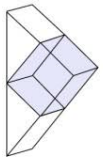
VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY



Autonomous Vehicle

Company: Delta Cognition

Participant:
Vo Thanh Nghia -
vothanhnghia270604@gmail.com

October 10, 2024

Contents

1 Information

1.1 Participant Information

Linkedin	Full name	Email
vtnghia	Vo Thanh Nghia	vothanhnghia270604@gmail.com

1.2 Source Code

Link to Source Code - Github

1.3 Abstract

This report outlines the development of an AI system aimed at improving the robustness of autonomous vehicles in rural and unfamiliar environments. The system is built around three primary tasks:

- Detecting known objects in video footage.
- Detecting novel objects in the video.
- Making collision-avoidance decisions during navigation.

Task	Research Approach Status	Implementation Status
Task 1: Detecting known objects in video footage	Done	Done
Task 2: Detecting novel objects in the video.	Done	Fail
Task 3: Making collision-avoidance decisions during navigation.	Done	Done

Emphasis is placed on the detection of novel objects and the decision-making process, as these are critical to enhancing the reliability and safety of autonomous vehicles in complex and unpredictable environments. The approach integrates object detection techniques and decision-making algorithms, with the goal of improving vehicle performance in challenging conditions.

2 Task 1: Detecting known object in video footage

2.1 Yolov10-Small

For the first task of detecting known objects in video footage, I employed the YOLO (You Only Look Once) model, which is one of the most efficient and widely-used object detection algorithms. YOLO is particularly well-suited for real-time applications because it processes the entire image in a single forward pass, making it both fast and accurate. It divides the image into a grid, predicting bounding boxes and class probabilities simultaneously, which allows it to detect multiple objects in a single frame.

I selected the YOLOv10-Small (YOLOv10-S) model due to its balance between speed and accuracy. YOLOv10 is a powerful and modern object detection architecture that comes in several variants to suit different application needs [1]:

- YOLOv10-N: Nano version for extremely resource-constrained environments.
- YOLOv10-S: Small version balancing speed and accuracy.
- YOLOv10-M: Medium version for general-purpose use.
- YOLOv10-B: Balanced version with increased width for higher accuracy.
- YOLOv10-L: Large version for higher accuracy at the cost of computational resources.
- YOLOv10-X: Extra-large version for maximum accuracy and performance.

Given the requirements of autonomous vehicle navigation, YOLOv10-Small was the most appropriate choice. This version offers a perfect trade-off between speed and detection accuracy, which is crucial for real-time processing. Autonomous vehicles must detect objects quickly to make timely decisions, such as avoiding obstacles, and YOLOv10-Small enables this with its ability to process video frames efficiently without compromising the quality of object detection.

By using pre-trained weights and follow the instruction of Mert [2] (this website needs VPN to access), I was able to achieve accurate object detection for various known objects, helping to enhance the overall robustness of the autonomous vehicle's perception system.

2.2 Crosswalk Detection

While YOLOv10-Small is highly effective for detecting a broad range of objects, it is limited to 80 predefined classes. Since the specific task of detecting crosswalks is crucial for safe autonomous vehicle navigation—especially for the third task of making collision-avoidance decisions—relying solely on YOLO's standard classes is insufficient. Therefore, I integrated a dedicated crosswalk detection model to fill this gap.

For this purpose, I chose a specialized crosswalk detection model from a GitHub repository [3], which demonstrated excellent performance on test data.

The decision to use this repository was based on its outstanding results, boasting

- A precision of 0.96
- A recall value of 0.93

Making it highly reliable for detecting crosswalks in diverse environments. These metrics indicate the model's ability to accurately identify crosswalks with minimal false positives and capture most

instances with minimal false negatives, both of which are critical for ensuring the autonomous vehicle makes safe decisions in real time.

2.3 Traffic light classification

In addition to detecting known objects and crosswalks, the ability to classify traffic light signals is essential for safe and efficient autonomous vehicle navigation. YOLOv10-Small's limitation to 80 predefined classes does not include the classification of traffic lights by their specific color (red, yellow, or green), which is crucial for the third task of making real-time collision-avoidance decisions based on traffic signals.

For this reason, I integrated a traffic light classifier to ensure the vehicle reacts appropriately at intersections, contributing to overall system robustness.

To achieve this, I used the traffic-light-classifier package from PyPi [4], which implements a robust probabilistic approach to classify traffic light signals. The classifier was developed using computer vision and machine learning techniques, incorporating extensive data cleaning, feature extraction, and a probabilistic metric for classification accuracy. This tool has been thoroughly tested, achieving an impressive 99.66 accuracy on the testing dataset, making it highly reliable for recognizing traffic light statuses in real time.

3 Task 2: Detecting novel objects in the video.

Unfortunately, I was unable to fully implement the second task of detecting novel objects in the video. However, during my research, I discovered several promising approaches that could be valuable for further development

3.1 Cooperative Foundational Models

In my future research on novel object detection (NOD), I intend to delve into the methodologies presented in the paper that focuses on transforming closed-set detectors into open-set detectors [5].

The innovative approach described in this work leverages the complementary strengths of foundational models like CLIP and SAM to enhance the detection capabilities for both known and novel object categories. Given the impressive results reported, including 17.42 mAP in novel object detection and 42.08 mAP for known objects on the LVIS dataset, I find their findings particularly relevant to my work.

I plan to thoroughly examine their proposed cooperative mechanism, which allows for the integration of CLIP's understanding of unseen classes with the localization capabilities of pre-trained models such as Mask-RCNN. Additionally, the ability to refine bounding boxes using SAM's instance mask-to-box properties intrigues me, as it offers a potential path to improve my own NOD efforts. I am also interested in the performance gains achieved when this approach is combined with state-of-the-art open-set detectors like GDINO.

To facilitate this exploration, I will review the paper in detail [6] and experiment with their code available at GitHub. By investigating their methodologies and results, I hope to gather insights that

will contribute to my understanding of novel object detection and potentially improve the robustness of my own models in recognizing and classifying both known and novel objects in real-time

3.2 Zero-Shot Object Detection

For my second approach to enhancing novel object detection, I plan to investigate Zero-Shot Detection (ZSD), which overcomes limitations in current Zero-Shot Learning (ZSL) methods that typically focus on recognizing a single unseen object category. ZSD seeks to simultaneously recognize and localize instances of novel categories without training examples, making it more suitable for complex real-world scenarios.

I am particularly interested in the new experimental protocol for ZSD based on the ILSVRC dataset, which addresses the rarity of unseen objects. The proposed end-to-end deep network that integrates visual and semantic information, along with the novel loss function utilizing meta-classes, represents a significant advancement in the field. Promising results from extensive experiments show substantial performance improvements over baseline models. I will thoroughly research their paper [7] and explore their GitHub repository [8] to implement these methods, aiming to enhance my own research on detecting and localizing novel objects in complex environments

4 Task 3:

In Task 3, I implemented collision avoidance decisions based on the detections from Task 1, which utilized the YOLO model to identify various objects in the environment.

The process detections function processes the detection results by examining the bounding boxes generated by the YOLO model.

For each detected object, it retrieves the class ID and corresponding class name. If the detected class matches any object in the stop list, the function prints a message instructing the vehicle to stop, whereas a match with the slow down list prompts a warning to slow down. This approach allows the system to make real-time decisions based on the detected objects, ensuring safer navigation through complex environments.

5 Reference

[1] Ultralytics. *Yolo Model Variants*. https://docs.ultralytics.com/models/yolov10/. (Visited on 10 Oct. 2024).

[2] Mert. *How to use YOLOv10 for Object Detection*. https://medium.com/@Mert.A/how-to-use-yolov10-for-object-detection-de9f47898db2. (Visited on 10 Oct. 2024).

[3] xN1ckuz. *Crosswalks-Detection-using-YOLO*. https://github.com/xN1ckuz/Crosswalks-Detection-using-YOLO. (Visited on 10 Oct. 2024).

[4] Shashank Kumbhare. *Traffic Light Classification*. https://pypi.org/project/traffic-light-classifier/. (Visited on 10 Oct. 2024).

[5] Rohit K Bharadwaj et al. *Enhancing Novel Object Detection via Cooperative Foundational Models*. https://arxiv.org/abs/2311.12068. (Visited on 10 Oct. 2024).

[6] Rohit K Bharadwaj et al. *Cooperative Foundational Models*. https://github.com/rohit901/cooperative-foundational-models. (Visited on 10 Oct. 2024).

[7] Shafin Rahman, Salman Khan, and Fatih Porikli. *Zero-Shot Object Detection: Learning to Simultaneously Recognize and Localize Novel Concepts*. https://arxiv.org/abs/1803.06049. (Visited on 10 Oct. 2024).

[8] Shafin Rahman, Salman Khan, and Fatih Porikli. *Zero-Shot Object Detection*. https://github.com/salman-h-khan/ZSD_Release. (Visited on 10 Oct. 2024).