN-1K	Super.IN	65.2
SODA10M	MDCRL	80.1(+14.9)

Tabel 1. Comparisons with different pretrained method.IN-1K means ImageNet 1K training images

Test Firstly, we only train single model ResNext101-32-16d in 10K images for training and validation to verify my method furtherly, testing the results on the challenge test server. In test server experiments, we use multi-scale test and soft-nms as defaut for getting a higher performance. The experimental results is shown in Table 2.

		Cascade RCNN 1x
Pre-trained Dataset	Method	mAP
SODA10M	MDCRL	83.55

Tabel 2. ResNext101-32-16d test performance with supervised learning

To get higher performance, we also train EfficientNet-v2 with the same config as ResNext101-32-16d. An ensemble model (ResNext101-32-16d and EfficientNet-v2) furthers improve the performance to 84.45. The experimental results is shown in Table 3.

		Cascade RCNN 1x
Pre-trained Dataset	Method	mAP
SODA10M	MDCRL	84.45

Tabel 3. ResNext101-32-16d and EfficientNet-v2 test performance with supervised learning

Finally, we use semi-supervised learning method, including Unbiased Teacher and Soft Teacher, getting the highest performance. An ensemble model further improve the score to 85.24. The experimental results is shown in Table 4.

		Cascade RCNN 1x
Pre-trained Dataset	Method	mAP
SODA10M	MDCRL& Semi-supervised learning	85.24

Tabel 4. Ensemble model test performance with MDCRL&semi-supervised learning

Table 5 shows the final results of ICCV 2021 Workshop SSLAD Track 1-2D object detection challenge.

User	Team Name	mAP
How	HRI_HOW	85.24
Newer2021	GDW-ML	81.79
IPIU-XDU		81.27

Tabel 5. Final test performance, our submission achives 85.24 and wins 1st place

4. Conclusion

In this work, we proposed an efficient framework collaboratively exploiting large-scale unlabeled data and a few labeled data to learn a robust detector. The proposed framework outperforms the state-of-the-art methods by a large margin on this challenge. The key contribution is that we proposed the self-supervised method MDCRL made the network had a well-initialized exploiting large-scale unlabeled data. Our solution outperforms the others significantly and wins the first place.

Due to time and GPUS constraint, we do not try more backbones, e.g. Transformer-based networks[19]. In the future, we will explore more excellent framework exploiting large-scale unlabeled data.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009

[2] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018

[3] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733, 2020.

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020.

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020

[6] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. arXiv preprint arXiv:2005.10243, 2020

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882, 2020

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. arXiv preprint arXiv:2011.10566, 2020

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497, 2015

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, pages 448–456. PMLR, 2015

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014

[13] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. International journal of computer vision, 104(2):154–171, 2013

[14] Cai, Z., & Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6154-6162).

[15] Han, J., Liang, X., Xu, H., Chen, K., Hong, L., Ye, C., ... & Liang, X. (2021). SODA10M: Towards Large-Scale Object Detection Benchmark for Autonomous Driving. arXiv preprint arXiv:2106.11118.

[16] Wei, F., Gao, Y., Wu, Z., Hu, H., & Lin, S. (2021). Aligning Pretraining for Detection via Object-Level Contrastive Learning. arXiv preprint arXiv:2106.02637.

[17] Liu, Y. C., Ma, C. Y., He, Z., Kuo, C. W., Chen, K., Zhang, P., & Vajda, P. (2021). Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480.

[18] Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., & Liu, Z. (2021). End-to-End Semi-Supervised Object Detection with Soft Teacher.arXiv preprint arXiv:2106.09018.

[19] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030