# Enhancing Monocular 3D Object Detection in Foggy Conditions: An Adapted MonoCon Approach for Autonomous Vehicles

Tung Do
*Department of ECE*
*University of Michigan, Ann Arbor*
Ann Arbor, USA
tungsdo@umich.edu

Xirong Liu
*Department of Mechanical Engineering*
*University of Michigan, Ann Arbor*
Ann Arbor, USA
xirongl@umich.edu

Rahul Swayampakula
*Department of Robotics*
*University of Michigan, Ann Arbor*
Ann Arbor, USA
rahulswa@umich.edu

***Abstract*—This paper explores advancements in monocular 3D object detection, a pivotal aspect of autonomous vehicle technology. We focus on enhancing detection accuracy and robustness in diverse weather conditions, specifically addressing the challenges in foggy scenarios. Implementing the MonoCon model [13], our methodology includes transfer learning, image augmentation techniques, and pre-processing strategies to improve visibility in foggy images. Challenges such as fluctuating Average Precision (AP) values and inefficient detection of distant or small vehicles in fog are addressed through a revised evaluation strategy and targeted image processing. Results showed an increase in AP from 7.05% to 17.67% for the normal dataset after training to more epochs and up to 25.82% for foggy conditions after training to 300 more epochs and applying CLAHE and blur. These findings underscore the model's adaptability and effectiveness in diverse environments.**

***Index Terms*—Monocular 3D Object Detection, Autonomous Vehicles, Deep Neural Networks (DNN), Deep Learning in Computer Vision**

## I. Introduction

Object detection is a crucial computer vision component, focusing on identifying and categorizing objects in images through 2D bounding boxes. The advancement of deep learning has significantly enhanced 2D object recognition, drawing extensive academic interest. Innovative models like Faster R-CNN [1], RetinaNet [2], and FCOS [3] have been instrumental in propelling the domain forward, contributing profoundly to varied applications, notably in autonomous vehicle technology.

While 2D information contributes significantly to object detection, it falls short in enabling intelligent systems to grasp the complexities of the three-dimensional world fully. This limitation is particularly evident in autonomous vehicles requiring precise 3D spatial understanding for safe navigation. Consequently, 3D object detection has garnered increasing importance, especially in robotics. Despite the state-of-the-art methods predominantly relying on detailed 3D data from LiDAR point clouds [4-6], the high cost of LiDAR systems poses a challenge. Thus, monocular 3D object detection, offering a more economical and straightforward deployment option, has emerged as a critical area of research.

Recent innovations in this domain focus on diverse approaches. For instance, MonoRUn emphasizes reconstruction and uncertainty propagation, enhancing localization accuracy in 3D space [7]. FCOS3D introduces an anchor-free method for object detection [8], while MonoEdge leverages local perspectives for detection [9]. MonoXiver, with its bounding box denoising technique [10], marks a significant advancement in detection accuracy. Moreover, methods like MonoNeRD utilize NeRF-like representations for continuous 3D geometry prediction [11], and innovative approaches in occupancy learning further enhance detection capabilities [12]. Inspired by these advancements and the foundational MonoCon model[13], this paper aims to refine monocular 3D object detection. We focus on leveraging monocular contexts as auxiliary learning tasks, drawing from the Cramer-Wold theorem to introduce effective representations in monocular 3D object detection. Our methodology involves a Deep Neural Network (DNN) based feature backbone, multiple regression head branches for essential parameter learning, and auxiliary context branches for performance enhancement. This approach, tested in the KITTI benchmark, demonstrates competitive accuracy and speed, confirming the feasibility and effectiveness of monocular 3D object detection in practical applications.

Our evaluation of the MonoCon model[13] was conducted using a large-scale dataset that included both clear and adverse weather conditions, focusing on normal and foggy weather scenarios to enhance autonomous vehicle safety. We achieved an Average Precision (AP) score of 25.82% in this benchmark's camera tracking category. This approach not only achieved a notable increase in AP under regular weather but also demonstrated significant enhancements in foggy conditions, validating the effectiveness of our adapted MonoCon model[13] in diverse environmental scenarios. The results highlight the model's adaptability and robustness in varying weather, underscoring its potential in real-world applications.

# II. Methodology

The core challenge of this project centers on 3D object detection using a Monocular RGB image. The objective is to accurately predict the objects' type, position, size, and orientation within these images. This task is inherently complex due to the lack of depth information typically available in stereo vision or LiDAR-based systems. We approach using MonoCon-based architecture, which effectively infers 3D spatial relationships from 2D data, relying heavily on the contextual and visual cues present in the RGB images to tackle this objective.

## A. Summary of the MonoCon

The MonoCon [13] framework employs an elegantly streamlined design comprising three fundamental components:

1) Feature Backbone: Like many 3D object detection networks, MonoCon [13] also has a deep learning-based feature backbone in the architecture. Given an input RGB image of dimensions $3 \times H \times W$, a feature backbone $f(;\Theta)$ is used to compute the output feature map F of dimensions $D \times h \times w$. We use the DLA network(DLA-34) in our implementation, as mentioned in the MonoCon paper.
2) 3D Bounding Box Regression Heads: These bounding box regression heads are responsible for estimating the 3D bounding box location of the object. The model regresses on several detection heads, which primarily perform the tasks of 2D bounding box center Heatmap, offset vector, depth uncertainty, shape dimensions, and observation angle. This module is useful for both training and inference.
3) Auxiliary Context Regression Heads: The approach leverages four distinct types of projection information from 3D bounding boxes as auxiliary learning tasks. These encompass heatmaps depicting the projected key points, offset vectors corresponding to the 8 projected corner points, dimensions of the bounding box, and quantization residuals of keypoint locations. These supplementary components are exclusively employed during the training phase, exerting a positive influence on the network's capacity to refine its estimations for improved bounding box detections.

## B. Summary of Our Approaches

The project employs a model architecture and loss function similar to the Monocon paper[13] . Further, we enhanced the training strategy and advanced image augmentation techniques and explored various modifications to the architecture to improve the performance of 3D object detection on the given test dataset.

# III. Experiments

In this section, we will describe the datasets utilized, the metrics employed, and the enhancements made to both the model and data to optimize the performance of 3D object detection.

## A. Datasets and Metrics

*1) **Datasets**:* We were supplied with a custom synthetic dataset capturing urban traffic scenes in standard KITTI dataset format [14]. This comprehensive database encompasses 1989 images of resolution 1242x375, including test and train images. The training subset consists of 1229 images earmarked for training and validation purposes. Ground positions and semantic labels are provided within this training set, serving as essential components for training and evaluating model performance during the validation phase.

The testing set comprises the remaining 760 images, serving as the benchmark for evaluating the trained model using Autograder. Unlike the training set, the testing set lacks ground truth semantic labels but includes images and poses for evaluation. Notably, the training set is further categorized into two distinct components: Normal and Bonus. The Normal set features synthetic images within clear scenarios, while the Bonus set introduces traffic scenarios in foggy environments, presenting a more challenging landscape for object detection.

*2) **Metrics**:* The primary metric for assessing our model's performance is the average precision (AP) on detections, aligning with the official KITTI 3D object detection evaluation protocol. Specifically, per the instructor's detail, we utilize the metric $AP_{3D}|R40$ @ IoU = 0.5 (moderate difficulty) for performance evaluation.

## B. Experiments

Further in this section, we shall discuss the series of experiments conducted to enhance the performance of our model in 3D object detection on our dataset.

Initially, our model underwent training with randomly initialized weights on the custom dataset, employing a batch size of 8 for 200 epochs. During the evaluation of the training dataset, we achieved a notable Average Precision (AP) value of 17.67 when assessed on the normal testing dataset. However, when subjected to the Bonus testing dataset, our model demonstrated a lower AP of approximately 5.

Upon analyzing the model's performance on the testing dataset, a noteworthy observation surfaced—our model exhibited challenges in detecting a significant portion of cars within foggy scenes, even with 2D detection. This underscores a critical limitation in our current detection capabilities, prompting the need for more effective training strategies to improve detections within such scenes. Recognizing the need for enhanced accuracy in detections, the conventional training approach for more significant epochs was considered. However, this approach carries the risk of overfitting, particularly given the distinctive nature of the bonus test dataset, which diverges significantly from the training dataset.

We adopted a strategic set of approaches to strike a balance between achieving better detection accuracy and preserving model generalization. Recognizing the divergence between the bonus test dataset and the training dataset, we implemented the following measures:

1) Transfer Learning: We adopted an iterative training approach, capitalizing on the strengths of the existing

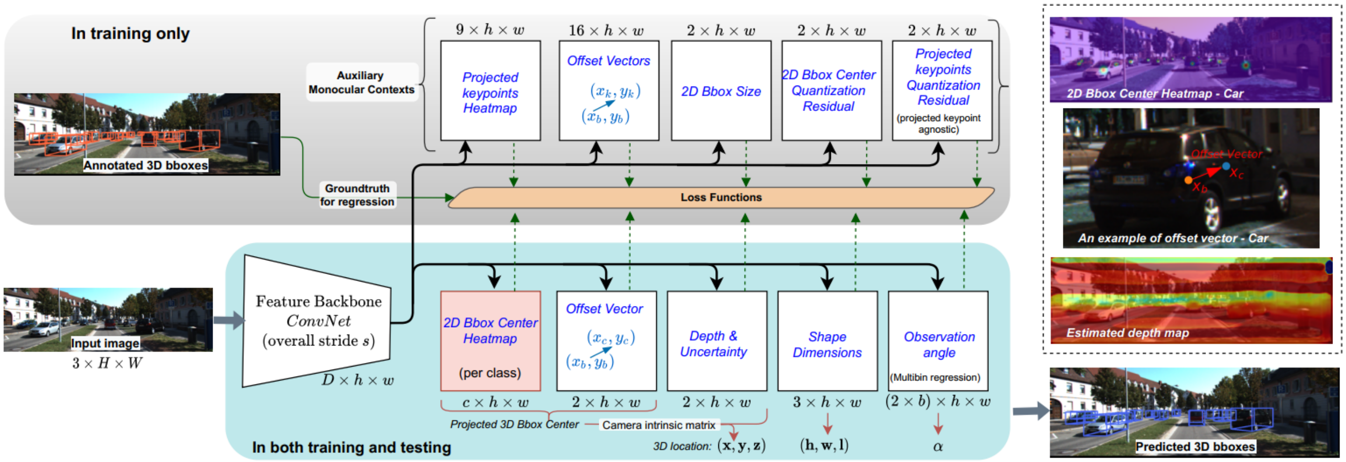

Fig. 1. MonoCon Architecture [13].

MonoCon model [13]. Our strategy involved continuing the training process using the best-performing weight file (.pth) as a starting point. This file was selected based on its superior performance metrics in previous epochs. By resuming training from this advanced state, we aimed to refine further and elevate the model's accuracy.
Leveraging pre-trained models as a starting point for our training process, capitalizing on the knowledge gained from a broader dataset. This aids in initializing our model with features useful for 3D object detection, reducing the risk of overfitting.

2) Image Augmentation Techniques: We implemented a series of image augmentation techniques to bolster the model's robustness and ability to generalize across diverse scenarios. These included:
    - Cropping: Adjusting the frame of the input images to present varied perspectives to the model.
    - Contrast Enhancement: As the training dataset is much darker with less visibility of objects, we amplified the contrast levels in the images to accentuate features, aiding the model to learn better feature detection and classification.
    - Blurring: Introducing blur effects to simulate real-world scenarios where images may not always be perfectly sharp, especially the foggy datasets with less clear vision of vehicles in front of them, thus preparing the model for a broader range of input conditions.
    - Rotation: Random flips in vertical and horizontal directions are increased to increase the robustness.

    These augmentations were carefully integrated into the preprocessing pipeline, ensuring a balance between realism and variance in the training dataset. This strategy enhanced the model's performance, particularly in challenging and diverse real-world scenarios, where lighting and clarity vary significantly.
3) Improving foggy images: We introduced the pre-processing step to the testing images to improve the detection quality. The visibility in the foggy images is too low for the object, which is relatively far away in the scene. We apply pre-processing to the image to enhance its quality before processing it through a neural network. We apply CLAHE to the input image to improve the contrast of the images, which helps in better detection.
4) Variation of backbone: We further enhanced our approach by substituting the backbone feature extractor in the architecture with alternative variants of the DLA. Remarkably, this modification did not significantly impact the detection quality compared to previous methods. Consequently, we decided to maintain the existing architecture for our continued efforts.
    In summary, we introduced several image enhancement modules for training and testing datasets and improved network initialization through pre-trained weights. This strategic augmentation optimized the network's performance on the test dataset and enhanced its overall generalizability.

## IV. Results and Evaluation

In this section, we delve into the outcomes of our experiments and elucidate the strides we've made in enhancing our network's performance across diverse datasets. Our approach involves a comprehensive comparison between our final evaluation and the training of the original Monocon[13]. Leveraging pre-trained weights and systematically assessing performance gains through augmentations form the crux of our methodology.

Consistency has been a cornerstone of our training regimen. We maintained a steadfast learning rate of 0.000225 throughout our experiments and a learning rate decay set at 1e-5. The incorporation of the DLA backbone in the majority of our trials proved instrumental in achieving optimal performance. Furthermore, we adhered to all other parameters in alignment with the official implementation of the Monocon [13] network.

TABLE I
RESULTS ON NORMAL TEST DATASET

| Model | mAP (%) |
|---|---|
| $\text{epochs}_{100}$ | 7.05 |
| $\text{epochs}_{200}$ | 10.23 |
| $\text{epochs}_{130}(best)$ | 17.67 |
| pre-train | 21.51 |

TABLE II
RESULTS ON FOGGY DATASET

| epochs | Augmentation / Processsing | mAp(%) |
|---|---|---|
| 130 | None | 7.89 |
| Pretrain | None | 10.1 |
| Pretrain | Clahe | 11.9 |
| Pretrain + 300 | Clahe | 16.2 |
| Pretrain + 300 | Blur and Clahe | 25.82 |

### A. Quantitative Analysis

Our benchmark for evaluation centers around the average mAP value, gauging the efficacy of object detections across our dataset. Initially, we subjected our model to a rigorous assessment on a Normal test dataset, undertaking approximately 100 epochs of training. This initial phase yielded an accuracy of around 7 AP on the normal dataset. After this, an extended training period of around 200 epochs was conducted, and the model selected for analysis exhibited the maximum AP instead of the final epoch model. Additionally, we tested the model on a Normal test dataset, utilizing a pre-trained model. The comparative results are detailed in Table I.

Remarkably, the pre-trained model showcased superior performance compared to its counterpart trained on synthetic datasets. This discrepancy is attributed to fewer training iterations and the pre-trained model's exposure to a more diverse and realistic range of sequences that are inherently challenging to predict. This infusion of robustness and generalization equips the network to excel in this scenario, underscoring the value of leveraging pre-trained models for improved performance.

Our primary focus is modifying our approach to enhance detection capabilities in foggy scenarios. We systematically assess the model's performance under various augmentation and training scenarios, and the outcomes are presented in Table II. Despite conducting tests with multiple settings, the table exclusively highlights salient configurations that led to a notable improvement in accuracy over the foggy dataset using our model.

This observation distinctly underscores the noticeable improvement in detection quality within foggy scenarios attributed to our augmentation techniques. Notably, the pre-train weights demonstrated commendable performance on normal test cases. Subsequent training from the pre-train stage ensures the model converges in the right direction. Augmentation and image enhancements then play a pivotal role in augmenting robustness and elevating the overall quality of detections.

### B. Qualitative results

Additionally, we conducted a qualitative evaluation of our results to debug the model and facilitate valuable experiments. Our model demonstrated exceptional performance in the normal dataset, accurately detecting most cars and their 3D bounding boxes.

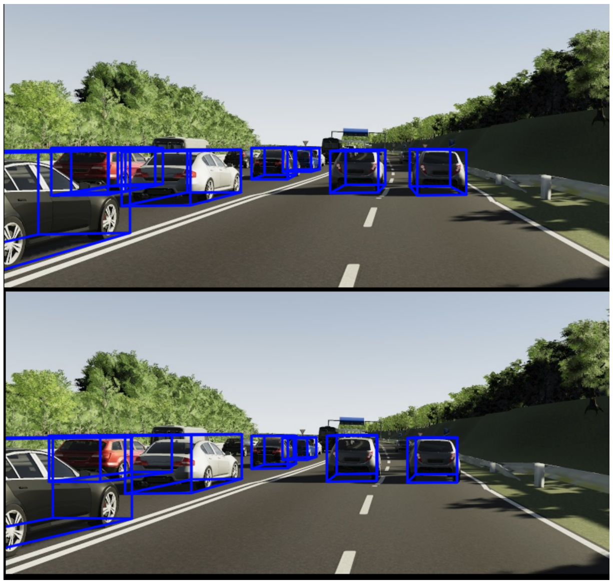

Fig. 2. (Top to bottom) Testing result on Normal dataset using 200 epochs and pretrain models

In foggy scenarios, the detection of objects positioned farther from the ego vehicle posed a considerable challenge. However, employing the earlier methods improved the model's capability to detect vehicles in fog at a reasonable accuracy, even under challenging visibility conditions. The results are shown in Figures 3 and 4.

Notably, as the vehicle traverses through the scene, the dynamic changes in visibility introduce variability in the results. To address this, we propose a few ideas, which will be elaborated upon in a subsequent section.

## V. CHALLENGES AND SOLUTIONS

Implementing our project presented unique challenges, particularly when adapting the MonoCon model [13] from the provided repository. Initially, testing the 'best.pth' file yielded improvements, but not to the desired extent. To overcome this, we extended training to 800 epochs. The principal challenge lies in training the model on a multi-bin architecture like Monocon, which optimizes multiple objective functions simultaneously. Our specific focus is on maximizing the mean average precision (AP) values for detection. This proves to be particularly challenging due to the inherent oscillations during training induced by the complexities of multi-objective optimization. To mitigate this, we revised our evaluation strategy from assessing every 10 to every 5 epochs. This adjustment allowed us to monitor the model's performance more frequently, enabling us to pinpoint and select the iteration exhibiting the highest AP more effectively.

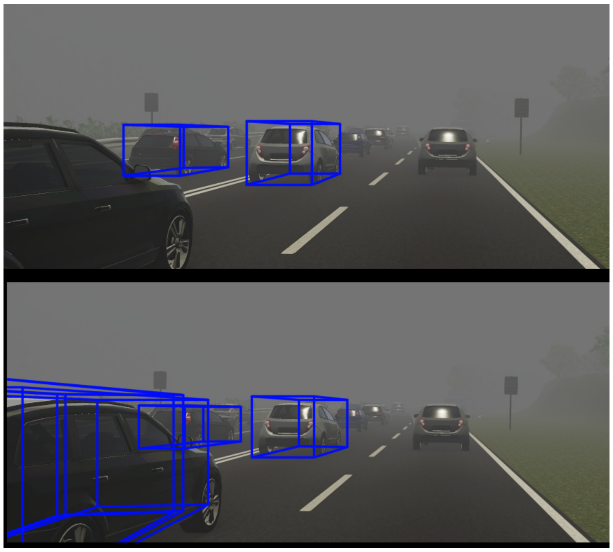

Fig. 3. (Top to bottom) Testing result on foggy dataset using 200 epochs and pretrain models

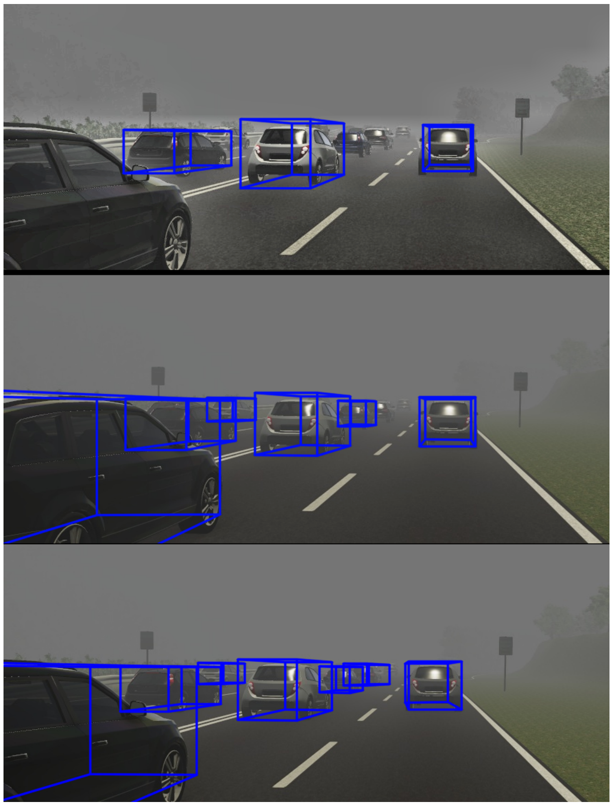

Fig. 4. (Top to bottom) Testing result on foggy dataset using augmentations, $pretrain + clahe$, $pretrain + 300 + clahe$, $pretrain + 300 + blur + clahe$

Additionally, we observed a peculiar inconsistency in the model's performance: when presented with two foggy images, the model could detect a specific vehicle in one picture but failed to detect the same vehicle in the other image, as illustrated in the picture below. This inconsistency in the detection error was highlighted and identified as a limitation in the model's ability to consistently recognize under varying environmental conditions like fog. The model likely lacked exposure to a broad range of scenarios, particularly those involving fog and distant or small vehicles, hindering its ability to generalize.

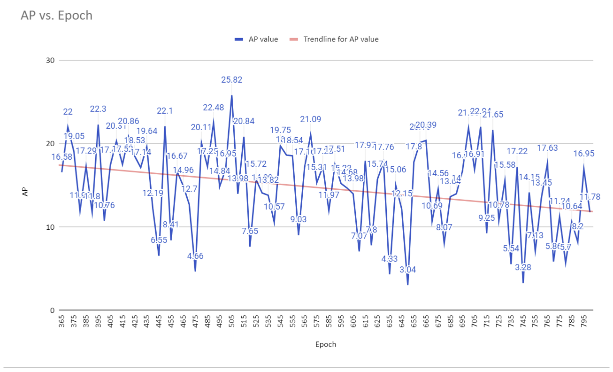

Fig. 5. Oscillations in the AP values

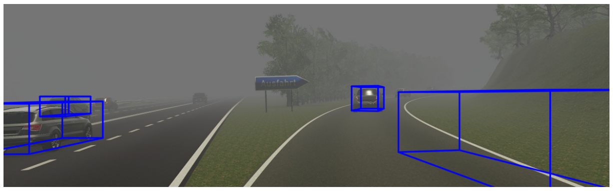

Fig. 6. Detection result for 619.png

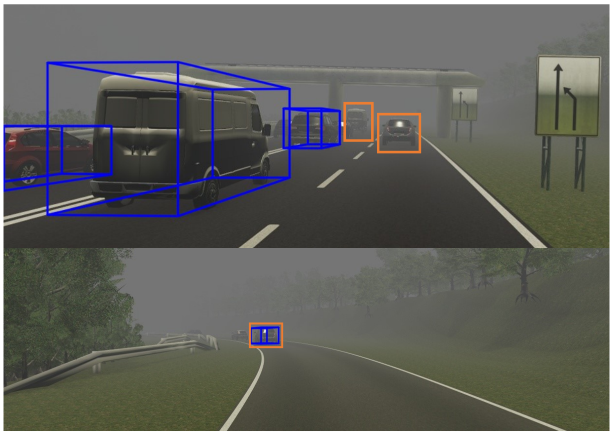

Fig. 7. Comparison of detection result under different foggy situations

Although effective in some situations, our current image processing techniques may not be suitable for all conditions, particularly those not well-represented in the training data. This highlights the importance of using more varied and comprehensive datasets for training to enhance the model's reliability across diverse conditions. We plan to implement these enhancements by incorporating more varied training data in the future to test and improve the model's performance across a broader range of conditions.

Furthermore, our strategic initiative involves substituting conventional enhancement methods with advanced deep learning-based approaches, such as GDIP (refer to the provided citation). GDIP leverages deep learning to enhance images captured in foggy scenes through the application of various image processing techniques. By adopting these innovative methods, we aim to fortify detections, rendering them more resilient across diverse environmental conditions.

These tailored approaches enabled us to tackle the specific requirements of monocular 3D object detection under diverse weather conditions, emphasizing the need for adaptability and robustness in model training and implementation.

## VI. Conclusions

This study made significant strides in monocular 3D object detection, particularly in challenging foggy conditions, enhancing the capabilities of autonomous vehicles. By implementing, adapting, and improving the MonoCon model, we achieved notable improvements in detection accuracy. Transfer learning and sophisticated image augmentation techniques like contrast enhancement and blurring resulted in a considerable increase in Average Precision (AP). Our results, showing an increase in AP from 7.05% to 17.67% in normal conditions and up to 25.82% in foggy conditions, underscore the effectiveness of our adaptations and training strategies. These findings highlight the model's adaptability and robustness in varied environmental scenarios, demonstrating its potential in real-world applications.

## References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.

[2] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980-2988.

[3] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9627-9636.

[4] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4490-4499.

[5] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 770-779.

[6] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12697-12705.

[7] H. Chen, Y. Huang, W. Tian, Z. Gao, and L. Xiong, "Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10379-10388.

[8] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 913-922.

[9] L. Yang et al., "Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3d object detection," IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[10] X. Liu, C. Zheng, K. B. Cheng, N. Xue, G.-J. Qi, and T. Wu, "Monocular 3D Object Detection with Bounding Box Denoising in 3D by Perceiver," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6436-6446.

[11] J. Xu et al., "Mononerd: Nerf-like representations for monocular 3d object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6814-6824.

[12] L. Peng et al., "Learning Occupancy for Monocular 3D Object Detection," arXiv preprint arXiv:2305.15694, 2023.

[13] X. Liu, N. Xue, and T. Wu, "Learning Auxiliary Monocular Contexts Helps Monocular 3D Object Detection,"

[14] Andreas Geiger, Philip Lenz, Raquel Urtasun. *Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite*. In: Conference on Computer Vision and Pattern Recognition (CVPR), 2012.