Enhancing Autonomous Navigation Few-Shot 3D Object Detection Using Depth Supervision

Ishita Mangla
Department of Computer Science
Stanford University
imangla@stanford.edu

Ishan Sabane
Department of Electrical Engineering
Stanford University
ishancs@stanford.edu

Tejas Y. Deo
Department of Mechanical Engineering
Stanford University
tejasdeo@stanford.edu

Motivation or Objective

Recent advancements in autonomous navigation and robotics require understanding of the surrounding 3D world by sensor data. Object detection for detecting cars, pedestrians, traffic signals etc. in the form of 2D/3D bounding boxes provides information crucial for collision free navigation for autonomous vehicles. Using only monocular camera images, YOLO, [1] SSD [2] and Fast RCNN [3] are promising for object detection. However, these approaches do not scale to novel categories without retraining on new object categories. Detecting novel or uncommon objects such as carts, animals, construction zone objects etc. is challenging as current models cannot adapt to new classes using few labelled images. Meta and transfer learning seem to be the suitable and promising approaches for solving this limitation. We propose to explore these methods for few shot object detection [4] on datasets such as KITTI [5], Waymo [6], Nuscenes [7] and FSOD [4]. Additionally, we plan to augment our model to understand 3D bounding boxes from the predicted 2D bounding boxes by using concepts from prior works in monocular 3D object detection [8, 9]. Such a model would enable novel 3D object detection using few shot with 2D bounding boxes and enable incremental learning to allow the model to learn new objects with time.

Related Work

State of the art methods in 2D object detection focus on either a Region Proposal Network enhanced by a Feature Pyramid network [10] followed by an object classifier [3] or design the model architecture such that the model outputs the bounding boxes and the image classes directly as the output [1] often called as the single shot object detectors

Recent works in few shot object detection [11, 12] focus on improving classifiers to improve the classification on the proposed regions. Other meta learning methods improve the Region Proposal Network Module by using attention mechanisms [13] to enable few shot object detection. Methods such as YOLO and SSD directly predict the bounding box parameters without using an intermediate object localization step. This enables image feature extraction followed by bounding box regression for training meta learning approaches [14]

Technical Outline

Our objective is to predict 3D bounding boxes on the query dataset images after meta-training for 2D bounding box prediction on the support set. To achieve this, we plan to use an optimization-based meta-learning method. This method involves an inner update to predict 2D boxes and an outer update to predict 3D bounding boxes. We intend to leverage color and depth images to enhance the accuracy of 3D localization estimates [8]. Considering the KITTI dataset, and a few others (as the KITTI dataset has very few classes), we plan to use the pre-trained DPT [15] model to extract depth images when color images are the only input available. In this scenario, the color images will include ground truth 2D annotations, which will be valuable for inner loop updates.

Our approach will utilize two robust backbones for obtaining embeddings of color and depth images, namely the pre-trained VGG-16 or VGG-19 [16], and the DPT model, respectively. These embeddings will be concatenated and passed through both a simple sequence model (LSTM) and a powerful model (Transformer) to regress 3D bounding box coordinates in the query dataset (consisting of 8 coordinates in 3D space). This will allow us to compare the effect of different types of sequence models on the object detection performance. We plan to perform ablation studies which involve effect of depth image, choice of image feature extractor and effect of object categories on performance.

Team Contributions

Since we are building the code-base from scratch, the three of us are responsible for processing and loading one dataset each. For the technical implementation, we have decided to meet up after the lectures to work on the model architecture, training, testing and evaluation setup together. The experiments would be distributed equally once we completed the coding sections.

References

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[2] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[3] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[4] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, 2020.

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[6] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[8] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017.

[9] Jason Ku, Alex D Pon, and Steven L Waslander. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11867–11876, 2019.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[11] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020.

[12] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 456–472. Springer, 2020.

[13] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4013–4022, 2020.

[14] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019.

[15] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.