Multiple-view Obscured Object Detection (MOOD)

Duke University MIDS Capstone

Authors: Robin Arun, Katie Hucker, Afraa Noureen, Jiayi Zhou

Abstract

This project explores methods to improve the detection of partially obscured objects in complex environments using multi-view data, focusing on traffic cones in the nuScenes dataset. Traffic cones pose unique challenges due to their small size, symmetry, and frequent occlusion. To address these, we compare the performance of a multi-view transformer-based DETR3D model against a single-view YOLO baseline. DETR3D processes images from multiple camera perspectives using camera extrinsic to predict 3D bounding boxes, aiming to enhance robustness under occluded conditions. Our methodology includes analyzing traffic cone visibility across four occlusion categories and evaluating detection performance under varying conditions. We examine how multi-view data can enhance detection accuracy for partially obscured objects and assess how varying levels of object obscurity influence detection performance. Additionally, we explore the minimum number of scene views required for reliable detection and identify which object features contribute most significantly to accurate predictions. By addressing these factors, our approach highlights the benefits of multi-view data and offers insights to optimize detection systems for real-world, high-stakes applications.

Table of Contents

Introduction

Problem Statement

Detecting partially obscured objects in complex environments is a significant challenge. This is especially important in the autonomous vehicle and military fields where identifying partially and fully hidden objects is critical to human life. A robust detection system must be able to overcome obstacles, such as manmade or natural debris that cause obscurement and block the visibility of objects of interest. Current single-view models have better performance when the object is very clear, however, when an object is obscured, the model can lose confidence in its prediction. Furthermore, if we can use multiple views of an obscured or visible object, which can enhance model performance. The multiple views must be co-registered with one another, that is, the model understands when two images are next to each other. This is crucial difference between a standard single-view object detector, because it provides scene contextualization to the image. It allows the model to use what is visible in the entire scene through multiple images, rather than what is just visible within one image.

Motivation

Hidden objects pose significant risks in critical environments, where not finding a partially obscured object can have high-stakes consequences. Some of these high stake's scenarios include autonomous car navigation, landmine detection, and search and rescue missions. Traditional object detection methods struggle in these scenarios, often requiring multiple data collects, with the possibility of still not finding the objects due to obscurement. Current multi-view methods are computationally heavy requiring 3D point clouds or rendered scenes. Therefore, this project introduces a '3D interpretation of scenes' to enhance detection accuracy without the heavy overhead of full 3D reconstruction, providing an efficient solution for real-world environments.

Goal

Develop a model that can accurately detect and classify partially obscured objects by integrating multiple images of a scene, leveraging the spatial relationships among these views for robust detection.

In addition to our main goal, we hope to answer four questions. The research questions will guide model development and evaluation. The key questions are: (1) What extent can multi-view data improve detection accuracy for partially obscured objects? (2) How does varying the level of object obscurity impact detection accuracy in multi-view models? (3) What is

the minimum number of scene views required to achieve reliable detection? (4) Which object features most significantly contribute to accurate detection?

Literature Review

We reviewed datasets and models relevant to solving our problem. The datasets focus on annotated image data with large sample sizes to support effective object detection model training. The models examined represent current approaches to multi-view computer vision, many tested in autonomous driving applications. We evaluate their suitability for our application and synthesize their limitations and strengths.

Relevant Datasets

This section includes descriptions of three datasets: PASCAL3D, CO3D, nuScenes. We will discuss the difference between the three datasets. These datasets are used throughout many computer vision and object detection applications. One dataset may fit a project better than another. Furthermore, we will analyze the differences including benefits and limitations.

PASCAL3D

Pascal3D set out to solve many problems at its time of publishing. It wanted to create a dataset which solved the following issues in data availability in the computer vision space. The datasets at time of publishing (2012), did not have clutter or occlusions. This leaves models unable to generalize to real scenes. In addition, the image sets had a small number of viewpoints and annotations.

After defining the problem, the data collectors focused on maintaining "dense and continuous viewpoint annotations." Following a traditional computer vision methodology, of the pinhole camera. 3D landmarks X are projected back to the 2D image x. They completed this with the use of 3D CAD models from Google CAD dataset. They project the 3D CAD model into the images. Stanford then annotates each object from the image in terms of azimuth, elevation, and distance in 3D. The visibility state is also labeled with the following categories: visible, self-occluded, occluded-by, truncated, unknown. In addition to the annotations, the user also now has viewpoints of 12 classes with distributions of the viewpoints. For example, the couch class has little to no viewpoints from behind it because typically a couch is on a wall. However, the boat class has a distribution from all angles. More specifically, in this dataset there is approximately equal 360-degree viewpoint distribution in 5 classes, one of which is the boat class. Stanford also cites the occlusion distribution in terms of thirds.

nuScenes

nuScenes is a comprehensive autonomous driving research dataset. It's got 1,000 20-second urban driving scenes collected in Boston and Singapore, featuring synchronized sensor data from 6 cameras, 1 LiDAR, 5 RADAR sensors, GPS, and IMU (Inertial Measurement Unit - helps with camera movement and orientation data), which helps capture orientation data for accurate scene understanding. They provide all cameras, LiDAR, RADAR and camera intrinsic and extrinsic data across dozens of objects per scene. The camera data expands over 1.4 million images at 2Hz (twice per second), including annotations from 23 object classes (incl. cars, pedestrians, bicycles) that were annotated by humans. Additionally, a 1.4 billion annotated LiDAR points, with segmentation provided for 32 classes (23 foreground, 9 background), also present opportunity for deep temporal and scene analysis.

nuScenes is a benchmark dataset for many pioneering research models on multi-view object detection, presenting experimentation opportunities like assessing whether the number of images improves detection accuracy, amongst other options as mentioned in the proposal above.

Common Objects in 3D

Common Objects in 3D (CO3D) is a dataset designed for learning category-specific 3D reconstruction and new-view synthesis using multi-view images of common object categories. CO3D facilitates advances in this field by providing a large-scale dataset composed of real multi-view images of object categories annotated with camera poses and ground-truth 3D point clouds. The CO3D dataset contains 1.5 million frames from nearly 19,000 videos capturing objects from 50 MS-COCO categories. The dataset is suitable for new-view synthesis methods, such as the seminal NeRF.

Current Models

There are many object detection models, however, there are few unique multi-view detection models. Many of the models are related to one another and use similar frameworks. We discussed a few models which we feel most relate to our problem and may be implemented in our methodology.

Occlusion-aware R-CNN

In the field of object detection, particularly in complex scenes, multiple models have been developed to address unique challenges. The Occlusion-aware R-CNN (OR-CNN) is one such model designed to significantly improve pedestrian detection in crowded

environments, where occlusion is a common challenge. OR-CNN introduces two key improvements: the aggregation loss (AggLoss) and the part occlusion-aware region of interest (PORoI) pooling unit. These features are what make the OR-CNN better at handling obscured objects more effectively than conventional methods (e.g., standard R-CNN). By using AggLoss, OR-CNN improves bounding box compactness, which reduces false positives in high-density pedestrian settings. Meanwhile, the PORoI pooling unit segments each detected pedestrian into visible parts, such as the head or torso, which increases detection reliability in scenes with partial occlusion or hidden objects. This approach is particularly beneficial in our use case, because it manages the obscurement level well. However, OR-CNN's limitations include computational demands and potential struggles in extremely dense crowds, especially if the model is used beyond pedestrian detection.

Neural Radiance Fields (NeRF)

In the area of scene synthesis, Neural Radiance Fields (NeRF) provide a sophisticated solution for generating 3D views from 2D images. NeRF represents a scene as a continuous 5D function, encoding spatial and directional information to produce synthetic views with high fidelity. The NeRF framework relies on a multi-layer perceptron (MLP) network, with hierarchical volume sampling and high frequency encoding to capture color and geometry variations accurately. This allows NeRF to synthesize views across different angles, which could be beneficial for our object detection efforts, potentially enhancing visibility across various perspectives of occluded objects. However, NeRF's performance relies on high-quality image datasets with well-defined camera poses, which could reveal limitations.

OccluBEV

OccluBEV takes a different approach by improving occlusion-aware object detection for autonomous driving applications. Occlusion in this definition is like obscurement previously defined. This model addresses the limitations of 2D images in representing occlusion in a 3D space by integrating point cloud data from both image and Bird's Eye View (BEV) perspectives. OccluBEV has a spatiotemporal mechanism that identifies and tracks objects across multiple frames, enhancing its ability to detect objects that may become occluded over time. Through feature fusion, OccluBEV constructs a more comprehensive 3D point cloud, effectively registering occluded objects. For scenarios involving a moving platform, such as ours, the use of sequential image frames is particularly advantageous. However, the dependency on LiDAR data and potential limitations with smaller objects highlight important considerations for its application in our specific context.

Detection Transformer for 3D (DETR3D)

DETR3D uses a transformer-based model to detect objects in 3D using multi-view images. Unlike traditional models, DETR3D directly predicts 3D bounding boxes from 2D images without requiring complex components like non-maximum suppression or anchor boxes. The model leverages transformer attention mechanisms, making it computationally efficient for handling high-resolution scenes. By encoding camera views into feature maps and employing attention-based queries, DETR3D efficiently identifies 3D object positions. DETR3D's reliance on multi-view images and avoidance of depth estimation align well with our data needs. However, its sole dependence on image inputs rather than LiDAR or point clouds may impact its depth prediction capabilities, particularly in cases where depth information is critical.

Few-Shot Object Detection and Viewpoint Estimation

Lastly, the Few-Shot Object Detection and Viewpoint Estimation model aims to detect objects with minimal labeled data while estimating viewpoint orientation. This approach is grounded in few-shot learning, allowing the model to generalize detection across unlabeled or sparsely labeled objects by learning category-agnostic features. Its viewpoint estimation component aids in inferring 3D object orientation based on multi-view geometric cues. This capability could be particularly valuable in applications where labeled data is limited or where generalizability to novel objects is required. Nonetheless, the trade-off in accuracy for generalizability raises concerns about performance consistency, although integrating this model with other 3D methods, like NeRF, could potentially enhance viewpoint estimation and scene synthesis outcomes.

In summary, each model provides unique capabilities for object detection and occlusion handling in complex scenes. While OR-CNN, OccluBEV, DETR3D, and NeRF address specific aspects of occlusion, view synthesis, and detection efficiency, the Few-Shot Object Detection and Viewpoint Estimation model introduces adaptability to scenarios with minimal labeled data. Together, these methodologies form a toolkit that can be adapted and potentially combined to tackle occlusion and visibility challenges in a comprehensive and flexible manner.

Data Overview

Our project explores how well different models can detect partially occluded objects in real-world settings, using the NuScenes dataset. NuScenes is a comprehensive resource for autonomous driving research, containing 1,000 urban driving scenes (each 20 seconds long) recorded in Boston and Singapore. The dataset includes data from six cameras, one LiDAR,

five RADAR sensors, GPS, and an IMU (Inertial Measurement Unit), providing highly accurate orientation and movement data.

With over 1.4 million images recorded at 2Hz and human-labeled annotations for 23 object classes (e.g., cars, pedestrians, bicycles), NuScenes has supported many pioneering research models in multi-view object detection. This rich dataset allows us to experiment with various conditions to assess and improve detection accuracy. For this project, we focus on traffic cones, a specific object class that presents unique challenges for detection, particularly under conditions of partial occlusion.

Traffic Cone Filtering

Among the 23 object classes in NuScenes, we chose traffic cones due to their specific properties that pose unique challenges in detection:

1. *Symmetry:* Traffic cones have a uniform shape, which can make distinguishing them from other objects difficult, particularly in dense or cluttered environments.
2. *Immobility:* Cones are static, unlike moving objects like cars and pedestrians, making them less predictable in appearance but critical for stationary object detection.
3. *Small Size:* Their relatively small physical footprint makes them susceptible to occlusion, especially in busy, urban driving scenes where they might be partially blocked by other objects.

These characteristics create a controlled yet challenging test case for analyzing model performance under partial occlusion in natural settings.

Specifications

1. *Dataset Composition:* Each NuScenes scene provides multi-sensor data with intrinsic and extrinsic calibration for accurate positioning, allowing us to assess model performance with realistic, synchronized views from multiple angles.
2. *Visibility Bins:* Traffic cone visibility in our study is divided into four bins based on the percentage of visible pixels across six camera views:

0-40% Visibility	Mostly obscured or invisible
40-60% Visibility	Partially visible
60-80% Visibility	Mostly visible, with minor occlusions
80-100% Visibility	Fully or nearly fully visible

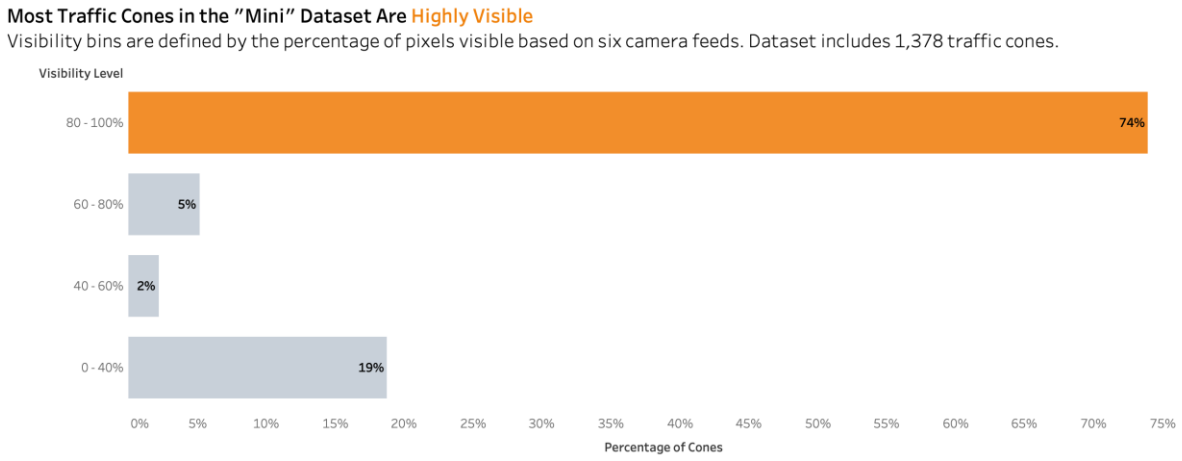
Discussion

To evaluate how model performance varies across different levels of occlusion, we analyzed the visibility distribution of traffic cones within the nuScenes dataset. Due to the large size of the full dataset and the computational limitations we face, we chose to conduct our experiments using the mini dataset, which is a smaller but representative subset of the full nuScenes collection. This allowed us to perform meaningful analyses within our resource constraints while maintaining relevance to the overall dataset's characteristics. The mini dataset mirrors the visibility distribution found in the complete nuScenes dataset, making it a practical choice for gaining insights into model performance.

Our analysis of the 1,378 traffic cones in the mini dataset revealed the following visibility distribution:

- 74% of cones fall into the 80-100% visibility range, indicating that most cones are nearly or fully visible.
- 19% of cones are in the 0-40% visibility range, representing cones that are largely hidden.
- 5% of cones have 60-80% visibility, while only 2% of cones are in the 40-60% visibility range, indicating partial occlusion.

Figure 1: Traffic cone distribution for dataset



This distribution highlights a significant visibility imbalance, with majority of the cones being highly visible (74%), while a much smaller proportion of cones experience partial occlusion (only 2% in the 40-60% range). The underrepresentation of partially visible cones suggests

that models may perform well when cones are fully visible but may struggle in real-world scenarios where occlusion is more common.

Given this imbalance, we are considering techniques to subset or balance the dataset, particularly focusing on increasing the number of cones in the 40-60% visibility range. This would allow for a more thorough evaluation of how well models can handle partial occlusion—an essential factor for real-world applications in autonomous driving.

By systematically analyzing how visibility impacts detection accuracy, our goal is to develop more robust models that can detect objects even when they are partially obscured. This work emphasizes the need for balanced datasets that reflect natural variations in visibility, which are essential for ensuring reliable object detection in urban environments.

Methodologies

To address this problem, we will evaluate the DETR3D model and compare it to a baseline model, YOLO, on single-view data, focusing on performance improvements as the amount of input information increases. This study will involve progressing from single view to multi-view data and from 2D to 3D representations. We will begin by establishing a baseline using YOLO for single-view object detection. Then, we will test DETR3D, which leverages multiple images of the same scene along with camera location and orientation to provide a complete 3D understanding of the environment. DETR3D detects objects across the entire 3D scene and projects these detections back onto the single-view input images. This progression will enable us to assess how DETR3D impacts the accuracy and efficiency of detecting partially obscured objects.

Single-View Detection with Enhanced Obscurement Handling

The conventional approach typically focuses on a baseline single view to simplify the analysis and avoid the complexity introduced by multi-view configurations. You only look once (YOLO) is a state-of-the-art, real-time object detection system that excels at recognizing and localizing multiple objects in an image [2]. Unlike traditional detection models, which process an image in multiple steps, YOLO's single neural network architecture streamlines the process by simultaneously predicting object locations and class labels in a single forward pass. YOLO is chosen for baseline for following reasons:

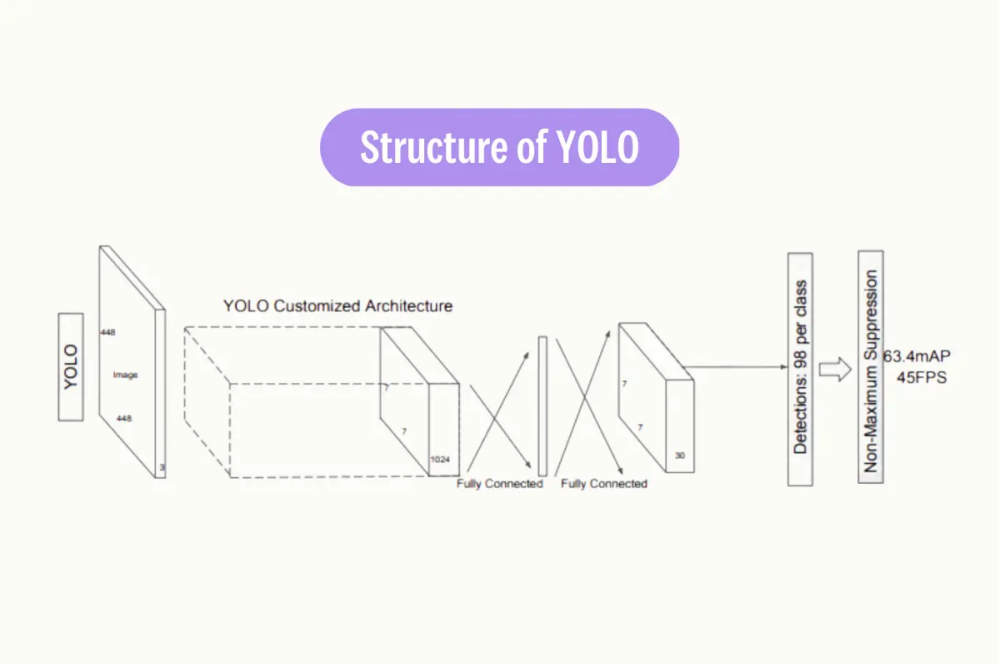
Simultaneous Detection and Classification: YOLO predicts both the class and the precise location of objects within an image, making it a comprehensive solution for real-time applications that require fast and accurate object recognition.

High Accuracy: YOLO's architecture allows it to achieve competitive accuracy, particularly excelling at detecting a variety of objects within an image while maintaining low error rates.

Fast Object Detection: Optimized for speed, YOLO processes images at impressive speeds, making it ideal for real-time applications where rapid response is critical.

Strong Performance on Small Objects: Unlike many other object detection algorithms, YOLO performs well even on small objects, enhancing its utility in tasks that involve detecting finer details within crowded scenes or cluttered environments.

Figure 2: Yolo Structure Diagram



Multi-View Detection using DETR3D

The second approach leverages multiple observation angles to enhance object detection with scene contextualization. Studies in autonomous vehicle research suggest that using Bird's Eye View (BEV) and 3D point clouds can enhance spatial understanding, especially under partial object occlusion. These point clouds or scene reconstruction methods provide detailed information about a scene. However, we wanted to explore methods which avoid this reconstruction component. This will save computational time through 2D images into 3D space projection, with DETR3D. This model implementation will offer robust solutions for multi-view detection without requiring explicit depth estimation or anchor boxes, simplifying the architecture while maintaining efficiency.

DETR3D

DETR3D processes images from various viewpoints and uses transformer-based attention to predict 3D bounding boxes directly from 2D inputs. DETR3D's efficiency, due to its lightweight architecture that eliminates non-maximum suppression and depth estimation, makes it a top performer on the nuScenes benchmark for autonomous driving. In scenarios with partial occlusion, this approach simplifies processing by extracting feature maps for each view and aligning them based on camera calibration data, allowing the model to leverage the combined feature maps from multiple views. DETR3D is chosen for the following details:

Direct Prediction of 3D Bounding Boxes: DETR3D bypasses the need for explicit depth estimation or LiDAR data by predicting 3D bounding boxes directly from 2D images. This reduces computational overhead and simplifies the detection pipeline.

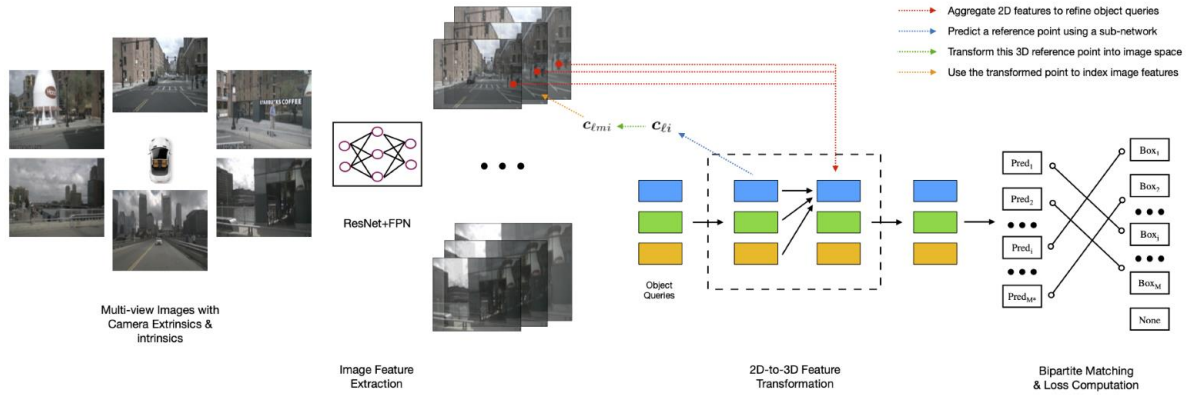
Reduced Computational Load: Given the need for multi-view image processing and synchronization with the annotation (labels and camera intrinsic and extrinsic) data, DETR3D made it most suitable for applications with a need for light-weightiness reducing inference time and computational resources required. DETR3D forgoes the need for a full 3D reconstruction of a scene and utilizes 2D feature maps instead.

Competitive Advantage: nuDETR3D achieves a mean average precision(mAP) of 41.2% on the nuScenes dataset, making it a state-of-the-art method in 3D object detection.

How DETR3D Works

The model extracts image features using a backbone (e.g., ResNet with FPN), and then transforms these 2D features into 3D space by aligning them with camera calibration data (intrinsics and extrinsics). This allows the model to perform object detection without relying on complex 3D point cloud data or explicit depth estimation. By aggregating features across multiple views, DETR3D can effectively capture spatial context, making it robust against challenges such as partial occlusions. The prediction process bypasses traditional detection methods like anchor boxes and non-maximum suppression, simplifying the pipeline and reducing computational overhead. For further technical details and implementation specifics, please see DETR3D paper [1].

Figure 3: DETR3D Model Framework



Planned Use of DETR3D

Evaluate DETR3D's effectiveness in detecting partially obscured objects, specifically traffic cones. The primary objective is to investigate whether multi-view image inputs offer significant improvements over single-image inputs in detection accuracy. By leveraging DETR3D's capability to predict 3D bounding boxes directly from 2D images, the project will compare detection performance under varying levels of occlusion and multi-view configurations. This approach aims to address key questions about the impact of multi-view inputs on detection accuracy and explore DETR3D's potential to streamline computational complexity while maintaining robust performance in challenging environments.

Conclusion and Future Directions

Both single-view and multi-view detection methods present promising avenues for improving the detection of partially occluded objects. Conventional single-view detection methods, such as occlusion-aware R-CNN, establish a reliable baseline, which can be further enhanced by incorporating techniques like NeRF to capture additional viewpoints. In parallel, multi-view detection models like OccluBEV and DETR3D highlight the advantages of 3D scene reconstruction, leveraging multiple observation angles to improve detection accuracy in environments with occluded objects. The choice between these methodologies depends on factors such as computational efficiency, the availability of multi-view data, and specific requirements for threat assessment accuracy in partially occluded scenes. As we move forward, the next steps include:

Train and test YOLO model on the full filtered dataset: The YOLO model was initially tested with a small sample of data, and now, with a complete subset focusing on traffic cones, we will proceed with training and testing the YOLO model using the prepared dataset.

DETR3D Model Implementation: The implementation of our multi-view model, DETR3D, is nearing completion. We plan to finalize it by the end of the semester. Once complete, it will offer a toolkit similar to the YOLO model.

Train and test DETR3D model on the full filtered dataset: After finalizing the DETR3D modeling code and testing it on the subset dataset, we will expand our tests to include the full filtered dataset.

Return to our driving questions: We will continue to investigate our stakeholder's core questions, focusing on the impact of down sampled data, how occlusion levels affect model performance, and enhancing model explainability through feature analysis.

Future research would focus on optimizing these models for various occlusion levels and object types, ultimately expanding their applications across diverse environments.

Resources

[1] DETR3D: Wang, Y., et al. "DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries." *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021*. Available at https://arxiv.org/abs/2110.06922.

[2] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). "You only look once: Unified, real-time object detection." *arXiv preprint arXiv:1506.02640*. Available at https://arxiv.org/abs/1506.02640