SDGDet: Semi-supervised Domain Generalization Learning for Road-side Corner Case 2D Detection

Jiahao Xu[1], Hancheng Ye[1], Bo Zhang[2], Tao Chen[1], and Jiayuan Fan[3]

[1] School of Information Science and Technology, Fudan University
[2] Shanghai AI Laboratory.
[3] Academy for Engineering and Technology, Fudan University
eetchen@fudan.edu.cn

Abstract. As a major factor affecting the safety of autonomous-driving perception systems, the corner case detection is crucial in recognizing various unexpected and dangerous situations. In this technical report, we introduce a semi-supervised domain generalization 2D corner case detection framework, SDGDet, which has won the third place in the 2D corner case detection SSLAD2022 Challenge at ECCV 2022. Previous existing 2D object detectors are often trained on a single dataset, but fail to recognize new classes from unseen scenes or novel instances from a known class. In contrast, our SDGDet proposes to integrate semi-supervised learning with domain generalization to detect unseen objects before. Specifically, we pre-train the baseline detector on a selected large-scale dataset with a proposed sampling strategy, to avoid the overfitting of the detection model on the target domain. Then, we develop a semi-supervised pseudo label updating strategy for novel corner case instances, and fine-tune the baseline model using both these pseudo-labelled instances and ground-truth labelled images of fused domains to iteratively improve the model's performance. The final experiment results verify the superiority of the proposed SDGDet in detecting the unseen instances or unknown classes, achieving a sum score of 2.83 on the leaderboard.

Keywords: Corner case detection, autonomous driving, pseudo-labeling, SDGDet

1 Introduction

Autonomous driving technologies aim at perceiving the surrounding environment, and controlling the vehicle's motion automatically, which can help to reduce the traffic accident and improve the comfort of driving. With the help of deep learning[15, 5] and large scale driving datasets, current road-side object detection models for autonomous driving application achieves promising performance in detecting common targets, such as pedestrians and vehicles. But it is still difficult to detect the corner cases, such as dogs on highways, baby strollers on roads, and over-turned trucks, which appear rarely but lead to potential safety risks. Considering this, the CODA[6] dataset is released as a real-world road corner case evaluation benchmark for object detection in autonomous driving.

Object detection, as a basic visual task, plays an important role in autonomous driving technologies. Detection algorithms can be generally divided into two categories according to whether there exists a region proposal process. The two-stage methods[2, 14] generate proposals at first then refine and classify these proposals, while the one-stage methods [9, 7] classify and regress the prediction boxes directly. Yolo[11–13, 1] series are one of the representative works of one-stage detection methods, which have wonderful detection accuracy and high efficiency. In this report, we use the yolov5 as our baseline detector.

Although training models on autonomous driving datasets directly achieves good performance, when the training and test dataset are from the same domain and contain similar categories. But when the road have corner case objects, such as animals, barriers, etc., which have not been met by the trained detector model, the detection performance on these corner cases will saturate a lot. The reason is that the learned model on the original source domain data, cannot well generalize to a new domain with changing environment and unseen corner case objects before.

To relieve the above out-of-domain corner case detection problem, we propose a simple-yet-effective semi-supervised domain generalization learning method, called SDGDet, to help the corner case detection. Specifically, apart from the autonomous driving datasets, we also pre-train the baseline detector using images and corresponding labels from Imagenet[3], by a proposed sampling strategy, to get a well-trained baseline model. Then, a semi-supervised pseudo label updating strategy is developed for the novel corner case instances that are missed in the original domain. These pseudo-labelled instances coupled with other ground-truth labelled instances in the same images, are then fused with more images in another dataset to fine-tune the well-trained baseline model. Such pseudo-labelling and model fine-tuning process can be iteratively done, to continuously improve the model's detection performance. The final experiment results verify the superiority of the proposed SDGDet in detecting the unseen instances or unknown classes, achieving a sum score of 2.83 on the leaderboard.

2 The Proposed Method

2.1 The Baseline Detection Model

We use yolov5l6 which is a powerful version on yolov5 as our baseline detector. With the help of CSPDarknet[16] and PANET[8] modules on the backbone and neck, yolov5l6 can make more accurate and efficient prediction than other algorithms on corner case detection task. In this track. we set the number of output classes for detection head as eight, where seven for common classes and one for the novel class.

2.2 Imagenet Random Sampling

Strengthening the model's ability on novel class detection is important on this task. Unfortunately, there are no annotations about novel classes on these two large scale autonomous driving datasets (ONCE[10], SODA10M[4]) that we can use. To make the model have good feature extraction capability for more objects, we thus sample more information that we need from a larger dataset: the Imagenet1k. Specially, we randomly pick about twelve classes as novel classes which may appear on the road, and train our models on sampled Imagnet, ONCE, and SODA10M. We test the trained model on validation set. Compared to train on ONCE and SODA10M directly, the result using the sampled Imagnet brings about 0.1 sum-score improvement.

2.3 Semi-supervised Pseudo-labelling

Fine-tuning the trained model on the validation set is straight-forward and can bring significant improvement. However, the label on validation set is not complete, and fine-tuning directly will hurt the performance. To mitigate this issue, we try to fine-tune the models via a semi-supervised way with two stages. We fine-tune the model on validation set firstly, and get the detection results on SODA10M. In this process, we use the pre-trained model to predict pseudo labels for only the novel unknown classes on SODA10M, and then combine these pseudo labels with SODA10M ground truth labels to get the hybrid SODA10M labels. We then fine-tune the model on the validation set and hybrid SODA10M set again to iteratively optimize the detection model performance for unknown classes detection.

2.4 Overall Procedure

As there are distortions on ONCE dataset, we thus first rectify the ONCE images with the camera matrix to make the bounding box labels more accurate firstly. Then we train the yolov5l6 model on the fused domain data of SODA10M, ONCE, and the sampled imagenet1k as described above. We further fine-tune the model with the hybrid SODA10M generated from pseudo-labelling strategy, then we ensemble all three models (fine-tuning on hybrid SODA10M set, fine-tuning on validation, and fine-tuning on hybrid SODA10M set and validation) with test time augment to report the final prediction results on CODA test set, and the whole procedure are also shown in Figure. 1.

3 Experiment

In this section, we first introduce the datasets and implementation details, then we report the experiment results.

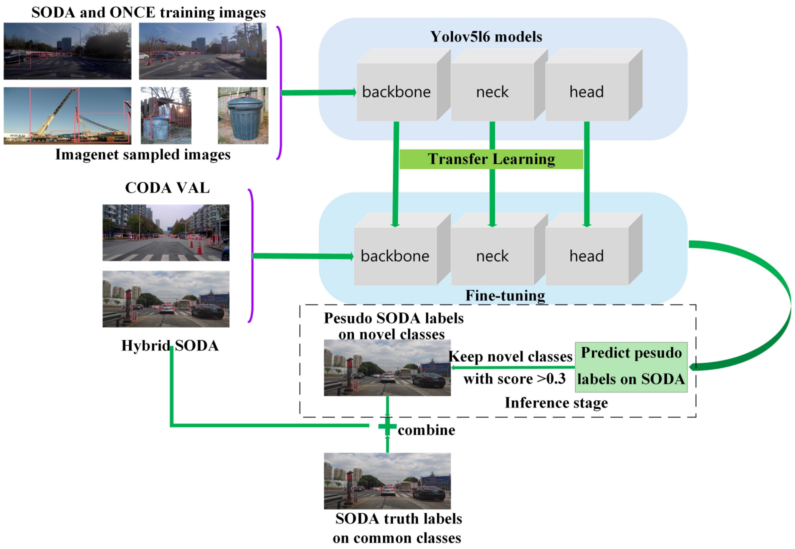
SODA and ONCE training images
Imagenet sampled images
Yolov5l6 models
backbone
neck
head
Transfer Learning
CODA VAL
backbone
neck
head
Fine-tuning
Hybrid SODA
Pesudo SODA labels on novel classes
Keep novel classes with score >0.3
Predict pesudo labels on SODA
Inference stage
combine
SODA truth labels on common classes

Fig. 1. The Overall Procedure

3.1 Datasets and Evaluation metrics

CODA is a corner case detection dataset, which consists of 9768 camera images with 80180 annotated objects from 43 object categories. The first seven classes are considered as common classes, and the rest are novel classes. CODA can be splited as validation set and test set, both of them have 4884 images. And the validation set has six common classes and twenty-three novel classes annotation, while the test set has the full classes annotation. SODA10M and ONCE are two large-scale real-world autonomous-driving datasets, which have six (pedestrian, cyclist, car, truck, tram, tricycle) and five (car, bus, truck, pedestrian, cyclist) classes labels.

We use the sum score of AP-common, AP-agnostic, AR-agnostic, AR-agnostic-corner for evaluation, where AP-common is the mAP over objects of common categories, AP-agnostic is the mAP over objects of all categories in a class-agnostic manner, AR-agnostic is the mAR over objects of all categories in a class-agnostic manner, and AR-agnostic-corner is the mAR over corner-case objects of all categories in a class-agnostic manner.

In this competition, we use SODA10M and ONCE training set, imagenet1k and CODA validation set for training, and CODA test for evaluation.

3.2 Implementation details

We pick yolov5l6 as our baseline. In all training stages, we reshape the image input size to 1960, with batch size 16. We choose SGD as our optimizer. And all other hyper-parameters are kept as default. In Imagenet sampling, we filter including ids of n02747177,n03126707,n03384352,n03496892,n03743016,n03785016, n03967562 ,n03976657,n04509417,n04604644,n06794110,n06874185 with their images and labels as novel labels. In Semi-supervised Pseudo-labelling, we set the score threshold as 0.3 to filter the reliable novel labels from the first stage of model's detection results. All experiments are run with eight Nvidia RTX 3090.

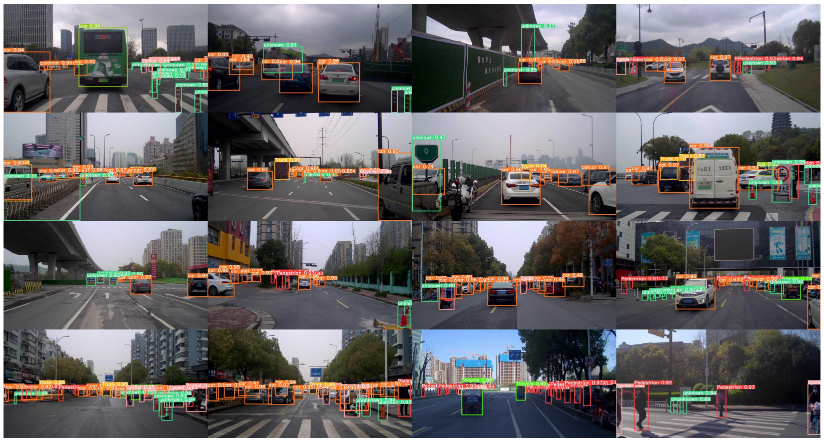

Fig. 2. detection results

3.3 Experiment result

The experiment result shows that our method achieves the third place on the leaderboard (with the sum of score 2.83), and the final leaderboard was shown in Table1. We also show the effectiveness about Semi-supervised Pseudo-labelling and the model ensemble in Table2, and some detection examples in Figure.2.

4 Conclusion

This report details the key techniques used in the ECCV2022 SSLAD Track 3 - Corner Case Detection competition. To help the model with good feature extraction capability on novel classes, we sample relative classes from Imagenet to enrich the datasets, and fine-tune the model using a developed semi-supervised domain generalization way. The experiment results show that our method can detect novel classes on the road more effectively. And with the muti-model ensemble as described last, we finally get the third place with an overall score of 2.83.

Table 1. The top eight results in the leatherboard, where AR-c means AR-agnostic-corner, AR-a means AR-agnostic, AP-a means AP-agnostic, and AP-c means AP-common.

usrname	Sum	AR-c	AR-a	AP-a	AP-c
gavin	3.09	0.80	0.85	0.78	0.66
IPIU-XDU	3.06	0.79	0.85	0.77	0.64
ours	2.83	0.76	0.81	0.70	0.55
Charles	2.83	0.73	0.79	0.72	0.58
BingDwenDwen	2.82	0.73	0.79	0.72	0.58
chenwei	2.82	0.74	0.79	0.72	0.58
wyndy	2.79	0.72	0.78	0.71	0.56
relax	2.77	0.71	0.78	0.71	0.57

Table 2. The direct finetune means that fine-tuning on CODA validation set. fintune with pseudo label means fine-tuning on hybrid labeled SODA10M and CODA validation set, and the model ensemble means the finally fused models.

method	Sum	AR-c	AR-a	AP-a	AP-c
direct finetune	2.63	0.71	0.76	0.66	0.50
fintune with pseudo label	2.80	0.75	0.81	0.70	0.54
model ensemble	2.83	0.76	0.81	0.70	0.55

References

1. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
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. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. 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)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017)
6. 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)
7. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)

8. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8759–8768 (2018)
9. 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)
10. Mao, J., Niu, M., Jiang, C., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., Yu, J., Xu, C., et al.: One million scenes for autonomous driving: Once dataset (2021)
11. 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)
12. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)
13. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
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. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H.: Cspnet: A new backbone that can enhance learning capability of cnn. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 390–391 (2020)