Domain Adaption Semi-Supervised Learning for Vehicle Detection from Drone Image

Youlkyeong Lee† Jehwan Choi and Kanghyun Jo

[1]Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan, South Korea
(Tel: +82-52-259-1664; E-mail: yklee00815, choijh1897@gmail.com, acejo@ulsan.ac.kr)

Abstract: In this study, we demonstrate semi-supervised vehicle detection based on domain adaptation according to the camera viewpoint. Since it is difficult to collect a large amount of data for training, datasets for object detection such as COCO and Pascal VOC are pre-trained and provided as a model. The presented dataset is primarily composed of images captured at eye level, which can result in differences in detection performance depending on changes in camera viewpoint. The collected drone data consists mostly of vehicles captured while filming roads with a drone. Additionally, the captured images show the top surface of the vehicle from a bird's eye view, which differs from images captured at eye level. The teacher-student framework is widely studied for improving the effective performance in semi-supervised object detection. The teacher model is trained using annotated drone data, while the student model generates pseudo-labels in the absence of annotation information. These pseudo-labels are generated using a model trained on COCO or Pascal VOC datasets from different viewpoints than the drone. Using the Efficient Teacher model employed in SSOD, the study adapts drone-view data captured from a bird's eye perspective to improve learning performance. Transfer learning with data trained on COCO-standard and VOC datasets in the experiment shows improved results compared to previous experiments.

Keywords: Semi-supervised Object Detection, Vehicle Detection, Pseudo Label, Domain Adaption

1. INTRODUCTION

Ongoing research in object detection [1-3] has led to significant performance improvements, primarily through supervised learning based on large amounts of annotated data. However, developing object detection models for specific tasks on a limited budget requires a cost-effective solution. Semi-supervised object detection (SSOD) proposes a teacher-student model that enhances learning performance with a small amount of data. By adopting unlabeled data [4] to generate pseudo-labels with ambiguous object position and class information, the student model is trained. The gradually improving pseudo-labels are leveraged to improve object detection performance. This training strategy reduces the cost of manual annotation while improving performance. Currently, SSOD is based on the Two-stage model Faster R-CNN [2], and many studies [5, 6, **?**, 6] are focusing on this approach. This model is advantageous for generating high-quality pseudo-labels through a separate network that guesses the arbitrary object position and multi-stage layers. In contrast, research on one-stage based SSOD is limited, but studies utilizing You Only Look Once (YOLO) series [7-9], which are dense prediction and anchor-based detectors, are more effective than Two-stage. One of issues for SSOD, pseudo-label inconsistency is a persistent issue in these models. Pseudo-labels are produced by using the student model trained with labeled data to extract the position and class of an arbitrary object from unlabeled data. The pseudo-label is then combined with ground-truth bounding boxes used in teacher model training. Therefore, producing high-quality pseudo-labels plays a crucial role in performance. Advancements in technology have made it possible to collect rich images from a bird's-eye view, such as with cameras attached to drones. This allows for the development of applications in various fields. This change in viewpoint has the potential to improve object detection performance by providing information about the top and side views of objects. Based on the aforementioned discussion, the goal of this study is to verify the detection performance of vehicles using drone flight images that are adapted to changes in viewpoint.

Fig. 1: First row is a dataset from COCO dataset and the second row is a collected drone dataset. Between datasets it is a different perspective view.

2. PROPOSED METHOD

2.1. Efficient Teacher

Efficient Teacher [10] constructs an SSOD framework using the YOLOv5 one-stage model. YOLOv5 generates high-density predictions, demonstrating better performance and speed than conventional two-stage models. The efficient teacher overcomes the challenge of generating low-quality pseudo labels in SSOD using the Pseudo Label Assigner method. The paper proposes two threshold values, high and low, for the pseudo label score (τ_1 and τ_2, respectively) to more precisely select high-quality

† Youlkyeong Lee is the presenter of this paper.

pseudo labels. Additionally, the paper reduces the number of parameters by changing the output of FPN from 5 to 3 using Dense Detector on the ResNet-50-FPN backbone. The objectness score is added to determine the presence of an object in the generated pseudo label of the bounding box.

2.2. Domain Adaption

Currently, the commonly used datasets for training object detection models are COCO and VOC, which mainly consist of images captured at the human eye level. However, these datasets show low detection rates for images with no distribution in the data used for training or images with different viewpoints, such as top-down views of objects. To improve performance in adapting to various viewpoint changes, the drone dataset $\mathcal{D}$ captured from a bird's eye view is used for training. $\mathcal{D}$ defines the drone data domain, where x is the data that exists in the domain, and $P(x)$ is the probability distribution for the data x, defined as $\mathcal{D}(x, P(x))$. The distribution of high-quality pseudo-labels generated from unlabelled images and the teacher model is learned using the Gradient Reverse Layer [11] to distinguish difficult-to-distinguish data from the prediction distribution generated by the student model trained on labeled images.

3. EXPERIMENT

Configuration Details: The images used in the learning process were resized to 960×960. A learning rate of 0.001 is utilized, and the optimizer employed is Adam. Focal loss is chosen as the loss function. Four NVIDIA A100 GPUs, each with 40GB of memory, are used, with a batch size of 8.

The collected drone video contains 15,878 frames, and the proposed classes include five categories: person, car, truck, bus, and motorcycle. Semi-supervised object detection typically evaluates performance on five different data conditions. As shown in Table 1, the dataset is divided into labeled and unlabeled sets, and the quantity of labeled data used for training is varied at 1%, 2%, 5%, and 10% of the entire dataset. Additionally, experiments are conducted with 30% and 50% of labeled data.

Table 1: Drone train and test dataset for vehicle state classification

Dataset	Label data	Unlabel data	Val
1%	**127**	**12,576**	**3,175**
2%	**254**	**12,449**	**3,175**
5%	**635**	**12,068**	**3,175**
10%	**1,270**	**11,433**	**3,175**
30%	**4,763**	**11,115**	**3,175**
50%	**7,939**	**7,939**	**3,175**

4. CONCLUSION

The ongoing project focuses on semi-supervised vehicle detection using a domain adaptation-based bird's eye view drone dataset. The collected drone dataset is divided into labeled and unlabeled data, and the project analyzes the detailed experimental results and semantic evaluations of the proposed domain adaptation method and experimental strategy.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the government(MSIT).(No.2020R1A2C200897212)

REFERENCES

[1] Ross B. Girshick. Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[2] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[4] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *ArXiv*, abs/1905.02249, 2019.

[5] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Neural Information Processing Systems*, 2019.

[6] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3131–3140, 2021.

[7] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *ArXiv*, abs/2004.10934, 2020.

[8] Glenn R. Jocher, Alex Stoken, Jiří Borovec, NanoCode, Ayushi Chaurasia, TaoXie, Liu Changyu, Abhiram, Laughing, tkianai, and yxNONG. ultralytics/yolov5: v5.0 - yolov5-p6 1280 models, aws, supervise.ly and youtube integrations. 2021.

[9] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 1 2023.

[10] Bowen Xu, Mingtao Chen, Wenlong Guan, and Lulu Hu. Efficient teacher: Semi-supervised object detection for yolov5. *ArXiv*, abs/2302.07577, 2023.

[11] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. *ArXiv*, abs/1409.7495, 2014.