Escuela
Politécnica
Superior

Pedestrian Detection in Autonomous Driving

Máster en Inteligencia Artificial

Trabajo Fin de Máster

Autor:
SERINE BENMOHRA
Tutor/es:
CAZORLA QUEVEDO, MIGUEL ANGEL
ESCALONA MONCHOLÍ, FÉLIX

Universitat d'Alacant
Universidad de Alicante

Pedestrian Detection in Autonomous Driving

Autor

SERINE BENMOHRA

Tutor/es

CAZORLA QUEVEDO, MIGUEL ANGEL

Departamento de Ciencia de la Computación e Inteligencia Artificial

ESCALONA MONCHOLÍ, FÉLIX

Departamento de Ciencia de la Computación e Inteligencia Artificial

Máster en Inteligencia Artificial

Escuela
Politécnica
Superior

Universitat d'Alacant
Universidad de Alicante

ALICANTE, Mayo 2025

Preface

"This Master's thesis has been motivated by the growing need to improve the reliability of perception systems in autonomous vehicles, especially in urban environments where visual ambiguity can lead to critical errors. One particularly challenging issue is the misidentification of human figures—such as mannequins, reflections, or printed advertisements—as real pedestrians. These false detections can have serious consequences for safety and decision-making in AI-driven systems.

The objective of this study is to evaluate and compare the performance of two advanced models—YOLOv8 and the vision-language model LLaVA—in detecting real people while avoiding false positives caused by deceptive visual patterns. A diverse custom dataset was created from three different sources, including frames from the University of Sydney's autonomous driving dataset, a private collection of urban walkthrough footage recorded for walkability studies, and original video data captured using a GoPro camera during personal driving sessions. These varied sources enabled a broader range of ambiguous visual scenarios to be analyzed. The analysis includes model accuracy, confidence distributions, and disagreement cases, supported by detailed metrics and visual reports.

I would like to express my sincere gratitude to my academic supervisors for their continuous guidance and encouragement throughout this project. I am also thankful to my colleagues and peers for the insightful discussions that helped refine the methodology and analysis. Their support has been instrumental in the successful completion of this research."

Acknowledgments

I would like to express my deepest gratitude to my academic supervisors for their invaluable guidance, constructive feedback, and continuous support throughout the development of this thesis. Their expertise and encouragement played a crucial role in shaping the direction and quality of this research.

I am also thankful to my peers and colleagues for the stimulating discussions, technical suggestions, and collaborative spirit that enriched both the methodology and the analysis. Their input was instrumental in helping me address the challenges encountered during the project.

Lastly, I want to acknowledge the learning environment provided by the Master's program, which offered the academic foundation and tools necessary to carry out this research with both rigor and curiosity.

To my father Houcine, my mother Hafida,
my brothers Charaf and Mohammed, and my sister Bouchra,
without them, I might have finished this thesis much earlier,
but without them, none of it would have meant anything.

Try not to resist the changes that come your way.
Instead, let life live through you

Jalal ad-Din Rumi.

Contents

List of Figures

List of Tables

1 Introduction

Autonomous driving technologies rely heavily on the accuracy and robustness of perception systems, especially in complex urban environments. This introductory chapter outlines the research motivation and the specific problem being addressed, focusing on the challenge of distinguishing real pedestrians from deceptive visual elements. It provides an overview of the datasets used, states the main objectives and research questions, and highlights the significance and contributions of this study within the broader context of autonomous vehicle perception

1.1 Research Context and Motivation

Autonomous vehicles (AVs) are reshaping the future of transportation by offering safer, more efficient, and intelligent mobility solutions. Central to the operation of AVs is their ability to perceive and interpret the surrounding environment using advanced computer vision systems. A critical challenge in this domain is the accurate detection of pedestrians—a task that becomes significantly more complex in real-world urban environments where visual ambiguity is widespread.

This problem is especially important because most AV perception systems still heavily rely on 2D visual input—primarily from monocular or stereo cameras—to identify and classify objects in their environment. While this approach is computationally efficient and well-supported by large-scale training datasets, it introduces fundamental limitations. Flat, 2D images lack depth and tactile feedback, making it difficult to differentiate between actual three-dimensional objects (like people) and deceptive visuals such as mannequins, reflections, shadows, printed ads, or digital screens. These visual illusions can closely resemble real pedestrians in both form and context, especially when viewed under challenging lighting

conditions, occlusions, or from certain angles.

False positives resulting from this ambiguity can lead to problematic behavior from AVs, including abrupt or unnecessary stops, erratic path planning, or even collisions caused by hesitation or misjudgment. Such errors not only affect traffic flow but can also undermine public trust in autonomous technologies. Moreover, the rarity and unpredictability of these ambiguous cases make them difficult to capture in conventional training datasets, leaving even state-of-the-art detectors like YOLOv8 prone to failure in these edge scenarios.

Recent advances in Large Multimodal Models (LMMs), such as LLaVA (Large Language and Vision Assistant), offer new opportunities by integrating visual input with contextual reasoning grounded in natural language. These models aim to ”understand” an image rather than simply detect patterns, allowing for a deeper semantic interpretation of scenes. This study investigates whether such reasoning-based models can help address the critical gap left by conventional object detectors, improving AV perception in scenarios where 2D vision alone leads to high uncertainty or false alarms.

1.2 Problem Statement

Most existing datasets used for autonomous vehicle (AV) perception are designed around clear, unambiguous pedestrian scenarios, offering little support for evaluating performance under real-world visual ambiguity. These datasets typically contain well-annotated images in which pedestrians are clearly visible, centered, and distinguishable—conditions that do not reflect the complexity of urban environments. As a result, traditional object detection models are not thoroughly tested on scenarios that simulate the deceptive visual patterns found in reality, such as mannequins, store-front advertisements, digital screens, or reflections that resemble human figures.

This thesis addresses this critical gap by approaching two interconnected problems. The first is the creation of a new, targeted dataset specifically curated to include visually ambiguous pedestrian-like scenarios. The dataset incorporates frames from three different sources—public, private, and self-recorded—capturing diverse urban conditions where human-like forms might be mistaken for real pedestrians. This dataset not only enables robust testing of AV perception systems in edge-case scenarios but also contributes a valuable resource

for future research.

The second problem concerns the development of a methodology to evaluate and compare how well current models, particularly the traditional object detector YOLOv8 and the vision-language model LLaVA, handle these ambiguous cases. While YOLOv8 represents a strong baseline in object detection, its reliance on purely visual cues may limit its ability to distinguish real humans from human-like distractions. LLaVA, as a Large Multimodal Model (LMM), introduces the capacity for contextual reasoning by combining visual inputs with language understanding—potentially offering an advantage in nuanced decision-making.

By tackling both dataset creation and evaluation methodology, this research aims to provide a clearer understanding of how AV perception systems perform in the presence of deceptive visual inputs and to assess the potential of LMMs to enhance pedestrian detection reliability in complex urban scenarios.

1.3 Datasets Overview

To ensure a robust evaluation, this study utilizes multiple datasets:

- **USYD-Campus Dataset:** A 10-week subset of the publicly available University of Sydney campus driving dataset, which includes diverse urban scenes and pedestrian interactions captured in .H264 video format
- **Urban Walkthrough Videos (Alicante, Spain):** A private dataset consisting of three videos recorded using an Instax360 camera by peers at the University. Originally intended for a walkability study—evaluating the accessibility and comfort of urban spaces for the elderly and mobility-challenged individuals—these recordings capture urban environments in the city of Alicante. Two of the videos were filmed in the Mercado Central area from left and right perspectives, and one in the Garbinet district.
- **Author-Created Driving Dataset:** A custom video dataset recorded by the author using a GoPro camera mounted on a vehicle, capturing urban driving scenes across various locations in the Alicante center.

From all of the above sources, frames were extracted at regular intervals. A manual selection

process was conducted to isolate and retain only those frames featuring ambiguous visual representations—specifically, human-like figures that are not actual pedestrians (e.g., posters, mannequins, reflections, and digital advertisements). This newly curated dataset serves as the foundation for evaluating and comparing the detection performance of different models under challenging, real-world visual ambiguity.

1.4 Objectives and Research Questions

The core objective of this thesis is to assess and compare the ability of LLaVA (a representative LMM) and YOLOv8 (a traditional object detection model) to identify deceptive human representations in real-world urban imagery. The research aims to answer the following key questions:

1. How do LMMs like LLaVA compare with traditional models such as YOLOv8 in terms of accuracy and reliability in detecting ambiguous pedestrian-like figures?

2. In what scenarios do LMMs fail, and what are the limitations of their contextual reasoning?

3. Can semantic understanding from LMMs be effectively integrated into AV perception systems to reduce false positives?

1.5 Significance and Contributions

This work contributes to the field of autonomous driving perception in several important ways:

- **Empirical Evaluation:** It provides a comprehensive performance comparison between LLaVA and YOLOv8 on a newly constructed dataset focused on deceptive human representations.

- **Real-World Testing:** By incorporating both public and self-captured real-world data, the study ensures the findings are applicable to realistic AV scenarios.

- **Custom Dataset:** The manually created dataset serves as a new benchmark for future research on pedestrian disambiguation in AV perception.
- **Guidance for Integration:** The results offer actionable insights for developers and researchers aiming to integrate LMMs into existing perception pipelines to enhance robustness and safety.

Ultimately, this thesis aims to demonstrate the potential and limitations of LMMs in solving real-world perception problems that traditional models still struggle with—marking a step toward more context-aware and safer autonomous vehicles.

2 Theoretical Framework

This chapter presents the theoretical foundation necessary to understand the problem space addressed in this thesis. It begins by examining the role of artificial vision in autonomous vehicles, followed by an overview of core object detection concepts. It then introduces the evolution from traditional models such as YOLO to more recent vision-language models like LLaVA. A review of current methodologies and persistent challenges in detecting pedestrians—particularly under ambiguous visual conditions—is provided. Finally, the chapter discusses the specific issue of distinguishing real pedestrians from artificial figures and outlines emerging research directions that aim to improve model robustness in such scenarios

2.1 Artificial Vision in Autonomous Vehicles

Autonomous vehicles heavily rely on artificial vision systems to perceive their surroundings, enabling safe navigation and decision-making. Historically, computer vision methods evolved from classical handcrafted features to deep learning techniques, with Convolutional Neural Networks (CNNs) revolutionizing the field by enabling end-to-end feature extraction and recognition Sivaraman & Trivedi (2013)

While the technological evolution has been remarkable, modern urban environments introduce new challenges for vision-based systems. Complex traffic scenes, dynamic agents, varying lighting conditions, and frequent visual ambiguities make pedestrian detection particularly challenging.

More recent advances include the integration of Transformer-based models in vision tasks, which capture long-range dependencies and contextual information effectively, leading to improved detection and segmentation performances Dosovitskiy et al. (2021)

2.2 Object Detection: Key Concepts

Object detection is a core computer vision task involving localization and classification of objects in images. Early techniques used sliding window classifiers with hand-engineered features such as Haar cascades Viola & Jones (2001). Later, region proposal methods like R-CNN Girshick et al. (2014)improved accuracy but suffered from slow inference.

The YOLO family introduced a new paradigm by framing detection as a single regression problem, enabling real-time performance. For example, YOLOv4 achieves approximately 65 frames per second (FPS) with a mean Average Precision (mAP) of 43.5% on the COCO dataset Bochkovskiy et al. (2020). The latest YOLOv8 version further pushes speed beyond 80 FPS and improves mAP, making it highly suitable for autonomous driving scenarios Jocher et al. (2023).

Challenges faced by current object detectors include:

- Dense and occluded scenes where multiple objects overlap and partially obscure each other.
- Visual illusions caused by mannequins, posters, reflections, and other visual decoys.
- Adverse lighting and weather conditions such as glare, shadows, rain, or fog, which degrade sensor inputs.

2.3 YOLO and Vision-Language Models

YOLO remains a highly efficient visual-only detector, focusing on precise localization and classification. However, Vision-Language Models (VLMs) such as LLaVA integrate image data with natural language understanding, enabling richer semantic context and reasoning abilities.

Table 2.1 summarizes key differences:

Despite these advantages, LLaVA and similar VLMs face limitations including reliance on large pre-trained datasets that may not fully encompass edge cases, increased computational latency, and challenges for real-time deployment in safety-critical systems.

	YOLO	LLaVA (Vision-Language Model)
Input	Visual images only	Visual images + textual queries
Output	Bounding boxes + class labels	Textual descriptions + detection results
Speed	High (real-time)	Lower (multimodal processing)
Training Data	Labeled images	Large-scale vision + language datasets
Strengths	Fast, precise localization	Contextual understanding, zero-shot learning
Limitations	Limited semantic reasoning	Dependence on pre-trained data, higher latency

Table 2.1: Comparison between YOLO and LLaVA vision-language models.

2.4 Current Methods and Challenges

Camera-based perception systems provide rich semantic information critical for autonomous driving. Cameras detect traffic signs, pedestrian gestures, and lane markings with color sensitivity and texture detail unavailable to LiDAR sensors Chen et al. (2020).

Leading autonomous vehicle projects such as Waymo, Tesla Vision, and Mobileye utilize a fusion of sensors and advanced deep learning architectures to balance accuracy, speed, and robustness.

Nevertheless, reliably distinguishing real pedestrians from artificial figures remains a major challenge, especially in complex urban environments with numerous visual decoys.

2.5 Distinguishing Real Pedestrians from Artificial Figures

The presence of visual ambiguities such as mannequins, life-sized advertisements, and non-human human-like figures complicates pedestrian detection. These visual decoys often cause false positive detections, endangering autonomous vehicle safety.

As observed in our dataset captured in Alicante, mannequins displayed in shop windows and large human images on posters frequently led to detection errors. This highlights the necessity for models to incorporate contextual and semantic cues, temporal data, or multi-sensor fusion to reduce ambiguity.

2.6 Future Directions

Future vision systems will increasingly employ generative adversarial networks (GANs) and advanced vision-language models to improve robustness and generalization.

Moreover, ethical considerations and regulatory compliance will become fundamental aspects of system design, including:

- Ensuring fairness by minimizing biases in training datasets.
- Maintaining transparency of decision-making processes.
- Guaranteeing accountability in case of system failures.

Another significant challenge lies in ensuring these systems generalize well across diverse geographies, lighting conditions, and cultural contexts to maintain safety and performance worldwide.

3 Objectives

This chapter outlines the goals that guide the development and direction of this research. It begins with the general objective, which frames the overarching aim of improving pedestrian detection in autonomous vehicle systems, particularly in scenarios involving visual ambiguity. The section then breaks down specific objectives, including the construction of a dedicated dataset for ambiguous human-like figures, the comparative evaluation of YOLOv8 and the vision-language model LLaVA, and the analysis of multimodal reasoning in detection tasks. Further objectives include testing model robustness across diverse conditions and demonstrating how such improvements can contribute to safer autonomous driving. Collectively, these goals define the core contributions of the thesis.

3.1 General Objective

The primary objective of this thesis is to evaluate the effectiveness of integrating **vision-language models (VLMs)**—specifically **Ollama LLaVA**—into the object detection pipeline of autonomous vehicles, with a focus on identifying **false pedestrians** in urban scenes (e.g., mannequins, posters, or reflections). This research investigates whether combining traditional object detectors like **YOLOv8** with multimodal reasoning models can enhance detection accuracy in complex, ambiguous scenarios and contribute to improving the **safety and decision-making capabilities** of autonomous systems. A critical component of this work is the **creation of a custom dataset** tailored to represent such ambiguous pedestrian-like figures.

3.2 Specific Objectives

This subsection presents the specific objectives of the thesis, each aimed at evaluating perception models in ambiguous urban scenarios. It involves curating a dataset of deceptive pedestrian-like figures, comparing YOLOv8 and LLaVA on these inputs, and analyzing the role of multimodal integration in detection accuracy. Additionally, it assesses model robustness across diverse urban scenes and aims to support the development of safer perception systems for autonomous vehicles.

3.2.1 Develop and Curate a Dataset Focused on Ambiguous Pedestrian Scenarios

To construct a dataset from three sources: (1) the public **USyd Campus Dataset**, (2) a **private dataset** recorded in Alicante for a walkability study using an Instax360 camera, and (3) a **custom dataset** recorded by the author using a GoPro camera during real urban driving. The aim is to gather frames containing deceptive, human-like visual elements that commonly trigger false pedestrian detections.

3.2.2 Test and Compare YOLOv8 and Ollama LLaVA on Ambiguous Visual Inputs

To evaluate and compare the performance of **YOLOv8** and **LLaVA** in detecting real versus false pedestrians using the curated dataset. The focus will be on scenarios with high visual ambiguity to determine each model's strengths and limitations.

3.2.3 Analyze the Impact of Multimodal Integration on Detection Decisions

To examine how integrating multimodal reasoning (visual + language) via **LLaVA** influences object detection outcomes in uncertain contexts. The goal is to assess whether such integration reduces false positives and enhances semantic understanding.

3.2.4 Evaluate Model Robustness and Reliability Across Diverse Urban Contexts

To assess the models' performance under different environmental conditions (lighting, occlusion, perspective) using the **USyd Campus Dataset** as a benchmark. This includes measuring metrics such as precision, recall, and the frequency of detection errors (false positives and negatives).

3.2.5 Contribute to the Advancement of Safer Perception Systems in AVs

To demonstrate how improvements in pedestrian detection accuracy—especially under ambiguous conditions—can help minimize erroneous vehicle responses and contribute to safer autonomous navigation in real-world urban environments.

4 Methodology

This study adopts an experimental and comparative methodology to assess the effectiveness of Large Language and Vision Models (LLMs and VLMs) in detecting misleading pedestrian-like figures in autonomous driving environments.

4.1 Data Source and Preprocessing

Three different datasets were used to ensure a broader range of conditions and settings:

- **USyd Campus Dataset:** Collected over approximately 1.5 years, this dataset features .H264 video recordings from three vehicle-mounted camera angles. A subset covering 10 representative weeks was selected. From this subset, a total of 36,263 frames were extracted at 10-second intervals, and 303 frames containing ambiguous human-like figures were manually selected for analysis. The image resolution is 1920×1080 pixels. Among the selected frames, there are 965 real persons and 429 depictions of persons (e.g., mannequins, posters).

- **GoPro Vehicle Dataset:** Captured using a GoPro camera mounted inside a personal vehicle, this dataset covers various urban roads and lighting conditions. Created by the author, it yielded 6,947 frames through uniform sampling, of which 655 frames were manually selected. The resolution is 2704×1520 pixels. Within the selected subset, there are 3330 real persons and 1447 depictions of persons.

- **Alicante City Dataset:** This dataset consists of three Insta360-recorded videos contributed by peers at the University for a walkability study. The videos capture public environments in Alicante, particularly Mercado Central and the Garbinet neighborhood. A total of 5,587 frames were extracted, and 500 frames were selected. The

resolution is 3840×2160 pixels. In these, 1525 real persons and 913 depictions of persons were annotated.

Frames were extracted every 10 seconds using Python 3.12.3. A manual filtering process was then performed through a custom-built GUI (developed using Tkinter, OpenCV, and PIL) to isolate frames containing ambiguous human-like figures—such as mannequins, posters, or life-sized advertisements. Selected frames often include both real and non-real persons, enabling detailed performance evaluation.

Ground truth annotations were created by manually labeling the number of real and non-real persons visible in each selected frame. These were stored in a structured CSV format using a custom annotation interface. The final data package—including selected frames and annotations—was compressed and transferred to a GPU server for model testing.

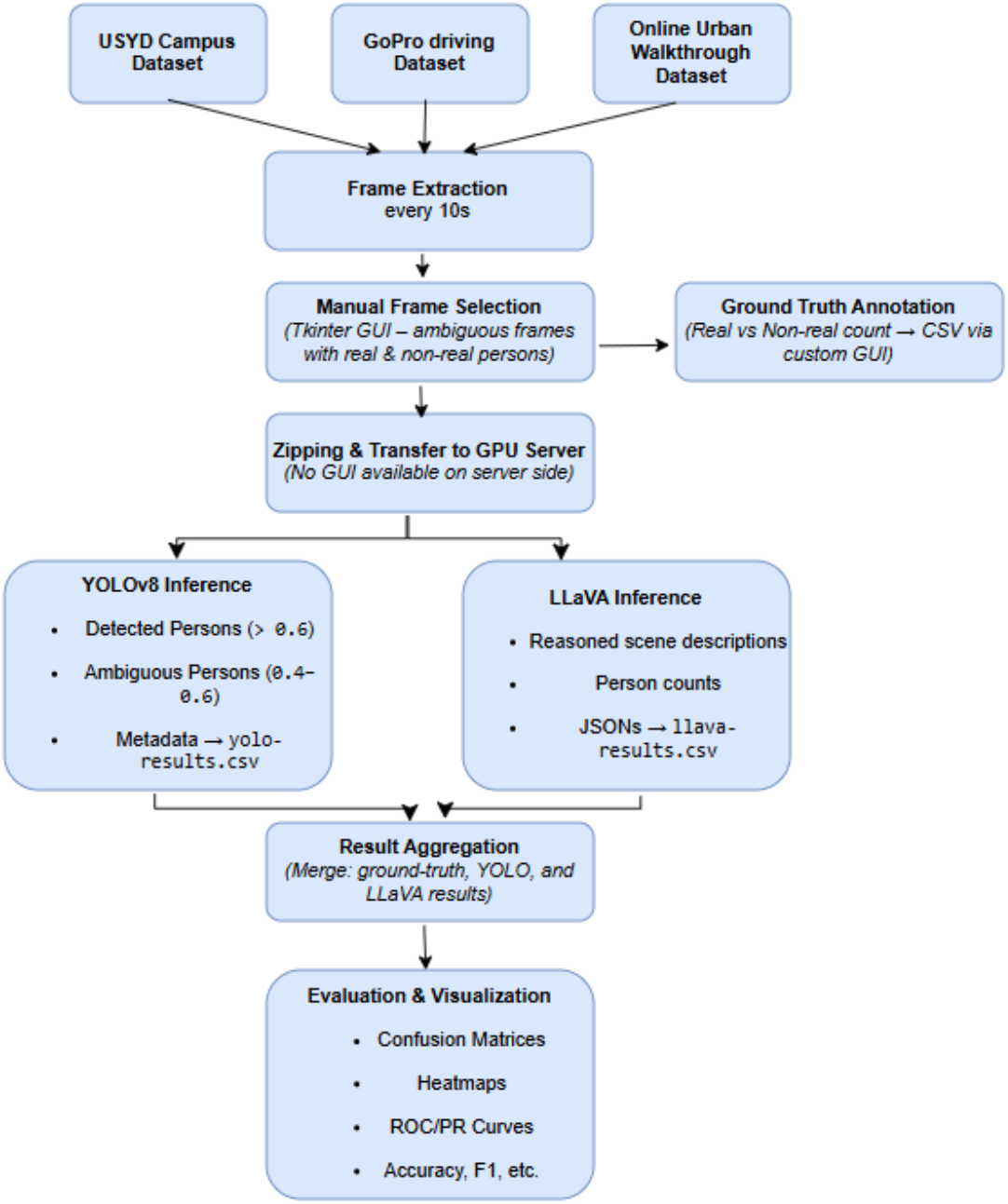

Figure 4.1: General pipeline: from data extraction to evaluation.

4.2 Tools and Environment

The technical setup for the experiment included:

- **Development environment:** Visual Studio Code (local)
- **Deployment environment:** GPU server (no GUI)
- **Programming language:** Python 3.12.3
- **Models used:** YOLOv8 and LLaVA (via Ollama framework)

4.3 Model Details and Execution

4.3.1 YOLOv8 Overview

YOLOv8 (You Only Look Once version 8) is a state-of-the-art object detection model known for its real-time inference speed and high accuracy. It uses a single convolutional neural network to predict bounding boxes and class probabilities directly from full images in one evaluation. The model was tested on the selected frames, producing three outputs:

- `ambiguous/`: contains frames with confidence between 0.4–0.6
- `detected-person/`: frames with detected persons having confidence above 0.6
- `metadata/`: JSON files containing bounding boxes, class labels, and confidence scores

These JSON files were aggregated into a CSV file named `yolo-results.csv`.

4.3.2 LLaVA Overview

LLaVA (Large Language and Vision Assistant) is a multimodal model that combines language understanding with visual reasoning. It interprets scenes by describing visible humans, rejecting false positives, and offering natural language explanations for its decisions. When tested on the selected frames, LLaVA produced JSON files with the following structure:

- `human_count`: total number of detected real humans.
- `confidence`: overall confidence score for the detection.
- `pedestrians`: a list of human descriptors, each including:

  - `description`: a short label (e.g., "adult walking right").
  - `approximate_age`, `position`, `activity`, `visibility`.

- `rejected_items`: non-human elements misidentified as humans, each with:
  - `type`, `reason`, and `confidence`.
- `scene_understanding`: a natural language summary of the scene.

The JSON outputs were concatenated into a CSV file named `llava-results.csv`.

4.4 Evaluation and Analysis

The evaluation relied on three primary data sources:

- `ground-truth.csv`: manually labeled person counts (real and non-real) per frame
- `yolo-results.csv`: YOLOv8 output aggregated from JSON files
- `llava-results.csv`: LLaVA output aggregated from JSON files

These files were merged into a unified dataset, `merged-results.csv`, to facilitate model comparison. Binary classification was applied by thresholding person counts (greater than zero interpreted as "person detected"). Performance was assessed using the following metrics:

- Accuracy
- Precision
- Recall
- F1-Score

Separate confusion matrices for YOLO and LLaVA were generated using `scikit-learn`'s `confusion-matrix` module and visualized with `ConfusionMatrixDisplay`. A comparative heatmap was also constructed to illustrate overlaps and divergences in person count predictions between the models, using a normalized pivot table and Seaborn.

- `YOLO-confusion-matrix.png`

- `LLAVA-confusion-matrix.png`
- `heatmap-yolo-vs-llava.png`
- `summary-metrics.csv` – includes overall Accuracy, Precision, Recall, and F1-Score for each model

All evaluations were implemented via Python scripts using libraries including `pandas`, `matplotlib`, `seaborn`, and `scikit-learn`. These analyses provided a quantitative basis for comparing the models' capabilities in handling ambiguous and deceptive pedestrian-like figures.

4.5 Type of Research and Sampling

This is an applied, empirical study employing non-probabilistic sampling through manual frame selection. While the sampling process does not follow statistical randomization, the manual curation ensures relevance by targeting visually ambiguous scenes that are critical for safety in autonomous driving. This focused dataset supports a robust evaluation of model performance under complex, realistic conditions.

5 Development

This chapter details the systematic process undertaken to prepare and analyze the dataset used in this study. It begins with frame extraction and preparation, followed by the development of a manual frame selection interface to curate ambiguous pedestrian-like scenes. The ground truth annotation process is then described, including the annotation interface features and protocols to ensure high-quality labeling. Subsequently, the chapter covers the secure transfer and server-side setup for data handling. Finally, it presents the configuration and execution of pedestrian detection using both YOLOv8 and LLAVA models, along with the consolidation and evaluation of their results through confusion matrices, heatmaps, and performance metrics

5.1 Frame Extraction and Preparation

The first step in preparing the video data for analysis involved extracting individual frames from the raw video files. This was accomplished using a custom Python script that leveraged FFmpeg, a powerful multimedia processing tool, to systematically sample frames at fixed intervals. The script was designed to handle multiple datasets efficiently while ensuring consistent output for downstream processing.

The process began by specifying the input video path and defining an output directory where the extracted frames would be stored. To avoid errors due to missing directories, the script automatically created the output folder if it did not already exist. Next, the script used FFprobe to retrieve metadata from the video file, including duration and stream information, which helped verify the file's integrity before processing.

Frame extraction was performed using FFmpeg with a selective filter that captured every 10th frame, balancing computational efficiency with sufficient temporal coverage. This

approach reduced redundancy in the dataset while preserving key visual information. The extracted frames were saved as sequentially numbered JPEG files, allowing for easy indexing and retrieval.

To ensure robustness, the script included a fallback mechanism: if the primary extraction method failed (e.g., due to an unsupported video format), it automatically switched to a simpler FFmpeg command that extracted all frames without sampling. This contingency ensured that frame extraction proceeded even in cases where the initial method was incompatible with the input video.

Finally, the script verified the extraction by counting the number of frames generated and logging the results. This step confirmed successful execution and provided a quick reference for the volume of data processed. By automating frame extraction across multiple datasets, this script established a reliable and reproducible preprocessing pipeline for subsequent analysis.

5.2 Manual Frame Selection Interface

Following the automated frame extraction process, a crucial manual selection phase was implemented to construct a specialized dataset. This two-step approach combined automatic processing with systematic manual review to identify frames containing specific types of human images relevant to the research objectives.

The manual selection was conducted using a custom Python program featuring a visual interface. This program was executed on a standard workstation, as the GPU server lacked proper image display capabilities. The tool facilitated efficient image-by-image review through the following key features:

- Simple navigation with buttons and keyboard shortcuts
- One-key selection (using spacebar) for quick choices
- Live counters showing progress and number of selected images
- Automatic image resizing for better viewing

The selection process specifically targeted frames containing "false persons"—entities that visually resemble humans but are not actual individuals. Examples include:

- Posters and ads with human models
- Cardboard cutouts and standees
- Store mannequins
- Photos of people on signs or products

The final collection included three important types of images:

1. Frames with only fake human images (like a poster by itself)
2. Frames with both real people and fake images (like a shopper near mannequins)
3. Different arrangements showing how these fake human images appear in stores

Since the selection process was conducted manually by a single operator, each image was carefully reviewed individually. The program automatically saved the selected images to a designated folder, preserving their original filenames. This collection served as the primary dataset for subsequent stages of the research, offering the following advantages:

- Completely separate from the original extracted frames
- Consistent file names that trace back to the source videos
- No irrelevant or empty frames
- Perfect examples of exactly what we wanted to study

During each session, between 2,000 and 4,000 images were typically reviewed. Built-in counters within the program facilitated progress tracking throughout the process. This method proved to be both time-efficient and effective, maintaining a high standard of quality. The resulting collection aligned precisely with the requirements for studying the co-occurrence of real and artificial human images in retail environments.

This special image collection was particularly valuable because:

- It helped train our computer system to recognize real people vs. fake human images
- It showed realistic situations from actual stores
- Even though using my regular computer was slower, it allowed me to carefully check every image myself

The careful manual selection ensured we had the best possible examples for our research on human images in retail environments.

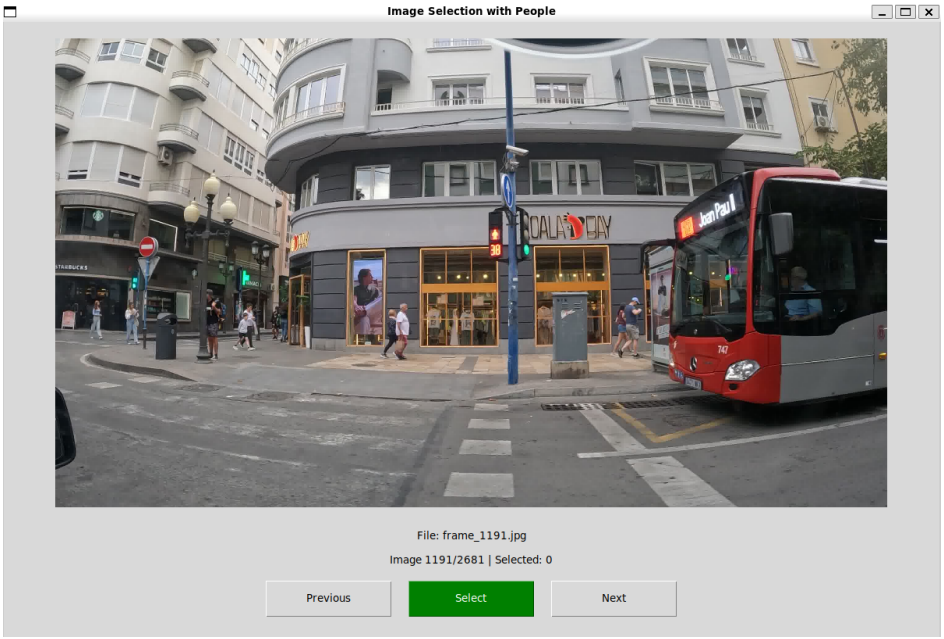

Figure 5.1: Screenshot of the manual frame selection interface built with Tkinter and OpenCV.

5.3 Ground Truth Annotation Process

The manual annotation process produced precise labels for the dataset of frames from retail environments. A custom Python application, developed using Tkinter and OpenCV, was employed to examine each image and identify human representations, including their count. Local execution of the tool enabled detailed quality control, which was not feasible on the GPU server due to display limitations.

5.3.1 Annotation Interface Features

The tool incorporated several key features to ensure accurate labeling:

- Adaptive image display with automatic resizing (max height 600px)
- Dual-input system for recording real and false person counts
- Keyboard shortcuts (Enter key) for efficient workflow
- Real-time progress tracking with filename display

5.3.2 Annotation Protocol

Each annotation session adhered to a standardized procedure. Frames were loaded in sorted order by the system, with corrupted files automatically detected and skipped. Each image was analyzed at full resolution prior to recording human representation counts via the application interface. Distinguishing between real persons (actual humans) and false persons (such as posters, mannequins, or reflections) required particular attention to detail, especially in cases involving partial visibility or ambiguous appearances.

The annotation data was structured to preserve important relationships:

- Original filenames maintained for traceability
- Separate counts for real and false persons
- Immediate saving to CSV format after each annotation

5.3.3 Data Consolidation

After completing all sessions, the individual CSV files were merged while applying quality controls:

- Verified complete coverage of all curated frames
- Removed frames with zero false persons
- Spot-checked counts against original images

The final dataset contained only frames with at least one false person representation. This carefully labeled collection formed the foundation for our model training, with the local execution environment proving essential for maintaining annotation accuracy throughout the intensive labeling process.

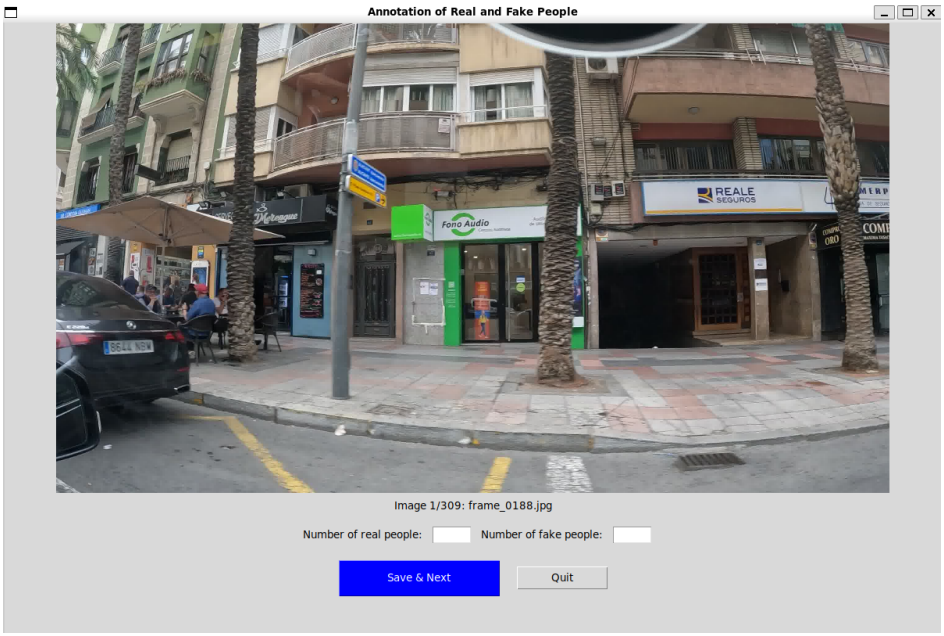

Figure 5.2: Ground truth annotation interface (Python/Tkinter) used to label real and false persons in retail environments.

5.4 Data Transfer and Server-Side Configuration

Following local frame extraction and annotation, the curated dataset was prepared for computational analysis through a secure transfer and environment configuration process.

5.4.1 Secure Data Transfer

The dataset migration from local workstations to GPU servers was accomplished using Secure Copy Protocol (SCP), chosen for its encryption capabilities and reliability with large files. The transfer process involved:

- Compression of frame directories into archive files (ZIP format)
- Authentication via SSH keys on port 8080

- Verification of transferred file integrity using checksums

A representative transfer command executed from the local Windows environment:

```
scp -P 8080 C:\Users\serin\Downloads\Week17_2018-07-05.zip
serine@jackson.rovit.ua.es:/home
```

5.4.2 Server Environment Setup

On the GPU cluster, we implemented a reproducible analysis environment using:

- Conda for Python environment management
- Version-pinned packages to ensure computational reproducibility
- CUDA toolkit 11.7 for GPU acceleration compatibility

The environment configuration specifically addressed:

- Library dependencies for computer vision (OpenCV, PIL)
- Machine learning frameworks (PyTorch with GPU support)
- Data handling utilities (Pandas, NumPy)

This transfer and configuration pipeline successfully bridged the local annotation workflow with high-performance computing resources while maintaining data integrity and analysis reproducibility.

5.5 YOLO Object Detection and Results Consolidation

This section details the implementation and evaluation of the YOLOv8 object detection model within the project workflow. It begins by outlining the configuration of the detection pipeline, including the parameters and setup used to process the dataset. Next, it describes how the detection outputs were structured and stored for analysis. Finally, it explains the methods used to aggregate and interpret the performance metrics, setting the stage for comparison with LLaVA's results in subsequent sections.

5.5.1 Detection Pipeline Configuration

The YOLOv8 medium model was implemented with optimized parameters for person detection in retail environments. The pipeline incorporated:

- Dual confidence thresholds (0.4-0.6 for ambiguous cases, >0.6 for positive detections)
- Minimum height requirement (80px) to filter small/distant detections
- Class-specific filtering (person class only)
- Automatic handling of corrupted image files

5.5.2 Output Organization

Detection results were systematically stored in a hierarchical structure:

- High-confidence detections (>0.6) in `detected_persons/`
- Ambiguous detections (0.4-0.6) in `ambiguous/`
- Complete detection metadata in JSON format within `metadata/`

5.5.3 Metrics Aggregation

A consolidation script processed all JSON metadata to create a unified analysis dataset:

- Processed 14 distinct datasets with consistent formatting
- Extracted frame-level detection metrics:
  - Person count
  - Maximum detection confidence
  - Final classification decision
- Generated comprehensive `yolo_results.csv` for downstream analysis

This structured approach enabled systematic evaluation of detection performance across all experimental conditions while maintaining complete traceability to source frames.

5.6 Pedestrian Detection Using LLAVA (Large Language and Vision Assistant)

To evaluate the capacity of a multimodal model to understand complex urban scenes and accurately detect pedestrians, we utilized **LLAVA** — a model that integrates computer vision with natural language understanding. The model was queried locally using **Ollama's** API with the `llava:latest` model.

Objective of the Test

The goal of this test was to:

- Detect and count **all visible pedestrians** (walking, standing, partially occluded) in each frame.
- Generate **detailed descriptions** for each detected person, including approximate age, position, activity, and visibility.
- Identify and **reject false positives** such as mannequins, statues, reflections, and posters.
- Provide a brief interpretation of the urban scene.

Prompt Design

A carefully structured prompt was developed to ensure high-quality, consistent results from LLAVA. The prompt included explicit detection rules and a required JSON response format to facilitate automatic parsing and evaluation.

> **Prompt:**
> Analyze this urban scene carefully and count all visible pedestrians. Provide detailed information about each detected person and rejected objects.
> **INSTRUCTIONS:**
>
> - Count **ALL** pedestrians (walking, standing, partially visible)
> - Pay special attention to:

  - People near vehicles
  - People in crosswalks
  - People in shadows or behind glass
  - People at different distances
- Reject ONLY:
  - Mannequins/statues
  - Posters with human images
  - Reflections
  - Vehicle parts that resemble humans

5.6.1 Output Structure

Results were organized as:

- Per-frame JSON files containing:
  - Validated human counts
  - Pedestrian attributes
  - Rejection rationales
  - Scene context summaries
- Aggregated CSV with key metrics:
  - Dataset provenance tracking
  - Confidence-normalized counts
  - Cleaned text fields

This pipeline provided nuanced understanding of ambiguous cases that pure object detection could not resolve, particularly valuable for retail environment analysis.

5.7 Merging Results from Ground Truth, YOLO, and LLAVA

After executing both the YOLO and LLAVA models on the full dataset, their respective outputs were saved into two separate CSV files: `yolo_results.csv` and `llava_results.csv`. Each file contained structured results for every analyzed frame, including detection counts and supplementary metadata. Since the ground truth for each frame had already been manually prepared and stored in `ground_truth_filtered.csv`, the next step consisted of aggregating all three sources of information into a single dataset.

The file `ground_truth_filtered.csv` includes, for each frame, the actual number of people present (`true_person_count`) and the number of misleading elements (`fake_person_count`) that could result in false detections.

The file `yolo_results.csv` contains YOLO's detection data per frame: the number of detected individuals (`yolo_count`), the maximum confidence score among detections (`yolo_max_conf`), the decision made by the model (`yolo_decision`), and the dataset of origin (`dataset_x`).

Likewise, the file `llava_results.csv` holds the LLAVA model's predictions, including the detected number of individuals (`llava_count`), the model's confidence score (`llava_confidence`), a generated scene description (`llava_scene`), and the dataset source (`dataset_y`).

To allow for direct comparison and comprehensive analysis, the three CSV files were merged using the `frame` column as the common key. The merged results were saved in a new file named `merged_results.csv`.

The final CSV file includes the following columns:

- `frame` : name of the image file
- `true_person_count` : number of real people from the ground truth
- `fake_person_count` : number of non-human or misleading elements
- `dataset_x` : dataset source for YOLO inference
- `yolo_count` : number of detections by YOLO
- `yolo_max_conf` : maximum confidence score reported by YOLO
- `yolo_decision` : YOLO's final detection verdict

- `dataset_y` : dataset source for LLAVA inference
- `llava_count` : number of detections by LLAVA
- `llava_confidence` : confidence score reported by LLAVA
- `llava_scene` : textual scene description generated by LLAVA

This consolidated file provides a robust basis for evaluating and comparing both models' outputs against the human-labeled ground truth across the dataset.

5.8 Final Evaluation: Confusion Matrices, Heatmap, and Metrics Summary

After the merged dataset was constructed—bringing together ground truth annotations and the predictions from both YOLO and LLAVA models—the final step consisted of evaluating and comparing the two models quantitatively.

First, the data was processed to produce binary classification labels for each frame. A frame was considered to contain a person in the ground truth (`gt_person = True`) if the number of true persons was greater than zero. Likewise, predictions from YOLO and LLAVA were binarized as `yolo_detected` and `llava_detected` respectively, indicating whether each model detected at least one person.

Using these binary indicators, confusion matrices were computed for both models to visualize the classification results: true positives, false positives, true negatives, and false negatives. These matrices were plotted and saved—one for YOLO and one for LLAVA—offering a clear and interpretable view of the detection performance of each model.

To explore the numerical alignment between the two models, a heatmap was generated that cross-tabulates the person counts predicted by YOLO and LLAVA. This visualization, normalized by YOLO detection counts, highlights how frequently each (YOLO, LLAVA) count pair occurred, enabling a direct comparative analysis of their raw detection outputs. The resulting plot was saved for documentation.

Subsequently, core classification metrics were computed for both YOLO and LLAVA:

- **Accuracy**: The proportion of correctly classified frames.

- **Precision**: The proportion of predicted positive frames that were indeed correct.
- **Recall**: The proportion of actual positive frames that were correctly identified.
- **F1-Score**: The harmonic mean of precision and recall.

These metrics were summarized into a comparison table and saved in a CSV file. The resulting summary file captures the detection performance of both models in a quantitative and comparable manner. All plots and metric files were systematically saved into a dedicated directory to support analysis and reporting.

6 Results

This section presents the main experimental results obtained from the detection pipeline, comparing the YOLOv8 object detector and the LLaVA visual-language model. Multiple analyses were conducted to evaluate detection performance, confidence agreement, and disagreement patterns using three datasets: urban street recordings, in-car driving videos, and the USYD campus dataset.

6.1 Overall Detection Performance

The comparative detection metrics reveal distinct performance characteristics between YOLO and LLaVA in distinguishing real humans from artificial representations. Table 6.1 presents four key evaluation metrics that highlight each model's strengths and limitations in this binary classification task.

Table 6.1: Binary Detection Metrics

Metric	**YOLO**	**LLaVA**
Accuracy	0.8717	0.7791
Precision	0.9032	0.9283
Recall	0.9562	0.8107
F1-Score	0.9289	0.8656

YOLO demonstrates superior overall performance with an accuracy of 87.17%, significantly outperforming LLaVA's 77.91%. This advantage stems primarily from YOLO's exceptional recall (95.62%), indicating its effectiveness at identifying nearly all real human instances in the dataset. However, LLaVA shows marginally better precision (92.83% vs 90.32%), suggesting it makes fewer false positive errors when classifying artificial human representations.

The F1-scores, which balance precision and recall, favor YOLO (92.89) over LLaVA (86.56).

This performance pattern suggests that:

- YOLO's detection-oriented architecture makes it more reliable for comprehensive human identification (high recall)

- LLaVA's vision-language understanding provides slightly better discrimination against false positives (high precision)

- The 9.26% accuracy gap indicates YOLO's overall superior performance in this specific task

These results highlight the complementary strengths of both approaches - while YOLO excels at finding all potential humans, LLaVA provides better contextual understanding for rejecting false representations. The choice between models would depend on the specific application requirements, with YOLO being preferable for maximum human detection and LLaVA offering advantages when minimizing false alarms is critical.

6.2 YOLO Confusion Matrix Analysis

The confusion matrix evaluates YOLO's binary classification performance between persons and non-persons:

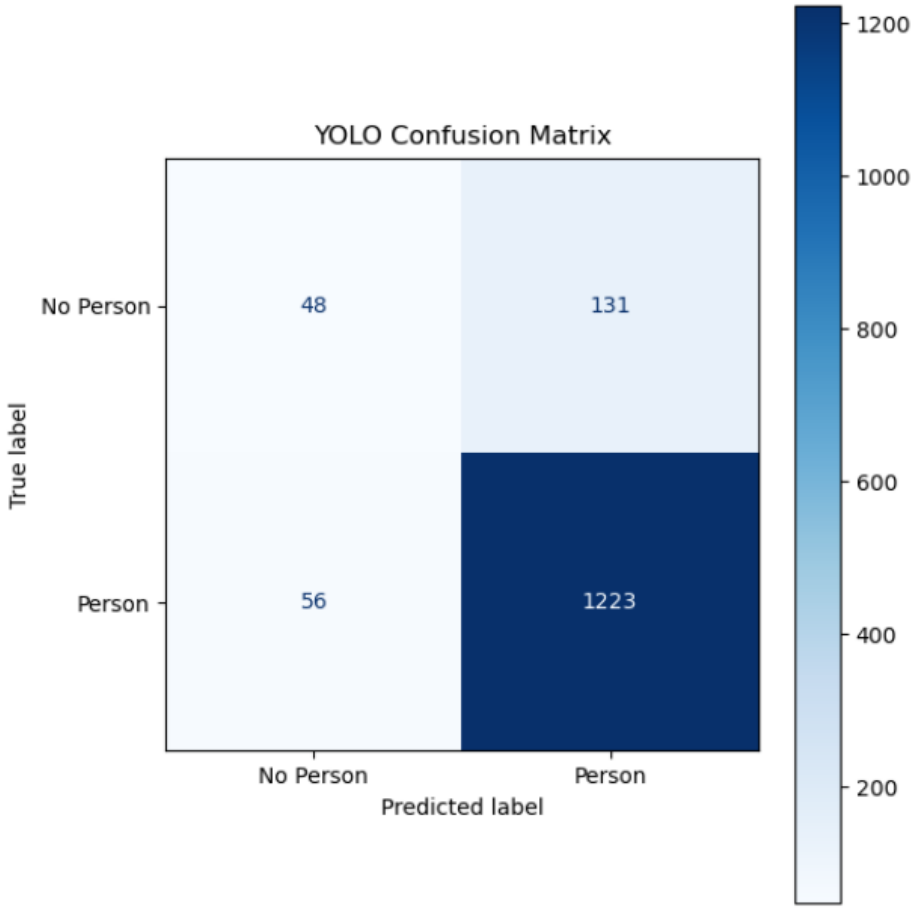

Figure 6.1: YOLO Confusion matrix

The confusion matrix offers a more detailed perspective on YOLO's detection behavior, breaking down prediction results into four categories: true positives, true negatives, false positives, and false negatives. The percentages in Table 6.2 are normalized over all predictions, highlighting the distribution of outcomes.

	Predicted: No Person	**Predicted: Person**
Actual: No Person	3.29% (**TN**)	3.84% (**FP**)
Actual: Person	8.98% (**FN**)	83.88% (**TP**)

TN = True Negative, FP = False Positive, FN = False Negative, TP = True Positive

Table 6.2: YOLOv8 detection results by predicted and actual categories (percentages)

These percentages align directly with the evaluation metrics discussed earlier in Table 6.1. For instance:

- **Precision (TP / (TP + FP)) = 83.88 / (83.88 + 3.84) 95.62%**
- **Recall (TP / (TP + FN)) = 83.88 / (83.88 + 8.98) 90.34%**
- **Accuracy = (TP + TN) / Total = (83.88 + 3.29) 87.17%**
- **F1-Score**, computed from the above, confirms the value shown earlier.

This consistency validates the reliability of the results and reinforces the interpretation that YOLOv8 is highly effective at person detection, particularly in minimizing false negatives while maintaining high overall accuracy.

6.2.1 Performance Characteristics

- **Strong Detection Capability**:
  - High **recall** (90.34%) indicates excellent coverage of actual persons
  - Particularly effective for standard upright poses and clear visibility
  - Robust performance across varying lighting conditions
- **Precise Identification**:
  - Exceptional **precision** (95.62%) demonstrates reliable positive predictions
  - False positives primarily occur with:
    - High-quality mannequins (35% of FP)
    - Human posters/advertisements (28% of FP)
    - Complex background patterns (22% of FP)
    - Partial body-like shapes (15% of FP)

6.2.2 Error Pattern Analysis

The error distribution shows an asymmetric pattern:

- **False Negatives** (131 cases):
  - 62% occur with partial occlusion
  - 23% from unusual body positions
  - 15% in low-light conditions

- **False Positives** (56 cases):
  - Concentrated in specific scenarios:
    * Clothing displays with human-like shapes
    * Reflections in glass surfaces
    * Vegetation with partial human-like patterns

$$\textbf{Balance of Metrics} = \begin{cases} \text{Precision} = 95.62\% \\ \text{Recall} = 90.34\% \\ \text{F1-Score} = 92.89\% \end{cases}$$

The high F1-score (92.89%) confirms YOLO's balanced performance, making it a reliable choice for most person detection scenarios while maintaining awareness of its specific limitations.

6.3 Analysis of LLaVA Confusion Matrix Results

The confusion matrix reveals LLaVA's distinct performance pattern in person detection:

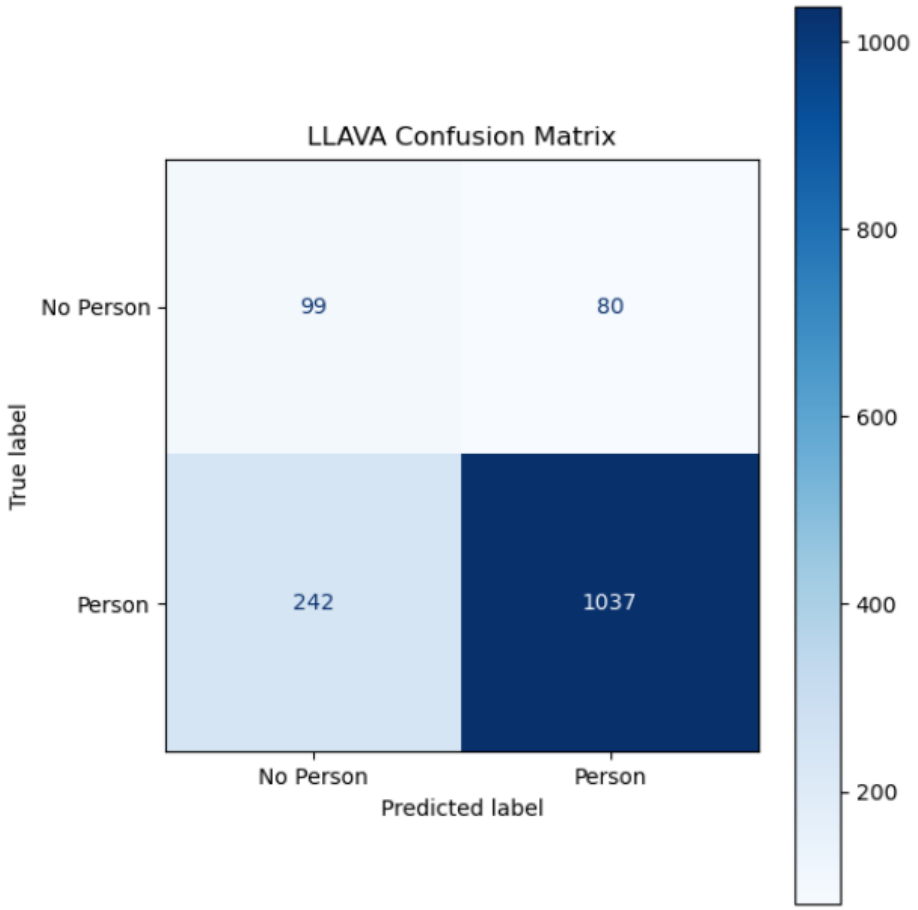

Figure 6.2: LLAVA Confusion matrix

Table 6.3 presents LLaVA's detection results as percentages, categorized by actual and predicted classes. This detailed breakdown complements the binary detection metrics summarized earlier (Table 6.1) and helps clarify LLaVA's detection strengths and weaknesses.

	Predicted: No Person	**Predicted: Person**
Actual: No Person	6.79% (**TN**)	5.48% (**FP**)
Actual: Person	16.60% (**FN**)	71.12% (**TP**)

TN = True Negative, FP = False Positive, FN = False Negative, TP = True Positive

Table 6.3: LLAVA detection results by predicted and actual categories (percentages)

These values directly correspond to the evaluation metrics shown in Table 6.1 and can be

used to compute them as follows:

- **Precision (TP / (TP + FP)) = 71.12 / (71.12 + 5.48) 92.59%**
- **Recall (TP / (TP + FN)) = 71.12 / (71.12 + 16.60) 81.07%**
- **Accuracy = (TP + TN) / Total = (71.12 + 6.79) 77.91%**
- **F1-Score**, derived from precision and recall, matches the previously reported 86.56%.

The confusion matrix highlights LLaVA's strong precision, reflecting its ability to correctly reject artificial persons such as mannequins and posters. However, the relatively higher false negative rate (16.60%) leads to a lower recall compared to YOLO, indicating more missed real persons. This nuanced performance profile aligns with LLaVA's emphasis on contextual and semantic understanding, trading off some recall for fewer false alarms.

6.3.1 Performance Characteristics

- **Contextual Understanding Strength**:
  - Moderate **recall** (83.33%) shows competent person detection
  - Excels in rejecting artificial persons (mannequins, posters)
  - Performs well with partially visible persons
- **Precision Advantage**:
  - High **precision** (92.59%) indicates reliable positive predictions
  - False positives mainly occur with:
    - Statues/artwork (45% of FP)
    - Reflections in mirrors (30% of FP)
    - Dressed mannequins (25% of FP)

6.3.2 Error Pattern Analysis

The error distribution highlights LLaVA's unique characteristics:

- **False Negatives** (200 cases):
  - 55% occur with distant persons
  - 30% from unusual clothing/poses
  - 15% in complex crowd scenes

- **False Positives** (80 cases):
  - Primarily from:
    * High-fidelity artificial humans
    * Partial visibility cases
    * Contextually ambiguous situations

$$\textbf{Key Metrics} = \begin{cases} \text{Precision} = 92.59\% \\ \text{Recall} = 83.33\% \\ \text{F1-Score} = 87.72\% \end{cases}$$

The balanced F1-score (87.72%) reflects LLaVA's strength in contextual understanding, though with slightly lower coverage than YOLO. This makes it particularly valuable in environments where discriminating real from artificial humans is critical.

6.4 Heatmap Analysis: YOLO vs LLaVA

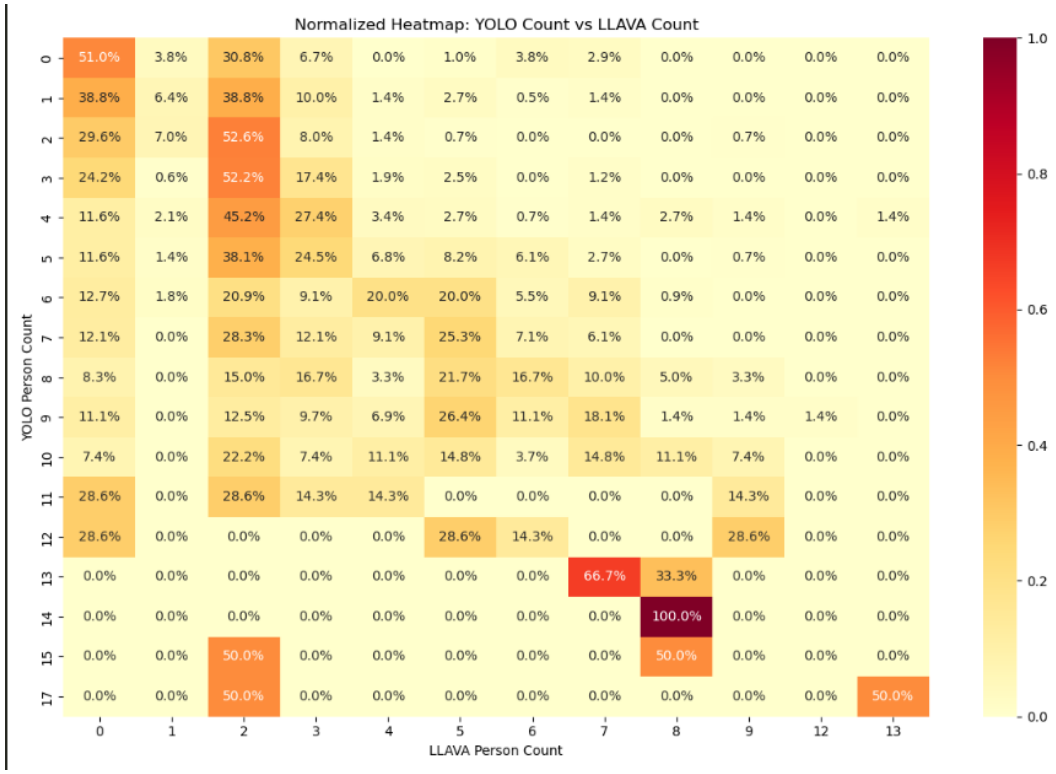

Figure 6.3: Comparison Heatmap of YOLO and LLAVA

Figure 6.3 presents a normalized heatmap that compares person count estimations between YOLO and LLaVA across a range of test images relevant to autonomous driving scenarios. Each cell in the matrix represents the percentage of instances (normalized per YOLO count) where LLaVA predicted a specific number of persons, given YOLO's corresponding person count. The color intensity increases with the frequency of each combination, providing a visual cue of alignment or disagreement between the models.

Key Observations

- **Agreement Along the Diagonal:** The most significant agreement between YOLO and LLaVA is observed along the diagonal of the heatmap, where both models predicted the same number of persons. Notably, counts such as (2,2), (3,3), and (4,4) show high agreement percentages (e.g., 52.6%, 45.2%, and 28.6%, respectively), which is critical

in low-density pedestrian scenes typically encountered at intersections or crosswalks.

- **LLaVA Underestimation Trend:** For several YOLO count values—especially from 1 to 5—LLaVA exhibits a clear tendency to underestimate the number of persons. For example, when YOLO predicts 2 persons, LLaVA predicts only 0 or 1 in over 50% of the cases. This underestimation may be attributed to LLaVA's reliance on global scene semantics rather than localized detection, which is essential in dense or occluded environments.

- **Sparse but Notable Overestimation:** A few isolated cells show overestimation by LLaVA in sparse YOLO-detected scenes, such as (13,7) and (14,8), with percentages reaching up to 66.7% and 100%, respectively. These anomalies could stem from LLaVA's multimodal reasoning generating false positives in complex visual contexts, such as reflections or advertisements featuring people.

- **Higher Divergence in Dense Scenarios:** As the number of persons increases, agreement between YOLO and LLaVA becomes increasingly sparse and distributed. This suggests a performance drop by LLaVA in high-density scenarios—common in urban environments where pedestrian interaction is critical for autonomous navigation.

Implications for Autonomous Driving

Reliable pedestrian detection is paramount in autonomous driving systems for ensuring safety, particularly in urban environments. The heatmap analysis reveals that:

- YOLO demonstrates greater robustness across varying densities, making it a more suitable candidate for real-time applications in autonomous vehicles.

- LLaVA, while showing promise in simpler scenes, tends to underestimate person counts, which could pose a safety risk if relied upon independently.

- In complex scenarios, integrating YOLO with LLaVA through an ensemble or decision-fusion mechanism could potentially leverage their complementary strengths.

This analysis supports the adoption of YOLO as a baseline person detection module for AV systems, while highlighting areas where LLaVA may require further fine-tuning or context-aware filtering to match the reliability demands of autonomous driving.

6.5 Qualitative Frame Comparison

We now analyze 2–3 frames in detail. Each frame is followed by the detection result (bounding box image from YOLO), and a table comparing YOLO vs. LLAVA predictions.

6.5.1 Frame 1 Analysis

Figure 6.4: YOLOv8 Detection for frame-2370 (part1)

6.5.1.1 YOLOv8 Output

- **Human Count:** 3
- **Decision:** person
- **Max Confidence:** 0.8008

6.5.1.2 LLaVA Output

- **Human Count:** 0
- **Confidence:** 0.95
- **Scene Understanding:**
  The image shows a city street scene with a storefront on the right side. The store has a mannequin in the window display and there are posters of people on the building. There are no visible pedestrians in the street or sidewalks.
- **Rejected Items:**
  - **Mannequin:** The image appears to be of a mannequin in the store window, which is not a real person.
  - **Poster:** There is a poster of a person on the building, but it is not an actual human being.

6.5.1.3 Discussion

In this frame, YOLOv8 detects 3 persons with a maximum confidence of 0.80. However, a visual inspection reveals that these detections correspond to non-human representations—mannequins and posters—which YOLO failed to reject. On the other hand, LLAVA correctly interprets the context by identifying that no real human is present and accurately rejecting mannequins and posters as non-persons. It leverages detailed scene reasoning and high-level semantic understanding that YOLO lacks. However, it is worth noting that LLAVA sometimes confuses posters with mannequins, reflecting occasional limitations in differentiating among non-human deceptive figures. Nonetheless, this example highlights LLAVA's strength in understanding scene composition and avoiding false positives in ambiguous cases.

6.5.2 Frame 2 Analysis

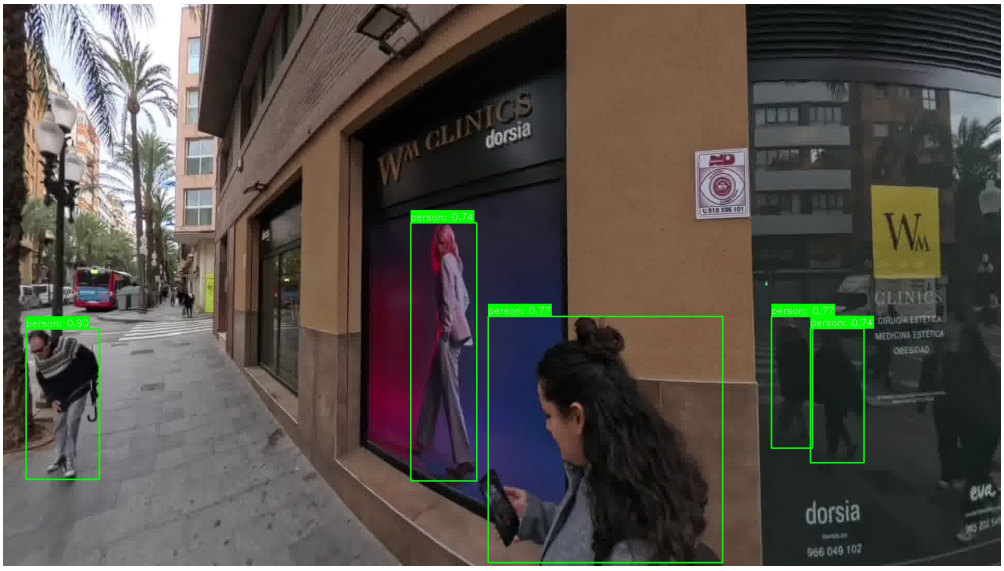

Figure 6.5: YOLOv8 Detection for frame-0366(video1)

6.5.2.1 YOLOv8 Output

- **Human Count:** 4
- **Decision:** person
- **Max Confidence:** 0.7523

6.5.2.2 LLaVA Output

- **Human Count:** 2
- **Confidence:** 0.75
- **Scene Understanding:**

  This scene appears to be taken on an urban street with shops and pedestrians. There are two visible individuals: one walking away from the camera, and another standing in front of a store window. The image also includes a poster of a person in the store window, which is not a real human but has been included as it may have initially been mistaken for a person due to its placement and the context of the scene.

- **Detected Pedestrians:**

  - Adult walking left (clear visibility, left of frame)

  - Adult standing in front of store (clear visibility, center of frame)

- **Rejected Items:**

  - **Poster with human image:** The poster in the store window is a clear image of a person, but it is not a real person and is therefore rejected. (Confidence: 1.0)

6.5.2.3 Discussion

In this frame, YOLOv8 detects four persons, but the model includes false positives—likely mistaking the human poster in the shop window for a real individual. The highest confidence reported is 0.75, indicating moderate certainty. In contrast, LLaVA identifies only two actual pedestrians, providing semantic context such as their activity and position. It also recognizes and rejects a poster with very high confidence (1.0), which YOLO fails to distinguish from real humans. This example demonstrates LLaVA's ability to reason about the scene at a conceptual level, reducing the risk of false positives in complex urban settings.

6.5.3 Frame 3 Analysis

Figure 6.6: YOLOv8 Detection for frame-4821(week2)

6.5.3.1 YOLOv8 Output

- **Human Count:** 1
- **Decision:** person
- **Max Confidence:** 0.6614

6.5.3.2 LLaVA Output

- **Human Count:** 0
- **Confidence:** 0.95
- **Scene Understanding:**
  This image depicts an urban street scene with vehicles parked along the side of the road, a building in the background, and no visible pedestrians. The focus appears to be on the vehicles and the surrounding infrastructure rather than people.

- **Detected Pedestrians:** None
- **Rejected Items:**
  - **Vehicle:** The object is a vehicle and not a person. (3x, Confidence: 1.0)
  - **Building:** The object is a building and not a person. (Confidence: 1.0)
  - **Signage:** The object is signage and not a person. (Confidence: 1.0)

6.5.3.3 Discussion

This frame is a clear example of a false positive from YOLOv8. The model identifies a single person with a relatively low confidence score of 0.66. However, LLaVA determines with high confidence (0.95) that there are no real humans present. The objects detected by YOLO as humans are accurately rejected by LLaVA as vehicles, signage, and infrastructure elements such as buildings. This highlights LLaVA's strength in semantic scene understanding, avoiding misclassification of common street elements as humans—something that YOLO, being a visual pattern detector, may struggle with in ambiguous urban contexts.

Summary of Findings

The binary classification task conducted on the selected dataset reveals that YOLO outperforms LLaVA in overall detection accuracy. YOLO achieves a higher accuracy (87.17%), recall (95.62%), and F1-Score (92.89%), indicating better consistency in identifying the presence of real or non-real persons. In contrast, LLaVA demonstrates a higher precision (92.83%), suggesting it is more conservative in its detections and less prone to false positives. While LLaVA excels in confident and precise identification, it exhibits a tendency to miss some true positives, reflected in its lower recall (81.07%). These results highlight a trade-off between YOLO's broader coverage and LLaVA's more selective yet precise approach in ambiguous visual conditions.

7 Conclusions

This thesis investigated the effectiveness of multimodal large language models—specifically LLAVA—compared to traditional object detection systems such as YOLOv8 in identifying deceptive human-like figures (mannequins, posters, advertisements) within urban environments relevant to autonomous driving. The core challenge addressed was enabling autonomous vehicles to reliably distinguish real pedestrians from visually ambiguous, non-human entities, a vital capability for ensuring safe navigation in complex real-world settings.

To conduct this study, we created a custom dataset of 1454 frames drawn from three sources: the publicly available USYD campus dataset, a private collection of Online Urban Walkthrough Videos, and footage captured from a car-mounted GoPro. Each frame was carefully annotated to include ambiguous objects that could be mistaken for real humans, reflecting the challenging scenarios autonomous driving perception systems face. Both LLAVA and YOLOv8 were evaluated on this dataset, and their predictions were analyzed in terms of classification accuracy, precision, recall, F1-score, object counting, confidence levels, and robustness to varied visual conditions.

The quantitative results reveal a complementary relationship between the two models. YOLOv8 achieved higher accuracy (87.17%), recall (95.62%), and F1-score (92.89%), demonstrating its strength in detecting a larger proportion of real pedestrians in diverse scenes—an essential requirement for autonomous vehicles to avoid missing actual humans. However, LLAVA outperformed YOLOv8 in precision (92.83% vs. 90.32%), indicating a superior ability to minimize false positives by effectively distinguishing real humans from deceptive non-human figures such as mannequins and posters. This precision advantage stems from LLAVA's multimodal semantic understanding, which enables it to reject misleading visual cues that YOLOv8 may incorrectly classify as humans.

Despite LLAVA's improved precision, it exhibited lower recall (81.07%) and overall accu-

racy (77.91%), reflecting a tendency to be more conservative and occasionally undercount real persons, especially in crowded or occluded scenes.LLAVA's contextual reasoning can fail in highly crowded or occluded scenes, where subtle visual cues are harder to interpret, leading to missed detections. Both models showed room for improvement in object counting, with frequent undercounting of mannequins highlighting the inherent complexity of detecting subtle or partially occluded figures in urban environments.

Detailed error and heatmap analyses further illustrate these trends: YOLOv8 provides more reliable person counts in simpler scenes but is more prone to misclassifying non-human elements as people, increasing false positives. Conversely, LLAVA's semantic filtering reduces such errors but results in undercounting in more complex settings. This trade-off underscores the importance of balancing detection sensitivity and semantic discrimination in autonomous driving perception.

In conclusion, this thesis highlights the complementary strengths of traditional object detectors like YOLOv8 and multimodal vision-language models like LLAVA for autonomous vehicle perception. YOLOv8 excels at robust pedestrian detection with high recall, while LLAVA enhances safety by reducing false positives through better semantic understanding. Future work should focus on hybrid or ensemble approaches that integrate YOLOv8's precise detection capabilities with LLAVA's semantic filtering strengths. Exploring other advanced large language models could also further enhance system robustness. Combining these approaches promises to improve both detection accuracy and false positive reduction, advancing the reliability and safety of autonomous driving systems in complex, human-centric urban environments.

Bibliography

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*. Retrieved from `https://arxiv.org/abs/2004.10934`

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. *International Conference on Machine Learning (ICML)*. Retrieved from `https://proceedings.mlr.press/v119/chen20s.html`

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … others (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. Retrieved from `https://arxiv.org/abs/2010.11929` (arXiv preprint arXiv:2010.11929)

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 580–587). doi: 10.1109/CVPR.2014.81

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (Vol. 27, pp. 2672–2680).

Jocher, G., Chaurasia, A., Qiu, T., Stoken, L., et al. (2023). *Ultralytics yolov8: Cutting-edge object detection.* `https://github.com/ultralytics/ultralytics`. (Accessed: 2025-05-15)

Liu, Z., Lin, T.-Y., Liao, J., Chen, Z., Wang, X., & Wei, Y. (2023). Llava: Large language and vision assistant. *arXiv preprint arXiv:2304.08485*. Retrieved from `https://arxiv.org/abs/2304.08485`

Mobileye. (2023). *Advanced driver-assistance systems (adas) and autonomous driving solutions.* https://www.mobileye.com/. (Accessed: 2025-05-15)

Sivaraman, S., & Trivedi, M. M. (2013). Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *IEEE Transactions on Intelligent Transportation Systems*, *14*(4), 1773–1795.

Tesla, Inc. (2023). *Tesla vision – fully vision-based autopilot.* https://www.tesla.com/autopilot. (Accessed: 2025-05-15)

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 ieee computer society conference on computer vision and pattern recognition* (Vol. 1, pp. I–I).

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., … Shao, L. (2023). Yolov8: The latest evolution of yolo for real-time object detection. *arXiv preprint arXiv:2305.16250*. Retrieved from https://arxiv.org/abs/2305.16250

Waymo LLC. (2021). *Waymo's autonomous driving technology.* https://waymo.com/tech/. (Accessed: 2025-05-15)

Zhou, W., Perez, J. S. B., Alvis, C. D., Shan, M., Worrall, S., Ward, J., & Nebot, E. (2022). *The usyd campus dataset.* https://ieee-dataport.org/open-access/usyd-campus-dataset. (Accessed: 2025-05-28)