AUTONOMOUS DRIVER ASSISTANT SYSTEM FOR UNSTRUCTURED ENVIRONMENTS

Thesis Submitted
in Partial Fulfillment of the Requirements for the
Degree of

DOCTOR OF PHILOSOPHY

in

Electronics and Communication Engineering

by

Shilpa Jain

(Roll No. 2K21/PHDEC/507)
Under the supervision of

Prof. S. Indu
(Professor)
Department of Electronics &
Communication Engineering
Delhi Technological University,
Delhi-110042

Prof. Nidhi Goel
(Professor)
Department of Electronics &
Communication Engineering
Indira Gandhi Delhi Technological
University for Women,
Delhi-110006



DEPARTMENT OF ELECTRONICS AND COMMUNICATION
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

October, 2026

© DELHI TECHNOLOGICAL UNIVERSITY, DELHI
2026

ALL RIGHTS RESERVED



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, **Shilpa Jain**, hereby declare that the work which is being presented in the thesis entitled "**Autonomous Driver Assistant System for Unstructured Environments**", in partial fulfilment of the requirements for the award of the Degree of Doctor of Philosophy, submitted in the **Department of Electronics and Communication Engineering**, Delhi Technological University, Delhi is an authentic record of my own work carried out during the period from **Jan 2022 to July 2026** under the supervision of **Prof. S. Indu**, Professor in the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India, and **Prof. Nidhi Goel**, Professor in the Department of Electronics and Communication Engineering, Indira Gandhi Delhi Technical University for Women, Delhi, India.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Place: Delhi
Date:

Ms. Shilpa Jain
Ph.D. Student
(Roll No. 2K21/PHDEC/507)

This is to certify that the student has incorporated all the corrections suggested by the examiner in the thesis, and the statement made by the student is correct to the best of our knowledge.

Signature of Supervisors **Signature of External Examiner**



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

Certified that **Shilpa Jain (2K21/PHDEC/507)** has carried out her research work presented in this thesis, entitled **"Autonomous Driver Assistant System for Unstructured Environments"**, for the award of Doctor of Philosophy from Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, under our supervision. The thesis embodies the results of original work, and studies are carried out by the student herself, and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Prof. S. Indu
(Professor)
Department of Electronics &
Communication Engineering
Delhi Technological University,
Delhi-110042

Prof. Nidhi Goel
(Professor)
Department of Electronics &
Communication Engineering
Indira Gandhi Delhi Technological
University for Women,
Delhi-110006

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

ACKNOWLEDGEMENT

The completion of this doctoral journey has been an enriching and rewarding experience, and I am deeply grateful to all those who have contributed to its successful accomplishment through their support, guidance, and encouragement.

First and foremost, I express my sincere gratitude to the Almighty for granting me the strength, wisdom, and perseverance to complete this research work.

I extend my heartfelt gratitude to my esteemed research supervisors, Prof. S. Indu, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, and Prof. Nidhi Goel, Department of Electronics and Communication Engineering, Indira Gandhi Delhi Technical University for Women, Delhi, for their invaluable guidance, continuous encouragement, insightful suggestions, and unwavering support. Their expertise, patience, and constant motivation have played a pivotal role in shaping this research and in my development as a researcher.

I also express my sincere appreciation to the faculty members and staff of the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, for providing a supportive academic environment.

I am deeply grateful to my family for their unconditional love, encouragement, and unwavering support throughout my academic journey. Their faith in me has been a constant source of strength. I also sincerely thank everyone who supported and encouraged me throughout this journey, making this thesis possible.

Place: Delhi
Date:

Ms. Shilpa Jain
2K21/PHDEC/507

ABSTRACT

Autonomous Driver Assistance Systems (ADAS) have emerged as a significant technological advancement in modern vehicle intelligence, enabling vehicles to perceive and interpret their surroundings for safer, more reliable operation. In real-world driving, however, perception must address multiple complementary tasks, including traffic object detection, traffic sign and traffic light recognition, lane detection, drivable area segmentation, and road defect identification. Developing an independent deep learning model for each of these tasks can lead to substantial computational and memory overhead, increased inference latency, and inefficient utilization of onboard resources. These limitations are particularly significant for embedded automotive platforms, where computational resources are constrained, and multiple perception functions must operate concurrently under real-time conditions. Furthermore, real-world, unstructured road environments introduce additional challenges, including irregular road geometry, heterogeneous traffic, adverse weather, poor illumination, damaged road surfaces, occlusions, and incomplete traffic infrastructure, which can further degrade the reliability of perception. These challenges necessitate the development of unified, computationally efficient, and robust perception frameworks that can simultaneously understand multiple aspects of complex driving environments in real time.

This research presents a comprehensive study on unified multi-task perception for autonomous driver assistance in unstructured environments. A novel unified multi-task deep learning architecture, namely **SmartDriveNet**, is developed to jointly perform seven critical perception tasks, including traffic object detection, lane and drivable-area perception, road-surface anomaly identification, and traffic sign and traffic light recognition, within a unified end-to-end architecture. The framework incorporates advanced feature extraction using Generalized ELAN (GELAN) enhanced with Programmable Gradient Information (PGI), a Density Adaptive Attention Mechanism (DAAM), multi-scale feature fusion, and balanced multi-task optimization to enhance perception accuracy and robustness

under heterogeneous traffic, adverse weather, and damaged road conditions. To support the development and evaluation of the proposed framework, the Indian Traffic and Adverse Weather Dataset (ITAWD), comprising 24,236 annotated images across 47 classes, was developed to represent diverse traffic scenarios, adverse weather, and challenging road conditions.

Building upon the multi-task perception approach, the research is further extended to autonomous defence vehicles through the development of **DefenceVisionNet**, a real-time unified perception framework designed for unstructured off-road and battlefield environments. The proposed architecture jointly performs defence–civilian entity detection (defence vehicles and personnel, including soldiers), drivable terrain segmentation, and trail and boundary detection using an E-ELAN-based shared backbone with multi-scale feature fusion through FPN and SPPF modules and task-specific decoder heads. To support realistic training and evaluation, the Tactical Vehicle Dataset (TVD), comprising 20,758 annotated images across 31 classes, was developed to capture diverse terrain conditions, camouflage, occlusions, degraded visibility, and varying operational scenarios. The framework further incorporates cascaded attention-guided feature refinement and adaptive multi-task optimization to enhance perception accuracy while maintaining real-time performance for autonomous defence applications.

Recognizing the growing importance of secure deployment for deep learning-based perception models, particularly in safety-critical applications such as VIP transportation, military logistics, emergency medical services, and police vehicles, where perception errors may lead to unsafe decisions and compromise human safety and mission-critical operations. These concerns emphasize the need for reliable mechanisms to safeguard the authenticity and integrity of perception models prior to deployment. Accordingly, this research investigates the protection and verification of deep learning models using post-quantum ML-DSA-based mechanisms. The research also demonstrates practical deployment through a low-cost Raspberry Pi-based ADAS prototype. Overall, the research contributes to the development of robust, secure, and deployable multi-task perception systems for complex and unstructured autonomous environments.

RESEARCH PUBLICATIONS

SCIE Indexed Journal Publications

J1: Jain, S., Indu, S. and Goel, N., 2025. SmartDriveNet: An integrated approach for robust driving perception in unstructured environments and adverse weather conditions. *Sādhanā*, Vol. 50, 344. Available at: https://doi.org/10.1007/s12046-025-03011-6. **(Accepted and Published)**

J2: Jain, S., Indu, S. and Goel, N., 2026. A Unified Multitask Deep Learning Framework for Autonomous Defence Vehicle Perception: DefenceVisionNet. *Defence Science Journal*, Vol. 76, No. 5, pp. 681–692. Available at: https://publicationsdrdo.in/index.php/dsj/article/view/21743/8884. **(Accepted and Published)**

J3: Jain S., Indu S. and Goel N., 2026. Q-DriveShieldNet: Secure multi-task perception for intelligent transportation systems. *Journal of Universal Computer Science (J.UCS)*. **(Accepted)**.

Published/Accepted International Conference Publications

C1: Jain, S., Indu, S. and Goel, N., 2023. Comparative Analysis of YOLO Algorithms for Intelligent Traffic Monitoring. Proceedings of the International Conference on Data Analytics and Computing (ICDAC 2022), Lecture Notes on Data Engineering and Communications Technologies, Vol. 175, pp. 159–168. Singapore, Springer. Available at: https://doi.org/10.1007/978-981-99-3432-4_13 **(Book Chapter Published)**

C2: Jain, S., Indu, S. and Goel, N., 2025. Real-time Embedded ADAS for Elderly Drivers with Vision-Based Interactive Display. International Conference on Emerging Electronics and Automation (E2A 2025).**(Accepted and Book Chapter Publication in Progress)**

C3: Jain, S., Indu, S. and Goel, N., 2026. Helmet Violation Detection and Automatic Number Plate Recognition of Motorcyclists Using Deep Learning. 68th National Convention of the Indian Institution of Industrial Engineering (IIIE) and 6th International Conference on Industrial Engineering (ICIE-2026). **(Accepted and Presented)**

Released Datasets

D1: Jain, S., 2025. Indian Traffic and Adverse Weather Dataset (ITAWD), *Roboflow Universe*. [Online]. Available at: `https://universe.roboflow.com/shilpa-mocpz/itawd` [Accessed: 03 April 2026] **(Open-Source Dataset)**

D2: Jain, S., 2026. Tactical Vehicle Dataset (TVD), *Roboflow Universe*. [Online]. Available at: `https://universe.roboflow.com/shilpa-o3j8x/tactical-vehicle-dataset-tvd-94dmv` [Accessed: 03 April 2026] **(Open-Source Dataset)**

D3: Jain, S., 2026. SafeRide: Helmet & License Plate Dataset (SHLPD), *Roboflow Universe*. [Online]. Available at: `https://universe.roboflow.com/tlsor/saferide-helmet-license-plate-dataset-shlpd-d4gwy` [Accessed: 03 April 2026]. **(Open-Source Dataset)**

D4: Jain, S., 2026. DefenceDrive Dataset, *Roboflow Universe*. [Online]. Available at: `https://universe.roboflow.com/shilpa-jain-workspace/defencedrive-dataset` [Accessed: 03 April 2026]. **(Open-Source Dataset)**

Communicated

J4 Jain, S., Indu, S. and Goel, N., 2026. Battlefield situational intelligence with quantum-resilient multi-task perception: QR-DefenceDrive. Submitted to *The Journal of Supercomputing*. **(Manuscript Under Review)**

TABLE OF CONTENTS

Candidate's Declaration ii

Certificate iii

Acknowledgement iv

Abstract viii

Table of Content xi

List of Tables xii

List of Figures xiv

List of Symbols, Abbreviations xv

1 Introduction 1
- 1.1 Background and Research Motivation . . . 1
- 1.2 Problem Statement and Formulation . . . 4
- 1.3 Research Gaps . . . 4
- 1.4 Research Objectives . . . 6
- 1.5 Research Methodology and Contributions . . . 8
  - 1.5.1 Deep Learning-Based Traffic Object Detection and Monitoring . . . 8
  - 1.5.2 Unified Multi-Task Driving Perception . . . 9
  - 1.5.3 Unified Multi-Task Perception for Autonomous Defence Vehicles . . . 11
  - 1.5.4 Secure Deployment and Real-Time Embedded ADAS Implementation 11
- 1.6 Thesis Organization . . . 12

2 Literature Review 15
- 2.1 Deep Learning-Based Perception task . . . 15
  - 2.1.1 Traffic Object Recognition . . . 16
  - 2.1.2 Traffic Sign Recognition . . . 18
  - 2.1.3 Traffic Light Recognition . . . 20
  - 2.1.4 Road Defect Detection: Potholes and Cracks . . . 21
  - 2.1.5 Drivable Area Segmentation . . . 23
  - 2.1.6 Lane Detection . . . 24
- 2.2 Multi-Task Learning for Autonomous Driving . . . 26
- 2.3 Datasets for Autonomous Driving Perception . . . 29
  - 2.3.1 Traffic Object Detection Datasets . . . 30
  - 2.3.2 Traffic Sign and Traffic Light Datasets . . . 30
  - 2.3.3 Road Damage Datasets . . . 31

2.3.4 Lane and Drivable Area Datasets . . . 32
2.3.5 Multi-Task Autonomous Driving Datasets . . . 32
2.4 Datasets for Autonomous Defence Systems . . . 33

3 Development of Deep Learning-Based Traffic Perception Frameworks 36
3.1 Evaluation of YOLO-Based Object Detection Models for Traffic Perception 36
3.1.1 Evolution of YOLO Architectures . . . 37
3.1.2 Traffic Vehicle Detection and Detection Refinement . . . 38
3.1.3 Evaluation Metrics . . . 38
3.1.4 Experimental Dataset and Traffic Video Processing . . . 39
3.1.5 Comparative Analysis of Results . . . 39
3.1.6 Integration of YOLOv4 with Deep SORT . . . 41
3.1.7 Vehicle Tracking and Counting . . . 42
3.1.8 Summary of the Important Results . . . 43
3.2 Helmet Violation Detection and Automatic Number Plate Recognition . . 43
3.2.1 Overall Architecture of the Cascaded Framework . . . 44
3.2.2 Dataset Preparation and Preprocessing . . . 44
3.2.3 Motorcycle and Rider Detection . . . 45
3.2.4 Helmet Compliance Classification . . . 46
3.2.5 License-Plate Localization . . . 47
3.2.6 Optical Character Recognition . . . 47
3.2.7 Implementation and Experimental Setup . . . 47
3.2.8 Detection Results . . . 48
3.2.9 License Plate Recognition Results . . . 48
3.2.10 Performance Analysis . . . 49
3.2.11 Comparitive Analysis . . . 49
3.2.12 Limitations . . . 50
3.2.13 Summary of the Important Results . . . 51

4 Development of a Unified Multi-Task Deep Learning Framework for Driving Perception 52
4.1 SmartDriveNet-Based Multi-Task Panoptic Driving Perception . . . 52
4.2 Methodology and System Development . . . 54
4.3 Development of the Indian Traffic and Adverse Weather Dataset . . . 55
4.3.1 Real-World Data Collection . . . 55
4.3.2 Selection and Integration of Existing Datasets . . . 56
4.3.3 Dataset Re-Annotation, Label Harmonization, and Class Organization 57
4.3.4 Data Preprocessing, Augmentation, and Dataset Partitioning . . . 58
4.4 Proposed SmartDriveNet Architecture . . . 59
4.4.1 Shared Encoder and Feature Extraction . . . 60
4.4.2 Task-Specific Decoder Heads . . . 60
4.4.3 Density Adaptive Attention Mechanism . . . 61
4.4.4 Architectural Rationale . . . 62
4.4.5 Multi-Task Loss Function . . . 62
4.5 Experimental Setup and Performance Evaluation . . . 63
4.5.1 Performance Assessment of Traffic Object Detection . . . 65
4.5.2 Performance Evaluation for Drivable-Area and Lane-Line Segmentation . . . 65

4.5.3 Performance Assessment of Traffic Sign Recognition 66
4.5.4 Performance Assessment of Pothole and Crack Detection 67
4.5.5 Performance Assessment of Traffic-Light Recognition 68
4.5.6 Ablation Analysis of Shared and Separate Decoders 69
4.5.7 Limitations and Future Scope 69
4.5.8 Summary of the Important Results 69

5 Development of a Unified Multi-Task Deep Learning Framework for Autonomous Defence Vehicle Perception 72
5.1 DefenceVisionNet-Based Multi-Task Perception for Autonomous Defence Vehicles 72
5.2 Methodology 74
5.2.1 Tactical Vehicle Dataset 74
5.2.2 Selection and Integration of Existing Datasets 75
5.2.3 Network Architecture 77
5.2.4 Multi-Task Loss Function with Cascade-Based Refinements . . . 82
5.3 Experimentation and Evaluation 85
5.3.1 Experimental Setup 85
5.3.2 Results and Evaluation 86
5.3.3 Summary of the Important Results 97

6 Secure Model Deployment and Embedded Autonomous Perception Systems 100
6.1 Secure Mullti-task Deep learning Autonomous Perception Systems 100
6.2 Embedded Realization of the Autonomous Driver Assistance System . . . 102
6.2.1 Hardware Architecture 103
6.2.2 Vision-Based Perception Pipeline 104
6.2.3 Lane Detection and Lane-Center Estimation 104
6.2.4 Traffic Sign and Traffic Light Recognition 105
6.2.5 LiDAR-Based Obstacle Detection and Collision Avoidance 106
6.2.6 Embedded Model Optimization and Deployment 107
6.2.7 Integrated Operational Workflow 107
6.3 Experimental Evaluation of the Embedded ADAS 108
6.3.1 Lane Tracking Performance 108
6.3.2 Traffic Sign and Traffic Light Detection Performance 108
6.3.3 Training Convergence and Generalization 109
6.3.4 Obstacle Detection Performance 109
6.4 Discussion 110
6.4.1 Summary of the Important Results 111

7 Conclusion, Future Directions, and Social Impact 113
7.1 Summary of the Work Done in the Thesis 113
7.2 Limitations of the Study 115
7.3 Future Scope and Social Impact 116
7.3.1 Future Scope/Directions 116
7.3.2 Social Impact 117

A Density Adaptive Attention Mechanism (DAAM) 119

B Multi-Task Gradient Balancing 120

List of Tables

1.1 Mapping of research objectives to the corresponding research gaps, outcomes, and publications. 7

2.1 Comparison of Multi-Task Learning Models for Autonomous Driving Perception 26
2.2 Comparison of datasets and repositories relevant to autonomous defence perception 34

3.1 Comparative characteristics of different YOLO versions. 37
3.2 Comparative Performance of YOLO Versions. 40
3.3 Performance of the Individual Stages of the Proposed Cascaded Framework. 49
3.4 Performance Comparison of Representative Object-Detection Architectures. 50

4.1 Datasets selected for the development of ITAWD 56
4.2 Comparative performance of the evaluated multi-task perception models on the ITAWD dataset [37] 65
4.3 Model Comparison for Traffic Sign Recognition across Various Classes. . 66
4.4 Comparative Performance Analysis for Crack and Pothole Detection. . . . 67

5.1 Comparison of datasets and repositories relevant to autonomous defence perception 76
5.2 Comparison of Model Parameters and Inference Speed. 87
5.3 Comparison of Defence Entity Detection Performance on the Tactical Vehicle Dataset (TVD). 90
5.4 Comparison of Drivable Terrain Segmentation Performance on the Tactical Vehicle Dataset (TVD). 91
5.5 Comparison of Trail and Boundary Detection Performance on the Tactical Vehicle Dataset (TVD). 93
5.6 Statistical Validation Across All Perception Tasks on the Tactical Vehicle Dataset (TVD). 94
5.7 Ablation Study of DefenceVisionNet on the Tactical Vehicle Dataset (TVD). 95

6.1 Performance comparison of the customized YOLOv8-nano with lightweight object-detection models. 109

List of Figures

1.1 Conceptual representation of the proposed unified panoptic driving perception framework integrating seven complementary perception tasks for comprehensive driving-scene understanding. 3
1.2 Overview of SmartDriveNet Panoptic Perception System. 9
1.3 Overview of DefenceVisionNet Panoptic Perception System. 10

2.1 Challenges in Traffic Object Recognition in Unstructured Roads and Adverse Environmental Conditions 16
2.2 Challenges Associated with Traffic Sign Recognition in Unstructured Roads and Adverse Environmental Conditions 19
2.3 Challenges in Traffic Light (TL) in Detection Unstructured Road and Challenging Environmental Conditions 20
2.4 Challenges in Potholes and Cracks Detection under Unstructured Environments and Adverse Weather Conditions. 22
2.5 Challenges in Lane and Drivable Area Segmentation under Unstructured Environments and Adverse Weather Conditions. 24
2.6 YOLOPv2 architecture 27
2.7 HybridNets architecture 28

3.1 Visual comparison of vehicle detection results obtained using different YOLO versions. 40
3.2 YOLOv4-based vehicle detection and Deep SORT multi-object tracking framework. 41
3.3 Proposed cascaded YOLO framework integrating helmet violation detection with ANPR. 45
3.4 License-plate image after brightness enhancement and preprocessing . . . 46
3.5 Representative detection results of the proposed framework under different traffic conditions. 48

4.1 Block Diagram of SmartDriveNet Panoptic Perception System 54
4.2 Comparison of image annotations: (a) Original image, (b) Annotation from the Indian Traffic Sign (ITS) dataset, and (c) Re-annotated image in the Indian Traffic and Adverse Weather Dataset (ITAWD). 57
4.3 Class imbalance in the ITAWD dataset. 58
4.4 SmartDriveNet: A Comprehensive Framework for Unified Visual Perception 59
4.5 Qualitative comparison of traffic perception performance for (a) input road scenes, (b) YOLOPv2, (c) fine-tuned HybridNets, and (d) SmartDriveNet. 64
4.6 Qualitative evaluation of traffic perception under adverse weather conditions, including snow, fog, and rain: (a) input images; (b) fine-tuned YOLOPv2; (c) fine-tuned HybridNets; and (d) SmartDriveNet. 64

4.7 Qualitative comparison of detection and segmentation results: (a) input road scenes containing potholes, cracks, traffic objects, and traffic lights; (b) predictions from the fine-tuned YOLOPv2; (c) predictions from the fine-tuned HybridNets; and (d) predictions obtained using the proposed SmartDriveNet. 64
4.8 Performance Comparison of Traffic Sign Models Based on performance Metrics 67
4.9 Performance-Based Comparison of Traffic Light Models. 68

5.1 Block diagram for the generation of TVD Dataset 75
5.2 DefenceVisionNet: A unified multi-task perception network 78
5.3 Visual evaluation of defence-entity detection for land-based defence vehicles in the TVD dataset: (a) input frame; (b) detection results from the fine-tuned YOLOv11 model; (c) detection results from the fine-tuned YOLOPv2 model; and (d) output generated by the proposed DefenceVisionNet. 88
5.4 Visual assessment of defence-entity detection for air-defence vehicles in the TVD dataset: (a) input frame; (b) results obtained using the fine-tuned YOLOv11 detector; (c) outputs generated by the fine-tuned YOLOPv2 model; and (d) predictions produced by the proposed DefenceVisionNet. . 89
5.5 (a) Per-class detection performance on TVD; (b) Performance under varying environmental conditions. 89

6.1 Overall architecture of the embedded autonomous driver assistance system integrating vision, LiDAR, embedded processing, vehicle control, and interactive feedback. 103
6.2 Processing flow of the vision-based perception module for lane detection, traffic light recognition, and traffic sign recognition. 104
6.3 Embedded lane-detection pipeline showing the input road image, perspective transformation, lane-center estimation, and corresponding control output. 105
6.4 Real-time visualization of traffic sign and traffic light detections on the embedded touchscreen interface. 106
6.5 Training and validation performance curves of the customized YOLOv8-nano traffic perception model. 110

List of Symbols

$\mathcal{L}_{\text{det}}$	Detection Loss
$\mathcal{L}_{\text{da}}$	Drivable-area Segmentation Loss
$\mathcal{L}_{\text{ll}}$	lane-line segmentation loss
$\mathcal{L}_{\text{Dice}}$	Dice loss
$\mathcal{L}_{\text{Focal}}$	Focal loss
$\mathcal{L}_{\text{Tversky}}$	Tversky loss
$\mathcal{L}_{\text{total}}$	Total loss
α, β, λ	Control parameters that balance the contributions of the loss components
B_p	Predicted bounding box
B_g	Ground-truth bounding box

CHAPTER 1
Introduction

1.1 Background and Research Motivation

Advanced Driver Assistance Systems (ADAS) [1] encompass a wide range of in-vehicle active safety technologies designed to improve road safety and assist drivers during vehicle operation. By enhancing the vehicle's awareness of its surroundings and supporting critical driving tasks, ADAS can reduce the likelihood of accidents and help mitigate the risk of fatalities and injuries. Achieving these capabilities requires an intelligent vehicle to continuously monitor and interpret its surrounding environment using cameras, sensors, and artificial intelligence (AI). The information obtained through these technologies enables the vehicle to understand the road scene, identify relevant objects and potential hazards, and support timely driving decisions. Reliable environmental perception is therefore a fundamental component of ADAS, as a vehicle must develop a comprehensive understanding of the driving environment rather than depend on a single source of information or an isolated perception task.

The evolution of deep learning, with convolutional neural networks (CNNs) at its core, has transformed vision-based perception by enabling models to learn rich, discriminative representations directly from visual data. This progress has substantially enhanced intelligent and autonomous vehicles' ability to analyze complex road scenes and derive meaningful environmental information [2]. Compared with relatively costly sensing technologies such as LiDAR and RADAR, camera-based perception offers a practical and economical alternative for obtaining rich semantic information from RGB imagery. A single RGB camera can capture rich visual and semantic information about the driving environment, enabling multiple perception tasks while maintaining effective scene understanding. Consequently, vision-based perception has become an important component of modern ADAS and autonomous driving systems, where accurate environmental understanding is fundamental to safe and reliable vehicle operation.

Over the past decade, deep learning has enabled significant advances in vision-based perception for intelligent vehicles, with researchers developing specialized models to address different aspects of the driving environment. Considerable progress has been reported across these individual perception tasks, including traffic sign recognition [3, 4], traffic light recognition [5, 6], pothole and pavement crack detection [7, 8, 9], lane detection [10, 11], drivable area segmentation [12, 13], and traffic-object detection. For object detection, a range of one-stage and two-stage deep learning approaches have been explored, with representative methods including Faster R-CNN [14], SSD [15], and successive

generations of YOLO models [16, 17, 18, 19]. These approaches are generally optimized for specific objectives, allowing them to achieve high performance on their respective tasks. However, using an independent deep neural network for every perception function can substantially increase computational demand, memory usage, and inference time. This becomes particularly challenging for embedded automotive platforms, where hardware resources are constrained and multiple perception tasks must be performed simultaneously under real-time requirements.

Multi-task learning (MTL) provides a suitable framework for achieving this integration by jointly optimizing multiple related perception tasks within a single network. In such architectures, a shared feature-extraction backbone learns common visual representations, while task-specific prediction heads generate outputs for individual perception objectives. This shared representation enables information learned from one task to benefit related tasks, making MTL particularly suitable for autonomous driving systems that require simultaneous interpretation of diverse elements within a road scene.

Several MTL architectures have been proposed to integrate different driving perception tasks within a single framework. MultiNet [20] was among the earliest unified architectures to jointly perform object detection and road segmentation using a shared encoder, demonstrating the feasibility of multi-task perception while exhibiting limited robustness in complex and unstructured scenes. DLT-Net [21] further improved multi-task feature learning by introducing context-aware information exchange among decoder branches, thereby enhancing detection and segmentation performance. Subsequently, YOLOP [2] proposed a unified end-to-end architecture that simultaneously identifies traffic objects, delineates lane boundaries, and segments the roadway into drivable and non-drivable regions, thereby striking a practical balance between predictive accuracy and real-time processing speed. HybridNets [22] enhanced this framework by incorporating optimized anchor design to improve feature fusion across multiple scales. More recently, YOLOPv2 [23] introduced architectural refinements that further improved computational efficiency and perception accuracy while maintaining real-time performance.

Despite these developments, current multi-task perception systems generally address only a limited subset of driving-related perception requirements, with most approaches concentrating on identifying road users, estimating lane boundaries, and determining the traversable road region. Other safety-critical capabilities, including the interpretation of traffic signs and signals, identification of potholes, and detection of pavement cracks, are often developed as independent perception modules rather than being integrated into a single comprehensive framework. Moreover, the majority of these models are developed and evaluated using structured urban datasets [24] [25], limiting their ability to generalize to complex real-world environments characterized by irregular road geometry, degraded lane markings, damaged road surfaces, heterogeneous traffic participants, severe occlusions, and adverse weather conditions. These limitations highlight the need for more comprehensive and robust multi-task perception frameworks capable of integrating a broader range of driving perception tasks within a common architecture, thereby providing a more complete representation of the driving environment and supporting coordinated perception across different aspects of the road scene in challenging operating conditions.

The challenges become even more pronounced in developing countries such as India, where autonomous vehicles encounter highly heterogeneous traffic comprising pedestrians, cyclists, motorcycles, auto-rickshaws, heavy vehicles, stray animals, poorly maintained roads, potholes, waterlogged surfaces, inconsistent traffic infrastructure, especially in adverse weather conditions like rain, fog, sandstorm, and snow. In addition,

varying illumination conditions, rain, fog, haze, shadows, and nighttime driving introduce significant domain shifts that adversely affect perception accuracy. Existing benchmark datasets such as BDD100K [24], IDD [26], and RDD2022 [27] have significantly advanced autonomous driving research; however, each primarily focuses on specific perception tasks or operating environments and therefore does not comprehensively represent the diversity and complexity encountered in real-world developing regions. Consequently, there remains a need for benchmark datasets and unified perception frameworks, that can address multiple perception tasks across diverse environmental and road conditions.

Reliable environmental perception is equally critical for autonomous defence systems [28, 29]. Unlike civilian transportation, defence platforms frequently operate in highly unstructured environments consisting of off-road terrain, vegetation, camouflage, degraded infrastructure, severe occlusions, and rapidly changing operational scenarios. Models trained primarily on civilian driving datasets often exhibit limited generalization under these conditions because of substantial domain differences between urban traffic scenes and battlefield environments. Furthermore, publicly available defence-oriented benchmark datasets containing comprehensive annotations for multiple perception tasks remain scarce, thereby restricting the development of robust and generalized autonomous defence perception systems [28, 29, 30]

Beyond achieving high perception accuracy, ensuring the security and trustworthiness of deployed deep learning models has become equally important for autonomous systems. Modern perception models are increasingly deployed through cloud services, edge computing platforms, and over-the-air software updates, exposing them to adversarial attacks, model tampering, and supply-chain vulnerabilities that can compromise system reliability and safety [31]. Existing protection mechanisms, such as deep neural network watermarking and fingerprinting, primarily support ownership verification and intellectual property protection but provide limited capability for real-time model integrity validation [32, 33]. More recently, post-quantum digital signature schemes have emerged as a promising solution for ensuring model authenticity and integrity while providing resilience against future quantum attacks [34]. However, their integration into lightweight, real-time autonomous perception frameworks remains largely unexplored.

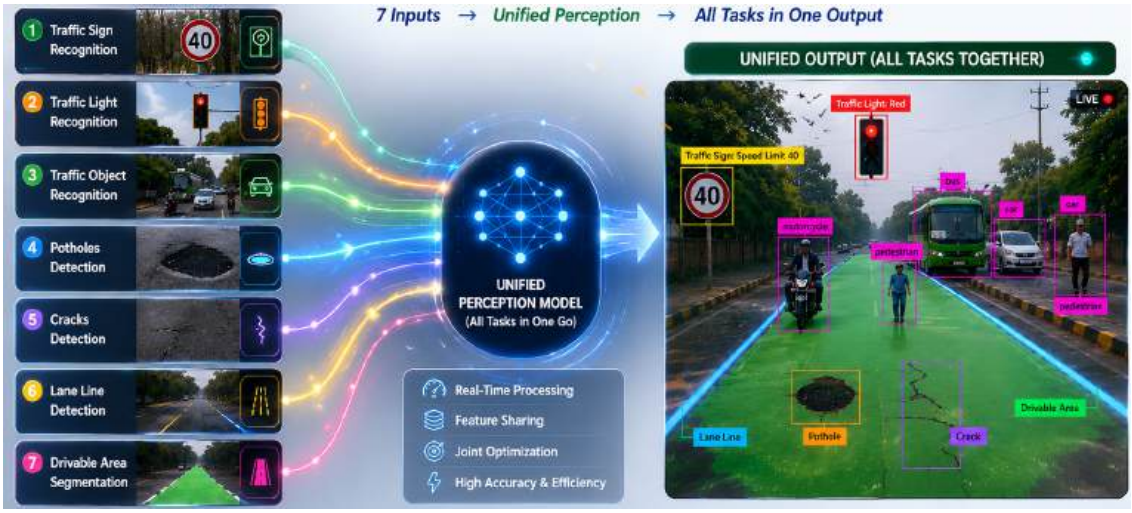

Figure 1.1: Conceptual representation of the proposed unified panoptic driving perception framework integrating seven complementary perception tasks for comprehensive driving-scene understanding.

1.2 Problem Statement and Formulation

The development of reliable environmental perception is a fundamental requirement for intelligent transportation and autonomous defence systems. However, existing perception approaches are often task-specific, computationally demanding, and limited in their ability to operate reliably in unstructured environments, adverse weather conditions, and defence scenarios. The availability of comprehensive datasets with unified annotations for multiple perception tasks is also limited, which restricts the development and evaluation of generalized multi-task perception models. In addition, the security and authenticity of deployed deep learning models remain important concerns, particularly for autonomous systems operating in connected and safety-critical environments.

Motivated by these challenges, this thesis formulates the problem as the development of a robust, computationally efficient, secure, and deployable unified perception system, as illustrated in Fig. 1.1. The proposed approach integrates multiple complementary perception capabilities within a shared multi-task architecture, enabling the joint interpretation of road users, traffic signs, traffic lights, lane boundaries, drivable regions, and road conditions. Such integration is particularly relevant to unstructured Indian road environments, where heterogeneous traffic, irregular road geometry, degraded infrastructure, road defects, and adverse weather conditions may occur simultaneously. Accordingly, the central research problem is to develop a unified deep learning framework that can provide comprehensive and efficient scene understanding while addressing the dataset, computational, environmental, and security constraints of practical autonomous systems.

The research addresses the identified challenges through the development of comprehensive benchmark datasets representing diverse and challenging real-world environments, followed by the design of multi-task deep learning frameworks capable of jointly performing complementary perception tasks within a unified architecture. For civilian driving environments, the research encompasses traffic-object, traffic-sign, traffic-light, lane, drivable-area, and road-defect perception under unstructured Indian road and adverse weather conditions. The research is further extended to autonomous defence environments through defence-entity detection, including defence vehicles and personnel, drivable-terrain segmentation, and trail or boundary detection under complex off-road conditions.s.

The formulated problem further incorporates the requirement for trustworthy deployment of deep learning models. Therefore, post-quantum digital signature-based model authentication and integrity verification are investigated to ensure that only authentic and untampered perception models are deployed for inference. Finally, the developed perception algorithms are implemented on resource-constrained embedded hardware to evaluate their practical real-time applicability. Thus, the overall problem formulation encompasses dataset development, multi-task perception model design, robustness to challenging environments, secure model verification, and embedded deployment, with the objective of developing accurate, efficient, secure, and practically deployable autonomous perception systems.

1.3 Research Gaps

Based on a comprehensive review of the existing literature, several unresolved challenges in intelligent autonomous navigation, particularly in unstructured and complex road scenarios, were identified. Recent deep learning studies have demonstrated considerable progress in

recognizing traffic entities, identifying road-surface defects, estimating lane boundaries, determining traversable regions, and supporting autonomous navigation. However, the existing solutions still exhibit shortcomings in achieving reliable performance across varying environments, maintaining computational efficiency for real-time deployment, and generalizing effectively beyond the conditions represented in their training data. Furthermore, most existing studies address individual perception functions or a limited combination of tasks rather than providing a comprehensive and unified scene-understanding framework. Based on the identified limitations and challenges in the existing literature, the following major research gaps are addressed in this research:

RG1: Challenges in Traffic Object Detection:

Existing traffic object detection approaches continue to face challenges in handling heterogeneous and densely populated traffic scenes, where multiple vehicle types, pedestrians, motorcycles, bicycles, buses, trucks, and other non-conventional road users may coexist. Variations in object scale, frequent occlusion, illumination changes, adverse weather, and complex backgrounds can significantly affect detection accuracy. Traffic sign recognition remains challenging when signs are small, distant, partially occluded, damaged, faded, or poorly illuminated, and can be further affected by viewpoint and weather variations and visual similarity between sign categories. Similarly, traffic light detection is affected by their small and distant appearance, partial occlusion, illumination variation, adverse weather, and complex surroundings. Furthermore, the limited representation of Indian traffic conditions in commonly used datasets restricts the generalizability of existing approaches to heterogeneous and unstructured Indian road environments. Therefore, a robust and efficient traffic perception framework is required to accurately detect and classify diverse traffic objects, signs, and signals under challenging real-world conditions while maintaining reliable real-time performance.

RG2: Challenges in Road Defect Detection:

Road defect detection remains challenging due to inter-class visual similarity, irregular defect geometries, and environmental interference. Existing models often struggle to distinguish between visually similar defects and surface conditions, such as waterlogging and wet roads, while accurately detecting irregularly shaped potholes and different types of cracks, including longitudinal, transverse, and alligator cracks, remains difficult. In addition, shadows, manholes, stains, and other surface irregularities can result in false detections and inaccurate segmentation. Therefore, a robust road defect detection framework is required to accurately identify and classify diverse road defects under varying environmental conditions while minimizing false positives and supporting real-time performance.

RG3: Challenges in Lane and Drivable Area Detection:

Existing lane and drivable area detection methods can become unreliable in unstructured and degraded road environments, particularly when lane boundaries are faded, partially visible, or obscured by mud, shadows, obstacles, and adverse environmental conditions. Moreover, existing approaches often emphasize accuracy, contextual modelling, or computational efficiency individually rather than addressing these requirements jointly. Their performance may degrade under variations in illumination, weather, road appearance, occlusion, and scene structure, while computationally intensive models can hinder real-time deployment on resource-constrained

embedded platforms. Lightweight models, on the other hand, may sacrifice fine boundary details and segmentation quality. These limitations highlight the need for a unified and computationally efficient framework capable of accurately and robustly identifying lane boundaries and drivable regions under diverse and challenging driving conditions.

RG2: Challenges in Safe Navigation:

Existing datasets and deep learning models predominantly focus on individual or limited perception tasks, resulting in fragmented environmental understanding. There is a lack of comprehensive datasets and unified multi-task frameworks that simultaneously integrate traffic object detection, road defect detection, lane boundary detection, and drivable area segmentation. Furthermore, existing approaches provide limited consideration of trajectory planning in the presence of uncertain or obstructed road segments. Thus, an integrated multi-task perception and navigation framework is required to achieve holistic scene understanding and support safe and reliable autonomous navigation in unstructured environments.

1.4 Research Objectives

The primary objective of this research is to develop robust, efficient, and unified deep learning frameworks for real-time environmental perception in autonomous civilian and defence systems. To accomplish this goal, the research has been systematically structured around four major research objectives. These objectives address key challenges associated with reliable perception in unstructured and dynamic environments, encompassing traffic object detection, road-defect detection, lane and drivable-area understanding, and integrated panoptic driving perception. The methodology adopted for each objective, along with its implementation, key findings, and resulting research contributions, is presented in the following subsections. The mapping of the research objectives to the corresponding research gaps, along with the resulting research outcomes and publications, is presented in Table 1.1.

Accordingly, the specific research objectives formulated for this research are as follows:

RO1: To design an efficient and accurate deep-learning model for traffic object detection, including traffic signs and signals, under diverse conditions of occlusion or illumination that can identify and track traffic objects with high speed and real-time precision.

RO2: To develop a deep neural network-based model that can be trained to do detailed semantic segmentation of road defects like muddy areas, cracks, potholes, etc., with proper labels, under varying illuminations and different environments, including shadowing, oil stains, or weather conditions.

RO3: To build a model with improved accuracy that can automatically detect, classify, and perform trajectory planning for drivable areas and lane boundaries even if the vision is obstructed by other obstacles.

RO4: To design an interactive dashboard that produces real-time semantic segmentation maps of traffic and irregularities on unstructured roadways, allowing for the building of an effective autonomous panoptic driving system capable of handling multiple tasks.

Table 1.1: Mapping of research objectives to the corresponding research gaps, outcomes, and publications.

Research Objective	Research Gap	Research Outcomes / Publications
RO1: To design an efficient and accurate deep-learning model for traffic object detection, including traffic signs and signals, under diverse conditions of occlusion or illumination that can identify and track traffic objects with high speed and real-time precision.	**RG1:** Traffic Object Detection	**J1:** SmartDriveNet **J2:** DefenceVisionNet **J3:** Q-DriveShieldNet **C1:** Comparative Analysis of YOLO Algorithms for Intelligent Traffic Monitoring **C3:** Helmet Violation Detection and Automatic Number Plate Recognition of Motorcyclists Using Deep Learning **D1:** ITAWD **D3:** SafeRide
RO2: To develop a deep neural network-based model that can be trained to do detailed semantic segmentation of road defects like muddy areas, cracks, potholes, etc., with proper labels, under varying illuminations and different environments, including shadowing, oil stains, or weather conditions.	**RG2:** Road Defect Detection	**J1:** SmartDriveNet **J2:** DefenceVisionNet **J3:** Q-DriveShieldNet **D1:** ITAWD
RO3: To build a model with improved accuracy that can automatically detect, classify, and perform trajectory planning for drivable areas and lane boundaries even if the vision is obstructed by other obstacles.	**RG3:** Lane and Drivable Area Detection **RG4:** Safe Navigation	**J1:** SmartDriveNet **J2:** DefenceVisionNet **J3:** Q-DriveShieldNet **D1:** ITAWD **D2:** TVD
RO4: To design an interactive dashboard that produces real-time semantic segmentation maps of traffic and irregularities on unstructured roadways, allowing for the building of an effective autonomous panoptic driving system capable of handling multiple tasks.	**RG1–RG4:** Integrated Panoptic Driving Perception	**J1:** SmartDriveNet **J2:** DefenceVisionNet **J3:** Q-DriveShieldNet **C2:** Real-time Embedded ADAS for elderly drivers with vision based Interactive **D1:** ITAWD **D2:** TVD

Note: The complete bibliographic details of the publications and datasets identified by the above codes are provided in the Research Publications.

1.5 Research Methodology and Contributions

The research contributions presented in this thesis are directed towards achieving the formulated research objectives for the development of an autonomous driver assistant system for unstructured environments. The proposed research systematically addresses the identified challenges in autonomous driving perception through a series of methodological and technological contributions. The work progresses from the investigation of a deep learning-based traffic object detection and specialized traffic monitoring to unified multi-task driving perception, extension of perception capabilities to challenging and defence-oriented environments, secure deep learning model deployment, and real-time embedded ADAS implementation. Each contribution is aligned with the corresponding research objective and builds upon the knowledge and outcomes of the preceding stage, resulting in a progressive development from fundamental perception algorithms to an integrated, robust, secure, and deployable autonomous driving perception system.

1.5.1 Deep Learning-Based Traffic Object Detection and Monitoring

The initial phase of this research focused on evaluating YOLO-based object detectors for their suitability in intelligent traffic surveillance and real-time transportation applications [35]. As object detection forms a core component of the perception pipeline in autonomous driving, a comparative investigation was conducted to examine the detection performance and computational requirements of different YOLO architectures. The analysis helped identify the balance between recognition accuracy and inference efficiency required for practical traffic-monitoring systems. Building on these findings, subsequent investigations addressed vehicle detection, tracking, and counting to support real-time traffic analysis. The scope of the research was then extended towards road-user safety by developing a deep learning framework for detecting helmet violations and performing automatic number plate recognition (ANPR) for motorcyclists [36]. These investigations established the foundation for the subsequent development of unified multi-task perception frameworks for autonomous driving in unstructured environments.

The methodology employed to accomplish this is outlined below:

- Review and analysis of existing deep learning-based object detection algorithms, with emphasis on YOLO-based architectures for real-time traffic perception and intelligent transportation applications.
- Experimental implementation and comparative evaluation of different YOLO versions using a standard object detection dataset containing diverse traffic objects.
- Analysis of the trade-off between detection accuracy and computational complexity to identify suitable YOLO architectures for real-time traffic monitoring.
- Integration of the selected object detection model with a multi-object tracking algorithm for vehicle-level detection, tracking, and counting.
- Extension of the traffic perception framework towards automated helmet violation detection for motorcyclists by identifying motorcycles, riders, and helmet compliance status.

- Development of a cascaded deep learning-based framework for automatic number plate recognition (ANPR), comprising motorcycle and rider detection, helmet classification, license plate localization, and Optical Character Recognition (OCR).
- Performance evaluation of the investigated models using parameters such as mean Average Precision (mAP), precision, recall, inference speed, and computational efficiency.

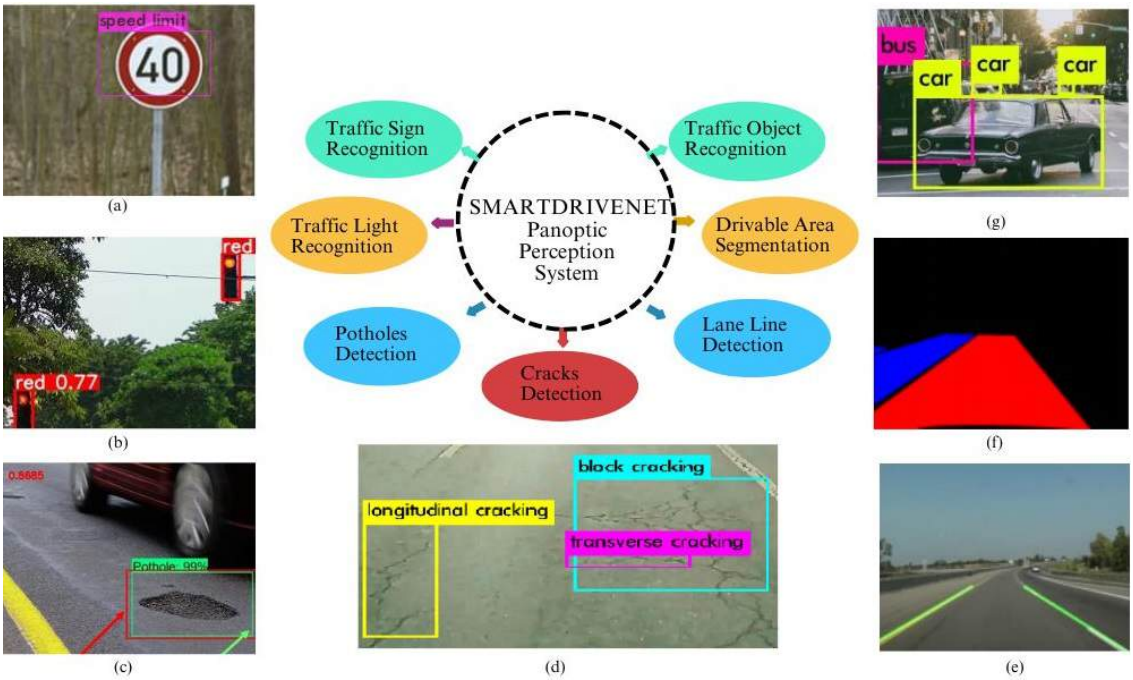

Figure 1.2: Overview of SmartDriveNet Panoptic Perception System.

1.5.2 Unified Multi-Task Driving Perception

The findings from the preliminary studies provided the basis for advancing the research from individual traffic-perception tasks towards an integrated multi-task driving perception system. Accordingly, **SmartDriveNet** [37] was developed as a unified deep learning framework for comprehensive environmental perception in unstructured road environments. The proposed camera-based system integrates seven key road-scene perception tasks, including traffic-object detection, traffic-sign and traffic-light recognition, pothole and crack detection, lane-line detection, and drivable-area segmentation. The objective was to develop a cost-effective and computationally efficient perception system capable of providing comprehensive road-scene understanding from a single visual input.

The methodological approach adopted for this contribution is briefly outlined below:

- A review of existing multi-task and driving perception frameworks was conducted to identify their capabilities and limitations in handling multiple traffic and road-scene perception tasks under unstructured environments, occlusions, illumination variations, and adverse weather conditions.
- A comprehensive Indian Traffic and Adverse Weather Dataset (ITAWD) [38], was developed by collecting images from real-world driving environments and integrating

relevant existing datasets. The resulting dataset comprises 24,236 images covering 47 classes and represents diverse traffic scenarios, traffic signs and signals, road conditions, and challenging environmental conditions.

- A unified multi-task deep learning architecture, **SmartDriveNet**, was developed to jointly address seven essential driving-perception tasks within a single framework. The architecture employs the Generalized Efficient Layer Aggregation Network (GELAN) to strengthen multi-level feature representation and information exchange, while Programmable Gradient Information (PGI) is incorporated to facilitate effective gradient flow and improve the optimization of the multi-task learning process.
- A Density Adaptive Attention Mechanism (DAAM) was integrated to enhance feature representation by adaptively emphasizing informative regions according to the characteristics and density of scene content, thereby improving perception under diverse and complex driving conditions.
- A task-aware training loss formulation and optimization strategy were developed to jointly train the multiple perception branches while maintaining an appropriate balance among the individual task objectives.
- The proposed framework was trained and evaluated on the **ITAWD** dataset using task-specific performance measures, including precision, recall, mean Average Precision (mAP), mean Intersection over Union (mIoU), lane IoU, and inference speed. Comprehensive comparative experiments were conducted to examine the effectiveness of **SmartDriveNet** in terms of detection and segmentation accuracy, computational efficiency, and robustness across diverse and challenging road-scene conditions.



Figure 1.3: Overview of DefenceVisionNet Panoptic Perception System.

1.5.3 Unified Multi-Task Perception for Autonomous Defence Vehicles

Building upon the unified driving perception framework developed in the preceding phase, the next major phase of the research extended multi-task perception to autonomous defence vehicles operating in highly unstructured and challenging environments. Defence-oriented scenarios introduce additional perception challenges, including severe occlusions, camouflage, complex backgrounds, heterogeneous defence and civilian entities, and variations in terrain and environmental conditions. To address these challenges, a unified multi-task deep learning framework, termed **DefenceVisionNet** [39], was developed for simultaneous perception of defence and civilian entities, drivable terrain, and trails or boundary regions. A dedicated Tactical Vehicle Dataset (TVD) [40], was also developed to support training and evaluation of the proposed framework. The methodology followed for this contribution is summarized below:

- Analysis of existing multi-task and autonomous driving perception approaches to identify their limitations in handling unstructured defence environments, particularly in the presence of camouflage, occlusions, complex backgrounds, and diverse operational scenarios.
- Development of the Tactical Vehicle Dataset (TVD) through the collection, integration, annotation, and preprocessing of defence and civilian vehicle imagery. The resulting dataset comprises 20,758 images representing 31 defence and civilian classes and provides a dedicated benchmark for evaluating perception models in defence-oriented environments.
- Development of a unified multi-task perception architecture, *DefenceVisionNet*, based on a shared E-ELAN encoder and three task-specific decoder heads. The decoder design incorporates C3k2, SPPF, and C2PSA modules to improve feature representation and robustness against scale variations, occlusions, camouflage, and complex backgrounds.
- A task-balanced multi-task training strategy was developed to jointly optimize the detection and segmentation objectives using the **TVD** dataset. The customized loss formulation was designed to regulate the contribution of individual tasks during training while preserving the computational efficiency required for real-time deployment. Joint learning enabled the network to exploit complementary feature representations across the perception tasks within a unified architecture.
- The proposed framework was comprehensively validated using Precision, Recall, mean Average Precision (mAP), Intersection over Union (IoU), Mean IoU, Pixel Accuracy, Boundary IoU, and inference speed (FPS). Comparative and ablation studies were conducted to quantify the contribution of the proposed architectural components and multi-task learning strategy, while also evaluating the robustness, computational efficiency, and real-time applicability of the framework for autonomous defence vehicle perception.

1.5.4 Secure Deployment and Real-Time Embedded ADAS Implementation

Building upon the deep learning-based perception frameworks developed in the preceding phases, the research was further extended towards the practical deployment of autonomous

perception systems, with emphasis on model security and real-time embedded implementation. The security aspect was investigated from the perspective of protecting trained deep learning models against unauthorized modification, replacement, and tampering, particularly in connected and distributed autonomous systems. In this context, a post-quantum secure model-deployment approach was explored using the Module-Lattice Digital Signature Algorithm (ML-DSA) for signing and verification of trained deep learning model artifacts. The approach is intended to provide model authenticity and integrity verification before deployment, thereby establishing a foundation for trustworthy deployment of deep learning-based perception systems. Since this work represents an ongoing research direction, detailed experimental results and performance analysis are not included in this chapter.

The research was subsequently extended towards the practical realization of autonomous driver assistance on resource-constrained embedded hardware. A low-cost real-time ADAS prototype was developed by integrating vision-based perception, LiDAR-based obstacle sensing, vehicle-control functions, and an interactive Human–Machine Interface (HMI). The system was designed to support essential driver-assistance functions, including lane following, traffic-sign and traffic-light recognition, obstacle detection, and real-time visualization and vehicle control. This implementation provided a practical platform for examining the feasibility of deploying lightweight perception and sensing functions in a cost-effective embedded ADAS system.

The methodology adopted for this research contribution is briefly summarized below:

- Investigation of security requirements for the deployment, distribution, and updating of trained deep learning models, with emphasis on model authenticity and integrity.
- Exploration of ML-DSA-based digital signatures for securing trained deep learning model artifacts and enabling integrity verification prior to deployment.
- Analysis of the computational and hardware requirements for implementing real-time driver-assistance functions on resource-constrained embedded platforms.
- Development and integration of lightweight vision-based perception, LiDAR-based obstacle sensing, vehicle-control, and Human–Machine Interface (HMI) modules for real-time environmental awareness and driver assistance.
- Experimental evaluation of the embedded ADAS prototype in representative driving scenarios to assess its perception capability, computational efficiency, and real-time operational feasibility.

1.6 Thesis Organization

The Thesis entitled **"Autonomous Driver Assistant System for Unstructured Environments"**, comprises seven chapters followed by a bibliography. The thesis is organized as follows:

Chapter 1: Introduction

This chapter will introduce the motivation and purpose of the outlined research topic. It will also contain the main idea for the development of the thesis.

Chapter 2: Literature Review

This chapter presents a comprehensive review of the existing literature related to vision-based autonomous perception systems. It begins by discussing recent advances in deep learning-based object detection algorithms, semantic segmentation, lane detection, drivable area segmentation, and road defect detection for intelligent transportation systems. Subsequently, the chapter reviews state-of-the-art unified multi-task learning frameworks developed for autonomous driving, highlighting their architectures, performance, and existing limitations.

The chapter further examines recent developments in autonomous defence vehicle perception, with particular emphasis on defence–civilian entity detection, drivable terrain segmentation, trail and boundary detection, and the challenges associated with unstructured off-road environments. In addition, recent benchmark datasets used for civilian and defence perception are critically reviewed to identify existing research gaps.

Chapter 3: Development of Deep Learning-Based Traffic Perception Frameworks

This chapter presents the development of deep learning-based traffic perception frameworks for intelligent transportation systems. It begins with a comparative evaluation of state-of-the-art YOLO object detection algorithms to analyze the trade-off between detection accuracy and computational efficiency for real-time traffic monitoring. Building upon the insights obtained from this study, the chapter further presents the development of a cascaded YOLOv11-based framework for helmet violation detection and automatic number plate recognition (ANPR) of motorcyclists. The proposed framework integrates motorcycle and rider detection, helmet classification, license plate localization, and Optical Character Recognition (OCR) for automated traffic law enforcement. Finally, the chapter discusses the key findings and research outcomes that established the foundation for the development of the unified multi-task driving perception framework presented in the subsequent chapter.

Chapter 4: Development of a Unified Multi-Task Deep Learning Framework for Driving Perception

This chapter presents the development of the proposed *SmartDriveNet*, a unified multi-task deep learning framework for robust driving perception in intelligent transportation systems. It describes the development of the Indian Traffic and Adverse Weather Dataset (ITAWD), the proposed network architecture, and the methodology adopted for jointly performing traffic object detection, traffic sign recognition, traffic light recognition, lane detection, drivable area segmentation, and road defect detection within a single framework. The chapter also presents the experimental setup, performance evaluation, comparison with state-of-the-art methods, and ablation studies to validate the effectiveness of the proposed framework. Finally, the research contributions are presented.

Chapter 5: Development of a Unified Multi-Task Deep Learning Framework for Autonomous Defence Vehicle Perception

This chapter presents the development of the proposed *DefenceVisionNet*, a unified multi-task deep learning framework for autonomous defence vehicle perception. It describes the development of the Tactical Vehicle Dataset (TVD), the proposed network architecture, and the methodology adopted for jointly performing defence–civilian entity detection, drivable terrain segmentation, and trail/boundary detection within a unified framework. The chapter further presents the experimental setup, quantitative and qualitative performance evaluation, comparison with state-of-the-art methods, and ablation studies to demonstrate the effectiveness of the proposed framework under challenging defence environments. Finally, the key research outcomes and contributions are summarized.

Chapter 6: Development of Secure and Embedded Autonomous Perception Frameworks

This chapter presents the final phase of the research, focusing on secure model deployment and real-time embedded implementation of autonomous perception systems. It provides a brief overview of the application of post-quantum cryptographic techniques, particularly the ML-DSA digital signature algorithm, to enhance the integrity and authenticity of deep learning models, as explored in the *QR-DefenceDrive* and *Q-DriveShieldNet* frameworks. The chapter further presents the design and implementation of a low-cost embedded Advanced Driver Assistance System (ADAS) prototype based on a Raspberry Pi platform, integrating vision-based perception, LiDAR-based obstacle detection, and an interactive Human-Machine Interface (HMI). Finally, the principal contributions and outcomes of the research are summarized.

Chapter 7: Conclusion, Future Directions, and Social Impact

This chapter summarizes the major contributions and key findings of the research presented in this thesis. It highlights the outcomes of the proposed unified multi-task perception frameworks, post-quantum-secure model-verification techniques, and embedded ADAS implementation. The chapter also discusses potential future research directions and outlines the broader social impact of the proposed work toward developing safe, secure, and intelligent autonomous transportation and defence systems.

CHAPTER 2
Literature Review

This chapter presents a comprehensive review of existing research on vision-based autonomous perception systems for civilian transportation and autonomous defence applications. It begins with a discussion of major deep learning-based perception tasks, including traffic-object detection, traffic-sign and traffic-light recognition, road-defect detection, lane detection, and drivable-area segmentation. The chapter subsequently reviews unified multi-task learning frameworks for autonomous driving, highlighting their architectural characteristics, capabilities, and limitations in challenging and unstructured environments.

The chapter further examines benchmark datasets developed for civilian and defence perception, with particular emphasis on datasets representing heterogeneous traffic, adverse weather, road defects, and tactical environments. Existing approaches for autonomous defence perception are also reviewed, including military and civilian entity detection and aerial-object detection, along with the challenges associated with unstructured operational environments. The review identifies key limitations and research gaps related to task coverage, robustness, generalization, computational efficiency, and dataset diversity, thereby establishing the motivation for the research presented in subsequent chapters.

2.1 Deep Learning-Based Perception task

Early vision-based autonomous driving systems relied primarily on manually designed image features and conventional machine learning algorithms. Although these approaches demonstrated the feasibility of automated interpretation of road scenes, their performance was highly dependent on the quality of the handcrafted features and the prevailing environmental conditions. The emergence of deep learning significantly transformed this research direction by enabling models to learn hierarchical and discriminative visual representations directly from training data.

With the rapid development of deep neural networks, learning-based approaches have subsequently been applied to a wide range of perception tasks in autonomous driving. These tasks enable the vehicle to interpret different aspects of its surrounding environment and collectively contribute to safe and reliable decision-making. Among the various components of autonomous driving, certain perception tasks have received particular attention because of their significant influence on the overall performance and safety of the vehicle. The research focuses on these core perception tasks, including traffic

object detection, traffic sign and traffic light recognition, lane detection, drivable-area segmentation, and road-defect identification, particularly the detection of potholes and pavement cracks.

Each of these tasks addresses a specific aspect of environmental perception and presents distinct challenges under varying traffic, illumination, weather, and road conditions. The relevant research developments, methodologies, and challenges associated with these perception tasks are discussed in the following subsections.

2.1.1 Traffic Object Recognition

Effective traffic-object recognition is a fundamental component of autonomous and driver-assistance systems, as it provides the information required for safe navigation and collision avoidance.



Figure 2.1: Challenges in Traffic Object Recognition in Unstructured Roads and Adverse Environmental Conditions

A perception system must distinguish a broad range of road users, such as cars, motorcycles, buses, trucks, bicycles, pedestrians, animals, and unconventional vehicles such as bullock carts. However, reliable recognition in real-world traffic remains challenging because objects may appear at very small scales or at large distances from the vehicle, while dense and heterogeneous traffic can result in frequent overlap and occlusion. Detection performance may further degrade under adverse weather, complex road environments, and low-contrast conditions, as illustrated in Fig. 2.1.

To improve the reliability of traffic-object recognition, deep learning-based detection methods have evolved considerably, with emphasis on accurate localization, effective feature learning, fast inference, and robustness across varying road conditions. Based on their detection mechanism, these methods are commonly divided into two-stage and single-stage detectors. Two-stage detectors first generate candidate regions and then perform classification and localization refinement, often achieving high accuracy with increased computational cost. Single-stage detectors, however, jointly perform object localization and classification, resulting in faster inference and making them well suited for real-time applications such as Advanced Driver Assistance Systems (ADAS) and autonomous driving.

Two-Stage Detectors

Two-stage detectors perform object detection through two sequential steps: generation of candidate regions followed by object classification and localization refinement. Ren et al. introduced Faster R-CNN [14], which employs a Region Proposal Network (RPN) to identify potential object regions from convolutional feature maps. Integrating the RPN with the detection network enables joint optimization and accurate object localization. However, the multiple processing stages increase computational cost and can limit its suitability for real-time traffic perception.

He et al. extended the Faster R-CNN framework to develop Mask R-CNN [41], incorporating an additional branch for instance-level segmentation along with object classification and bounding-box regression. Although this provides pixel-level object representations and richer scene information, the additional computation increases processing requirements. Performance can also be affected by occlusion and complex road conditions.

Overall, two-stage detectors provide strong localization and detailed object understanding but generally involve higher computational complexity and inference latency. These limitations have encouraged the development of computationally efficient single-stage detectors for real-time ADAS applications.

Single-Stage Detectors

Single-stage detectors formulate object detection as a unified prediction task, directly estimating object classes and bounding-box coordinates without a separate region proposal stage. This streamlined design reduces computational overhead and supports faster inference, making these detectors well suited to real-time traffic perception and autonomous driving applications.

Liu et al. proposed the Single Shot MultiBox Detector (SSD) [15], which carries out object classification and localization within a single forward pass. The method uses feature maps of different resolutions together with predefined default boxes, allowing objects of varying sizes to be detected efficiently. This multi-scale design provides a practical trade-off between detection accuracy and computational cost.

The YOLO series advanced the single-stage detection approach by jointly predicting object locations and class probabilities in a single inference process. YOLOv5 [42] provided a practical and efficient implementation with improvements in network design and training strategies, supporting real-time object detection across different model scales. Its flexible architecture contributed to the widespread adoption of YOLO models in computer vision and intelligent transportation applications.

YOLOv8 [17] introduced further architectural refinements, including improvements to feature extraction and the detection head, together with an anchor-free detection strategy. These changes enhanced the overall efficiency and flexibility of the framework compared with earlier YOLO generations.

YOLOv9 [18] introduced Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN). PGI was designed to improve the preservation and utilization of useful training information, while GELAN provides an efficient architecture for feature aggregation and representation learning.

YOLOv10 [43] introduced an end-to-end detection framework with a focus on reducing inference overhead. Its architectural design incorporates consistent dual assignments and a lightweight classification head, while eliminating the dependence on conventional

non-maximum suppression during inference. These changes further improve the balance between detection accuracy and computational efficiency.

More recently, YOLOv11 [19] continued the evolution of the YOLO architecture with improvements in backbone and neck design, feature extraction, and model scalability. These refinements provide different model configurations suited to varying computational requirements while retaining the real-time detection characteristics of the YOLO family.

Overall, the development from SSD to successive YOLO generations reflects a continuous emphasis on efficient feature representation, multi-scale detection, improved architectural design, and reduced inference overhead. These developments have established single-stage detectors as important components of real-time computer vision systems, including traffic perception and ADAS applications.

Research Gap

The reviewed object detection methods demonstrate significant progress in balancing detection accuracy with computational efficiency. Two-stage approaches such as Faster R-CNN and Mask R-CNN provide detailed object localization and richer representations but generally involve greater computational requirements. In contrast, SSD and successive YOLO architectures offer faster and more streamlined detection, making them more appropriate for real-time ADAS deployment.

Despite these advances, important challenges remain for reliable traffic-object recognition in unstructured road environments. Most existing approaches are primarily evaluated on general-purpose or relatively structured benchmark datasets, such as PASCAL VOC [44] and MS COCO [45]. Consequently, their effectiveness in heterogeneous and unconstrained traffic environments is not fully established.

Indian road environments present additional complexity because multiple vehicle types, pedestrians, motorcycles, bicycles, buses, trucks, and non-conventional road users can coexist within densely populated and irregular traffic scenes. Variations in object scale, frequent occlusion, illumination changes, adverse weather, and irregular road structures further increase the difficulty of reliable detection.

Therefore, a need remains for an efficient traffic-object detection framework that can maintain real-time inference while providing robust recognition across diverse object categories and challenging environmental conditions. Addressing these limitations is essential for developing generalized deep learning-based perception systems for unstructured Indian road environments.

2.1.2 Traffic Sign Recognition

Traffic sign recognition is a fundamental perception function in ADAS, enabling vehicles to identify regulatory, warning, mandatory, and informational signs for safer navigation and decision-making. A reliable recognition system must identify signs across variations in scale, viewpoint, illumination, weather, and visibility, while also handling occlusion, physical damage, fading, and visually similar sign categories. These conditions become particularly challenging in heterogeneous and unstructured traffic environments, as depicted in Fig. 2.2. Yen *et al.* [46] investigated traffic sign recognition under occlusion using a convolutional neural network-based framework that incorporated regional masks and local feature extraction. Their approach demonstrated the potential of combining localized visual information for recognizing partially occluded signs. However, its performance



Figure 2.2: Challenges Associated with Traffic Sign Recognition in Unstructured Roads and Adverse Environmental Conditions

remained affected by certain occluded samples, indicating the need for more effective integration of local and global contextual information.

Qin *et al.* [47] addressed traffic sign recognition under visibility degradation by developing the NZ-Traffic-Sign 3K benchmark and examining detection methods based on Faster R-CNN and YOLOv5. The study provided a useful benchmark for assessing recognition performance under challenging conditions. Nevertheless, damaged, poorly visible, and degraded signs continue to present difficulties for reliable recognition, particularly when important visual features are partially missing.

Zhu *et al.* [4] investigated YOLOv5-based traffic sign recognition and compared its performance with SSD using the GTSRB [48] dataset. The study demonstrated the effectiveness of modern single-stage detectors for traffic sign recognition. However, the evaluation was predominantly based on non-Indian traffic sign datasets, limiting the assessment of their applicability to the diverse sign structures and road conditions encountered in India.

In contrast, Megalingam *et al.* [49] proposed an RM R-CNN-based framework specifically targeting Indian traffic signs. The study constitutes a significant contribution toward bridging the gap in India-specific traffic sign recognition research. However, reliable recognition remains challenging in the presence of occlusion, illumination variations, and visually similar signs, which can lead to confusion between different sign categories.

Ayadande *et al.* [3] further explored YOLO-based traffic sign recognition for deployment in real-time road safety systems. Their work demonstrates the potential of lightweight single-stage detection frameworks for rapid recognition. However, the evaluation provides limited evidence regarding robustness under highly diverse traffic conditions, adverse weather, and other challenging real-world driving scenarios.

Research Gap

The reviewed literature indicates substantial advances in traffic sign detection and recognition. Nevertheless, existing methods still face significant challenges in accurately identifying signs that are partially occluded, physically damaged, faded, distant, or captured under insufficient illumination. Recognition can also be affected by variations in viewpoint, weather conditions, and visual similarity between different sign categories. Fur-

thermore, many existing methods are evaluated using datasets collected outside India, such as GTSRB and GTSDB, which may not adequately represent the appearance, distribution, and environmental conditions of Indian traffic signs.

Although efforts such as the Indian traffic sign recognition framework of Megalingam *et al.* [49] address this dataset and geographic limitation, robust recognition under unstructured traffic conditions remains insufficiently explored. In particular, the simultaneous effects of occlusion, illumination variation, sign degradation, scale differences, and adverse weather require further investigation. Therefore, there is a need for a traffic sign recognition framework that can provide accurate and computationally efficient recognition while maintaining robustness across diverse and challenging Indian road environments.

2.1.3 Traffic Light Recognition

Traffic light recognition constitutes a key component of the perception system in autonomous vehicles and Advanced Driver-Assistance Systems (ADAS). It allows a vehicle to detect traffic lights and accurately interpret their operational states, thereby supporting safe vehicle movement, appropriate navigation, and effective driving decisions. A reliable system should accurately localize traffic lights and classify their states, including red, yellow, green, and directional arrow signals, while maintaining stable predictions across consecutive video frames. This task is particularly important at intersections, where incorrect interpretation of a signal state can adversely affect vehicle planning and safety.

Traffic light recognition remains challenging because signals often appear as small and distant objects in road scenes, resulting in limited visual information for detection and classification. Their appearance can also be affected by partial occlusion from vehicles and roadside objects, illumination changes, reflections, nighttime conditions, and adverse weather. Rain, fog, and snow may further reduce visibility and contrast, as depicted in Fig. 2.3.

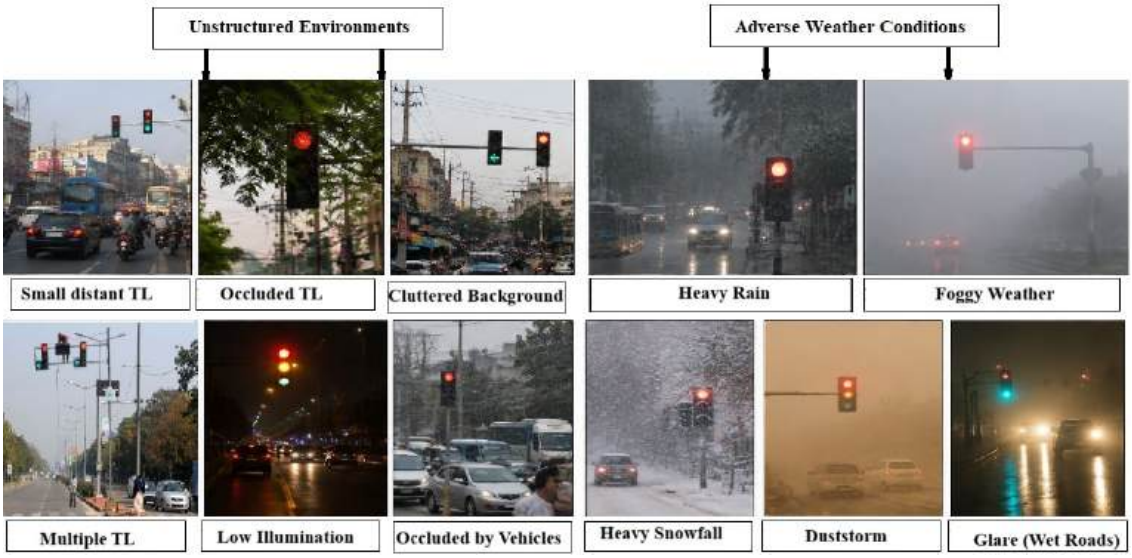

Figure 2.3: Challenges in Traffic Light (TL) in Detection Unstructured Road and Challenging Environmental Conditions

To address these challenges, researchers have explored various deep learning-based solutions. Wang and Zhou [50] proposed a real-time traffic light recognition framework that

combines high dynamic range (HDR) imaging with CNN-based classification, saliency analysis, and temporal trajectory tracking. The integration of spatial and temporal information improved recognition reliability and reduced false detections. However, the approach is closely associated with its imaging configuration and has limited validation in highly heterogeneous and unstructured traffic environments.

Chuang et al. [51] proposed a traffic light detection framework incorporating self-attention, pyramidal feature fusion, and a parallel residual Bi-FPN to strengthen multi-scale feature representation. These mechanisms are particularly useful for detecting traffic lights of different apparent sizes. Nevertheless, the method was primarily evaluated using established traffic datasets and structured road scenarios, leaving its robustness in complex and heterogeneous traffic environments insufficiently explored.

Garg et al. [6] investigated the robustness of traffic light detection in adverse weather by augmenting the Bosch Small Traffic Light Dataset with synthetically generated snowfall and assessing the performance of YOLOv7, YOLOv8, and Faster R-CNN. Although the study demonstrated improved detection under snowy conditions, its evaluation primarily focused on snowfall and did not comprehensively address other adverse conditions, such as heavy rain, dust storms, fog, poor illumination, or combinations of multiple environmental factors.

Behrendt et al. [52] explored a deep learning-based framework for traffic light detection, tracking, and classification, highlighting the importance of combining these perception stages for reliable video-based recognition. However, broader validation under diverse real-world traffic conditions remains necessary.

Research Gap: Despite the progress achieved by existing approaches, several limitations remain. Most methods are evaluated on specific benchmark datasets or relatively structured road environments and therefore may not adequately represent the complexity of unstructured traffic scenarios. In particular, the combined effects of small and distant traffic lights, illumination variation, partial occlusion, adverse weather, complex backgrounds, and heterogeneous traffic remain insufficiently addressed. Furthermore, existing studies often focus on individual environmental challenges rather than evaluating robustness across multiple conditions. The identified limitations emphasize the necessity of developing a robust and generalizable traffic light recognition framework capable of sustaining accurate real-time detection and signal-state classification across diverse environmental and road conditions.

2.1.4 Road Defect Detection: Potholes and Cracks

Potholes and pavement cracks are major road-surface defects, as shown in Fig.2.4, that affect vehicle safety, ride comfort, and infrastructure quality. In India, poor road conditions and potholes are significant road-safety concerns [53]. To automate road inspection, researchers have explored semantic segmentation, object detection, stereo vision, and deep-learning-based approaches. Yarram *et al.* [9] proposed a cascaded semantic-segmentation framework for city-scale road auditing. Although effective for large-scale defect localization, the method can experience confusion between visually similar conditions such as wet roads and water accumulation. Rithish *et al.* [54] evaluated YOLOv4 for road irregularity detection and showed its potential for unconstrained road scenes; however, performance can decrease under vehicle occlusion and strong illumination.

Park *et al.* [7] compared the YOLOv4, lightweight YOLOv4-tiny, and YOLOv5s detectors for real-time pothole detection. While lightweight detectors provide faster inference,

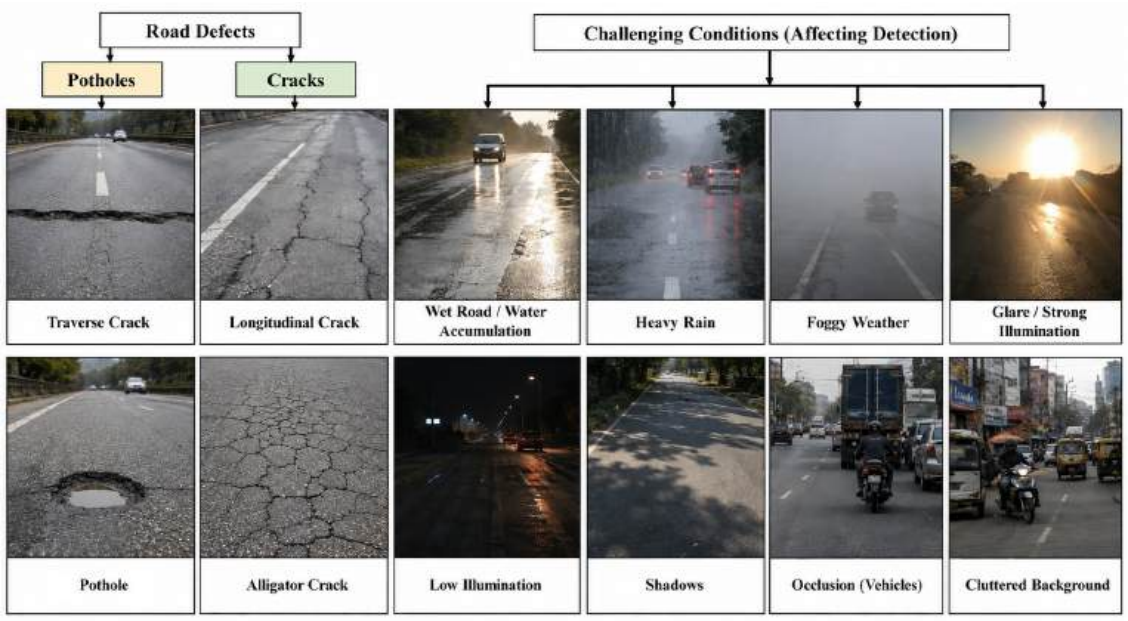

Figure 2.4: Challenges in Potholes and Cracks Detection under Unstructured Environments and Adverse Weather Conditions.

the study mainly focused on potholes and used a limited dataset, leaving crack detection and challenging environmental conditions less explored. Fan *et al.* [55] incorporated stereo disparity analysis combined with three-dimensional road-surface reconstruction to improve pothole detection. However, the use of depth information increases sensing and computational requirements. Similarly, Dhiman and Klette [56] investigated computer-vision and learning-based pothole detection, but real-time deployment and robustness under varying visual conditions remain challenges.

For pavement cracks, Yang *et al.* [57] proposed a feature-pyramid-based hierarchical network (FPHBN) to capture crack features across multiple spatial scales. Nevertheless, thin cracks remain difficult to detect under shadows, low contrast, and illumination changes. Salcedo *et al.* [58] combined road segmentation and object detection for road-damage identification and maintenance prioritization, but further evaluation is required in complex traffic and adverse conditions. Wang *et al.* [8] proposed an improved YOLOv5-based model for pavement-crack detection in complicated scenarios; however, its generalization across different road surfaces, weather conditions, and deployment environments remains limited.

Research Gap: Despite substantial progress in deep learning-based pothole and pavement crack detection, existing research remains largely focused on individual road-defect categories and is often validated using limited or task-specific datasets. Although methods based on object detection, semantic segmentation, stereo vision, and YOLO architectures have shown promising results, their robustness in complex, unstructured road environments has not been adequately established. Factors such as occlusion, inadequate illumination, shadows, glare, rain, fog, background clutter, and visually similar road surfaces can adversely affect detection accuracy. In addition, small potholes and fine cracks are particularly challenging because of their limited spatial extent and weak visual contrast. Stereo-vision and multi-stage solutions may further increase computational complexity, creating challenges for real-time deployment. Therefore, a unified and computationally efficient framework is required to detect multiple types of road defects

while maintaining reliable accuracy and real-time performance under diverse road and environmental conditions

2.1.5 Drivable Area Segmentation

Drivable area segmentation is a fundamental perception task in autonomous driving, as it enables a vehicle to identify regions of the scene that can be safely traversed. Unlike object detection, which focuses on individual road users and obstacles, segmentation provides pixel-level information about navigable regions and their boundaries. This capability is especially important in unstructured environments where lane markings may be absent, road boundaries may be unclear, and the driving surface can be influenced by shadows, vegetation, pedestrians, parked vehicles, and other visual disturbances.

Early deep learning research showed that encoder–decoder architectures provide an effective basis for pixel-wise prediction. Ronneberger *et al.* introduced U-Net [59], in which the encoder learns hierarchical semantic features while skip connections preserve spatial information during decoding. This architecture has subsequently been applied to road and drivable-region segmentation because of its ability to retain detailed boundary information. Qiao and Zulkernine [12] investigated deep-learning approaches for drivable area detection and demonstrated the applicability of convolutional neural networks for identifying navigable road regions. However, models developed on relatively controlled datasets may show reduced robustness when road appearance and scene structure vary considerably.

Several segmentation architectures have focused on improving spatial representation and contextual understanding. SegNet [60] uses an encoder–decoder structure in which pooling indices from the encoder guide decoder upsampling, thereby reducing the computational and parameter requirements of the decoding process. PSPNet [61] employed pyramid pooling to aggregate contextual features across multiple spatial resolutions, while Feature Pyramid Networks (FPN) [62] exploit multi-resolution feature representations to combine information from different levels. These multi-scale strategies are relevant to drivable-area segmentation because the visible road region can vary significantly with camera position, perspective, and scene geometry.

Attention-based methods have further improved the selection of informative features in complex road scenes. Li *et al.* [13] developed TA-Unet by integrating triplet attention within a U-Net architecture to enhance the representation of informative spatial and channel features. Such mechanisms can help distinguish important road features from surrounding scene content. Nevertheless, weak road boundaries, shadows, occlusions, and visually similar non-drivable regions can still lead to segmentation errors.

Computational efficiency is also important for practical autonomous-driving applications. Baheti *et al.* proposed Eff-UNet [63], which combines an EfficientNet-based encoder with a U-Net-style decoder to obtain effective feature representations with reduced computational requirements. Lightweight encoder–decoder architectures are therefore attractive for embedded platforms with limited processing capacity. However, reducing computational complexity may introduce trade-offs in fine boundary recovery and robustness under challenging conditions.

Multi-scale contextual information has also been addressed using atrous convolution. DeepLab-based models utilize dilated convolutions to expand the receptive field while preserving detailed spatial information in the feature maps. DeepLabv3 [64] introduced atrous spatial pyramid pooling to capture context at different sampling rates, whereas

DeepLabv3+ [65] combined this strategy with an encoder–decoder design for improved boundary refinement. These approaches are useful for road-region segmentation because both local surface characteristics and wider scene context are required to determine whether a region is drivable.

Domain variation remains another important challenge. A segmentation model trained on data from a particular geographical region, weather condition, or illumination setting may not generalize effectively to substantially different visual domains. Image-to-image translation techniques such as CycleGAN [66] have therefore been explored to reduce domain discrepancies without requiring paired training images. Although these methods can support domain adaptation, generated images may not always preserve the scene characteristics necessary for accurate pixel-level segmentation.

Research Gap: Despite advances in deep-learning-based drivable-area segmentation, their performance may degrade under changes in illumination, weather, road appearance, occlusion, and scene structure, while computationally intensive models can limit real-time deployment on embedded platforms. Lightweight approaches may also sacrifice fine boundary details. These limitations motivate the development of a unified, computationally efficient segmentation framework that provides accurate, robust drivable-area predictions under diverse, challenging driving conditions.

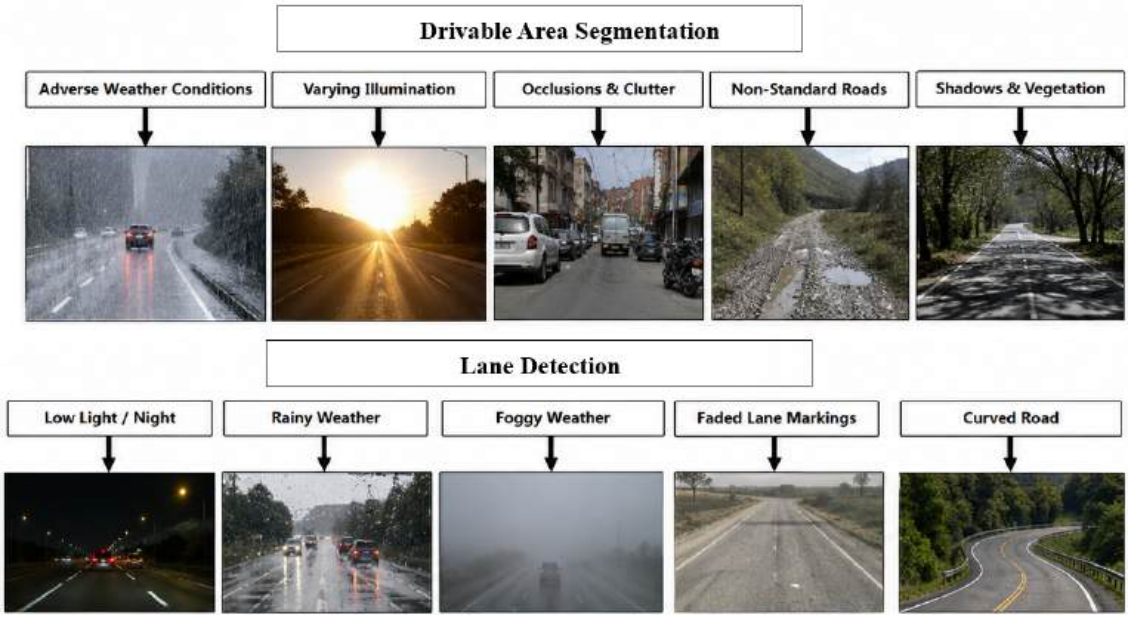

Figure 2.5: Challenges in Lane and Drivable Area Segmentation under Unstructured Environments and Adverse Weather Conditions.

2.1.6 Lane Detection

Lane detection is an essential component of autonomous driving and Advanced Driver-Assistance Systems (ADAS), enabling vehicles to identify road lane boundaries and supporting applications such as lane-keeping, vehicle positioning, and departure warning. The objective is to identify lane markings and estimate their geometry from road images. However, reliable lane detection remains challenging due to variations in lane-marking visibility, including faded, discontinuous, curved, or partially occluded markings. Variations in illumination, weather, road appearance, and surrounding traffic further increase

the difficulty, particularly on unstructured roads where standard lane markings may be absent or irregular.

Wang *et al.* introduced LaneNet [67], which formulates lane detection as a combination of pixel-level lane classification and subsequent lane-instance localization. This approach demonstrated the effectiveness of deep-learning-based feature extraction for identifying lane structures, but its performance can decrease when lane markings are incomplete or when the road environment differs substantially from the training data. Hou *et al.* proposed Self Attention Distillation (SAD) [68] to improve lightweight lane detection networks by transferring attention information within the network without requiring additional supervision. Although the method improves feature representation while maintaining a compact architecture, achieving a suitable balance between detection accuracy and real-time inference remains challenging.

Tabelini *et al.* developed LaneATT [69], an attention-based lane detection framework that uses global image features to improve lane representation and localization. The method provides effective lane detection while targeting real-time operation; however, maintaining consistent performance under crowded traffic, occlusion, and complex road conditions remains difficult. Oguz *et al.* [11] investigated a one-dimensional deep-learning approach for lane detection, aiming to reduce computational requirements while maintaining robustness under adverse conditions. Nevertheless, variations in road structure and lane appearance can still affect its performance across diverse environments.

Other approaches have explored structural and geometric information to improve lane detection efficiency and reliability. Qin *et al.* proposed a row-based lane detection approach [70], which treats lane localization as the selection of lane positions across predefined image rows. By exploiting global features and structural constraints, the method reduces the computational burden associated with conventional pixel-level prediction. Sultana *et al.* [71] developed a vision-based lane detection and tracking framework that combines intensity-based edge processing with geometric constraints and temporal tracking. The approach improves lane verification and maintains lane position across consecutive frames; however, severe weather, heavy occlusion, and degraded or non-standard lane markings can still reduce detection reliability.

Research Gap: Although recent lane detection methods have achieved notable improvements through enhanced feature extraction, attention mechanisms, structural constraints, and efficient network designs, their reliability in unstructured driving scenarios remains insufficiently established. Most existing approaches are developed and assessed using structured road datasets with relatively well-defined lane markings, limiting their applicability to environments where markings may be faded, fragmented, absent, or partially occluded. Additional factors, including irregular road geometry, poor illumination, adverse weather, heavy traffic, and background clutter, further complicate accurate lane localization and continuity estimation. Moreover, there remains a trade-off between detection accuracy and computational efficiency: compact networks may sacrifice fine-grained lane information, whereas more sophisticated architectures can impose higher computational costs and inference latency. These limitations highlight the need for a lightweight yet robust lane detection framework that can preserve lane structure and continuity while providing accurate and efficient perception across diverse road, traffic, illumination, and weather conditions.

2.2 Multi-Task Learning for Autonomous Driving

Multi-task learning (MTL) is an effective strategy for autonomous-driving perception, enabling multiple related tasks to be learned within a common neural network. Instead of using separate models for individual perception functions, MTL typically shares a feature-extraction backbone with task-specific prediction branches. This reduces computational redundancy, memory usage, and inference time while allowing related tasks to benefit from complementary visual information. Hence, MTL is well suited to real-time autonomous-driving and ADAS applications, where multiple perception tasks must operate simultaneously under limited computational resources.

Table 2.1: Comparison of Multi-Task Learning Models for Autonomous Driving Perception

Model	Key Architectural Approach	Research Gap / Limitation
MultiNet [20]	Employs a unified encoder–decoder architecture with shared features for classification, vehicle detection, and road segmentation. The three tasks are processed in a single forward pass, reducing computational redundancy and supporting real-time inference.	Limited task coverage and primarily evaluated on the KITTI dataset, with limited investigation in diverse and unstructured road environments.
DLT-Net [21]	Uses a joint architecture for object detection, drivable-area detection, and lane detection with task-aware information exchange through context tensors.	Task interaction increases architectural complexity, while its effectiveness depends on the correlation among the individual tasks.
YOLOP [2]	Uses a CSPDarknet-based shared encoder with SPP–FPN–PAN feature processing and three task-specific branches. The architecture promotes efficient feature extraction and multi-task prediction.	Performance may degrade under illumination changes, occlusion, adverse weather, faded lane markings, and irregular road structures.
HybridNets [22]	Employs an EfficientNet-B3 backbone with weighted BiFPN for bidirectional multi-scale feature fusion and a unified segmentation decoder for related tasks.	Feature fusion and unified decoding introduce additional computational requirements and require a balance between accuracy and efficiency in complex traffic scenes.
YOLOPv2 [23]	Uses an E-ELAN-based feature extractor with improved multi-scale processing and a YOLOv7-inspired detection framework. Segmentation branches operate at different feature levels to recover spatial details.	Evaluation is mainly done on structured driving datasets, with limited investigation under highly unstructured roads, adverse conditions, and heterogeneous traffic.

An early contribution to multi-task perception for autonomous driving was MultiNet [20], introduced by Teichmann *et al.*. The proposed architecture integrates scene understanding, vehicle recognition, and road segmentation into a common framework. By combining a common feature-extraction backbone with dedicated modules for individual tasks, the network can generate results for all three functions simultaneously in one inference cycle. This design reduces computational overhead and makes the approach suitable for real-time driving applications. Its performance was evaluated using the KITTI benchmark, demonstrating the potential of unified multi-task perception systems. However, the framework addresses a relatively limited set of perception tasks and its evaluation is

mainly based on conventional driving scenes, leaving its robustness in highly diverse and unstructured environments less explored.

Qian *et al.* introduced DLT-Net [21] to jointly perform traffic-object detection, drivable-area detection, and lane-line detection. In addition to sharing visual features, the architecture facilitates information exchange between task branches through context tensors. Such task-aware interaction enables information learned by one perception task to support the prediction of related tasks. Although this approach demonstrates the benefits of inter-task information sharing, the effectiveness of such interaction depends on the degree of correlation among the tasks and may introduce additional architectural complexity.

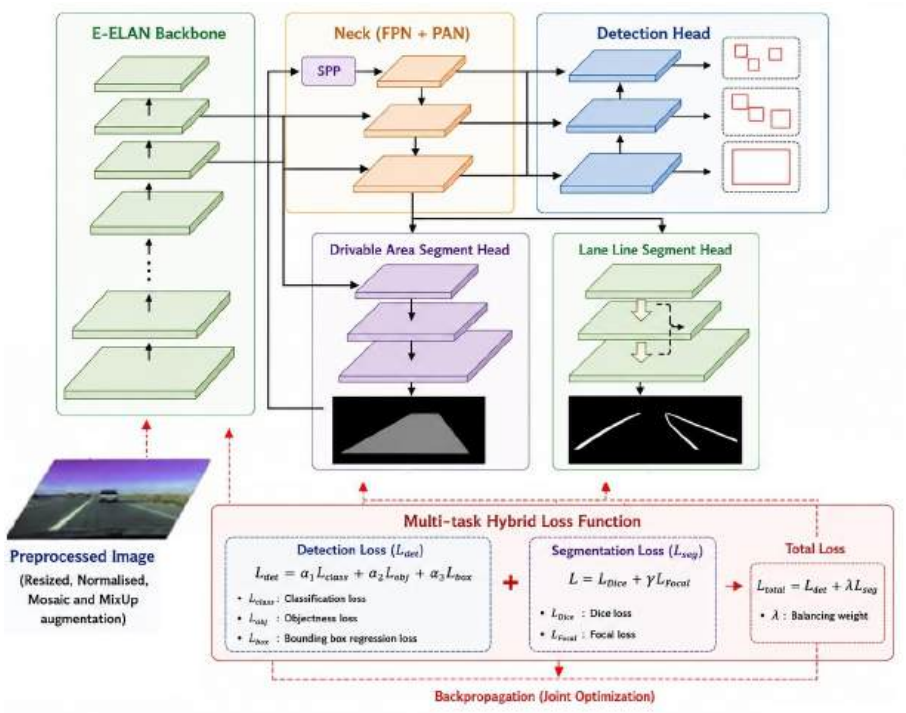

Figure 2.6: YOLOPv2 architecture

YOLOP [2] introduced a unified panoptic driving perception framework that simultaneously performs traffic-object detection, drivable-area segmentation, and lane-line segmentation within a single network. Its architecture consists of a CSPDarknet backbone, an SPP–FPN based neck, and task-specific prediction heads. The CSPDarknet backbone extracts hierarchical features while reducing redundant gradient information through partial feature-map processing. The SPP module captures contextual information at multiple scales, while the FPN and PAN structures enhance multi-scale feature fusion for subsequent task prediction. Separate heads are employed for object detection, drivable-area segmentation, and lane detection, allowing multiple perception outputs to be generated from a shared feature representation. This unified design reduces computational redundancy compared with independent task-specific models while supporting real-time perception.

The three tasks in YOLOP are jointly optimized using a multi-task loss, which can be expressed as

$$\mathcal{L}_{\text{YOLOP}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{da}} + \mathcal{L}_{\text{ll}}, \tag{2.1}$$

where $\mathcal{L}_{\text{det}}$, $\mathcal{L}_{\text{da}}$, and $\mathcal{L}_{\text{ll}}$ represent the detection, drivable-area segmentation, and lane-line segmentation losses, respectively. Thus, the three perception objectives are optimized simultaneously using the shared feature representation.

Building upon YOLOP, YOLOPv2 [23] introduced several architectural modifications, as shown in Fig.2.6, to improve feature extraction, multi-scale representation, segmentation quality, and inference efficiency. The CSPDarknet backbone was replaced by an E-ELAN-based backbone, which provides enhanced feature aggregation while maintaining efficient gradient propagation. The network retains the SPP–FPN feature-processing structure, with PAN-based aggregation supporting the object-detection branch. YOLOPv2 also modifies the connections between the shared features and task-specific decoders. The drivable-area branch uses relatively shallow features followed by upsampling to preserve spatial information, whereas the lane branch uses deeper fused features and deconvolution to recover fine lane structures. Mosaic and MixUp augmentation are additionally employed during training to increase the diversity of driving scenes.

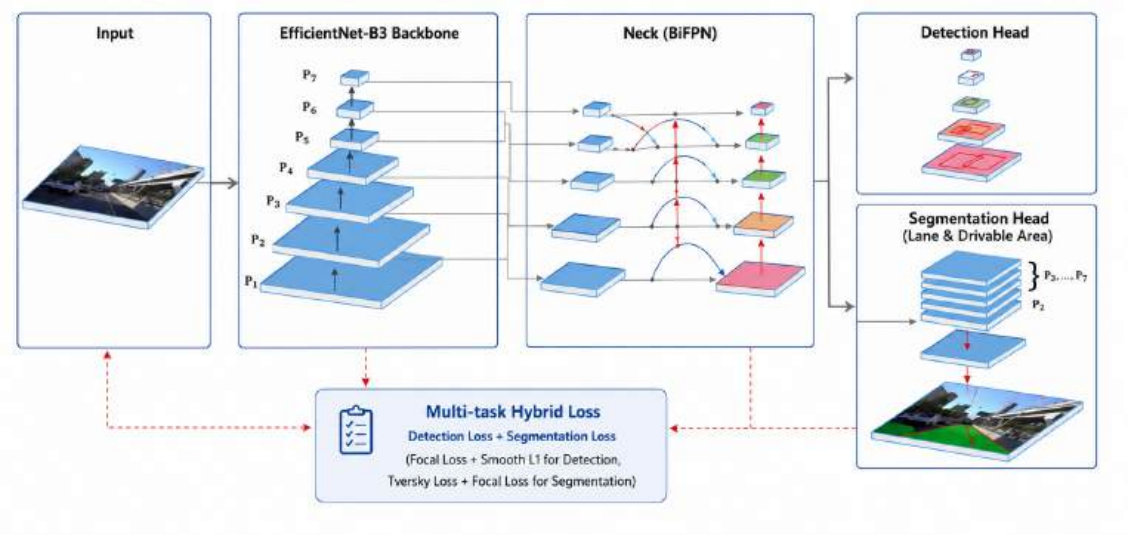

Figure 2.7: HybridNets architecture

The loss formulation of YOLOPv2 is more structured than that of YOLOP, with separate objectives for object detection and pixel-level segmentation. The detection loss incorporates classification, objectness, and bounding-box regression, while the segmentation loss combines Dice and Focal losses to handle class imbalance and improve region-level agreement. The overall multi-task objective can be expressed as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda \left(\mathcal{L}_{\text{Dice}} + \gamma \mathcal{L}_{\text{Focal}} \right), \tag{2.2}$$

where $\mathcal{L}_{\text{det}}$ comprises classification, objectness, and bounding-box regression losses, while $\mathcal{L}_{\text{Dice}}$ and $\mathcal{L}_{\text{Focal}}$ represent the Dice and Focal losses used for segmentation. The parameters λ and γ balance the contributions of the segmentation and Focal loss components, respectively. This joint optimization enables YOLOPv2 to learn object-level and pixel-level representations within a unified real-time perception framework.

Overall, YOLOPv2 extends YOLOP through an E-ELAN-based feature extractor, improved multi-scale feature processing, redesigned segmentation branches, and enhanced data augmentation. These modifications improve feature representation and fine-grained spatial understanding while retaining the real-time multi-task perception capability of the original framework.

HybridNets [22] further developed unified driving perception by replacing the CSPDarknet backbone of YOLOP with EfficientNet-B3, as shown in Fig.2.7, providing hierarchical

feature extraction with a favorable balance between accuracy and computational cost. The resulting feature maps are subsequently processed through a Bi-directional Feature Pyramid Network (BiFPN), which integrates information across different resolution levels through learnable top-down and bottom-up pathways. This multi-scale aggregation allows the model to preserve both high-level contextual information and fine spatial details, improving its ability to recognize vehicles, objects, and road-related features of varying sizes. For object detection, HybridNets employs a multi-scale anchor-based detection head with automatically customized anchor configurations for different feature levels. The detection objective combines Focal losses for classification and object confidence with Smooth L1 loss for bounding-box regression. For segmentation, HybridNets uses a unified decoder for drivable-area and lane-line prediction, with a hybrid Tversky–Focal loss to address pixel-level class imbalance. The overall multi-task objective is formulated as

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{det}} + \beta \left(\mathcal{L}_{\text{Tversky}} + \lambda \mathcal{L}_{\text{Focal}} \right), \tag{2.3}$$

where $\mathcal{L}_{\text{det}}$ comprises Focal losses for detection within a unified network. These modifications improve feature sharing, multi-scale representation, and computational efficiency while maintaining real-time inference. However, the increased feature-fusion complexity and reliance on shared representations still require an appropriate balance between computational cost and task-specific accuracy, particularly in highly heterogeneous and visually challenging driving environments.

Research Gap Overall, existing multi-task perception frameworks demonstrate the benefits of shared representations, task interaction, and multi-scale feature learning for efficient autonomous driving. However, most approaches address a limited number of perception tasks and are mainly evaluated on conventional driving datasets. Their robustness in highly unstructured environments remains insufficiently explored, where irregular road boundaries, faded lane markings, occlusions, adverse weather, illumination variations, heterogeneous traffic, and degraded road surfaces can simultaneously affect multiple tasks. Moreover, increasing model complexity may improve accuracy but introduce additional computational overhead. Therefore, there is a need for a lightweight and robust multi-task perception framework that can support diverse perception tasks while preserving task-specific information and maintaining reliable real-time performance in challenging and heterogeneous environments.

2.3 Datasets for Autonomous Driving Perception

The effectiveness of deep learning-based perception systems largely relies on the availability of diverse and representative datasets that adequately capture real-world driving conditions. Dataset characteristics such as scene diversity, environmental conditions, object categories, annotation structure, image resolution, and supported perception tasks directly influence the performance and generalization of trained models. Accordingly, several datasets have been developed for autonomous driving, addressing perception tasks such as traffic-object detection, traffic-sign and traffic-light recognition, lane detection, drivable-area segmentation, and road-damage detection.

Existing datasets vary in their geographic coverage, scene characteristics, environmental conditions, annotation formats, and task composition. Some provide large-scale driving data with annotations for multiple perception tasks, while others are designed primarily

for specific perception objectives. These datasets provide important benchmarks for developing and evaluating deep learning-based autonomous-driving systems.

The following subsections review representative datasets for vehicle and road-object detection, traffic-sign and signal identification, road-damage detection, and lane and drivable-area perception. Particular emphasis is placed on datasets relevant to Indian road environments, heterogeneous traffic, and challenging weather and illumination conditions.

2.3.1 Traffic Object Detection Datasets

Traffic-object detection datasets provide the visual data required to identify and localize vehicles, pedestrians, cyclists, and other road users. The Microsoft Common Objects in Context (MS COCO) dataset [45] is a widely adopted benchmark for general-purpose object recognition and detection. It comprises approximately 328,000 images containing more than 2.5 million labeled object instances across 91 object categories. Owing to its large scale and diversity, the dataset is commonly used for training, pre-training, and assessing object detection models. Nevertheless, MS COCO was developed for general computer vision applications rather than intelligent transportation systems. Consequently, it does not adequately capture several characteristics of real driving environments, including traffic-specific interactions, road layouts, spatial relationships among vehicles and infrastructure, and varying environmental conditions.

The KITTI dataset [72] was specifically developed for autonomous driving research and provides synchronized camera and range-sensing data collected from a vehicle operating in urban environments. Its benchmarks include stereo visual-odometry sequences covering approximately 39.2 km, and more than 200,000 3D object annotations captured in complex traffic scenes containing multiple vehicles and pedestrians. It has been widely used for vehicle, pedestrian, and cyclist detection, localization, and scene understanding. Although KITTI established an important benchmark for autonomous-driving perception, its scenes are comparatively structured and its geographic coverage is limited. Consequently, it does not adequately represent the highly heterogeneous traffic and irregular road environments encountered in developing regions such as India.

The Detection in Adverse Weather Nature (DAWN) dataset [73] provides another important source of challenging traffic scenes. It includes 1,000 images collected from real traffic environments, covering challenging situations such as fog, rainfall, snowfall, and sandstorms. The dataset provides bounding-box annotations for vehicles, enabling the evaluation of detection algorithms under conditions where poor visibility can affect detection accuracy and reliability.

Overall, traffic-object datasets provide important resources for developing robust detection systems, but their suitability for unstructured Indian road scenes depends on their geographic diversity, representation of local traffic conditions, environmental variability, and coverage of relevant object categories.

2.3.2 Traffic Sign and Traffic Light Datasets

Traffic-sign and traffic-light datasets are essential for enabling vehicles to interpret regulatory information and traffic-signal states. The German Traffic Sign Recognition Benchmark (GTSRB) [48] is a widely used benchmark for traffic-sign classification. It contains more than 50,000 images representing 43 traffic-sign classes. Although GTSRB provides a standardized benchmark for recognition algorithms, its traffic signs and visual

environment are primarily representative of Germany. Consequently, its direct applicability to Indian traffic-sign recognition is limited.

Indian-specific traffic-sign datasets have been developed to overcome this geographical limitation. Megalingam *et al.* [49] developed an India-specific traffic sign dataset with 6,480 images and 7,056 sign instances covering 87 categories, using images acquired from real-world road scenes in India and supplemented with additional images obtained from public sources, providing variation in traffic scenes and sign appearance. The dataset provides substantial coverage of Indian traffic-sign categories. Such India-specific data are particularly relevant because Indian traffic signs may exhibit different designs, placement, appearance, and environmental degradation compared with European datasets.

For traffic-light detection and recognition, the **Bosch Small Traffic Lights Dataset (BSTLD)** [52] provides a dedicated benchmark for traffic-light perception in driving scenarios. The dataset contains 5,000 manually annotated images for training and a separate video sequence comprising 8,334 frames for evaluation. Its annotations support the detection of small and distant traffic-light instances under varying visual conditions. The dataset is primarily used for evaluating traffic-light detection and recognition, particularly for challenging small-object perception in real-world driving scenes.

Overall, these datasets provide valuable benchmarks for developing and evaluating traffic-sign and traffic-light perception models across diverse recognition and detection scenarios.

2.3.3 Road Damage Datasets

Road-surface defects such as potholes and pavement cracks are important considerations for vehicle safety, road maintenance, and autonomous navigation. Several datasets have been developed to support automated detection and segmentation of pavement damage.

The **Road Damage Detection Dataset (RDD)** [27] is a widely used benchmark for road-damage detection. The dataset comprises 26,336 road images from India, Japan, and the Czech Republic, along with more than 31,000 annotated road-damage instances. The dataset comprises four categories of pavement defects: longitudinal-cracks, transverse-cracks, alligator-cracks, and potholes. Its multi-country composition provides variation in pavement appearance and road conditions, making it useful for evaluating road-damage detection models.

The **Crack500 dataset** [74] is a pavement-crack dataset developed for fine-grained crack detection and segmentation. It contains 500 high-resolution road images with pixel-level annotations of pavement cracks. The dataset includes cracks with variations in shape, width, orientation, and visual appearance, making it suitable for evaluating segmentation methods that require precise localization of thin and irregular crack structures.

The **Pothole-600 dataset** [75] is specifically designed for pothole detection and contains 600 annotated road images. The dataset includes potholes with variations in size, shape, position, and surrounding road conditions, providing a benchmark for object-detection approaches targeting pavement failures. Its focused annotations support the development and evaluation of automated pothole detection systems.

These datasets provide complementary resources for road-defect perception. RDD offers broad coverage of multiple pavement-damage categories, while Crack500 and Pothole-600 provide focused datasets for crack segmentation and pothole detection, respectively. Together, they represent important benchmarks for developing deep learning-based road-damage perception systems.

2.3.4 Lane and Drivable Area Datasets

Lane and drivable-area perception provides essential geometric information for autonomous driving by enabling vehicles to understand road structure, estimate lane boundaries, and identify regions suitable for safe movement. Lane detection focuses on identifying the position and geometry of lane markings, whereas drivable-area segmentation determines the traversable regions of a scene. These tasks become particularly challenging in unstructured environments where lane markings may be faded, discontinuous, occluded, or absent, and road boundaries may be poorly defined.

The **CULane dataset** [76] is a large-scale lane detection benchmark comprising 133,235 frames collected from over 55 hours of driving videos in Beijing. The dataset is divided into 88,880 training, 9,675 validation, and 34,680 test images. The data set includes normal scenes and challenging categories representing conditions such as crowded traffic, nighttime, shadows, and other difficult scenarios. Lane boundaries are manually annotated using cubic splines, including lanes that are partially occluded or not directly visible but can be inferred from contextual information. CULane primarily focuses on four major lane markings relevant to practical driving applications.

The **TuSimple dataset** [77] is another widely used lane detection benchmark containing 6,408 highway images from the United States at a resolution of 1280×720 pixels. It includes 3,626 training, 358 validation, and 2,782 test images, with variations in environmental and weather conditions. An extended version also provides additional segmentation annotations. However, the dataset primarily represents structured highway environments and has limited geographic and traffic diversity.

Overall, CULane and TuSimple provide detailed lane annotations but are primarily task-specific.

2.3.5 Multi-Task Autonomous Driving Datasets

Multi-task autonomous-driving datasets provide annotations for multiple perception objectives within a common driving scene. Such datasets are particularly useful for developing unified perception systems because different tasks can be learned from shared visual representations while maintaining task-specific outputs.

The **Berkeley DeepDrive 100K (BDD100K) dataset** [24] is a large-scale benchmark developed for diverse street-scene understanding. It includes 100,000 driving videos, each approximately 40 seconds long, recorded across more than 50,000 trips. The dataset covers diverse road scenes, weather conditions, illumination levels, and times of day. It provides annotations supporting multiple perception tasks, such as traffic-object detection, lane detection, and drivable-area segmentation. The detection annotations provide bounding boxes for major road-object classes, while the lane and drivable-area annotations support road structure and navigability analysis. These characteristics make BDD100K an important benchmark for multi-task autonomous-driving perception.

The **Indian Driving Dataset (IDD)** [26] was developed to represent the complexity of real-world Indian road environments. It contains approximately 10,000 annotated images covering 34 semantic classes, collected from driving sequences in Hyderabad and Bengaluru. The dataset represents heterogeneous road users, dense traffic, irregular lane structures, varying road widths, and unconstrained driving conditions. Its annotations support object detection and semantic segmentation, providing useful information for road-scene understanding in Indian traffic environments. IDD therefore provides an

important India-specific benchmark for evaluating perception models under complex and heterogeneous driving conditions.

Overall, BDD100K and IDD provide valuable multi-task resources, with BDD100K offering large-scale and diverse driving scenes and IDD representing heterogeneous Indian road conditions, but it does not comprehensively cover traffic signals and road defects.

Research Gaps: The review indicates that existing autonomous-driving datasets generally provide task-specific or partially multi-task annotations, with no single benchmark comprehensively covering traffic objects, traffic signs, traffic lights, lanes, drivable areas, potholes, and pavement cracks. Furthermore, many widely used datasets do not adequately represent heterogeneous Indian traffic and irregular road environments, while adverse-weather datasets typically focus on specific environmental conditions. Road-defect datasets also primarily address pavement conditions rather than their integration with broader driving perception. In addition, variations in annotation schemes, environmental conditions, and task coverage make direct integration of multiple datasets challenging. These observations motivate the development of the Indian Traffic and Adverse Weather Dataset (ITAWD) as a unified benchmark for multi-task perception in complex and unstructured Indian driving environments.

2.4 Datasets for Autonomous Defence Systems

Autonomous defence systems require reliable visual perception for operating in complex and dynamic environments. Depending on the application, a defence perception system may need to identify military vehicles, personnel, aircraft, helicopters, civilian vehicles, and other objects while operating under varying viewpoints, illumination, occlusion, and background conditions. The availability of sufficiently diverse and well-annotated datasets is therefore important for developing and evaluating deep learning-based defence perception systems.

Several publicly available repositories provide military imagery that can be used as a source of visual data. The **Defense Visual Information Distribution Service (DVIDS)** [78] is a large repository of military photographs and videos collected across different operational, training, and public-affairs contexts. Its diverse visual content can support military object recognition and defence scene analysis. However, the available media are generally not supplied with standardized annotations for supervised deep learning, requiring additional labeling and preprocessing before model training.

Wikimedia Commons [79] provides another publicly accessible source of military-vehicle imagery from different countries, viewpoints, and operational contexts. Such images can be useful for increasing visual diversity and supporting military-vehicle classification. However, the content is collected from heterogeneous sources and consequently varies in image quality, viewpoint, background, and contextual information. Standardized annotations required for multi-task computer vision are also generally unavailable.

More structured datasets have been developed specifically for military vehicle perception. The **Defense Military Vehicle Detection and Classification (DMVDC)** dataset [80] contains 5,899 annotated images of military vehicles captured under different viewing conditions, including partially occluded targets. It provides a useful benchmark for military-vehicle detection and classification, but its category coverage and annotation scope remain focused on vehicle-level perception.

Defence perception also increasingly involves aerial and satellite imagery, where

Table 2.2: Comparison of datasets and repositories relevant to autonomous defence perception

Dataset / Repository	**Images**	**Classes**	**Annotation / Content**	**Primary Application**
DVIDS [78]	Large repository	Not standardized	Military photographs and videos from operational, training, and public-affairs contexts	Military object recognition and scene analysis
Wikimedia Commons [79]	Large collection	Not standardized	Heterogeneous military imagery with varied viewpoints, backgrounds, and image quality	Military-vehicle classification and visual diversity
DMVDC [80]	5,899	Vehicle-focused	Annotated military-vehicle images captured under different viewing conditions and partial occlusion	Military-vehicle detection and classification
xView [81]	Large-scale	60	∼1 million annotated instances in overhead imagery at about 0.3 m ground sampling resolution	Aerial object detection and surveillance
DOTA-v1.0 [82]	2,806	15	188,282 instances with oriented quadrilateral bounding-box annotations	Oriented aerial-object detection
DOTA-v2.0 [82]	11,268	18	More than 1.79 million annotated instances with oriented object annotations	Large-scale aerial and remote-sensing detection

objects may appear at small scales and arbitrary orientations. The **xView** dataset [81] is a large-scale overhead-imagery benchmark containing approximately one million annotated object instances across 60 classes at a ground sampling resolution of about 0.3 m. It includes diverse small, fine-grained, and multi-type objects and is useful for studying large-scale object detection in overhead scenes [81]. Although xView is not exclusively a military dataset, its overhead perspective and coverage of vehicles and other infrastructure make it relevant to defence and surveillance applications.

The **DOTA (Dataset for Object Detection in Aerial Images)** benchmark [82] is a large-scale resource for aerial-image object detection. DOTA-v1.0 comprises 2,806 high-resolution images with 188,282 annotated objects spanning 15 categories. Each object is represented using an oriented quadrilateral bounding box, allowing the representation of objects with different orientations and shapes. Later versions further increased the scale and category coverage, with DOTA-v2.0 containing 11,268 images, 18 categories, and more than 1.79 million annotated instances. These characteristics make DOTA particularly useful for evaluating object detection in aerial and remote-sensing scenarios.

Research Gap: Overall, existing defence-related datasets span a range of sources, including military image repositories, ground-level vehicle datasets, and aerial or satellite benchmarks. However, most are primarily designed for specific detection or classification objectives, while their annotation structures and object categories vary substantially. In particular, a common dataset integrating diverse tactical objects and supporting multi-class perception in ground-level operational environments is still limited. This motivates the development of the **Tactical Vehicle Dataset (TVD)** in this research to provide a dedicated data resource for multi-class defence perception.

CHAPTER 3

Development of Deep Learning-Based Traffic Perception Frameworks

---

This chapter presents the preliminary stage of the research, focusing on the development and evaluation of deep learning-based traffic perception frameworks. The work focuses on two complementary aspects of traffic perception. First, state-of-the-art YOLO-based object detection algorithms are comparatively evaluated to investigate their detection performance and computational efficiency for real-time traffic monitoring. The findings from this evaluation provide the basis for selecting an appropriate detection architecture for subsequent traffic perception applications.

Second, a cascaded YOLOv11-based framework is developed for automated helmet violation detection and automatic number plate recognition (ANPR) of motorcyclists. The proposed framework integrates motorcycle and rider detection, helmet classification, license plate localization, and optical character recognition (OCR) into a sequential perception pipeline. This enables the system to identify helmet violations and automatically extract vehicle registration information from traffic scenes.

The studies presented in this chapter establish the preliminary methodological foundation for the subsequent development of a unified multi-task driving perception framework. The findings from these studies provide insights into the selection and application of deep learning architectures for traffic perception in diverse and challenging road environments.

3.1 Evaluation of YOLO-Based Object Detection Models for Traffic Perception

The first stage of the research focused on investigating and evaluating deep learning-based object detection frameworks to enable efficient real-time traffic monitoring [35]. Accurate detection of vehicles is a fundamental requirement for traffic perception because subsequent tasks such as vehicle tracking, counting, traffic density estimation, and adaptive traffic signal management depend on reliable object localization and classification.

YOLO-based detectors were chosen for their single-stage architecture and real-time efficiency. They directly predict object location, class, and confidence through a unified neural network, avoiding separate region-proposal and classification stages.

In this research, four generations of YOLO, namely YOLOv1, YOLOv2, YOLOv3, and YOLOv4, were comparatively investigated under traffic monitoring conditions. The objective was to identify the evolution of detection capability across the architectures and

determine a suitable detector for subsequent multi-object tracking and traffic analysis.

3.1.1 Evolution of YOLO Architectures

The different YOLO versions introduced progressive improvements in feature extraction, localization accuracy, small-object detection, and computational efficiency, as shown in Table 3.1.

YOLOv1 established the basic single-stage object detection framework and provided high inference speed. However, it suffered from inaccurate object localization and relatively low recall, particularly for objects with complex spatial relationships. The architecture used a 24-layer convolutional network followed by fully connected layers.

YOLOv2 introduced several improvements, including batch normalization, a higher-resolution classifier, multi-scale training, and anchor boxes. These modifications improved the overall detection capability; however, the model continued to face difficulties in detecting small and closely positioned objects. YOLOv2 employed the Darknet-19 architecture.

YOLOv3 further improved the detection of objects at different scales by introducing predictions at three different feature-map resolutions. For a 416×416 input, predictions were generated at 13×13, 26×26, and 52×52 scales. This multi-scale representation improved the detection of smaller objects that could be difficult to identify using earlier YOLO versions. YOLOv3 used the Darknet-53 backbone with residual connections.

YOLOv4 incorporated a combination of training and architectural improvements, including the Bag of Freebies, Bag of Specials, and Spatial Pyramid Pooling. These improvements were intended to enhance detection accuracy while retaining high inference speed. The model employed the CSPDarknet53 backbone and demonstrated improved localization and real-time detection performance.

Table 3.1: Comparative characteristics of different YOLO versions.

YOLO Version	**Major Features**	**Major Limitation**	**Backbone**
YOLOv1	IoU, Non-Maximum Suppression, high inference speed	Inaccurate localization and lower recall	GoogleNet, 24-layer convolutional architecture
YOLOv2	Batch normalization, anchor boxes, multi-scale training	Difficulty with small and closely positioned objects	Darknet-19
YOLOv3	Multi-scale detection and improved feature connections	Increased model size and anchor-box dependency	Darknet-53
YOLOv4	Bag of Freebies, Bag of Specials, Spatial Pyramid Pooling	Higher computational complexity than earlier versions	CSPDarknet53

3.1.2 Traffic Vehicle Detection and Detection Refinement

The proposed traffic perception pipeline processes frames obtained from a surveillance video, with each frame independently supplied to the selected YOLO detector for vehicle detection. For each input image, the detector identifies the vehicles by estimating their spatial extent, corresponding categories, and prediction confidence. The overall detection pipeline can be broadly divided into feature extraction, feature refinement, and final prediction. Initially, the input frame is transformed by the backbone into multi-level feature representations containing progressively richer spatial and semantic information. These representations are subsequently refined through the intermediate feature-processing layers before being delivered to the detection head, where the bounding-box coordinates, object categories, and associated confidence values are estimated.

The vehicle categories considered in the traffic monitoring application include car, bus, truck, and motorbike. The detector may produce several candidate boxes for a single object. To eliminate redundant detections, Non-Maximum Suppression (NMS) is employed during post-processing. The method ranks the candidate boxes according to their confidence scores and preserves the most reliable prediction while discarding other boxes whose Intersection over Union (IoU) with the selected box is greater than the specified threshold. The Intersection over Union (IoU) between two bounding boxes, B_p and B_g, is defined as

$$IoU = \frac{|B_p \cap B_g|}{|B_p \cup B_g|}. \tag{3.1}$$

A higher IoU indicates greater spatial overlap between the two bounding boxes. IoU is used during NMS to identify highly overlapping predictions and is also an important measure for evaluating object localization accuracy against the corresponding ground-truth boxes.

3.1.3 Evaluation Metrics

The performance of the YOLO-based detectors was assessed using two primary metrics, namely mean Average Precision (mAP) and Frames Per Second (FPS). The first metric represents the detection capability of the model, while the second reflects its computational efficiency. These measures were selected to evaluate the trade-off between detection accuracy and real-time processing capability.

Mean Average Precision

Mean Average Precision (mAP) measures the overall object detection accuracy across the evaluated object classes. For each class, the confidence scores associated with the predicted bounding boxes are used to obtain the corresponding Precision–Recall relationship. The resulting Average Precision (AP) for each class is then averaged over all evaluated classes.

If C represents the total number of object classes, the mAP can be expressed as

$$mAP = \frac{1}{C} \sum_{c=1}^{C} AP_c, \tag{3.2}$$

where AP_c denotes the Average Precision obtained for the c-th class.

In this study, detection performance is evaluated using $mAP@50$, corresponding to an IoU threshold of 0.50. A detection is considered correct if the predicted class matches the

ground truth and the bounding box achieves an IoU of at least 0.50. The AP values across all classes are averaged to obtain the final $mAP@50$, with higher values indicating better detection performance.

Frames Per Second

Frames Per Second (FPS) quantifies the inference rate of the detector, representing the number of frames processed per second and calculated as

$$FPS = \frac{N}{T}, \tag{3.3}$$

where N is the number of processed frames and T is the total inference time in seconds.

A higher FPS indicates faster processing and is desirable for real-time traffic monitoring applications. In this study, FPS is considered together with mAP to assess whether improvements in detection accuracy are achieved without compromising the real-time processing capability of the detector.

3.1.4 Experimental Dataset and Traffic Video Processing

The MS COCO [45] dataset was used as a common benchmark for evaluating different YOLO versions. It provides annotated images covering various object categories, including traffic-related classes such as cars, buses, trucks, and motorcycles. A common dataset was maintained across all four YOLO implementations to ensure consistent experimental conditions and enable a reliable comparison of detection performance. For the traffic monitoring experiment, video input obtained from a surveillance camera was first processed to extract individual frames, which were subsequently supplied to the respective YOLO detectors for vehicle detection and localization. The detected vehicles were classified with corresponding confidence scores, after which Non-Maximum Suppression (NMS) was applied to eliminate overlapping predictions. Detection performance was assessed using mean Average Precision (mAP) and *Frames Per Second (FPS)*, where mAP denotes the overall detection performance, while FPS represents the inference speed of the detector.Precision and recall were also considered for evaluating the quality of the detected objects.

In addition, the detectors were assessed under different traffic conditions, particularly variations in vehicle density, illumination, and occlusion, to examine their suitability for practical traffic monitoring. Low- and high-density traffic were considered to evaluate the effect of closely positioned vehicles, while different illumination conditions were used to examine robustness to changes in image brightness. Partial or complete vehicle occlusion was also considered as it can affect accurate object localization in congested scenes. The refined detections were subsequently passed to the multi-object tracking stage. GPU resources were used for model training, validation, and testing, with experiments conducted on Google Colab.

3.1.5 Comparative Analysis of Results

The comparative performance of the four YOLO versions is summarized in Table 3.2. The results indicate a progressive improvement in detection performance across successive YOLO versions. YOLOv1 achieved an mAP of 63.50% at an inference rate of 45.00 FPS.

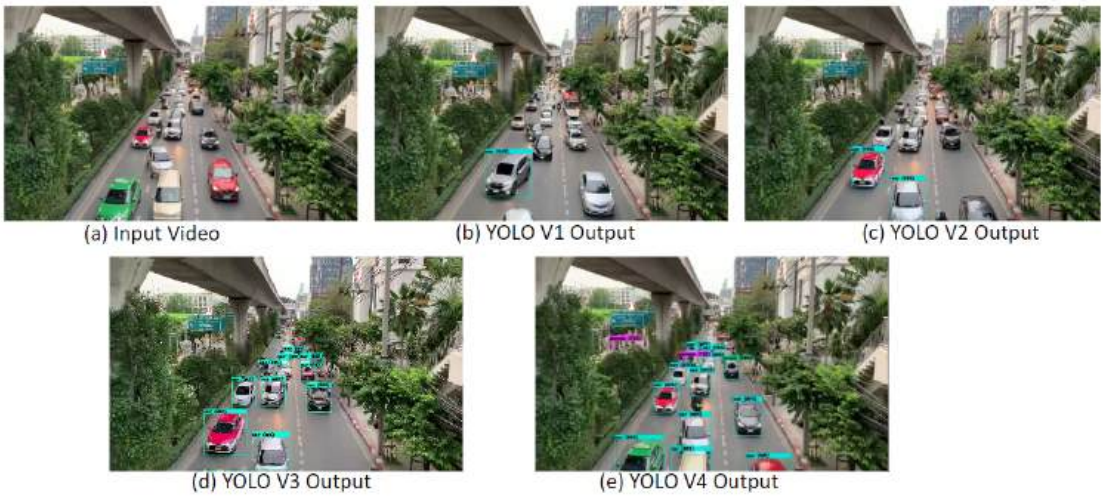

Figure 3.1: Visual comparison of vehicle detection results obtained using different YOLO versions.

The introduction of architectural and training improvements in YOLOv2 increased the mAP to 78.60%, while also improving the processing speed to 67.00 FPS.

YOLOv3 further improved the detection accuracy, achieving an mAP of 88.09%. However, its inference speed decreased to 51.26 FPS compared with YOLOv2. This indicates that the improvement in detection capability was accompanied by an increase in computational requirements. The results therefore demonstrate that detection accuracy and processing speed do not necessarily improve simultaneously with every architectural modification.

Table 3.2: Comparative Performance of YOLO Versions.

Model	**FPS**	**mAP (%)**
YOLOv1	45.00	63.50
YOLOv2	67.00	78.60
YOLOv3	51.26	88.09
YOLOv4	**90.00**	**89.88**

YOLOv4 provided the best overall performance among the four evaluated models. It achieved an mAP of 89.88% and an inference speed of 90.00 FPS. Compared with YOLOv1, YOLOv4 improved the mAP by 26.38 percentage points while increasing the processing speed from 45.00 FPS to 90.00 FPS. Compared with YOLOv3, the mAP improvement was 1.79 percentage points, while the inference speed increased by 38.74 FPS.

The results indicate that YOLOv4 provides the most favourable balance between detection accuracy and computational efficiency among the evaluated versions. Its combination of high mAP and high inference speed makes it particularly suitable for

real-time traffic monitoring. Therefore, the comparative evaluation establishes YOLOv4 as the preferred detector among the four versions considered in this experimental study.

3.1.6 Integration of YOLOv4 with Deep SORT

Although object detection provides the location and class of vehicles in individual frames, it does not inherently maintain the identity of an object across consecutive frames. For traffic monitoring applications, this temporal association is essential because the system must determine whether a vehicle detected in the current frame corresponds to a vehicle detected previously.

Therefore, Deep SORT [83] was integrated with YOLOv4 to perform multi-object tracking. The combination enables each detected vehicle to be assigned a persistent identity and followed throughout the video sequence. Deep SORT enhances object tracking by combining motion cues with appearance features to associate objects across consecutive frames.

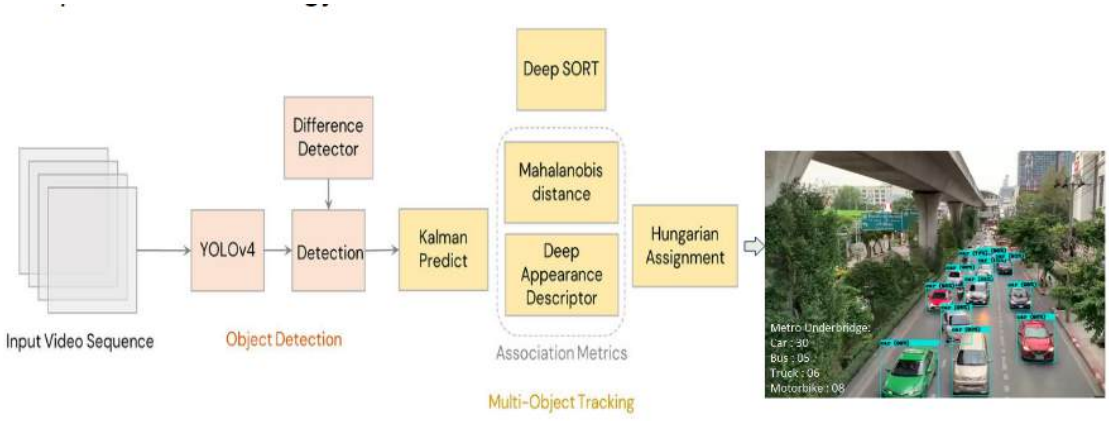

Figure 3.2: YOLOv4-based vehicle detection and Deep SORT multi-object tracking framework.

Kalman Filter-Based Motion Prediction

Deep SORT uses a Kalman filter to estimate the current state of a tracked vehicle based on its previous state and current observations. The filter maintains information such as the estimated position and velocity of the target and recursively updates the state as new detections become available.

The prediction stage is particularly useful when a vehicle is temporarily difficult to detect because of occlusion or changes in appearance. Instead of relying exclusively on the current detector output, the tracker can use the previously estimated motion of the object to predict its expected location.

Mahalanobis Distance

The association process uses the Mahalanobis distance to measure the compatibility between a predicted track and a newly detected object. The Mahalanobis distance takes the uncertainty of the Kalman filter prediction into account and therefore provides a more appropriate association measure than a simple Euclidean distance in this tracking context.

This enables detections to be associated with existing tracks according to both their predicted spatial position and the uncertainty associated with that prediction.

Deep Appearance Descriptor

Motion information alone may not be sufficient when several vehicles are close to each other. Deep SORT therefore incorporates deep appearance features extracted from the detected objects. These features provide additional information about the visual characteristics of each vehicle and help distinguish visually different vehicles occupying nearby regions of the image.

The combination of motion and appearance information improves the reliability of vehicle identity association across successive frames.

Hungarian Assignment

The Hungarian algorithm then determines the optimal assignment between existing tracks and newly detected objects based on the computed association metrics. The algorithm determines an appropriate correspondence between predicted tracks and current detections based on the calculated association costs.

3.1.7 Vehicle Tracking and Counting

The integration of YOLOv4 with DeepSORT [83] enables the detected vehicles to be continuously monitored across successive video frames. After YOLOv4 generates the vehicle detections, Deep SORT associates the detections with existing tracks using motion and appearance information. Each newly established track is assigned a unique identity, which is retained as long as the corresponding vehicle can be reliably associated with subsequent detections.

Figure 3.2 illustrates the output of the YOLOv4–Deep SORT framework for vehicle detection, tracking, and counting. The detected vehicles are enclosed by bounding boxes and associated with persistent tracking identities. The trajectories of the vehicles are maintained as they move through consecutive frames, allowing individual vehicles to be distinguished from newly entering vehicles. The resulting persistent identities are subsequently used for vehicle counting within the predefined counting region.

For vehicle counting, a virtual counting region is defined at a selected location in the traffic scene. The trajectory of each tracked vehicle is monitored with respect to this region. When a vehicle crosses the predefined counting boundary, its unique track identity is recorded and the vehicle count is incremented. Since the counting operation is based on the persistent identity generated by Deep SORT, multiple detections of the same vehicle across consecutive frames do not result in repeated counting.

The resulting vehicle count provides a frame-wise estimate of the traffic volume within the monitored region. This information can be used to characterize variations in traffic demand and identify changes in vehicle flow over time. Thus, the YOLOv4–Deep SORT framework extends frame-level object detection to vehicle-level tracking and counting, providing useful traffic information for subsequent traffic monitoring and management applications.

3.1.8 Summary of the Important Results

This chapter presented the development and evaluation of a deep learning-based framework for vehicle detection, tracking, and counting in traffic scenes. Four YOLO architectures, namely YOLOv1, YOLOv2, YOLOv3, and YOLOv4, were comparatively evaluated using mAP and FPS, considering variations in vehicle density, illumination, and occlusion. The major findings are summarized as follows:

1. **YOLOv4 yielded the highest overall detection performance among the evaluated models.** YOLOv1, YOLOv2, YOLOv3, and YOLOv4 achieved mAP values of 63.50%, 78.60%, 88.09%, and 89.88%, respectively. YOLOv4 also achieved the highest processing speed of 90.00 FPS, providing the most favourable accuracy–speed trade-off for real-time traffic monitoring.

2. **YOLOv4 was employed as the detector for multi-object tracking**, with its predicted object locations, class labels, and confidence values passed to the Deep SORT tracker. Deep SORT integrates Kalman-filter-based motion estimation, appearance features, and data association to preserve vehicle identities across consecutive frames.

3. **The integrated YOLOv4–Deep SORT framework enabled reliable vehicle counting.** Vehicle counts were obtained by monitoring persistent tracks as they crossed a predefined counting region, thereby reducing repeated counting of the same vehicle across successive video frames. The framework consequently provides useful vehicle-level information for traffic monitoring and management.

4. **The developed framework provides a foundation for subsequent road-safety perception tasks.** The combined detection, tracking, and counting capabilities support the progression toward specialized applications, leading to the subsequent development of a cascaded YOLOv11-based framework for helmet violation detection and automatic number plate recognition (ANPR).

3.2 Helmet Violation Detection and Automatic Number Plate Recognition

The next stage of the research [36] extends the traffic perception capabilities developed in the preceding section toward automated road-safety monitoring and vehicle identification. In this, a dedicated deep learning framework was developed to perform helmet violation detection and Automatic Number Plate Recognition (ANPR) of motorcyclists.

Helmet compliance is an important component of motorcycle safety because motorcyclists are particularly vulnerable to serious head injuries during crashes. Conventional enforcement generally relies on traffic personnel or manual examination of surveillance footage, which requires considerable human effort and becomes increasingly difficult as the volume of traffic and number of surveillance cameras increase. An automated vision-based approach can continuously examine traffic video and identify potential violations without requiring every frame to be manually inspected. Such a system can further associate a detected violation with the registration number of the relevant vehicle, thereby providing a more complete basis for subsequent verification and enforcement.

The developed framework addresses this requirement through a cascaded architecture based on YOLOv11. Instead of attempting to perform all perception tasks simultaneously on the complete traffic scene, the framework divides the problem into a sequence of specialized stages. These stages perform motorcycle and rider detection, helmet-status classification, license-plate localization, and Optical Character Recognition (OCR), respectively. The output generated by one stage is used to define the input region for the following stage. Consequently, the computationally more expensive license plate detection and OCR operations are restricted to motorcyclists identified as potentially violating the helmet requirement.

The problem is challenging because the relevant objects can occupy only a small portion of a surveillance frame. In addition, riders and license plates may be partially obscured by other vehicles or riders, while variations in illumination, viewing angle, traffic density, and image resolution can affect the reliability of detection and recognition. These challenges are further pronounced in crowded urban traffic, where multiple motorcycles may occur within the same scene and several riders may be present on a single motorcycle.

To address these conditions, YOLOv11 models were employed as task-specific detectors within the cascaded framework. The first stage identifies the motorcycle and associated rider regions. The rider regions are subsequently examined to determine helmet compliance. Riders classified as helmeted are removed from the subsequent processing path, whereas non-helmeted riders are retained for vehicle identification. The corresponding motorcycle region is then examined for license-plate localization, after which the detected plate region is processed by an OCR module to obtain the registration characters.

The developed framework therefore establishes a processing sequence from traffic-scene perception to safety-rule violation identification and vehicle recognition. This approach provides a more targeted solution than applying license-plate detection and OCR indiscriminately to every vehicle in a traffic scene. The following subsections describe the overall architecture, dataset preparation, individual processing stages, implementation, experimental evaluation, and limitations of the developed framework.

3.2.1 Overall Architecture of the Cascaded Framework

The proposed helmet violation detection and ANPR system follows a sequential cascaded architecture in which multiple YOLOv11-based detection models are integrated with an OCR module. The overall processing pipeline is illustrated in Fig. 3.3.

3.2.2 Dataset Preparation and Preprocessing

The **SafeRide: Helmet and License Plate Dataset (SHLPD)** [84] was prepared to support the sequential detection stages of the proposed helmet-violation and license plate recognition framework. The dataset contains images of motorcycles, riders, helmeted and non-helmeted riders, and vehicle registration plates captured under varied traffic and imaging conditions. The images were organized according to the specific detection requirements of each stage of the proposed cascade. Roboflow was employed for dataset management, annotation verification, preprocessing, augmentation, and generation of the final datasets used during model development. The preprocessing procedure included resizing the input images to the selected network resolution and applying image enhancement operations where required. Data augmentation was incorporated to increase the variability of the training samples and improve model generalization under changes in

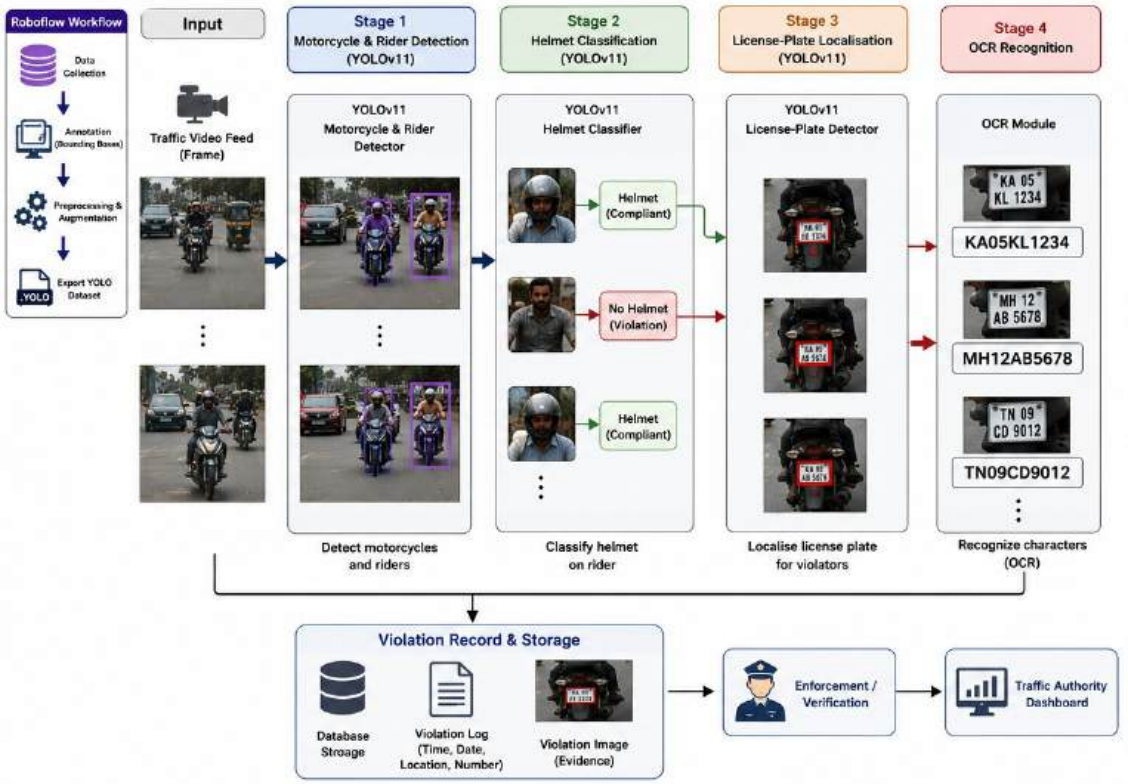

Figure 3.3: Proposed cascaded YOLO framework integrating helmet violation detection with ANPR.

viewpoint, illumination, scale, and traffic configuration.

The augmentation pipeline included operations such as image rotation, horizontal flipping, scaling, and brightness variation. These operations were selected to expose the models to a broader range of visual conditions while retaining the semantic characteristics of motorcycles, riders, helmets, and license plates. Following preprocessing and augmentation, the datasets were exported in a YOLO-compatible format and divided into training, validation, and testing subsets.

The resulting datasets were subsequently used to train the task-specific YOLOv11 models independently. This organization ensures that each model learns the visual characteristics relevant to its assigned task while the overall cascade combines their outputs during inference.

3.2.3 Motorcycle and Rider Detection

The first stage of the framework performs motorcycle and rider localization from the input traffic-surveillance frame. A YOLOv11-based object detector is used to identify the motorcycles and riders present in the scene and to generate their corresponding bounding boxes.

The purpose of this stage is to establish the candidate regions required by the subsequent helmet-compliance analysis. Processing the complete traffic image at this stage allows multiple motorcycles and riders to be identified within a single frame. The resulting rider regions are then extracted and forwarded to the helmet-classification stage.

The localization of both the motorcycle and rider is important because the relationship

between these objects provides contextual information for the subsequent stages. In dense traffic scenes, multiple motorcycles can occur close to one another, making independent analysis of the complete frame more challenging. Restricting subsequent processing to the detected candidate regions helps reduce the influence of unrelated objects and background regions.

3.2.4 Helmet Compliance Classification

The second stage determines the helmet status of the detected riders. The rider regions obtained from the preceding stage are supplied to a YOLOv11-based detector trained to distinguish between helmet-compliant and non-compliant cases.

Unlike full-frame helmet detection, the proposed approach analyzes localized rider regions. This reduces the spatial search area and enables the model to focus on the head and upper-body region of the rider. Each detected rider is therefore assigned a helmet-status category based on the output of the second-stage model.

Riders classified as helmet-compliant are excluded from further processing. For riders classified as non-helmeted, the corresponding vehicle information is retained and forwarded to the license-plate localization stage. This decision mechanism forms the principal filtering operation of the cascaded architecture and prevents unnecessary license-plate and OCR processing for compliant riders.

The helmet-classification stage is therefore not treated as an independent classification output alone; rather, its result controls the subsequent processing path of the complete system. This makes the framework suitable for violation-oriented surveillance, where computational resources are preferentially allocated to potentially non-compliant cases.

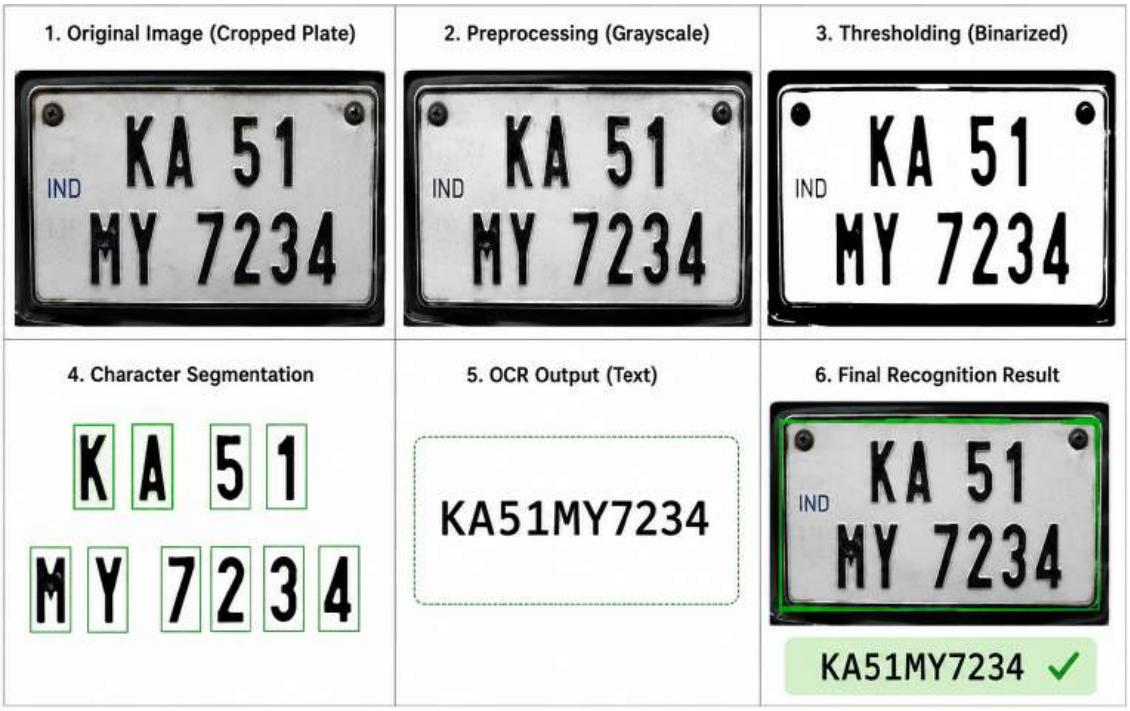

Figure 3.4: License-plate image after brightness enhancement and preprocessing

3.2.5 License-Plate Localization

The third stage is activated only for riders identified as non-helmeted. The corresponding motorcycle or vehicle region is processed using a dedicated YOLOv11-based license-plate detector. The objective of this stage, as illustrated in Fig. 3.4, is to localize the registration plate within the vehicle region and obtain a bounding box surrounding the plate.

License plates generally occupy a substantially smaller image area than the vehicle itself. Their appearance can also vary because of viewing distance, camera perspective, illumination, motion blur, and partial occlusion. Therefore, restricting the search to the relevant vehicle region provides a more focused input to the license-plate detector.

After detection, the coordinates of the license-plate bounding box are used to crop the plate from the corresponding image. The resulting plate image constitutes the input to the OCR stage. Thus, the output of the third stage provides the spatial localization required for subsequent character recognition.

3.2.6 Optical Character Recognition

The final stage, as shown in Fig. 3.4, converts the localized license-plate image into an alphanumeric registration string. The cropped plate image obtained from the previous stage is first subjected to suitable image preprocessing to improve character visibility. Depending on the quality of the input image, operations such as brightness adjustment, contrast enhancement, and noise reduction can be applied before character recognition.

The preprocessed plate image is then supplied to the OCR module, which identifies the characters appearing on the registration plate. The resulting registration string is associated with the corresponding helmet-violation event. Consequently, the final system output contains both the evidence of the detected violation and the recognized vehicle registration information.

The integration of OCR with the preceding detection stages transforms the system from a helmet-compliance detector into a vehicle-identification framework. This enables the detected violation to be associated with a specific registration number, subject to successful plate localization and character recognition.

3.2.7 Implementation and Experimental Setup

The proposed framework was implemented and experimentally assessed using Google Colaboratory, a cloud-based Python platform offering GPU-accelerated computing. The YOLOv11 framework was employed for model development, while Roboflow was used for organizing and annotating the SafeRide dataset. Model training, validation, and testing were conducted within the same experimental environment. Python-based tools, including Matplotlib and OpenCV, were utilized to generate performance plots, evaluation results, and qualitative visualizations.

For comparative evaluation, all models were trained and tested using a common experimental protocol. The same dataset configuration, preprocessing strategy, training settings, and evaluation criteria were maintained across the experiments to ensure consistency and enable meaningful performance comparisons.

The complete system follows a four-stage sequential processing pipeline comprising motorcycle and rider detection, helmet-status classification, license-plate localization, and OCR-based character recognition. The latter two stages are conditionally executed only when a rider is detected as not wearing a helmet. Consequently, license-plate detection

and character recognition are avoided for helmet-compliant riders, reducing redundant processing and improving the computational efficiency of the overall system.

3.2.8 Detection Results

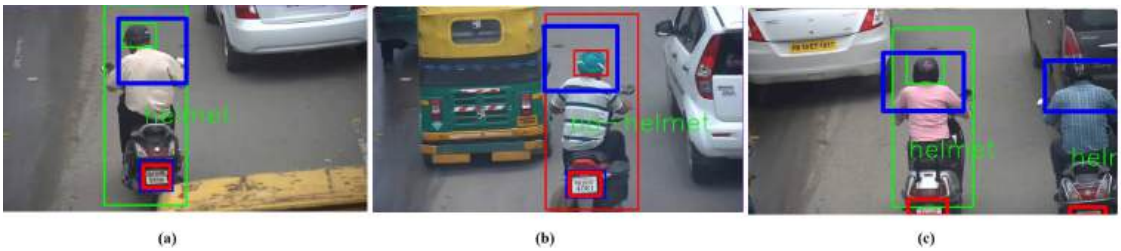

Figure 3.5: Representative detection results of the proposed framework under different traffic conditions.

Three classes—Motorcycle, Helmet, and Non-Helmet—were used for the evaluation. Fig. 3.5 shows representative detection outputs in which motorcycles and riders are successfully localized and classified. The YOLOv11-based framework demonstrates its ability to distinguish helmeted and non-helmeted riders across diverse traffic situations, including scenes with partial occlusion and multiple riders or vehicles.

The test examples cover a range of traffic densities, from relatively unobstructed roads to crowded urban scenes. In both conditions, the model reliably identifies individual motorcycles, riders, and helmet violations without significant confusion between neighboring vehicles. Helmeted riders are generally localized with well-defined bounding boxes, including cases where the rider is partially obscured by another person. In congested scenes, the detector continues to maintain separate detections for closely positioned motorcycles and riders rather than combining adjacent instances.

The comparatively lower recall than precision as shown in Table 3.3, indicates that the model tends to miss a limited number of actual helmet violations rather than generating a large number of false violation detections. Although this behavior reduces false alarms, improving recall remains important for practical enforcement applications. In particular, additional and more diverse training samples representing motorcycles carrying multiple riders could help the model better handle such relatively under-represented scenarios.

3.2.9 License Plate Recognition Results

For riders classified as helmet-rule violators, the corresponding license-plate area was extracted from the detection output and forwarded to the OCR stage. Before recognition, the cropped plate images underwent basic enhancement operations, including brightness correction and contrast adjustment, to make the alphanumeric characters more distinct. Fig. 3.4 illustrates a representative license-plate image following preprocessing. The OCR component was able to recover the vehicle registration number from the processed image, indicating the suitability of the proposed ANPR pipeline for automated traffic-monitoring and enforcement applications.

3.2.10 Performance Analysis

The quantitative performance of the individual components of the proposed cascaded framework is summarized in Table 3.3. Precision, recall, and mean Average Precision (mAP) were used to evaluate the object-detection stages, while recognition accuracy was considered for the OCR component.

Table 3.3: Performance of the Individual Stages of the Proposed Cascaded Framework.

Stage	**Task**	**Precision (%)**	**Recall (%)**	**Performance (%)**
Stage 1	Motorcycle and Rider Detection	96.2	94.8	96.5 (mAP)
Stage 2	Helmet/No-Helmet Classification	95.4	93.7	95.1 (mAP)
Stage 3	License-Plate Detection	94.8	92.9	94.5 (mAP)
Stage 4	OCR Recognition	–	–	96.3 (Accuracy)
Overall System	**Helmet Violation Detection and ANPR**	**95.1**	**93.3**	**94.7 (mAP)**

As shown in Table 3.3, the motorcycle and rider detection stage achieved a mAP of 96.5%, with precision and recall values of 96.2% and 94.8%, respectively. This indicates that the first stage provides reliable candidate regions for the subsequent helmet-compliance analysis.

The helmet classification stage obtained a mAP of 95.1%, with precision of 95.4% and recall of 93.7%. The results demonstrate the ability of the second-stage detector to distinguish helmet-compliant and non-compliant rider regions. The license-plate detection stage achieved a mAP of 94.5%, with precision and recall values of 94.8% and 92.9%, respectively. The slightly lower recall of this stage can be associated with the relatively small size of license plates and the presence of partial occlusions or image-quality variations.

The OCR component achieved an overall recognition accuracy of 96.3% for the evaluated license-plate samples. This result indicates that the localized and preprocessed plate regions provide sufficiently informative inputs for extracting the alphanumeric registration information.

Considering the integrated system, the proposed framework achieved an overall precision of 95.1%, recall of 93.3%, and mAP of 94.7%. These results demonstrate that the individual stages can be effectively combined to provide an end-to-end pipeline for helmet violation detection and vehicle registration recognition.

3.2.11 Comparitive Analysis

To examine the effectiveness of the proposed YOLOv11 architecture, its performance was compared with representative object-detection architectures, as shown in Table 3.4.

The proposed YOLOv11-based framework records the highest precision, recall, and mAP@50 among the listed methods, while also maintaining a high inference rate of 125 FPS.

Table 3.4: Performance Comparison of Representative Object-Detection Architectures.

Method	**Precision (%)**	**Recall (%)**	**mAP@50 (%)**	**FPS**
Faster R-CNN	91.8	90.5	91.2	18
YOLOv5	90.9	89.0	89.8	105
YOLOv7	92.1	90.4	91.3	110
YOLOv8	93.0	91.4	92.4	115
YOLOv9	93.8	92.0	93.1	119
YOLOv10	94.4	92.8	93.9	123
Proposed YOLOv11	**95.1**	**93.3**	**94.7**	**125**

The comparison indicates a progressive improvement in detection performance across the more recent YOLO architectures. The proposed YOLOv11 configuration achieves precision of 95.1%, recall of 93.3%, and mAP@50 of 94.7%. The corresponding inference speed of 125 FPS demonstrates that the selected architecture can support high-throughput traffic-frame processing.

The relatively high inference rate compared with the two-stage Faster R-CNN architecture further supports the selection of a YOLO-based detector for the proposed application. This is particularly relevant for traffic-surveillance systems, where timely processing of successive video frames is required.

3.2.12 Limitations

Although the developed framework demonstrates encouraging performance, some limitations remain. Performance can be affected by severe occlusion, very low illumination, motion blur, adverse weather, and low-resolution surveillance imagery. The recognition of license plates can be particularly sensitive to plate size, viewing angle, blur, and partial obstruction. Similarly, unusual rider configurations, such as motorcycles carrying multiple passengers, may be under-represented in the available training data and can therefore result in missed detections.

Another limitation arises from the sequential nature of the cascade. Although the selective processing mechanism reduces unnecessary computation in the later stages, an error in an earlier stage can propagate through the subsequent stages. For example, if a non-helmeted rider is not detected or is incorrectly classified as helmeted, the corresponding license plate will not be processed. Improving the robustness of the early detection and classification stages is therefore important for further increasing the overall system reliability.

Future evaluation on larger and more diverse real-world traffic datasets, including adverse weather, severe occlusion, low-light conditions, and different camera viewpoints,

would provide a more comprehensive assessment of deployment robustness. Integration with multi-object tracking and multi-camera surveillance could further improve the association between riders, motorcycles, detected violations, and registration numbers.

3.2.13 Summary of the Important Results

This section presented a cascaded YOLOv11-based framework for automated helmet-violation detection and ANPR. The pipeline performs rider detection, helmet-status assessment, license-plate localization, and Optical Character Recognition (OCR), with subsequent stages activated only for relevant detections.

1. Preparation and preprocessing of the **SafeRide Helmet and License Plate Dataset (SHLPD)** to support the cascaded helmet-violation detection and automatic number plate recognition framework.

2. **The detection stages achieved strong performance for rider and helmet-status identification.** Motorcycle and rider detection obtained 96.2% precision, 94.8% recall, and 96.5% mAP, while helmet-status detection achieved 95.4% precision, 93.7% recall, and 95.1% mAP.

3. **The ANPR module accurately localized and recognized vehicle registration information.** License-plate detection achieved 94.8% precision, 92.9% recall, and 94.5% mAP, with OCR attaining 96.3% recognition accuracy.

4. **The integrated cascade provided effective and selective processing.** The complete framework achieved 95.1% precision, 93.3% recall, and 94.7% mAP, with plate detection and OCR applied only to riders classified as non-helmeted, thereby avoiding unnecessary processing for compliant riders.

5. **The framework extends traffic perception toward application-specific road-safety monitoring.** Building upon the preceding YOLOv4–Deep SORT framework for detection, tracking, and counting, it establishes the basis for **SmartDriveNet**, which advances toward unified multi-task perception of traffic objects, signs, lights, potholes, cracks, lanes, and drivable areas in unstructured driving environments.

CHAPTER 4

Development of a Unified Multi-Task Deep Learning Framework for Driving Perception

---

*This chapter presents **SmartDriveNet**, a unified multi-task deep learning framework for robust driving perception in unstructured environments and adverse weather conditions. It describes the development of the **Indian Traffic and Adverse Weather Dataset (ITAWD)** and the proposed architecture for seven perception tasks: traffic-object, traffic-sign, traffic-light, pothole, and crack detection, as well as lane-line and drivable-area segmentation. The chapter further covers training strategy, experimental evaluation, comparative analysis, and ablation studies.*

4.1 SmartDriveNet-Based Multi-Task Panoptic Driving Perception

A reliable understanding of the surrounding road scene is essential for autonomous vehicles and ADAS to operate safely and efficiently. In practical driving scenarios, particularly on unstructured roads, the perception module must simultaneously recognize diverse road users, interpret traffic lights and signs, identify lane structures, determine the drivable region, and detect road-surface abnormalities such as potholes and cracks. Reliable execution of these tasks is essential for supporting safe navigation, decision-making, and vehicle control.

Recent advances in deep learning have enabled effective solutions for a broad range of autonomous driving perception tasks, including traffic sign recognition [48], traffic light detection [5, 6], pothole and pavement crack detection [7, 8, 9], lane detection [10, 11], and drivable-area segmentation [12, 13]. Traffic-object detection has also been extensively studied using established two-stage and one-stage architectures, including Faster R-CNN [14], SSD [15], and successive YOLO-based detectors [16, 17, 18, 19]. Despite the effectiveness of these task-specific approaches, conventional perception systems generally employ separate networks for individual tasks, resulting in increased computational and memory requirements. Since each network independently extracts features from the same input image, redundant computation can occur, while sequential or parallel execution of multiple models may further increase inference latency. Moreover, independently trained models have limited ability to exploit complementary information across perception tasks. For example, object-level information can assist traffic sign and traffic light interpretation, whereas lane structures and road geometry can provide useful contextual

cues for estimating the drivable region. These limitations motivate the development of unified multi-task perception architectures that can share visual representations while simultaneously addressing multiple driving-perception tasks. Multi-task learning provides an effective alternative by enabling multiple perception tasks to be learned within a unified deep learning framework. A shared feature-extraction network can reduce redundant computation and improve computational efficiency, while task-specific prediction branches can preserve the characteristics required by individual tasks. However, the development of a robust multi-task perception system remains challenging because different tasks have different output representations, spatial characteristics, and optimization requirements. In addition, real-world driving scenes contain substantial variations in object scale, traffic density, illumination, weather, occlusion, road geometry, and surface conditions. These variations can introduce task interference and adversely affect the robustness of a unified perception network.

To address these challenges, this research develops **SmartDriveNet** [37], a unified camera-based multi-task panoptic driving perception framework for complex and unstructured driving environments. The proposed system integrates seven complementary perception capabilities within a single architecture, namely traffic-object, traffic-sign, traffic-light, pothole, and crack detection, as well as lane-line and drivable-area segmentation. By jointly learning these complementary tasks, SmartDriveNet aims to develop a richer understanding of the surrounding driving environment while reducing the computational redundancy associated with independently deployed task-specific networks.

The overall development of SmartDriveNet is therefore guided by four principal requirements: (i) unified perception of multiple complementary driving tasks, (ii) efficient feature extraction with reduced computational redundancy, (iii) robust representation learning under varying traffic, illumination, and environmental conditions, and (iv) availability of a diverse dataset representative of Indian road scenarios. These requirements collectively motivate the proposed SmartDriveNet framework, which seeks to provide comprehensive, efficient, and robust visual scene understanding for autonomous driving and ADAS applications, particularly in unstructured and challenging road environments.

The key contributions of the proposed SmartDriveNet framework can be outlined as follows:

1. A cost-effective, camera-based seven-task panoptic driving perception framework, termed SmartDriveNet, is developed to jointly perform traffic-object, traffic-sign, traffic- light, pothole, and crack detection, as well as lane-line and drivable-area segmentation within a unified architecture. The traffic-object recognition task encompasses a diverse range of road users and objects, including cars, buses, trucks, motorcycles, rickshaws, pedestrians, and animals.

2. The network incorporates **GELAN** for efficient feature extraction and feature aggregation and integrates **Programmable Gradient Information (PGI)** to improve gradient propagation and feature learning during multi-task optimization.

3. The **Density Adaptive Attention Mechanism (DAAM)** is integrated to improve the network's ability to adapt to variations in object density and scene complexity, thereby strengthening feature representation under diverse driving conditions.

4. A carefully designed **multi-task loss formulation and training strategy** is employed to balance the optimization of the seven perception objectives and reduce the possibility of one task dominating the learning process.

5. The **Indian Traffic and Adverse Weather Dataset (ITAWD)** [38] is developed, comprising **24,236 images across 47 classes**. The dataset encompasses a broad range of traffic situations, road users, transportation infrastructure, roadway surface conditions, and adverse environmental scenarios through the integration of newly acquired real-world data with relevant existing datasets.

4.2 Methodology and System Development

The development of SmartDriveNet involves a sequence of interconnected stages, beginning with visual data acquisition and progressing through dataset development, data preparation, network design, and model optimization. The overall methodology is designed to establish a camera-based perception system capable of simultaneously addressing multiple driving-perception tasks under diverse road and environmental conditions. An overview of the complete development pipeline is illustrated in Fig. 4.1.

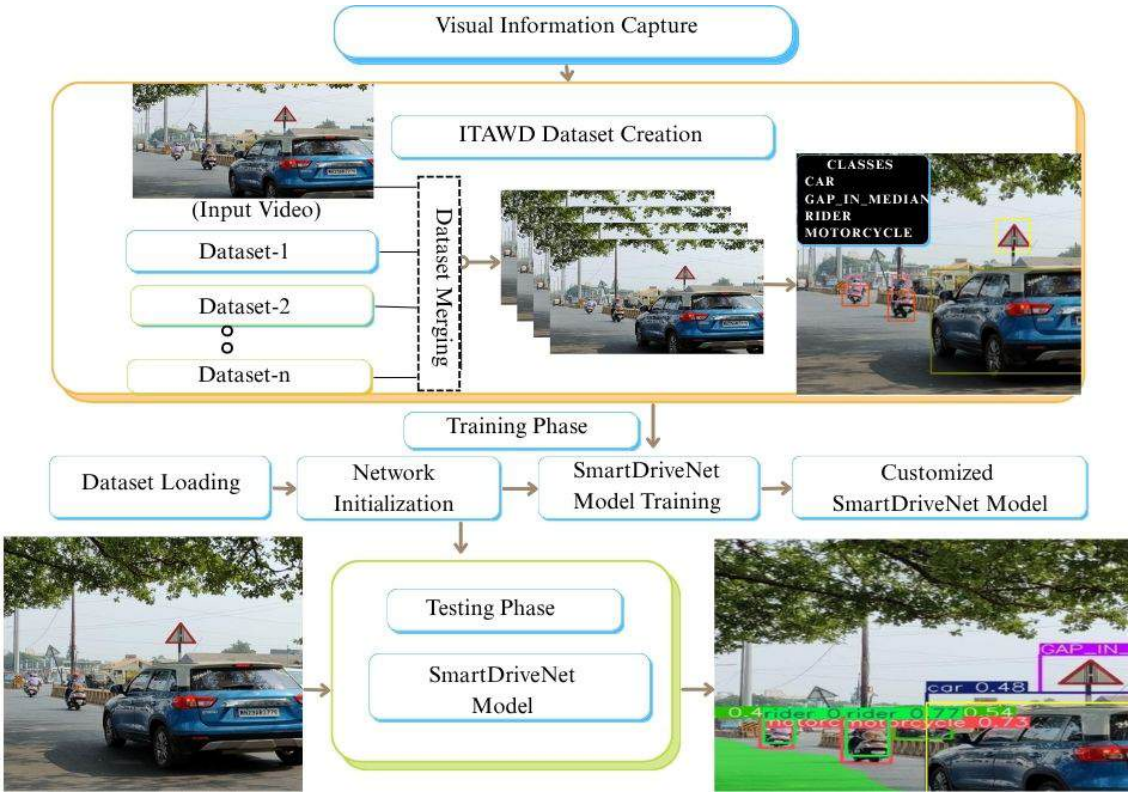

Figure 4.1: Block Diagram of SmartDriveNet Panoptic Perception System

The proposed methodology consists of four major stages: (i) visual information acquisition, (ii) development and preparation of the Indian Traffic and Adverse Weather Dataset (ITAWD), (iii) design of the SmartDriveNet architecture and formulation of the multi-task learning objective, and (iv) training, testing, and evaluation of the developed network. The training stage incorporates balanced loss formulation, optimization techniques, and learning-rate scheduling to ensure effective learning across multiple tasks and stable convergence.

4.3 Development of the Indian Traffic and Adverse Weather Dataset

The development of a representative dataset is a fundamental requirement for a unified multi-task perception system. The review presented in Chapter 2 showed that existing public datasets differ substantially in terms of geographic coverage, perception categories, and annotation structure, and environmental conditions. In particular, no single dataset provides the complete combination of traffic objects, traffic signs, traffic lights, road-surface defects, lane structures, and drivable regions required by the proposed seven-task framework. Furthermore, many existing benchmarks are developed primarily from structured road environments and therefore provide limited representation of the heterogeneous traffic conditions encountered on Indian roads.

To address these limitations, the **Indian Traffic and Adverse Weather Dataset (ITAWD)** [38] was developed as part of this research. The dataset contains 24,236 images distributed across 47 classes. It represents diverse traffic participants, road-surface conditions, and environmental variations relevant to the proposed SmartDriveNet framework. Beyond variations in object appearance and scale, the dataset captures diverse traffic densities, illumination conditions, road geometries, and adverse weather scenarios, including fog, snow, dust storms, and heavy rain.

The dataset was constructed using two complementary sources: (i) real-world Indian driving scenes collected specifically for this research and (ii) carefully selected publicly available datasets. Combining these sources enabled the dataset to capture a broader range of traffic scenarios, environmental conditions, and perception categories than could be obtained from either source alone.

4.3.1 Real-World Data Collection

Real-world driving data were collected using a camera mounted inside a vehicle during approximately **30 km of driving sequences in Mumbai**. The collected routes included urban streets, highways, flyovers, and other heterogeneous road environments. This acquisition strategy was selected to obtain visual samples representative of the conditions in which the proposed perception system is intended to operate.

The recordings included both daytime and nighttime scenes, thereby introducing variations in illumination and visibility. The collected sequences also represented different traffic densities, vehicle types, road geometries, and road-surface conditions. Particular attention was given to challenging scenarios, including wet roads and water-affected areas, as well as scenes containing road-surface irregularities of different sizes and appearances.

During dataset preparation, the captured frames were inspected for visual quality. Images affected by excessive glare, severe motion blur, or inadequate visual information were excluded where appropriate. This filtering helped maintain the quality of the dataset while preserving the natural variability of real-world driving scenes.

The real-world component provides an important foundation for ITAWD. It was subsequently complemented with selected public datasets to increase the diversity of traffic signs, traffic lights, adverse-weather conditions, and road-surface defects.

4.3.2 Selection and Integration of Existing Datasets

As established through the dataset review presented in Chapter 2, no single publicly available dataset provides the complete set of perception categories and environmental conditions required by the proposed seven-task framework. Therefore, a set of complementary datasets was selected for ITAWD development based on their relevance and their ability to address specific gaps in the real-world data collected in this research.

Five existing datasets were selected for integration: the Indian Traffic Sign Dataset [49], Detection in Adverse Weather Nature (DAWN) [73], Indian Driving Dataset (IDD) [26], Road Damage Detection Dataset (RDD) [27], and Bosch Small Traffic Lights Dataset (BSTLD) [52]. Their selection was guided primarily by task and environmental complementarity rather than dataset size alone. The Indian Traffic Sign Dataset and IDD enhance the representation of Indian traffic signs, road users, and road environments, while DAWN contributes additional variability under adverse weather conditions. RDD provides samples of potholes and diverse pavement-crack patterns, whereas BSTLD supplements traffic-light samples, particularly for small and distant objects. Thus, these datasets complement the real-world Mumbai driving data by expanding its task and environmental diversity. Their respective contributions to the development of ITAWD are summarized in Table 4.1.

Table 4.1: Datasets selected for the development of ITAWD

Dataset	**Contribution to ITAWD**	**Key Characteristics**
Indian Traffic Sign Dataset [49]	Traffic-sign recognition	6,480 images with 7,056 annotated traffic signs across 87 categories, providing India-specific traffic-sign characteristics.
IDD [26]	Traffic-object and road-scene perception	The dataset contains approximately 25.8 GB of data, with 10,004 annotated images across 34 categories, representing diverse and congested Indian urban traffic conditions in cities like Hyderabad and Bengaluru.
DAWN [73]	Adverse-weather perception	1,000 real-world traffic images captured under fog, rain, snow, and sandstorm conditions, providing challenging visibility scenarios.
RDD [27]	Pothole and crack detection	26,336 road images containing over 31,000 annotated damage instances, including potholes and various pavement-crack types, collected from multiple countries, including India.
BSTLD [52]	Traffic-light recognition	Annotated traffic-light images representing multiple signal categories, with emphasis on small and distant traffic-light instances.



Figure 4.2: Comparison of image annotations: (a) Original image, (b) Annotation from the Indian Traffic Sign (ITS) dataset, and (c) Re-annotated image in the Indian Traffic and Adverse Weather Dataset (ITAWD).

4.3.3 Dataset Re-Annotation, Label Harmonization, and Class Organization

The external datasets selected for ITAWD development were originally created for different perception objectives and consequently employed different class definitions, annotation formats, and labeling policies. Although these datasets provide complementary information, their original annotations were generally restricted to the primary objectives for which the datasets were developed. Consequently, relevant objects present in an image were left unannotated. For example, an ITS image originally annotated only for traffic signs was re-examined and additional relevant road objects were annotated according to the unified ITAWD class structure. Fig. 4.2 illustrates the original and re-annotated versions. To obtain a consistent multi-task training dataset, the external datasets were systematically re-annotated and harmonized before their integration into ITAWD.

The re-annotation process was carried out using the following steps:

1. **Dataset Upload:** Pre-annotated images and their corresponding labels were imported into the annotation environment for inspection and further processing.

2. **Annotation Format Conversion:** The original annotations were converted into a common format compatible with the object-detection framework used in this research. Where required, annotations were converted using the available label-conversion utilities or manually recreated.

3. **Annotation Refinement:** Existing bounding boxes were examined and adjusted wherever necessary to improve localization accuracy. Additional relevant objects that were present but absent from the original annotations were identified and labeled. Incorrect, incomplete, or inconsistent labels were also corrected.

4. **Label Standardization:** Class names, class identifiers, and annotation conventions were unified across the different source datasets. This ensured that visually similar categories were represented consistently within the final ITAWD class structure.

5. **Final Dataset Export:** After verification and harmonization, the re-annotated samples were exported in the required format for integration, training, and evaluation of the proposed SmartDriveNet framework.

Following re-annotation and label harmonization, the external datasets were integrated with the real-world Mumbai driving data. The resulting ITAWD was organized into a unified class structure supporting the seven perception capabilities of SmartDriveNet. The final dataset contains **24,236 images across 47 classes**, providing a diverse combination of Indian traffic characteristics, road-surface conditions, and environmental variations for training and evaluating the proposed framework.

4.3.4 Data Preprocessing, Augmentation, and Dataset Partitioning

Prior to model development, all images were resized to 640×640 **pixels**, providing a suitable balance between preserving details of small objects, such as traffic signs and lights, and maintaining computational efficiency for real-time inference. The preprocessing stage involved image resizing, annotation conversion, class-label normalization, and visual-quality checking, with severely corrupted or visually inadequate images excluded.

To enhance the generalization of SmartDriveNet, training samples were augmented through scaling, horizontal flipping, rotation, and mosaic augmentation [85]. In addition, **targeted data augmentation** using value-guided adaptive techniques was employed to improve the representation of under-represented classes [86], thereby avoiding unnecessary augmentation of already dominant categories.

The resulting dataset of 24,236 images was divided into **training (87.49%)**, **validation (8.34%)**, and **test (4.17%)** subsets. The training set was used for network parameter optimization, while the validation set supported convergence monitoring and model configuration during development. The independent test set was retained for the final performance evaluation. The split was designed to preserve representative coverage of the major traffic and environmental conditions present in the dataset.

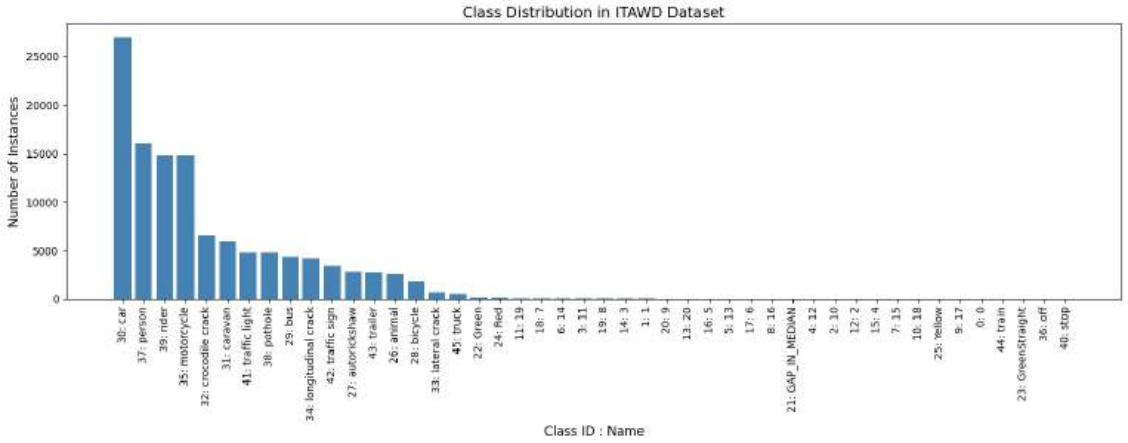

Figure 4.3: Class imbalance in the ITAWD dataset.

Analysis of Class Distribution

An assessment of the ITAWD dataset was performed to examine the distribution of instances among its 47 object categories. The analysis reveals a substantial variation in class frequency. Common categories such as *car*, *person*, and *motorcycle* contain 27,058, 16,030, and 14,776 instances, respectively. In contrast, several semantically important categories have very few samples, including *stop* with 3 instances, *off* with 5 instances, and *train* with 7 instances. The overall class-wise distribution is illustrated in Figure 4.3.

The considerable difference in sample availability can affect model optimization, as classes with a larger number of instances may contribute more strongly to the learning process than sparsely represented categories. Consequently, rare classes may be more difficult to recognize reliably, particularly under varying traffic and environmental conditions. To reduce the influence of this imbalance, several complementary strategies were incorporated into the training process:

- **Focal Loss** [23] was employed to increase the contribution of difficult training examples and reduce the dominance of easily classified samples.
- **Dynamic loss weighting**, motivated by the GradNorm approach [87], was used to adapt the relative contribution of different learning objectives during optimization.
- **Targeted data augmentation** [86] was applied preferentially to underrepresented categories. This strategy increases the diversity of samples available for infrequent classes while avoiding excessive augmentation of already well-represented categories.

The class distribution analysis therefore provides important insight into the composition of ITAWD and supports the use of imbalance-aware training strategies. These measures were adopted to improve the model's ability to learn both frequently occurring and sparsely represented categories, thereby promoting more consistent detection performance across diverse driving environments.

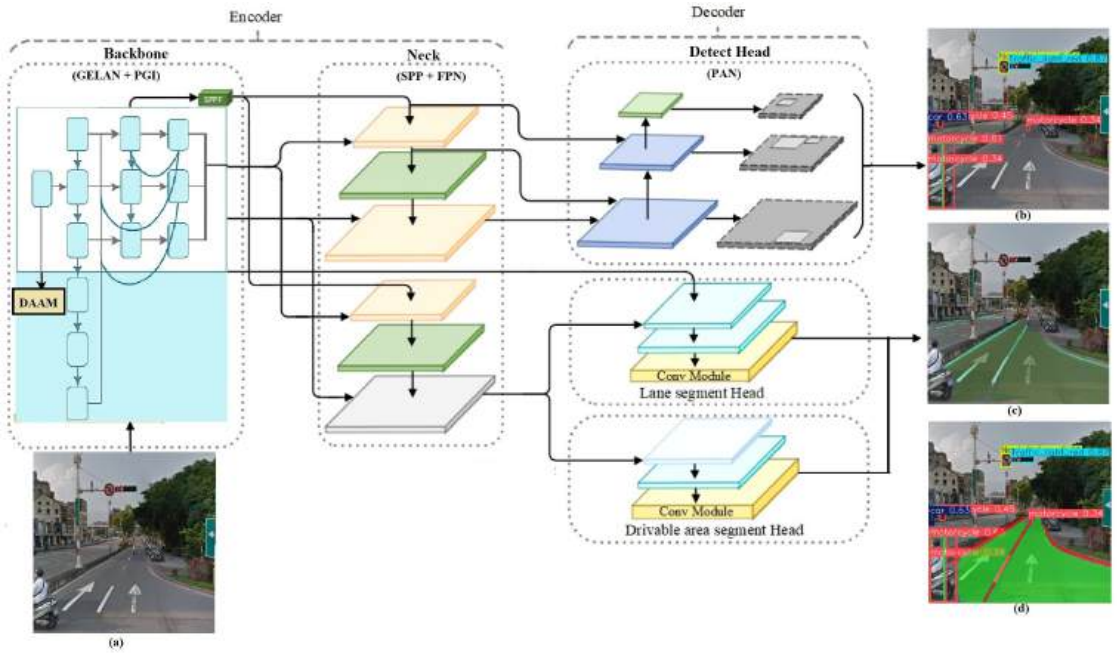

Figure 4.4: SmartDriveNet: A Comprehensive Framework for Unified Visual Perception

4.4 Proposed SmartDriveNet Architecture

SmartDriveNet is developed as a unified multi-task panoptic driving perception network in which a common feature-extraction pipeline is shared across multiple perception objectives.

The design is motivated by the multi-task architectures, particularly YOLOPv2 [23], while introducing modifications in the backbone, attention mechanism, and task-decoder organization to address the requirements of heterogeneous and unstructured Indian road environments.

As shown in Fig. 4.4, the network receives a camera image as input and processes it through a shared feature-extraction stage followed by task-specific prediction branches. The architecture contains three principal processing paths: a detection head, a lane-segmentation head, and a drivable-area segmentation head. The detection branch jointly handles traffic objects, traffic signs, traffic lights, potholes, and cracks, whereas the two segmentation branches independently estimate lane structures and the drivable region.

Unlike architectures that use a common segmentation decoder for lane and drivable-area prediction, SmartDriveNet employs separate decoders for these two tasks. This separation is motivated by their different spatial and contextual requirements. Lane detection depends strongly on fine-grained spatial information and thin structural features, whereas drivable-area segmentation benefits from broader contextual and region-level information. The use of independent heads therefore allows each task to utilize features at an appropriate level of abstraction. The principal architectural components of SmartDriveNet are presented in the following subsections.

4.4.1 Shared Encoder and Feature Extraction

The shared encoder is responsible for extracting hierarchical visual representations that are subsequently utilized by all perception heads. SmartDriveNet employs the **GELAN** as the principal feature-extraction structure, together with **PGI**, following the design principles introduced in YOLOv9 [18].

GELAN enables information from different computational paths to be efficiently aggregated while maintaining a suitable balance between feature representation and computational cost. The use of grouped convolution within the backbone further promotes diverse feature learning by allowing different groups of channels to process complementary information.

PGI is incorporated to improve the flow and utilization of gradient information during network optimization. This is particularly useful in the proposed multi-task setting, where the shared encoder receives learning signals from several heterogeneous objectives. The combination of GELAN and PGI therefore provides the common feature representation required by the detection and segmentation branches while limiting unnecessary computational complexity.

Following the backbone, multi-level features are progressively combined within the network neck. The **Spatial Pyramid Pooling (SPP)** module enhances contextual representation by aggregating information over different receptive-field sizes, while the **Feature Pyramid Network (FPN)** combines features with different levels of semantic information. The resulting multi-scale feature representation is important for the proposed application because the input scenes contain objects with substantially different spatial dimensions, ranging from small traffic lights and signs to large vehicles and road regions.

4.4.2 Task-Specific Decoder Heads

The shared features are delivered to three independent decoder paths designed according to the requirements of the corresponding perception tasks. This arrangement reduces

representational interference while retaining the computational advantages of a shared encoder.

The first branch is the **detection head**, which jointly performs traffic-object, traffic-sign, traffic-light, pothole, and crack detection. A multi-scale detection strategy is employed to improve the localization of objects with different sizes. The detection pathway combines FPN and **Path Aggregation Network (PAN)** features so that both high-level semantic information and lower-level spatial details can contribute to object prediction.

The second branch is dedicated to **lane detection**. Lane structures are generally thin and spatially continuous, requiring preservation of fine image details. Therefore, the lane decoder receives features from the later stages of the FPN where richer semantic information is available while retaining sufficient spatial resolution for identifying lane boundaries.

The third branch performs **drivable-area segmentation**. In contrast to lane detection, this task requires region-level understanding of the road surface and surrounding scene. The drivable area decoder is therefore connected to an earlier feature stage, followed by an upsampling operation. This design preserves spatial information while avoiding unnecessary deep processing that may not provide additional benefit for broad-area segmentation.

The separation of the lane and drivable-area decoders allows the network to learn task-specific representations instead of forcing both segmentation objectives to use an identical feature pathway. At the same time, the common encoder enables all three branches to benefit from shared information concerning road geometry, objects, and scene context.

4.4.3 Density Adaptive Attention Mechanism

Driving scenes can exhibit substantial variation in the number, distribution, and scale of visible objects. A relatively open road may contain only a few objects, whereas a dense urban scene may contain multiple vehicles, pedestrians, motorcycles, traffic signs, and other road users within a limited image region. To account for these variations, SmartDriveNet incorporates the **Density Adaptive Attention Mechanism (DAAM)** [88].

DAAM is positioned after the stem layer and before the shared encoder. Consequently, the attention mechanism operates on early feature representations rather than directly on the raw image. This placement allows informative spatial and channel-level characteristics to be refined before the features undergo deeper hierarchical processing.

The mechanism aggregates information across multiple feature scales and generates adaptive attention responses for the input scene. Informative regions receive greater emphasis, while less useful responses are suppressed. Such feature refinement is beneficial for the proposed multi-task framework because different driving elements may occur at different scales and densities within the same image.

DAAM also employs lightweight operations to maintain computational efficiency. This characteristic is important for real-time ADAS applications, where improvements in feature representation must be achieved without imposing excessive computational overhead. The detailed algorithmic procedure of DAAM, including its feature processing and attention computation steps, is provided in Appendix A. The principal functions of DAAM in SmartDriveNet can be summarized as follows:

- **Multi-scale feature aggregation:** combines information from different feature scales to represent both local details and broader scene context.

- **Adaptive feature weighting:** assigns greater importance to informative spatial and channel features according to the characteristics of the input scene.
- **Lightweight attention computation:** improves feature discrimination while maintaining the computational efficiency required for real-time perception.
- **Shared feature refinement:** provides refined representations to the subsequent task-specific branches, allowing the detection and segmentation tasks to benefit from common contextual information.
- **Robust scene representation:** improves feature adaptation to variations in object density, occlusion, illumination, and environmental conditions.

Thus, DAAM serves as an early feature-refinement stage that complements the hierarchical representation learned by the GELAN-based encoder.

4.4.4 Architectural Rationale

The overall architecture of SmartDriveNet is designed around the principle of sharing information where tasks are complementary while maintaining task-specific processing where their requirements differ. The shared encoder reduces repeated feature extraction and therefore limits the computational redundancy associated with independent networks. In contrast, the independent decoder heads allow each perception objective to learn features appropriate to its output representation.

The combination of GELAN and PGI provides an efficient shared representation, while DAAM performs adaptive refinement of early features. SPP, FPN, and PAN contribute multi-scale contextual and spatial information, and the three task-specific heads transform these shared representations into the required detection and segmentation outputs.

This organization provides a unified framework for seven perception capabilities while maintaining a clear separation between detection, lane estimation, and drivable-area segmentation. The contribution of the individual architectural components is subsequently examined through comparative experiments and ablation analysis.

4.4.5 Multi-Task Loss Function

Since SmartDriveNet simultaneously optimizes detection and segmentation objectives, an appropriate loss formulation is required to prevent one group of tasks from dominating the training process. The overall training objective combines the detection and segmentation components, similar to YOLOPv2 [23], as

$$L_{\text{total}} = \alpha L_{\text{det}} + \beta L_{\text{seg}}, \tag{4.1}$$

where L_{det} denotes the combined detection loss, L_{seg} represents the segmentation loss, and α and β control their relative contributions.

The detection objective includes the losses associated with classification, localization, and confidence prediction. Focal Loss is additionally used to increase the contribution of difficult examples, which is particularly useful for under-represented and hard-to-detect categories in ITAWD.

For drivable-area segmentation, **Cross-Entropy loss** is employed for pixel-level classification. Lane segmentation presents a different challenge because lane markings

occupy a relatively small portion of the image compared with the background. Therefore, **Focal Loss** is adopted for the lane-segmentation objective to place greater emphasis on difficult lane pixels.

The relative contribution of detection and segmentation is controlled through dynamic loss weighting. The coefficients are adjusted according to the learning behaviour of the corresponding objectives, reducing the risk that a rapidly converging or numerically dominant task suppresses the learning of other tasks. This strategy follows the general principles of adaptive multi-task optimization, including GradNorm [87].

For the initial training configuration, the loss coefficients were set to $\alpha = 1.0$ and $\beta = 0.8$. These values were selected through validation-based experimentation to obtain an appropriate balance between detection accuracy and segmentation quality. A sensitivity analysis was also performed around these values to verify that the training process remained stable under moderate changes in the loss weights.

4.5 Experimental Setup and Performance Evaluation

The experimental study was conducted on an NVIDIA Tesla V100 GPU using PyTorch 1.12. SmartDriveNet was trained for 350 epochs with a batch size of 20 using the Adam optimizer. The initial learning rate was set to 0.01 and was controlled through a cosine annealing schedule. A three-epoch warm-up period was included at the beginning of training to support stable optimization during the initial learning phase. All images were resized to 640×640 pixels before being supplied to the network. The augmentation and class-imbalance handling procedures described in the preceding sections were also incorporated during training.

The validation subset was used to monitor convergence and guide model selection, whereas the independent ITAWD test subset was reserved exclusively for the final evaluation. This separation ensured that the test data did not influence model development or configuration selection.

To investigate the influence of decoder architecture, three SmartDriveNet configurations were examined. The first configuration employed a shared decoder and was trained and evaluated directly on ITAWD. The second configuration used separate decoders and was initialized using BDD100K [24], followed by fine-tuning on ITAWD. The third configuration also incorporated separate decoders but was trained entirely from scratch using ITAWD. This experimental design enabled the effects of decoder structure and training strategy to be studied under comparable conditions.

For comparative analysis, SmartDriveNet was evaluated against the YOLOPv2 [23] and HybridNets [22] frameworks. Both their original models and corresponding fine-tuned versions were included in the comparison. The original models provide a measure of direct cross-dataset generalization, whereas the fine-tuned variants allow the benefit of adapting the established architectures to the ITAWD domain to be examined.

Model performance was quantified using established detection and segmentation measures. Precision, recall, and mAP@50 were considered for detection-related outputs, while IoU, mIoU, and pixel accuracy were used for segmentation assessment where applicable. In addition to accuracy-related measures, inference speed was recorded to evaluate the practical suitability of the models for real-time deployment. Thus, the experimental analysis considers both predictive effectiveness and computational efficiency.

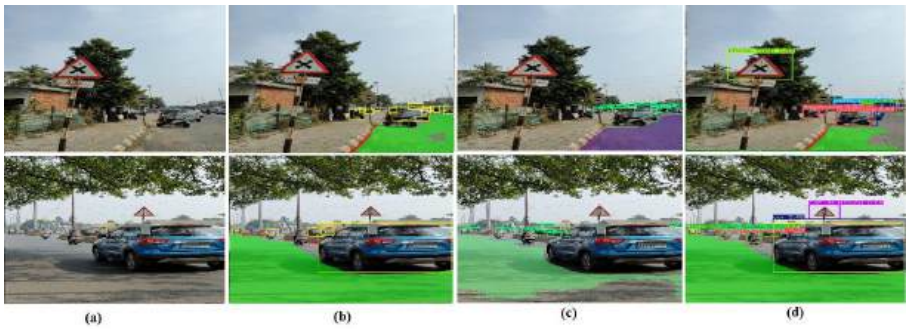

Figure 4.5: Qualitative comparison of traffic perception performance for (a) input road scenes, (b) YOLOPv2, (c) fine-tuned HybridNets, and (d) SmartDriveNet.

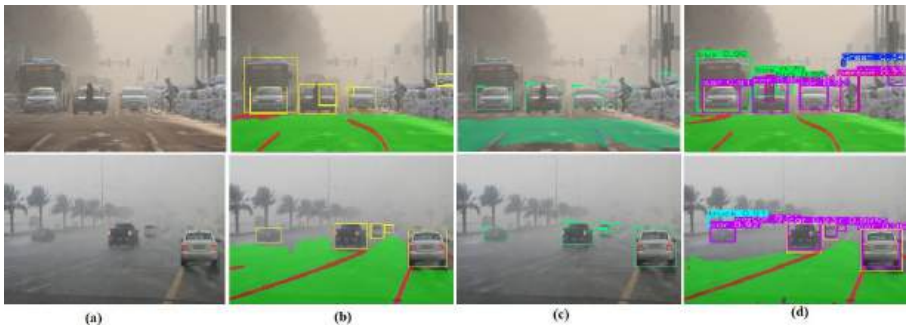

Figure 4.6: Qualitative evaluation of traffic perception under adverse weather conditions, including snow, fog, and rain: (a) input images; (b) fine-tuned YOLOPv2; (c) fine-tuned HybridNets; and (d) SmartDriveNet.

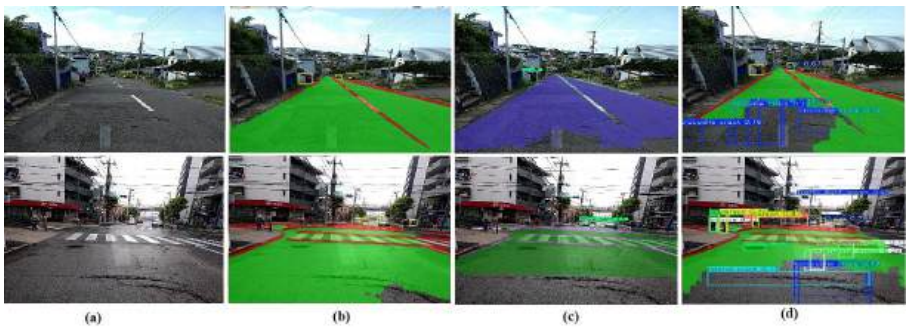

Figure 4.7: Qualitative comparison of detection and segmentation results: (a) input road scenes containing potholes, cracks, traffic objects, and traffic lights; (b) predictions from the fine-tuned YOLOPv2; (c) predictions from the fine-tuned HybridNets; and (d) predictions obtained using the proposed SmartDriveNet.

Table 4.2: Comparative performance of the evaluated multi-task perception models on the ITAWD dataset [37]

Model	Speed (FPS)	Traffic Object Detection Results		Drivable Area Segmentation Results	Lane Segmentation Results	
		mAP@50 (%)	Recall (%)	Drivable mIoU (%)	Accuracy (%)	Lane IoU (%)
HybridNets [22]	28	77.3	92.8	90.5	85.4	31.60
Fine-Tune HybridNets	28	76.8	92.1	88.3	82.2	29.9
YOLOPv2 [23]	91	83.4	91.1	93.2	87.31	27.25
Fine-Tune YOLOPv2	91	82.7	89.9	92.7	86.9	27.1
Custom SmartDriveNet (Shared Decoders)	91	83.2	92.4	92.6	86.4	28.0
Fine-Tuned SmartDriveNet (Separate Decoders)	91	83.7	92.9	93.7	87.5	32.3
Custom SmartDriveNet (Separate Decoders)	91	**84.5**	**93.1**	**93.9**	**87.9**	**33.1**

4.5.1 Performance Assessment of Traffic Object Detection

The evaluation of traffic object detection performance relies on key metrics such as mAP@50 (mean Average Precision at 50% Intersection over Union) and recall. The mAP@50 metric assesses the model's ability to accurately detect traffic objects based on a predefined overlap threshold, while recall measures the model's effectiveness in capturing all true positive instances. A balanced evaluation of detection performance is achieved by considering both metrics.

SmartDriveNet achieves an mAP@50 of 84.5% and a recall of 93.1% for traffic-object detection. The performance improvement can be associated with the complementary roles of GELAN, PGI, and DAAM. GELAN improves the aggregation of features from different network paths, strengthening multi-scale representation and consequently helping the model detect objects with substantially different sizes. This is particularly useful for identifying both nearby vehicles and small or distant road users.

PGI contributes to the stability and quality of gradient propagation during optimization. Its effect is particularly relevant in the multi-task setting, where the shared encoder receives learning signals from several heterogeneous objectives. Improved gradient information helps preserve useful representations during training and supports higher recall for partially visible or difficult objects.

4.5.2 Performance Evaluation for Drivable-Area and Lane-Line Segmentation

Drivable area segmentation performance is assessed using the mean Intersection over Union (mIoU) metric, which quantifies the overlap between predicted segmentation masks and ground truth annotations. For segmentation, SmartDriveNet obtains a drivable-area mIoU of 93.9%. A higher mIoU value indicates a more precise segmentation of drivable regions. The multi-scale representation generated through GELAN supports the identification of both broad road regions and local boundaries, while DAAM enhances responses around informative road regions and boundaries. DAAM further refines the encoded representation by assigning greater attention to informative spatial regions. This is beneficial in crowded Indian traffic scenes, where numerous objects and background structures occur simultaneously. Consequently, relevant object features can be emphasized while less informative responses are suppressed. PGI supports stable optimization of the

shared representation, which is important when detection and segmentation objectives are trained jointly.

Lane line segmentation is evaluated using two key metrics: accuracy and lane IoU. Accuracy measures the overall correctness of lane predictions, while lane IoU assesses the overlap between predicted lane regions and ground truth, providing a more fine-grained evaluation of lane detection precision. The lane branch achieves a lane IoU of 33.1% and an accuracy of 87.9%. GELAN contributes to the extraction of fine and multi-scale features required for thin and elongated lane structures, while DAAM helps emphasize lane-related spatial patterns in the presence of vehicles, road clutter, and other visual distractions. The use of a dedicated decoder further prevents the lane task from competing directly with the broader contextual requirements of drivable-area segmentation.

Table 4.3: Model Comparison for Traffic Sign Recognition across Various Classes.

Class	Faster R-CNN		Mask R-CNN		YOLOv5		YOLOv8		SmartDriveNet	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
Bus-Stop	85.0%	88.0%	88.0%	90.0%	89.0%	91.0%	93.0%	92.0%	**98.0%**	**95.0%**
Do-Not-Turn	86.0%	87.0%	87.0%	89.0%	88.0%	90.0%	91.0%	93.0%	**96.0%**	**94.0%**
Do-Not-Turn-Left	87.0%	88.0%	88.0%	90.0%	89.0%	91.0%	93.0%	94.0%	**97.0%**	**96.0%**
Do-Not-Turn-Right	85.0%	86.0%	86.0%	88.0%	87.0%	89.0%	92.0%	90.0%	**96.0%**	**95.0%**
Enter-Left-Lane	90.0%	92.0%	91.0%	93.0%	92.0%	95.0%	94.0%	96.0%	**97.0%**	**96.0%**
Gap-In-Median	85.0%	86.0%	86.0%	88.0%	87.0%	89.0%	92.0%	90.0%	**95.0%**	**93.0%**
No-Parking	86.0%	87.0%	87.0%	89.0%	88.0%	91.0%	93.0%	92.0%	**95.0%**	**94.0%**
Parking	85.0%	86.0%	87.0%	88.0%	88.0%	90.0%	92.0%	91.0%	**96.0%**	**95.0%**
T-Intersection	88.0%	89.0%	90.0%	91.0%	89.0%	92.0%	94.0%	93.0%	**98.0%**	**96.0%**
U-Turn	85.0%	86.0%	87.0%	89.0%	86.0%	88.0%	91.0%	90.0%	**95.0%**	**94.0%**
Cross-Roads	86.0%	87.0%	88.0%	89.0%	87.0%	89.0%	93.0%	91.0%	**96.0%**	**94.0%**
Pedestrian-Crossing	84.0%	85.0%	85.0%	87.0%	86.0%	88.0%	91.0%	89.0%	**95.0%**	**90.0%**

4.5.3 Performance Assessment of Traffic Sign Recognition

Traffic-sign recognition was evaluated on the ITAWD test set using precision and recall. Table 4.3 compares SmartDriveNet with Faster R-CNN, Mask R-CNN, YOLOv5, and YOLOv8 across the evaluated traffic-sign categories.

SmartDriveNet achieves an overall precision of 96.09% and recall of 94.64%. GELAN contributes to this performance by providing multi-scale feature representations, which are important because traffic signs can appear at substantially different sizes and distances. This improves the representation of small and distant signs.

PGI supports stable gradient propagation during the joint training process, helping preserve useful feature information for difficult sign instances. DAAM complements these features by emphasizing spatial regions containing informative sign characteristics. This is particularly useful when signs are partially occluded or embedded in visually complex traffic scenes.

The class-wise comparison is presented in Table 4.3, while Figure 4.8 provides a visual comparison of the average performance.

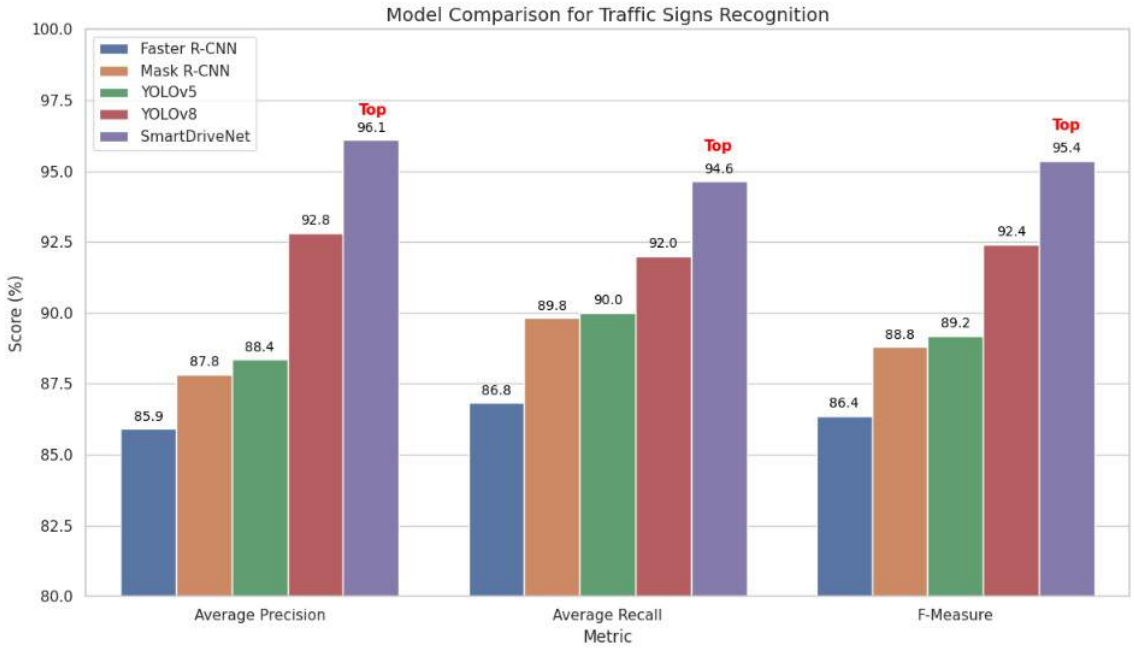

Figure 4.8: Performance Comparison of Traffic Sign Models Based on performance Metrics

4.5.4 Performance Assessment of Pothole and Crack Detection

SmartDriveNet was evaluated for pothole and pavement-crack detection using the ITAWD test set. The results are compared with SSD, Faster R-CNN, YOLOv5, and YOLOv8 in Table 4.4.

For crack detection, SmartDriveNet achieves an average mAP of 85.00%, compared with 82.00% for YOLOv8 and 80.21% for YOLOv5. The model obtains 82.00%, 78.10%, and 80.20% for Type I, Type II, and Type III cracks, respectively. GELAN contributes to

Table 4.4: Comparative Performance Analysis for Crack and Pothole Detection.

	Average mAP	**Crack Detection (IoU=50%)**			**Pothole Detection (%)**		
Model		**Type I**	**Type II**	**Type III**	**Precision**	**Recall**	**mAP@50**
SSD	41.44%	53.50%	17.80%	55.30%	43.86%	38.43%	40.54%
Faster R-CNN	53.06%	71.60%	45.30%	70.80%	59.23%	51.91%	54.75%
YOLOv5	80.21%	80.10%	76.23%	76.85%	86.88%	76.16%	80.29%
YOLOv8	82.00%	81.00%	76.30%	77.10%	90.99%	79.71%	84.07%
SmartDriveNet	**85.00%**	**82.00%**	**78.10%**	**80.20%**	**95.38%**	**83.56%**	**88.17%**

crack detection by aggregating features across different scales and computational paths, which helps represent both fine crack structures and larger pavement patterns. PGI supports the learning of these fine-grained features by maintaining effective gradient propagation during optimization. DAAM further emphasizes informative pavement regions, helping distinguish crack structures from visually similar road-surface patterns.

For pothole detection, SmartDriveNet achieves a precision of 95.38%, recall of 83.56%, and mAP@50 of 88.17%. DAAM is particularly useful for this task because potholes can occupy irregular regions and may appear under varying illumination, occlusion, and traffic density. By adaptively emphasizing informative spatial regions, DAAM supports more effective pothole localization. GELAN contributes through multi-scale representation, while PGI supports stable learning of the associated detection features.

The results demonstrate that the combination of multi-scale feature extraction, improved gradient propagation, and adaptive spatial refinement is beneficial for road-surface defect detection.

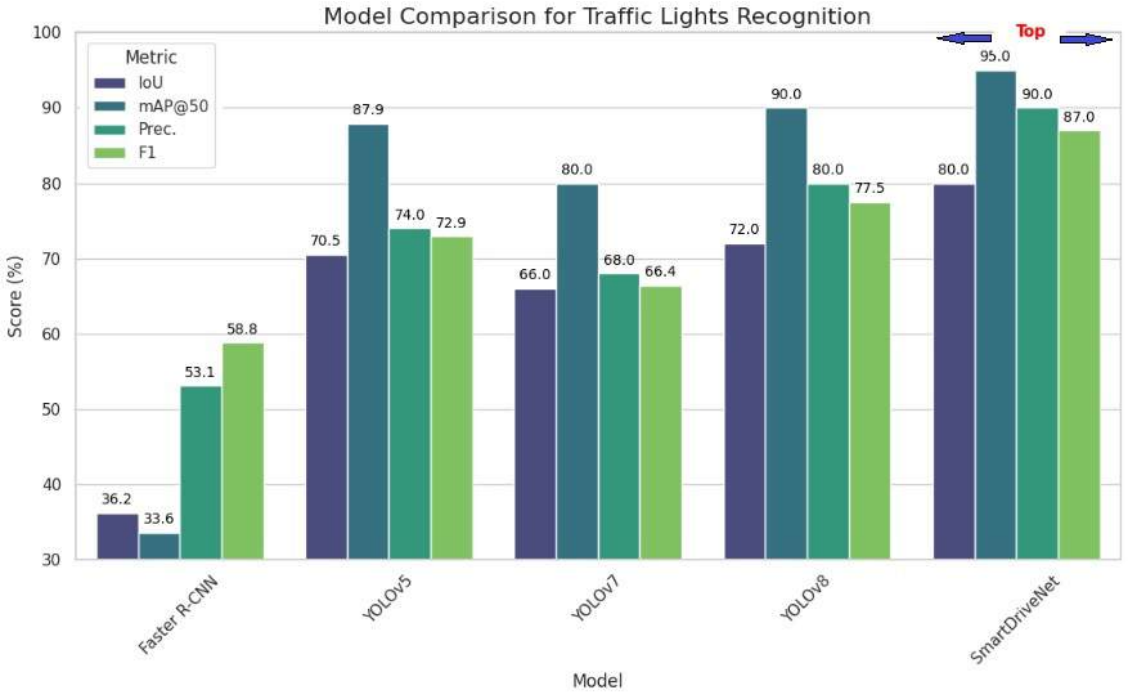

Figure 4.9: Performance-Based Comparison of Traffic Light Models.

4.5.5 Performance Assessment of Traffic-Light Recognition

Traffic-light recognition was evaluated using average IoU, mAP@50, precision, recall, and F1-score. SmartDriveNet achieves an average IoU of 80.0%, mAP@50 of 95.0%, precision of 90.0%, recall of 85.0%, and an F1-score of 87.0%, as shown in Figure 4.9.

The strong traffic-light performance is partly attributed to GELAN, which improves multi-scale representation of small and distant objects. This is important because traffic lights may occupy only a small portion of the input image.

PGI improves the reliability of feature learning during training by supporting effective gradient propagation across the shared architecture. DAAM then enhances the representation by assigning greater attention to relevant regions containing traffic-light features. This is useful under conditions involving glare, low illumination, partial occlusion, or complex backgrounds.

The combined effect of these components allows SmartDriveNet to maintain reliable traffic-light detection while simultaneously supporting the other perception tasks.

4.5.6 Ablation Analysis of Shared and Separate Decoders

An experimental analysis was performed to determine how the decoder design influences the multi-task performance of SmartDriveNet. Two architectural variants were considered, namely, a shared-decoder design and a separate-decoder design. To ensure a fair comparison, both configurations were trained and tested on the ITAWD dataset using identical experimental conditions.

The results reported in Table 4.2 show that the shared-decoder variant obtains 83.2% mAP@50 and 92.4% recall for traffic-object detection, together with 92.6% drivable-area mIoU and 28.0% lane IoU. When individual decoders are assigned to the segmentation tasks, the performance increases to 84.5% mAP@50, 93.1% recall, 93.9% drivable-area mIoU, and 33.1% lane IoU, respectively.

The notable gain in lane IoU suggests that task-specific decoders can better preserve the feature characteristics required for lane and drivable-area segmentation. Despite this architectural modification, the inference throughput remains unchanged at 91 FPS. This indicates that the use of independent decoders improves the overall perception performance while retaining the real-time processing capability of the network.

Based on these observations, the separate-decoder configuration was selected for the final SmartDriveNet architecture. The results further indicate that limiting feature sharing between the segmentation branches can provide more task-specific representations and improve the effectiveness of the overall multi-task perception framework.

4.5.7 Limitations and Future Scope

Although SmartDriveNet provides effective multi-task perception for autonomous driving, some limitations remain. The model can be less reliable when traffic signals exhibit unusual colors, such as brown, that differ from the standard red, yellow, and green representations. Performance may also decline under severe environmental or road conditions. For example, heavily snow-covered scenes can affect the detection of faded lane markings, while shadows and wet road surfaces under low illumination may occasionally be confused with potholes, leading to false detections.

These limitations indicate several directions for future development. The training data can be expanded with more samples containing non-standard signal appearances, severe weather, degraded road surfaces, and low-light conditions. Condition-aware feature adaptation and targeted augmentation could further improve robustness in such challenging scenarios. The framework can also be extended to additional perception tasks, including vehicle trajectory prediction and pedestrian intention estimation, to provide richer information for higher-level driving decisions.

For practical deployment on resource-constrained edge platforms, techniques such as pruning, quantization, and other model-compression methods can be investigated to reduce computational and memory requirements. Finally, evaluation across a wider range of geographic, weather, traffic, and road conditions would be valuable for assessing the generalization, scalability, and reliability of SmartDriveNet in real-world autonomous driving applications.

4.5.8 Summary of the Important Results

This chapter presented the development and experimental evaluation of **SmartDriveNet**, a unified multi-task deep learning framework for driving perception in heterogeneous and

unstructured road environments. The framework integrates seven perception capabilities, namely traffic-object detection, traffic-sign recognition, traffic-light recognition, pothole detection, crack detection, lane detection, and drivable-area segmentation within a single network. To support the development of the proposed framework, the **Indian Traffic and Adverse Weather Dataset (ITAWD)** was developed by combining real-world Indian driving data with selected publicly available datasets. The major findings obtained in this chapter are summarized as follows:

1. The **ITAWD dataset** was developed to provide a diverse perception benchmark for Indian road environments. The final dataset contains **24,236 images across 47 classes**, incorporating traffic objects, traffic signs, traffic lights, road-surface defects, and scene-level driving information under varying illumination, traffic density, road conditions, and weather conditions.

2. The proposed **SmartDriveNet** integrates seven perception tasks within a unified architecture, thereby reducing the need for independently deployed task-specific networks. The framework uses a shared feature-extraction network together with task-specific prediction branches for detection, lane segmentation, and drivable-area segmentation.

3. The proposed architecture incorporates **GELAN, PGI, and DAAM** to improve feature extraction, gradient propagation, and adaptive feature refinement, respectively. Multi-scale feature processing using **SPP, FPN, and PAN** further supports the representation of objects and structures with substantially different spatial characteristics.

4. The final SmartDriveNet configuration achieves **84.5% mAP@50** and **93.1% recall** for traffic-object detection. For drivable-area segmentation, the model obtains a **93.9% mIoU**, while lane detection achieves **33.1% IoU** and **87.9% accuracy**. Comparative evaluation against HybridNets and YOLOPv2, along with their fine-tuned variants, demonstrates the effectiveness of the proposed SmartDriveNet architecture for the heterogeneous perception requirements considered in this research. The results particularly demonstrate the advantage of the proposed framework in jointly handling road-surface defects, traffic infrastructure, and conventional traffic-object perception.

5. For **traffic-sign recognition**, SmartDriveNet achieves an overall **96.09% precision** and **94.64% recall** across the evaluated traffic-sign categories. The results indicate effective recognition of signs appearing at different scales and within visually complex road scenes.

6. For **traffic-light recognition**, the proposed model achieves a **95.0% mAP@50**, together with an average IoU of **80.0%**, precision of **90.0%**, recall of **85.0%**, and F1-score of **87.0%**.

7. For **road-surface defect detection**, SmartDriveNet achieves an average **85.00% mAP** for crack detection. The model obtains **82.00%**, **78.10%**, and **80.20%** mAP for Type I, Type II, and Type III cracks, respectively. For pothole detection, the model achieves **95.38% precision**, **83.56% recall**, and **88.17% mAP@50**.

8. The **decoder ablation study** demonstrates the benefit of using separate decoders for lane and drivable-area segmentation. Compared with the shared-decoder configuration, the separate decoder configuration improves traffic-object mAP@50 from **83.2% to 84.5%**, recall from **92.4% to 93.1%**, drivable-area mIoU from **92.6% to 93.9%**, and lane IoU from **28.0% to 33.1%**.

9. The separate-decoder configuration maintains an inference speed of approximately **91 FPS**, indicating that the improvement in task-specific segmentation performance is achieved without reducing the reported real-time processing capability.

Overall, the results establish that SmartDriveNet can provide comprehensive multi-task driving perception within a single real-time-capable framework. The combination of a shared feature extractor, multi-scale feature processing, adaptive feature refinement, and separate task-specific decoders enables the network to address diverse perception requirements while maintaining a reported inference speed of 91 FPS. The developed ITAWD dataset further provides an India-specific foundation for evaluating such unified perception systems under heterogeneous and challenging road conditions.

CHAPTER 5

Development of a Unified Multi-Task Deep Learning Framework for Autonomous Defence Vehicle Perception

---

This chapter presents the development of the proposed DefenceVisionNet, a unified multi-task deep learning framework for autonomous defence vehicle perception. It describes the development of the Tactical Vehicle Dataset (TVD), the proposed network architecture, and the methodology adopted for jointly performing defence–civilian entity detection, drivable terrain segmentation, and trail/boundary detection within a unified framework. The chapter further presents the experimental setup, quantitative and qualitative performance evaluation, comparison with state-of-the-art methods, and ablation studies to demonstrate the effectiveness of the proposed framework under challenging defence environments. The chapter further presents the training methodology, experimental evaluation, comparative assessment, and ablation analysis of the proposed frameworks.

5.1 DefenceVisionNet-Based Multi-Task Perception for Autonomous Defence Vehicles

Autonomous defence vehicles require reliable perception capabilities to operate safely in environments characterized by unstructured terrain, limited infrastructure, and adverse weather conditions. Unlike conventional road environments, defence scenarios may contain poorly defined paths, irregular terrain, changing surroundings, and limited navigational infrastructure. In such conditions, the perception system must distinguish defence vehicles and defence personnel from civilian vehicles and persons, identify regions suitable for vehicle traversal, and determine trails or terrain boundaries that can assist navigation. Accurate perception of these elements is essential for supporting autonomous mobility, surveillance, reconnaissance, and situational awareness. However, reliable recognition becomes challenging when relevant targets are partially occluded, visually similar, camouflaged, or observed under poor illumination and adverse environmental conditions.

Traditional vision-based perception approaches typically rely on task-specific deep learning architectures, with each perception requirement handled independently. For defence-oriented applications, dedicated models have been developed for detecting military vehicles [89], while separate deep learning methods have been proposed for the detection and classification of military personnel [90]. In addition to entity detection, other perception requirements, such as drivable terrain segmentation [60],[61] and trail or

boundary detection [68],[69], [11], are commonly treated as separate tasks. Although task-specific networks can achieve effective performance for their respective objectives, deploying multiple models increases the overall computational and memory requirements of the perception system and may introduce additional inference latency. Since each model independently processes the same camera input and performs feature extraction, considerable computational redundancy can arise. Moreover, independently trained models have limited ability to exploit complementary information across perception tasks. For example, information regarding the location and characteristics of defence vehicles and personnel can provide useful contextual cues for terrain interpretation, while drivable terrain and boundary information can contribute to a more comprehensive understanding of the operational scene.

Multi-task learning provides an effective strategy for addressing these limitations by enabling multiple perception objectives to be learned jointly within a common architecture. A shared feature extractor can learn reusable representations from the input image, reducing repeated computation and improving resource utilization, while task-specific branches can retain the characteristics required for individual prediction objectives. However, the direct application of existing multi-task perception approaches to defence environments remains challenging. Many existing frameworks have been developed primarily for structured urban driving scenarios, where road layouts, lane markings, traffic infrastructure, and environmental conditions are relatively well defined. Defence environments, in contrast, can exhibit substantial variations in terrain, target appearance, illumination, weather, viewpoint, object scale, occlusion, clutter, and camouflage. Such differences can result in domain variation and degrade the performance of models trained or designed for conventional driving environments.

These challenges motivate the development of a unified perception framework specifically designed for autonomous defence vehicles operating in complex and visually challenging environments. To address the limited availability of defence-oriented perception data, the **Tactical Vehicle Dataset (TVD)**[40] is developed to provide diverse defence and civilian categories under varying terrain, viewpoint, illumination, weather, and visibility conditions. Building on this dataset, the proposed **DefenceVisionNet**[39] integrates defence-civilian entity detection, drivable terrain segmentation, and trail or boundary detection within a single feed-forward multi-task architecture. The entity detection branch is designed to distinguish defence-related vehicles and personnel from relevant civilian entities, while the segmentation and boundary branches provide complementary information regarding traversable terrain and navigational structures. Joint learning of these tasks enables the framework to develop shared visual representations while retaining task-specific processing for reliable prediction.

The development of DefenceVisionNet is therefore guided by four principal requirements: (i) joint perception of complementary defence-related visual tasks, (ii) efficient feature extraction and real-time inference with reduced computational redundancy, (iii) robust representation learning under occlusion, camouflage, clutter, illumination changes, and adverse weather conditions, and (iv) evaluation using a diverse defence-oriented dataset. These requirements form the basis for the proposed shared encoder, task-specific decoder architecture, attention-guided feature refinement, and balanced multi-task optimization strategy. Collectively, the framework is intended to provide efficient and robust visual perception for autonomous defence vehicles operating across diverse and unstructured environment.

The major contributions of the DefenceVisionNet framework are summarized as

follows:

1. A unified, camera-based multi-task perception framework, termed DefenceVisionNet, is developed to jointly perform defence-civilian entity detection, drivable terrain segmentation, and trail/boundary detection within a single architecture. The entity detection capability encompasses defence-related vehicles, soldiers, and civilian persons in challenging scenes.

2. An efficient shared feature-extraction architecture is developed using an E-ELAN-inspired backbone with multi-scale feature processing. The network incorporates C3k2, SPPF, and C2PSA-based components to improve feature representation and spatial feature selection under conditions involving occlusion, clutter, camouflage, and low visibility.

3. Attention-guided mechanisms are incorporated to strengthen feature refinement for difficult visual conditions. Contextual attention is used to improve defence entity discrimination, whereas attention-guided feature refinement supports the preservation of spatial information required for drivable terrain and boundary perception.

4. A customized multi-task loss and training strategy is developed to balance the optimization of defence entity detection, drivable terrain segmentation, and trail/boundary detection. Cascade-based refinement and adaptive task weighting are incorporated to improve learning stability and emphasize difficult samples and task-specific requirements.

5. The Tactical Vehicle Dataset (TVD) is developed for evaluating defence-oriented perception, comprising 20,758 images across 31 defence and civilian classes. The dataset includes diverse military vehicles and personnel under variations in terrain, viewpoint, illumination, weather, occlusion, and visual appearance, providing a dedicated benchmark for evaluating perception in challenging defence environments.

5.2 Methodology

This section presents the methodology adopted for the proposed defence-oriented perception framework. First, the Tactical Vehicle Dataset (TVD) is described, including its acquisition, preprocessing, annotation, augmentation, and class distribution. Subsequently, the DefenceVisionNet architecture is introduced as a unified multi-task learning framework based on an encoder–decoder structure. The network jointly performs defence entity detection, drivable terrain segmentation, and trail/boundary detection through a combined multi-task optimization objective.

5.2.1 Tactical Vehicle Dataset

Most existing object detection datasets have been developed for structured civilian road environments, primarily focused on detecting and recognizing conventional civilian vehicles. Datasets such as BDD100K [24] and Cityscapes [25] provide extensive annotations for vehicles in well-defined road scenarios; however, they are not designed to represent defence-specific vehicles and operational environments. Consequently, they lack adequate

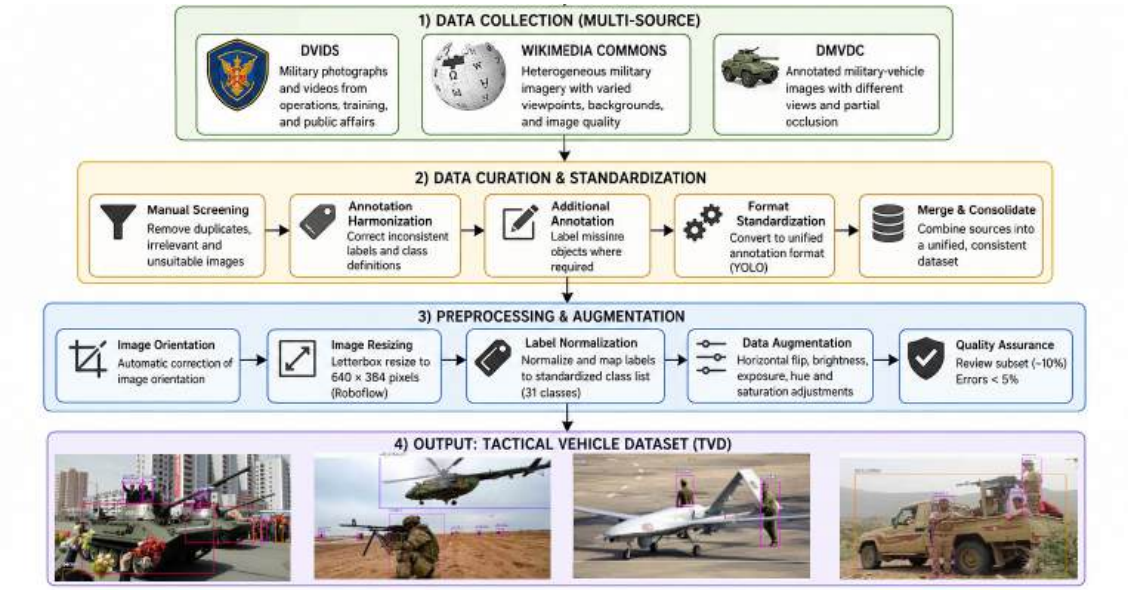

Figure 5.1: Block diagram for the generation of TVD Dataset

representation of tactical vehicles and the distinctive visual conditions associated with defence applications, including camouflage, partial visibility, occlusion, and challenging operational terrain. Similarly, available military datasets often focus on a limited number of object categories and are collected under relatively restricted environmental or viewpoint conditions.

To address these limitations, the Tactical Vehicle Dataset (TVD) [40] was developed, as shown in Fig.5.1, as a large-scale, multi-class dataset for defence-oriented visual perception. The dataset contains 20,758 images covering 31 categories of tactical vehicles. It includes diverse military platforms, such as armoured personnel carriers, tanks, helicopters, missile carriers, and other tactical vehicles. The images represent a broad range of terrain types, camera viewpoints, illumination conditions, weather conditions, and levels of target visibility. Such diversity provides a challenging evaluation environment for assessing the robustness of DefenceVisionNet, particularly in the presence of camouflage, occlusion, and low-visibility conditions.

5.2.2 Selection and Integration of Existing Datasets

The dataset investigation presented in Chapter 2 indicates that existing publicly available resources do not adequately represent the range of defence-related objects and visual conditions required for comprehensive tactical scene perception. Existing datasets are generally designed for specific vehicle categories or individual recognition tasks, while military image repositories provide broader visual content but lack consistent annotations. To overcome these limitations, multiple complementary sources were considered during the construction of the Tactical Vehicle Dataset (TVD).

The TVD incorporates imagery obtained from the Defense Visual Information Distribution Service (DVIDS) [78], Wikimedia Commons [79], and the Defense Military Vehicle Detection and Classification (DMVDC) dataset [80]. These sources were chosen to provide complementary visual characteristics rather than simply increase the number of images. DVIDS contributes military imagery covering a variety of operational and

training contexts, whereas Wikimedia Commons provides additional representations of military vehicles from different countries, viewpoints, and backgrounds. The DMVDC dataset provides previously annotated military-vehicle images, including samples in which targets are partially obscured.

The integration of these sources increases the representation of different tactical vehicle categories, object scales, viewpoints, backgrounds, illumination conditions, and levels of target visibility. As the original sources differ in image quality, annotation format, and class definitions, the collected imagery was subjected to a systematic quality-assessment and standardization process. This involved manual screening to remove duplicate and unsuitable samples, correction of inconsistent annotations, and labelling of additional objects where required. The data were subsequently harmonized and preprocessed to obtain a consistent set of ground-truth annotations. The resulting standardized collection forms the Tactical Vehicle Dataset (TVD) used for training and evaluating the proposed defence perception framework. The detailed annotation, label harmonization, preprocessing, and dataset preparation procedures are presented in Chapter 4. The principal characteristics and contributions of the selected data sources are summarized in Table 5.1.

Table 5.1: Comparison of datasets and repositories relevant to autonomous defence perception

Dataset / Repository	**Images**	**Classes**	**Annotation / Content**	**Primary Application**
DVIDS [78]	Large repository	Not standardized	Military photographs and videos from operational, training, and public-affairs contexts	Military object recognition and scene analysis
Wikimedia Commons [79]	Large collection	Not standardized	Heterogeneous military imagery with varied viewpoints, backgrounds, and image quality	Military-vehicle classification and visual diversity
DMVDC [80]	5,899	Vehicle-focused	Annotated military-vehicle images captured under different viewing conditions and partial occlusion	Military-vehicle detection and classification

To assess annotation quality, an independently reviewed subset corresponding to approximately 10% of the TVD was examined. The assessment found annotation errors in fewer than 5% of the reviewed images, indicating a high level of annotation consistency within the dataset.

The preprocessing pipeline included label normalization, correction of inconsistent class labels, automatic image orientation, and image resizing. The images were resized to 640×384 pixels using letterbox resizing through the Roboflow platform. Data augmentation was applied during training to increase sample diversity and improve model robustness to variations in real-world conditions. The augmentation procedures included

horizontal flipping and photometric transformations involving brightness, exposure, hue, and saturation adjustments.

Dataset Class Distribution Analysis

The class distribution of the TVD follows a long-tailed pattern, which is frequently observed in datasets collected from real-world defence environments. A relatively small number of categories, including persons, soldiers, tanks, cars, and command or utility vehicles, account for a substantial proportion of the available annotations. These categories occur more frequently because they are commonly observed across surveillance scenarios, operational environments, and different camera viewpoints.

In contrast, several defence-specific categories, including helicopters, jets, Humvees, and other specialized platforms, contain considerably fewer samples. This imbalance can influence the learning behaviour of the detection model by causing greater emphasis on frequently occurring classes. Consequently, the model may become biased toward dominant categories while providing weaker recognition of less represented but operationally important targets.

The limited number of samples for minority classes also makes it more difficult to learn discriminative visual features under challenging conditions, particularly when targets are small, partially occluded, camouflaged, or observed at different scales. In a multi-task learning setting, class imbalance may further affect optimization because frequently occurring classes can contribute disproportionately to the gradients during training.

To mitigate these effects, several strategies are incorporated into the proposed framework. Focal loss is employed to increase the contribution of difficult and underrepresented samples during optimization. Class-balanced training is used to reduce the influence of frequency imbalance, while cascade-based adaptive weighting provides additional emphasis on challenging instances, such as small or partially occluded targets. These strategies are intended to improve recognition across both dominant and minority classes while maintaining stable multi-task optimization.

5.2.3 Network Architecture

The proposed DefenceVisionNet is designed as a unified multi-task perception network for autonomous defence vehicles operating in unstructured and visually challenging environments. The framework simultaneously addresses three complementary perception tasks: defence– civilian entity detection, drivable terrain segmentation, and trail/boundary detection. Instead of deploying an independent network for each task, DefenceVisionNet employs a shared encoder–neck structure followed by dedicated task-specific decoder heads. This configuration allows the computationally intensive feature extraction stages to be shared among the perception tasks while preserving specialized processing for the individual prediction requirements.

The overall architecture is organized into three principal components: (i) a shared encoder and feature-fusion neck for extracting and combining multi-scale visual representations, (ii) three task-specific decoder heads for entity detection, drivable terrain segmentation, and trail/boundary detection, and (iii) attention-guided modules for enhancing task-relevant features under challenging environmental conditions. The architecture is illustrated in Fig. 5.2. The shared representation generated by the encoder provides

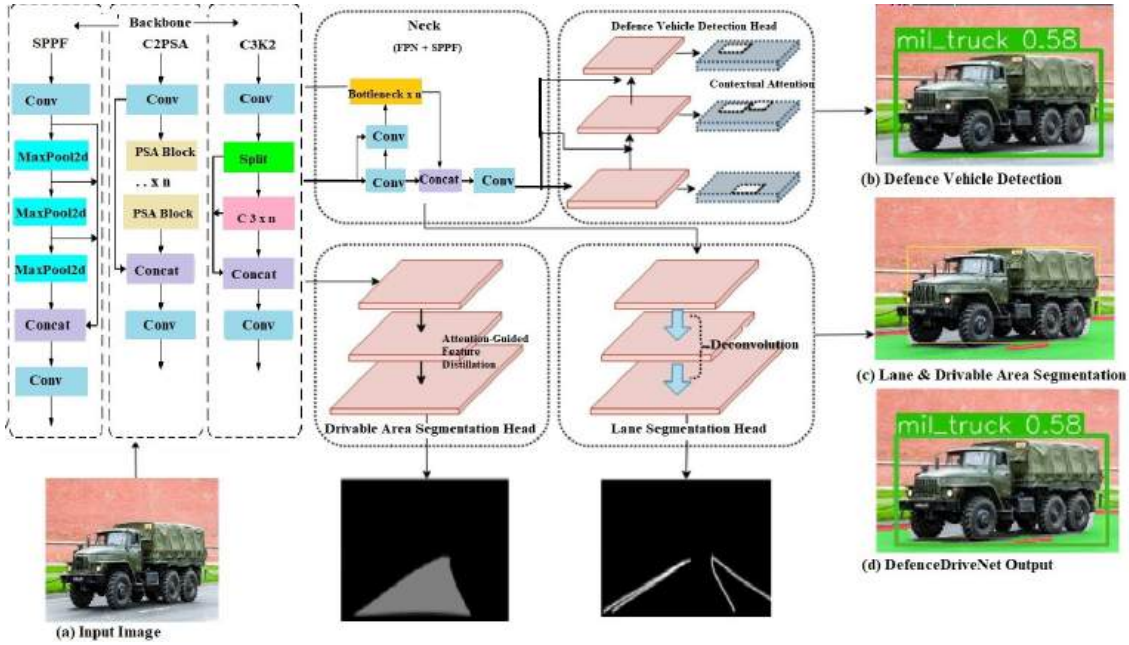

Figure 5.2: DefenceVisionNet: A unified multi-task perception network

common semantic information to all three branches, whereas the individual decoders transform these representations according to the spatial and semantic requirements of their respective tasks.

A unified architecture is particularly advantageous for autonomous defence perception because the same input image is required by all perception tasks. Independent networks would repeatedly perform similar low-level and intermediate-level feature extraction operations, resulting in unnecessary computational overhead. In DefenceVisionNet, these operations are performed once within the shared encoder, and the resulting feature maps are subsequently reused by the task-specific branches. At the same time, the use of separate decoder heads avoids forcing different tasks to share their final prediction representations. This separation helps preserve task-specific information and reduces the possibility of negative interaction between objectives having different prediction characteristics.

The network is therefore designed to achieve a balance between feature sharing and task specialization. Low and high-level representations generated by the encoder–neck provide common contextual information, while the individual heads selectively process these representations to generate object-level, region-level, and boundary-level predictions. Such a design is suitable for defence environments in which targets may be small, partially occluded, visually similar to their surroundings, or affected by camouflage, illumination variations, clutter, and adverse weather.

Shared Encoder and Feature-Fusion Neck

The shared encoder constitutes the primary feature extraction component of DefenceVision-Net. Its purpose is to transform the input image into hierarchical feature representations containing both spatial and semantic information that can be utilized by the three perception branches. The encoder follows an encoder–neck configuration in which progressively deeper layers capture increasingly abstract semantic representations, while the feature-fusion neck combines information across multiple spatial resolutions.

The backbone is inspired by the design principles of Efficient ELAN (E-ELAN),

particularly the use of feature splitting, grouped processing, and feature aggregation to improve information flow through the network. These operations facilitate the learning of diverse representations without requiring a proportional increase in computational complexity. Such characteristics are important for autonomous defence vehicles, where perception must operate with limited computational resources while maintaining sufficient representational capacity for complex scenes.

The backbone incorporates C3k2 blocks for efficient feature processing and channel interaction. These blocks facilitate repeated feature transformation and reuse while maintaining a comparatively lightweight architecture. The resulting hierarchical feature maps contain information at different spatial resolutions, allowing the network to represent both relatively large defence platforms and smaller or partially visible targets.

A Spatial Pyramid Pooling–Fast (SPPF) component is incorporated to increase the contextual receptive field of the network. By aggregating features over multiple effective spatial scales, SPPF enables the network to obtain contextual information beyond the local neighbourhood of individual convolutional operations. This is particularly useful in defence scenes where the appearance of an entity can depend on its surrounding terrain, nearby objects, or scene configuration.

The encoder further employs C2PSA-based feature processing to improve spatial feature selection. The self-attention mechanism enables the network to model relationships between spatial regions and selectively emphasize informative features. This property is useful for scenes containing clutter, partial occlusion, camouflage, and low illumination, where discriminative information may occupy only a small portion of the input image. By refining the spatial representation before it is passed to the subsequent stages, the network can retain relevant information while suppressing less useful background responses.

The multi-resolution features generated by the backbone are subsequently processed by the feature-fusion neck. The neck combines Feature Pyramid Network (FPN)-based hierarchical fusion with SPPF-based contextual aggregation. The FPN facilitates the propagation of semantic information from deeper layers to higher-resolution feature maps, thereby combining fine spatial details with stronger high-level representations. This multi-scale fusion is particularly important for defence perception, where the same scene may contain large vehicles in the foreground and small or distant entities in the background.

Consequently, the shared encoder–neck generates a set of multi-resolution feature maps containing complementary spatial and semantic information. These feature maps form the common representation for the three task-specific branches. Sharing this representation reduces duplicated computation and enables the different perception tasks to benefit from common visual cues while retaining independent prediction pathways.

Task-Specific Decoder Heads

Following the shared encoder–neck, DefenceVisionNet employs three dedicated decoder heads corresponding to the three perception objectives. Each head is optimized for the characteristics of its associated task, while receiving the multi-scale representations generated by the shared feature extraction stages. The separation of the heads is intended to maintain task-specific feature processing and reduce interference among object detection, semantic segmentation, and thin-structure detection.

Defence–Civilian Entity Detection Head The entity detection head is responsible for identifying and localizing defence and civilian entities present in the scene. It operates on

the multi-scale feature maps produced by the FPN–SPPF neck so that objects of different sizes can be represented effectively. This multi-scale design is important for defence scenarios, where a target may appear either as a large nearby vehicle or as a small distant object.

The detection branch performs both localization and category prediction. In addition to conventional feature processing, contextual attention is introduced to strengthen discriminative information associated with target regions. The attention-guided representation is intended to improve the distinction between visually similar targets and their surroundings, particularly when vehicles or personnel are partially hidden, camouflaged, or embedded within cluttered backgrounds. The resulting detection head therefore provides object-level information that complements the terrain and boundary information generated by the other branches.

Drivable Terrain Segmentation Head The drivable terrain branch generates dense pixel-level predictions corresponding to regions that can potentially be traversed by the autonomous vehicle. Unlike object detection, which predicts discrete instances, this task requires preservation of fine spatial information across the image. The segmentation decoder therefore utilizes intermediate high-resolution features from the feature-fusion neck.

An attention-guided feature refinement mechanism is incorporated within this branch to emphasize terrain-related information while suppressing irrelevant visual variations. Such refinement is beneficial in unstructured environments where shadows, vegetation, surface irregularities, weather effects, and illumination changes may otherwise produce misleading responses. The refined features are progressively upsampled to recover spatial resolution and generate a dense drivable terrain map with well-defined region boundaries.

The use of intermediate feature maps provides a suitable balance between semantic representation and spatial detail. Deeper features provide context regarding the overall scene, whereas higher-resolution features retain local terrain structures. Their combined utilization enables the segmentation head to distinguish traversable regions in complex off-road environments.

Trail and Boundary Detection Head The trail/boundary branch is designed to identify elongated and relatively thin structures that can provide useful navigational cues in the absence of conventional road markings. Such structures may correspond to visible trails, terrain boundaries, or other scene cues that indicate the potential direction or extent of vehicle movement.

This branch receives semantically rich features from the deeper stages of the feature-fusion neck, allowing subtle boundary patterns to be interpreted within their broader scene context. Since trail and boundary structures can occupy only a small number of pixels, the decoder progressively restores spatial resolution using deconvolution-based upsampling operations. Additional refinement layers are used to improve the continuity and spatial quality of the predicted structures.

The trail/boundary head is particularly important for defence environments because conventional lane markings may be absent. The ability to recover thin navigational structures from visually ambiguous terrain provides information that complements the drivable-area prediction and object detection outputs.

Attention-Guided Feature Refinement

Defence environments present substantial visual ambiguity due to camouflage, occlusion, clutter, illumination changes, and adverse weather. Under such conditions, convolutional feature extraction alone may not sufficiently distinguish task-relevant regions from visually similar background structures. To improve feature selectivity, DefenceVisionNet incorporates attention-guided mechanisms at appropriate stages of the task-specific processing pathways.

Rather than applying a single attention mechanism uniformly to all tasks, the proposed framework uses task-oriented attention strategies. This is motivated by the fact that the relevant visual information is different for object detection, terrain segmentation, and trail/boundary detection. For example, entity detection benefits from context surrounding an object, whereas terrain segmentation requires precise spatial information and boundary detection depends strongly on fine elongated structures.

Two principal attention-guided mechanisms are therefore incorporated: (i) attention-guided feature distillation for drivable terrain segmentation and (ii) contextual attention for defence entity detection. These mechanisms operate on the feature representations associated with their respective task branches.

Attention-Guided Feature Distillation for Drivable Terrain Segmentation

The attention-guided feature distillation mechanism is integrated into the drivable terrain segmentation pathway to improve the selection of terrain-relevant spatial features. The objective is to retain information associated with potentially traversable regions while reducing the influence of visually distracting structures.

In unstructured environments, terrain appearance can vary substantially because of shadows, surface texture, vegetation, dust, rain, and illumination changes. These factors may introduce features that are visually prominent but provide limited information about traversability. The attention-guided mechanism consequently assigns greater importance to informative feature responses and suppresses less relevant representations before dense prediction.

The refined representation is subsequently passed through the segmentation decoder, where spatial resolution is progressively recovered. This process enables the network to combine semantic understanding with detailed spatial information and thereby generate more consistent drivable-region predictions.

Contextual Attention for Defence Entity Detection

Contextual attention is incorporated into the entity detection branch to improve target discrimination by considering both object-specific features and their surrounding visual context. This is especially important for defence imagery because military vehicles and personnel may share similar colours, textures, and shapes with elements of the surrounding environment.

The contextual attention mechanism enhances feature responses associated with relevant target regions while reducing responses originating from confusing background structures. Consequently, the detection branch can make more informed predictions when only a portion of an object is visible or when the target is embedded within clutter or camouflage. The mechanism also complements the multi-scale feature representation

provided by the encoder–neck, allowing contextual information to support detection across different object sizes and viewpoints.

Overall, the attention-guided components enhance the task-specific representations without eliminating the benefits of the shared feature space. The resulting architecture combines common feature extraction with task-dependent refinement, enabling DefenceVisionNet to exploit cross-task information while maintaining specialized processing for each perception objective.

Overall Information Flow

The complete processing pipeline of DefenceVisionNet begins with a camera image that is supplied to the shared encoder. The backbone progressively extracts hierarchical representations, while the neck aggregates information across multiple spatial scales. The resulting feature maps are then distributed to the three task-specific decoder heads.

The detection branch produces bounding-box and class predictions for defence and civilian entities. In parallel, the terrain segmentation branch generates a pixel-wise representation of the traversable region, while the trail/boundary branch predicts thin navigational structures. The three outputs therefore provide complementary scene information: entity-level understanding identifies relevant objects, terrain segmentation describes where the vehicle can potentially move, and trail/boundary detection provides structural cues for navigation.

The shared encoder minimizes redundant feature extraction, whereas the independent decoder heads preserve task-specific representations. Attention-guided refinement further improves the ability of the network to focus on relevant information in visually challenging scenes. This combination of shared representation learning, multi-scale feature fusion, task-specific decoding, and selective attention forms the core design of the proposed DefenceVisionNet framework.

5.2.4 Multi-Task Loss Function with Cascade-Based Refinements

The training of DefenceVisionNet is formulated as a joint optimization problem in which the three perception tasks are learned simultaneously. These tasks include defence–civilian entity detection, drivable terrain segmentation, and trail/boundary detection. A combined objective is used to coordinate the learning of the individual task branches while maintaining an appropriate balance between their contributions to the overall optimization. The overall loss is obtained by combining the individual task losses with task-specific weighting factors, similar to YOLOPv2 [23]. The total training loss is defined as

$$L_{\text{total}} = \alpha L_{\text{det}} + \beta L_{\text{drive}} + \gamma L_{\text{lane}} \tag{1}$$

where L_{det}, L_{drive}, and L_{lane} represent the loss functions associated with defence–civilian entity detection, drivable terrain segmentation, and trail/boundary detection, respectively. The coefficients α, β, and γ control the relative contribution of the three objectives during optimization. In the proposed framework, $\alpha = 1.0$, $\beta = 0.8$, and $\gamma = 0.6$ are used. These values were selected empirically by monitoring validation performance and training stability. The sensitivity analysis indicated that assigning excessive importance to the detection objective could reduce the learning effectiveness of the segmentation tasks. Therefore, the weighting coefficients were selected to provide a suitable balance between object-level and dense prediction objectives.

The combined loss formulation enables the shared encoder to receive supervision from all three tasks. Consequently, the learned feature representation is influenced by complementary object, terrain, and boundary information rather than being optimized exclusively for a single perception objective.

Defence Entity Detection Loss

The defence–civilian entity detection branch must simultaneously localize target objects and determine their corresponding categories. This becomes particularly challenging when objects are small, partially occluded, camouflaged, or surrounded by visually similar structures. To address these requirements, the detection objective combines localization, objectness, and classification losses. The detection loss is expressed as

$$L_{\text{det}} = \lambda_{\text{box}} L_{\text{CIoU}} + \lambda_{\text{obj}} L_{\text{BCE}} + \lambda_{\text{cls}} L_{\text{Focal}} \tag{2}$$

where L_{CIoU} denotes the Complete Intersection over Union (CIoU) loss used for bounding-box regression, L_{BCE} represents the Binary Cross-Entropy (BCE) loss for objectness prediction, and L_{Focal} denotes the focal loss used for category classification. The corresponding weighting factors are $\lambda_{\text{box}} = 5.0$, $\lambda_{\text{obj}} = 1.0$, and $\lambda_{\text{cls}} = 1.0$.

The CIoU component encourages accurate alignment between predicted and ground-truth bounding boxes by considering their overlap and geometric relationship. The BCE term supervises the prediction of whether a candidate region contains an object. Focal loss gives greater emphasis to difficult classification examples and reduces the dominance of frequently occurring categories. This is particularly relevant for the TVD because the dataset exhibits a long-tailed class distribution in which several defence categories have substantially fewer samples than the dominant classes.

The relatively larger value assigned to λ_{box} emphasizes accurate localization during optimization. This is important for autonomous defence perception because imprecise object localization can affect subsequent interpretation of the operational scene. The combined formulation consequently provides simultaneous supervision for spatial localization, object presence, and category recognition.

Drivable Terrain Segmentation Loss

The drivable terrain branch performs dense pixel-level classification of the input scene into traversable and non-traversable regions. Since the task is formulated as binary segmentation, Binary Cross-Entropy-based cross-entropy supervision is employed to optimize the predicted terrain map. The corresponding loss evaluates the agreement between the predicted pixel probabilities and their ground-truth labels.

The attention-guided feature refinement introduced in the segmentation branch complements the loss-based supervision by directing the decoder towards spatially relevant terrain features. This is particularly useful in unstructured defence environments, where shadows, vegetation, surface irregularities, dust, and adverse illumination can produce visual patterns that are difficult to distinguish from genuinely traversable areas. The combination of dense supervision and attention-guided features helps maintain spatial consistency in the predicted drivable regions and supports the recovery of meaningful terrain boundaries.

Trail and Boundary Detection Loss

Trail and boundary detection differs from terrain segmentation because the target structures are generally sparse, narrow, and spatially discontinuous. In defence environments without clearly defined roads or lane markings, these structures may provide important navigational cues, but their visual appearance can vary considerably with terrain, illumination, occlusion, and visibility.

Focal loss is therefore employed to address the substantial imbalance between informative trail/boundary pixels and the dominant background. The loss reduces the relative influence of easily classified background samples and places greater emphasis on difficult positive and hard-to-classify regions. This formulation supports the learning of fine-grained trail and boundary structures that might otherwise be overwhelmed by the much larger number of background pixels.

The use of focal supervision is particularly beneficial for preserving weak or partially visible boundary structures. When combined with the high-level semantic representations and progressive spatial refinement of the trail/boundary decoder, the loss facilitates more consistent prediction of thin navigational structures under challenging visual conditions.

Cascade-Based Loss Refinements

Although the basic multi-task objective provides joint supervision for the three perception branches, the difficulty of individual samples and the reliability of task predictions can vary considerably during training. Defence imagery may contain targets with different degrees of occlusion, visibility, scale, and background complexity. Similarly, terrain and trail boundaries may become ambiguous under changing illumination or adverse weather. A uniform optimization strategy may therefore assign insufficient emphasis to difficult samples or overemphasize tasks that are already performing satisfactorily.

To address this issue, DefenceVisionNet incorporates cascade-based refinement strategies that adapt the optimization process according to sample difficulty and task uncertainty. Two complementary mechanisms are considered: difficulty-aware cascade weighting and uncertainty-guided adaptive task weighting.

Difficulty-Aware Cascade Weighting

The difficulty-aware weighting mechanism is intended to increase the training emphasis on challenging samples. Its formulation is motivated by IoU-sensitive loss strategies, in which the quality of predicted localization is used as an indicator of sample difficulty. In the context of DefenceVisionNet, samples with inaccurate localization or weak spatial correspondence are treated as more difficult examples and receive greater attention during optimization.

This mechanism is particularly relevant to defence imagery because objects may be partially concealed, visually blended with the background, or observed at substantially different scales. Instead of allowing easily recognized samples to dominate the training process, difficulty-aware weighting encourages the network to devote additional learning capacity to cases that remain poorly predicted.

The cascade-based formulation consequently complements the focal loss used for classification by considering the quality of the overall prediction during refinement. This provides an additional mechanism for improving the representation of difficult target instances and supporting more robust localization in complex scenes.

Uncertainty-Guided Adaptive Task Weighting

The second refinement mechanism dynamically considers the relative uncertainty associated with the different perception tasks. In multi-task learning, the three objectives may exhibit substantially different levels of difficulty and convergence behaviour. A fixed weighting scheme may therefore become suboptimal during different stages of training. Uncertainty-guided task weighting addresses this issue by adapting the effective contribution of individual task losses according to their estimated uncertainty.

A task exhibiting greater uncertainty is given appropriate consideration during optimization, preventing its contribution from being suppressed by objectives that converge more rapidly or produce stronger gradients. Conversely, excessive emphasis on an already well-learned objective can be avoided. This adaptive strategy is useful for DefenceVisionNet because object detection, terrain segmentation, and trail/boundary prediction operate on different output representations and exhibit different levels of visual ambiguity.

The uncertainty-guided mechanism therefore provides an additional level of balancing beyond the fixed coefficients in Eq. (1). The fixed coefficients establish the overall relative importance of the three tasks, while adaptive weighting allows the optimization process to respond to changing task difficulty during training.

Combined Multi-Task Optimization

The combination of task-specific losses and cascade-based refinements provides a hierarchical optimization strategy for DefenceVisionNet. The primary multi-task objective ensures that the shared encoder learns representations useful for all three perception tasks, while the individual loss components provide specialized supervision for localization, classification, terrain segmentation, and trail/boundary prediction. Difficulty-aware weighting further emphasizes challenging samples, whereas uncertainty-guided weighting regulates the relative influence of the individual tasks during optimization.

This formulation is intended to prevent any single task from dominating the learning process and to improve the utilization of the shared feature representation. It also provides greater emphasis on difficult defence scenarios involving occlusion, camouflage, clutter, small targets, ambiguous terrain, and weak boundary structures. Consequently, the combined optimization strategy supports improved training stability and generalization while maintaining a unified learning framework for defence-oriented multi-task perception.

5.3 Experimentation and Evaluation

This section presents the experimental methodology adopted to train and evaluate the proposed DefenceVisionNet. It describes the computational environment, optimization strategy, training configuration, and partitioning of the Tactical Vehicle Dataset (TVD) used for model development and evaluation. The performance of the proposed framework is subsequently assessed against representative state-of-the-art multi-task perception approaches under structured, unstructured, and adverse-weather conditions.

5.3.1 Experimental Setup

All experiments were conducted using an NVIDIA Tesla V100 GPU with PyTorch version 1.12. The Adam optimizer was employed for training the proposed DefenceVisionNet

because of its adaptive parameter-update mechanism and its ability to provide stable optimization in the presence of varying gradient magnitudes. This property is particularly beneficial for the proposed multi-task framework, where defence–civilian entity detection, drivable terrain segmentation, and trail/boundary detection generate heterogeneous optimization objectives.

The network was trained for 350 iterations using a batch size of 20. A cosine-annealing learning-rate schedule with warm restarts was adopted to regulate the learning rate during optimization. The gradual variation of the learning rate facilitates smooth convergence, while periodic restarts help the optimization process move away from potentially stagnant regions of the parameter space. The initial learning rate was set to 0.01, with warm restarts incorporated during the early stages of training to promote stable convergence.

For experimental evaluation, the Tactical Vehicle Dataset (TVD) was divided into training, validation, and testing subsets using a class-aware random splitting strategy. The splitting procedure was designed to maintain the representation of different categories across the three subsets and thereby reduce the potential influence of class imbalance on model evaluation. A total of 18,024 images, corresponding to approximately 87% of the dataset, were assigned to the training subset. The validation subset comprised 2,011 images (approximately 10%), while the remaining 723 images (approximately 3%) were reserved for independent testing.

The training subset was used to optimize the network parameters, whereas the validation subset was used to monitor the training behaviour and assess model configurations during development. The independent test subset was retained for the final performance assessment. This separation ensures that the reported results provide an indication of the generalization capability of DefenceVisionNet on images that were not directly used for network parameter optimization.

5.3.2 Results and Evaluation

A comprehensive quantitative and qualitative evaluation of the proposed DefenceVisionNet was conducted to assess its effectiveness for the three targeted perception tasks, namely defence–civilian entity detection, drivable terrain segmentation, and trail/boundary detection. The evaluation was performed using the Tactical Vehicle Dataset (TVD), with particular emphasis on both task-specific performance and the ability of the framework to operate under diverse defence-related environmental conditions. In addition to accuracy-related measures, computational efficiency and inference speed were examined to determine the suitability of the proposed framework for real-time autonomous defence applications.

Model Parameters and Inference Speed

The computational complexity and inference efficiency of DefenceVisionNet were compared with representative multi-task perception models, as summarized in Table 5.2. The comparison considers the number of trainable parameters and the inference speed for an input resolution of 640 pixels. As shown in Table 5.2, DefenceVisionNet contains 41.2 million parameters, which is higher than the parameter count of the compared multi-task models. The additional parameters are primarily associated with the attention-guided feature processing and the dedicated decoder branches required for the three perception objectives. Despite this increase in model capacity, the proposed framework achieves an

Table 5.2: Comparison of Model Parameters and Inference Speed.

Model	Input Size	Params (M)	Speed (FPS)
YOLOP	640	7.9	49
HybridNets	640	12.8	28
YOLOPv2	640	38.9	91
DefenceVisionNet	**640**	**41.2**	**91.2**

inference speed of 91.2 FPS at an input resolution of 640 pixels.

The obtained inference rate substantially exceeds the commonly adopted real-time threshold of 30 FPS and demonstrates the computational feasibility of the proposed framework for real-time perception. Notably, DefenceVisionNet achieves a speed comparable to YOLOPv2 while providing a defence-oriented multi-task architecture with specialized processing for entity detection, drivable terrain segmentation, and trail/boundary detection. This efficiency is enabled by sharing the backbone and feature-fusion stages among the three tasks, thereby avoiding repeated feature extraction from the same input image. Furthermore, lightweight task-specific decoder branches help limit the additional computational cost introduced by the multi-task design.

The results indicate that the proposed architecture provides a favourable balance between representational capacity and inference efficiency. Although the inclusion of attention mechanisms and task-specific processing increases the total parameter count, the resulting computational overhead remains sufficiently low to support real-time operation. This characteristic is particularly relevant for autonomous defence vehicles, where perception outputs must be generated rapidly to support timely navigation and scene interpretation.

Deployment Memory and Edge Inference Considerations

In addition to inference speed, the memory requirement of the proposed DefenceVisionNet was considered from the perspective of deployment on resource-constrained platforms. With 41.2 million parameters, the model requires approximately 165 MB of memory for storing the parameters when standard 32-bit floating-point (FP32) representation is used. The parameter-memory requirement can be substantially reduced through lower-precision representations. In particular, FP16 mixed-precision inference reduces the parameter memory to approximately 82 MB, while INT8 quantization can reduce it further to approximately 41 MB.

These reductions are particularly relevant for autonomous defence systems, where perception models may need to operate on embedded computing platforms with constrained memory and processing resources. Deployment-oriented techniques such as numerical quantization, mixed-precision computation, layer fusion, and TensorRT-based inference optimization have been shown in previous edge-deployment studies to reduce computational and memory overhead while preserving an acceptable level of model accuracy.

Accordingly, the proposed DefenceVisionNet can potentially benefit from similar deployment optimizations. Applying FP16 inference, INT8 quantization, layer fusion,

and hardware-specific acceleration may further reduce the resource requirements and inference latency of the network. Such optimization is expected to improve the suitability of DefenceVisionNet for deployment on embedded edge-computing platforms used in autonomous defence vehicles, particularly where computational resources and power availability are limited.

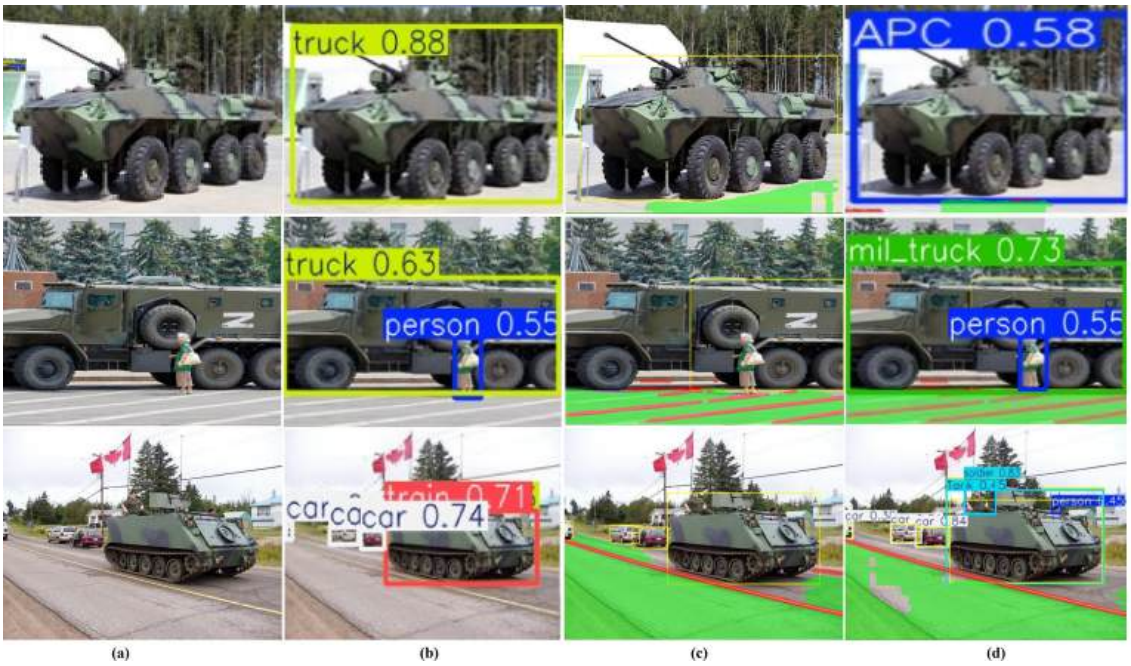

Figure 5.3: Visual evaluation of defence-entity detection for land-based defence vehicles in the TVD dataset: (a) input frame; (b) detection results from the fine-tuned YOLOv11 model; (c) detection results from the fine-tuned YOLOPv2 model; and (d) output generated by the proposed DefenceVisionNet.

Evaluation of Defence Entity Detection

The defence entity detection capability of the proposed DefenceVisionNet was evaluated for identifying and localizing both defence-related vehicles and personnel in the Tactical Vehicle Dataset (TVD). The proposed framework was compared with conventional single-task object detectors, including YOLOv11 and Faster R-CNN, as well as representative multi-task perception architectures such as MultiNet, DLT-Net, YOLOP, YOLOPv2, and HybridNets. For consistency, all comparison models were evaluated on the same TVD test set.

Detection performance was assessed using mean Average Precision at an intersection-over-union threshold of 0.50 (mAP@50) and Recall. The mAP@50 metric reflects the ability of a detector to correctly localize and classify target objects, whereas Recall measures the proportion of ground-truth instances successfully identified by the model. Considering both measures provides a more comprehensive assessment of detection quality, particularly for defence scenes containing small, occluded, or partially visible targets.

The quantitative comparison is presented in Table 5.3. Models originally developed or trained predominantly for conventional civilian driving environments show a noticeable reduction in performance when evaluated on TVD. This behaviour can be attributed to the substantial difference between civilian road scenes and defence environments, including

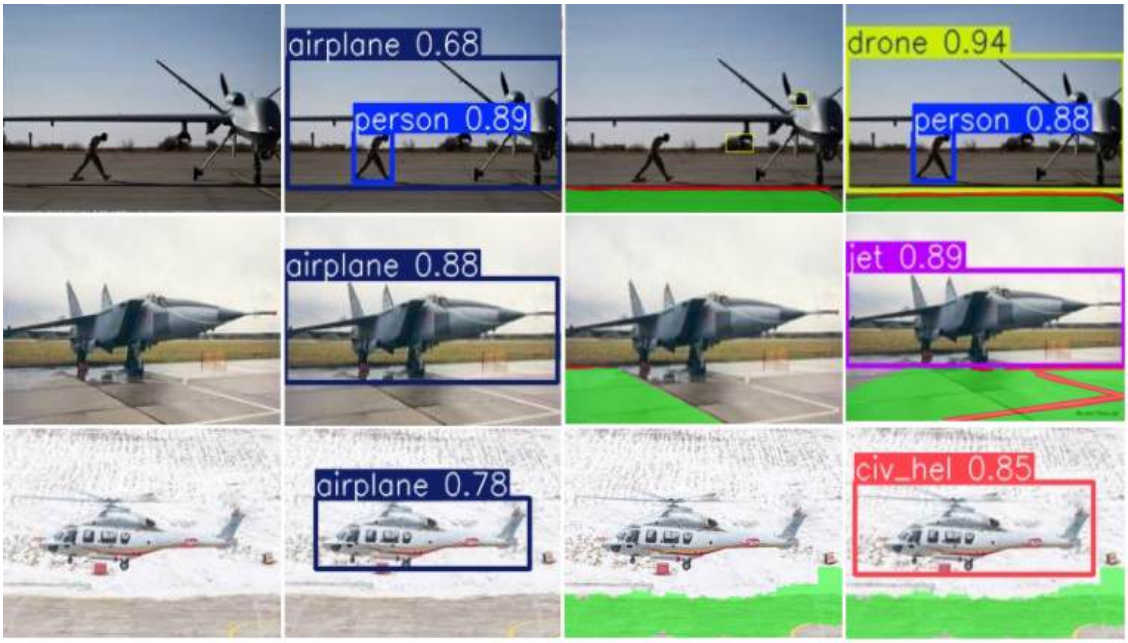

Figure 5.4: Visual assessment of defence-entity detection for air-defence vehicles in the TVD dataset: (a) input frame; (b) results obtained using the fine-tuned YOLOv11 detector; (c) outputs generated by the fine-tuned YOLOPv2 model; and (d) predictions produced by the proposed DefenceVisionNet.

variations in object appearance, terrain, viewpoint, background complexity, and visibility. The domain discrepancy becomes particularly evident for military objects and personnel that are not adequately represented in conventional road-driving datasets.

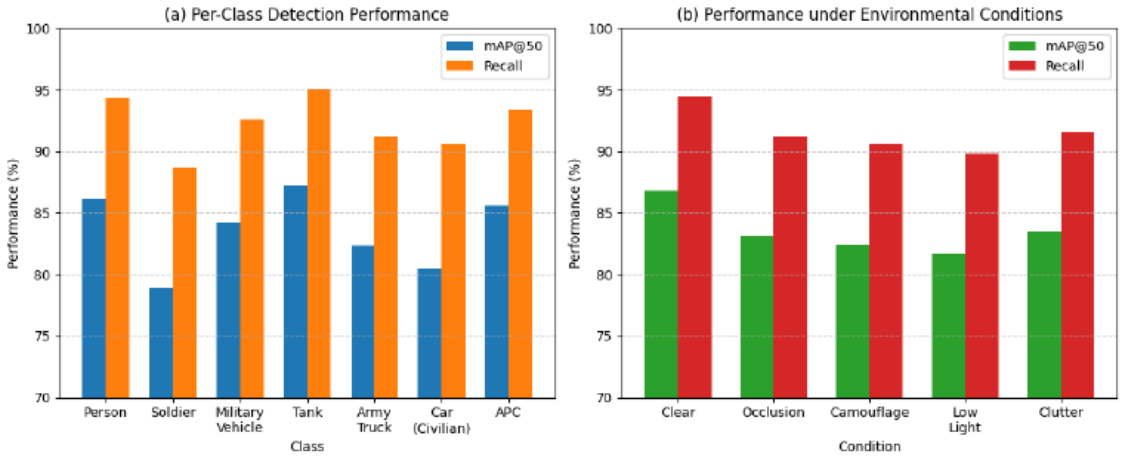

Figure 5.5: (a) Per-class detection performance on TVD; (b) Performance under varying environmental conditions.

Qualitative observations further support these findings. As illustrated in Fig. 5.3, the fine-tuned baseline models exhibit difficulties in distinguishing visually similar categories and in maintaining reliable predictions under challenging scene conditions. Typical errors include confusion between jets and commercial aircraft and between armoured personnel carriers and trucks. Reduced confidence and missed detections are also observed when target objects are affected by camouflage, background clutter, partial occlusion, or

non-uniform illumination.

The class-wise and condition-wise performance of DefenceVisionNet is presented in Fig. 5.5. The class-level analysis indicates stronger detection performance for frequent and visually distinctive categories, particularly Tank and Person. Comparatively lower performance is observed for categories such as Soldier and Car. The reduced performance of these categories can be associated with factors such as smaller target dimensions, partial occlusion, visual similarity with neighbouring classes, and variations in appearance. The difference between the Person and Soldier categories also demonstrates the difficulty of distinguishing fine-grained classes with similar human appearance.

The environmental analysis shows that the detector performs most effectively under clear viewing conditions. Its performance gradually decreases as the level of visual complexity increases, particularly in the presence of occlusion, camouflage, low illumination, and clutter. The most pronounced degradation is observed under low-light and camouflage-dominated conditions. These observations are consistent with the challenging characteristics of the TVD discussed in Section 5.2.2 and demonstrate the influence of both class imbalance and environmental variability on defence-oriented object detection.

Table 5.3: Comparison of Defence Entity Detection Performance on the Tactical Vehicle Dataset (TVD).

Model	**Speed (FPS)**	**mAP@50 (%)**	**Recall (%)**
MultiNet	8.6	57.1	78.6
DLT-Net	9.3	65.2	86.9
Faster R-CNN	8.8	61.3	77.8
YOLOP	41	73.4	86.5
HybridNets	28	74.8	89.6
YOLOPv2	91	79.2	88.3
YOLOv11s	90	80.6	89.9
DefenceVisionNet	**91.2**	**84.5**	**93.2**

As shown in Table 5.3, DefenceVisionNet achieves the highest detection performance among the evaluated methods, with an mAP@50 of 84.5% and a Recall of 93.2%, while maintaining an inference speed of 91.2 FPS. The proposed framework therefore provides an improvement of approximately 3.9 percentage points in mAP@50 over YOLOv11s and approximately 4.2 percentage points over YOLOPv2. The Recall improvement is also notable, with DefenceVisionNet exceeding YOLOv11s by approximately 3.3 percentage points and YOLOPv2 by approximately 4.9 percentage points.

The improvement is particularly relevant for difficult defence conditions, where successful identification of partially visible, small-scale, or camouflaged entities is more important than performance on clear scenes alone. The proposed architecture combines multi-scale feature extraction, contextual attention, and task-specific processing to improve the representation of defence entities while retaining real-time inference capability. The separate task-specific branches also limit the interference that can occur when fundamentally

different prediction objectives are forced into a common prediction head.

Overall, the results demonstrate that training the proposed framework specifically for the defence-oriented characteristics of TVD provides advantages over models primarily designed for conventional road perception. The combination of improved detection accuracy, higher recall, and an inference rate of 91.2 FPS indicates that DefenceVisionNet is capable of providing reliable real-time entity perception for autonomous defence vehicles operating across diverse and challenging environments.

Evaluation of Drivable Terrain Segmentation

The drivable terrain segmentation capability of the proposed DefenceVisionNet was evaluated on the Tactical Vehicle Dataset (TVD) using the mean Intersection over Union (mIoU) metric. The mIoU provides a region-level measure of segmentation quality by evaluating the overlap between the predicted drivable-terrain region and its corresponding ground-truth mask. A higher mIoU indicates more accurate identification of navigable terrain and, consequently, provides a stronger basis for downstream navigation and path-planning operations.

The quantitative comparison of the evaluated models is presented in Table 5.4. The results demonstrate considerable variation in segmentation performance across the different architectures. MultiNet and DLT-Net achieve mIoU values of 70.8% and 70.4%, respectively, indicating limited capability to represent the complex spatial characteristics of unstructured defence environments. In comparison, PSPNet achieves a substantially higher mIoU of 88.7%, demonstrating the benefit of dedicated semantic segmentation architectures for identifying navigable regions.

The multi-task perception models, namely YOLOP, HybridNets, and YOLOPv2, further improve the segmentation performance, obtaining mIoU values of 90.6%, 89.8%, and 92.4%, respectively. Their improved performance can be attributed to the sharing of spatial and semantic representations across multiple perception tasks, which facilitates more effective localization of the drivable region. However, these models remain less effective when the terrain contains irregular surfaces, ambiguous boundaries, or weak visual separation between drivable and non-drivable regions.

Table 5.4: Comparison of Drivable Terrain Segmentation Performance on the Tactical Vehicle Dataset (TVD).

Model	**Drivable mIoU (%)**
MultiNet	70.8
DLT-Net	70.4
PSPNet	88.7
YOLOP	90.6
HybridNets	89.8
YOLOPv2	92.4
DefenceVisionNet	**94.1**

Among all the evaluated approaches, DefenceVisionNet achieves the highest drivable-terrain segmentation performance, with an mIoU of 94.1%. This result represents an improvement of 1.7 percentage points over YOLOPv2, which provides the second-highest mIoU of 92.4%. Compared with YOLOP and HybridNets, the proposed framework improves the mIoU by 3.5 and 4.3 percentage points, respectively. The results indicate that the proposed architecture is better suited to extracting and preserving the spatial characteristics required for reliable terrain segmentation in defence-oriented environments.

The improved segmentation capability is primarily supported by the attention-guided feature refinement mechanism and skip-connected upsampling pathways incorporated within DefenceVisionNet. These components facilitate the retention of fine spatial information during feature reconstruction and improve the delineation of drivable regions. This is particularly beneficial in challenging TVD scenes characterized by uneven terrain, waterlogged surfaces, weak road boundaries, and poorly defined navigable paths. By producing more accurate terrain boundaries while maintaining contextual information, the proposed framework provides a more reliable representation of the navigable region for autonomous vehicle navigation.

Overall, the results in Table 5.4 demonstrate that DefenceVisionNet provides superior drivable terrain segmentation compared with both conventional segmentation networks and existing multi-task perception frameworks. The achieved 94.1% mIoU highlights the effectiveness of the proposed attention-guided spatial refinement strategy for handling the irregular and visually complex terrain encountered by autonomous defence vehicles.

Evaluation of Trail and Boundary Detection

The trail and boundary perception capability of the proposed DefenceVisionNet was evaluated using Accuracy and Boundary IoU. Accuracy represents the overall correctness of the predicted trail regions, whereas Boundary IoU provides a more stringent assessment by measuring the overlap between the predicted boundaries and the corresponding ground-truth boundaries. The latter is particularly important for autonomous defence vehicles, as accurate boundary localization is essential for identifying navigable paths in environments where road markings are absent or poorly defined.

The comparative results obtained on the Tactical Vehicle Dataset (TVD) are reported in Table 5.5. The performance of the lightweight segmentation networks, including ENet, SCNN, and ENet-SAD, is relatively limited. ENet achieves an Accuracy of 34.12% and a Boundary IoU of 14.64%, while SCNN provides only a marginal improvement, with 35.79% Accuracy and 15.84% Boundary IoU. ENet-SAD further increases the values to 36.56% and 16.02%, respectively, but the overall performance remains inadequate for complex defence terrain.

The multi-task perception frameworks exhibit substantially better performance than the lightweight segmentation baselines. YOLOP achieves 70.50% Accuracy and 26.20% Boundary IoU, while HybridNets increases these values to 85.40% and 31.60%, respectively. YOLOPv2 provides a higher Accuracy of 87.31%; however, its Boundary IoU of 27.25% indicates that improved regional classification does not necessarily translate into equally accurate boundary localization. This difference highlights the importance of explicitly preserving fine structural information when detecting trails and terrain boundaries.

Among all the evaluated approaches, DefenceVisionNet achieves the best overall performance, obtaining an Accuracy of 91.8% and a Boundary IoU of 32.4%. Compared with YOLOPv2, the proposed framework improves Accuracy by 4.49 percentage points

Table 5.5: Comparison of Trail and Boundary Detection Performance on the Tactical Vehicle Dataset (TVD).

Model	**Accuracy (%)**	**Boundary IoU (%)**
ENet	34.12	14.64
SCNN	35.79	15.84
ENet-SAD	36.56	16.02
YOLOP	70.50	26.20
HybridNets	85.40	31.60
YOLOPv2	87.31	27.25
DefenceVisionNet	**91.8**	**32.4**

and Boundary IoU by 5.15 percentage points. It also exceeds HybridNets by 6.4 percentage points in Accuracy and 0.8 percentage points in Boundary IoU. These results demonstrate the advantage of the proposed architecture in simultaneously identifying trail regions and maintaining their structural boundaries.

The improved performance can be attributed to the dedicated trail and boundary detection head, which operates on deeper FPN-based feature representations. The use of attention-guided feature refinement enables the network to emphasize relevant trail structures while suppressing irrelevant background information. In addition, skip-connected feature reconstruction helps retain spatial details that may otherwise be lost during progressive downsampling and upsampling. These characteristics are particularly useful in defence scenarios involving off-road terrain, forest trails, desert regions, vegetation, debris, shadows, and poorly demarcated routes.

Overall, the results demonstrate that DefenceVisionNet provides more reliable trail and boundary perception than the considered single-task and multi-task baseline methods. The combination of high regional Accuracy and improved Boundary IoU indicates that the proposed framework can better preserve trail continuity and boundary structure, which are important for autonomous navigation across complex and unstructured defence environments.

Statistical Validation Across Tasks

The reliability and reproducibility of the proposed DefenceVisionNet were investigated through three independent training runs performed with different random seed values. The performance obtained from each run was recorded for defence entity detection, drivable terrain segmentation, trail and boundary detection, and inference speed. The complete statistical evaluation is presented in Table 5.6.

As observed from Table 5.6, the proposed framework exhibits only minor performance variations across the three independent training configurations. For defence object detection, the mAP@50 values obtained for Seeds 1, 2, and 3 are 84.3%, 84.7%, and 84.5%, respectively. These results correspond to an overall performance of 84.5 ± 0.2%. The corresponding Recall values are 92.9%, 93.4%, and 93.3%, resulting in a mean Recall

Table 5.6: Statistical Validation Across All Perception Tasks on the Tactical Vehicle Dataset (TVD).

Configuration	Speed (FPS)	Defence Object Detection (%)		Drivable Terrain Segmentation (%)	Trail and Boundary (%)	
		mAP@50	Recall	Drivable mIoU	Accuracy	Boundary IoU
Seed 1	90.9	84.3	92.9	93.8	91.5	32.2
Seed 2	91.4	84.7	93.4	94.3	92.0	32.5
Seed 3	91.3	84.5	93.3	94.1	91.8	32.4
Mean ± Std	91.2 ± 0.2	84.5 ± 0.2	93.2 ± 0.3	94.1 ± 0.2	91.8 ± 0.2	32.4 ± 0.12

of $93.2 \pm 0.3\%$. The slightly larger variation in Recall can be associated with the sensitivity of this metric to difficult detection instances, particularly small targets, partially occluded objects, and objects with weak visual distinction from their surroundings.

The drivable terrain segmentation results remain highly consistent across the three training configurations. The obtained mIoU values are 93.8%, 94.3%, and 94.1% for Seeds 1, 2, and 3, respectively. The resulting mean performance is $94.1 \pm 0.2\%$. The small standard deviation indicates that the learned representation of navigable terrain is largely unaffected by changes in the random initialization. This consistency is desirable for autonomous navigation, where reliable terrain estimation is required for subsequent path planning and vehicle control.

For trail and boundary detection, the model achieves Accuracy values of 91.5%, 92.0%, and 91.8% across the three runs, giving an overall value of $91.8 \pm 0.2\%$. The Boundary IoU values are 32.2%, 32.5%, and 32.4%, with a mean of $32.4 \pm 0.12\%$. The small fluctuation in Boundary IoU is expected because thin and elongated trail boundaries are particularly sensitive to minor spatial displacements between the predicted and reference boundaries. Consequently, small changes in localization can produce measurable variations in the boundary overlap metric even when the overall trail prediction remains consistent.

The computational performance also shows negligible variation across the independent runs. The measured inference speeds of 90.9, 91.4, and 91.3 FPS result in an average of 91.2 ± 0.2 FPS. This demonstrates that the proposed architecture maintains a stable computational throughput in addition to its consistent perception performance.

Overall, the low standard deviations across all evaluated metrics demonstrate that DefenceVisionNet provides stable and reproducible results across different random initializations. The consistent behaviour of the detection, terrain segmentation, and trail and boundary branches indicates that the proposed multi-task learning framework is not strongly dependent on a particular training seed. Furthermore, the stable inference speed confirms that the model retains its real-time processing capability across independent runs. These findings provide additional evidence of the robustness of DefenceVisionNet for unified perception in complex and unstructured defence environments.

Ablation Studies

A comprehensive ablation analysis was performed on the Tactical Vehicle Dataset (TVD) to determine the contribution of the principal architectural and training components incorporated into DefenceVisionNet. The study considers two decoder configurations. In the first configuration, the drivable terrain and trail/boundary tasks utilize a common decoder, while the defence entity detection branch retains its dedicated detection head. In the second configuration, independent decoders are employed for the individual perception

tasks. The latter arrangement is intended to reduce feature competition between tasks and allow each decoder to learn task-specific representations more effectively.

The ablation experiments progressively examine the influence of backbone fine-tuning, Mosaic and MixUp augmentation, attention-guided feature refinement, multi-task loss optimization with cascade-based refinements, and separate task-specific decoders. The complete results are reported in Table 5.7. The baseline corresponds to the shared-decoder configuration without attention-guided refinement or task-aware loss optimization.

Table 5.7: Ablation Study of DefenceVisionNet on the Tactical Vehicle Dataset (TVD).

Configuration	**Speed (FPS)**	**Defence Object Detection (%)**		**Drivable Terrain (%)**	**Trail/Boundary (%)**	
		mAP@50	**Recall**	**Drivable mIoU**	**Accuracy**	**Boundary IoU**
Baseline (Shared decoder)	90.0	83.5	91.2	92.6	87.4	32.0
+ Backbone fine-tuning	90.4	83.7	91.5	93.0	88.6	32.1
+ Mosaic & MixUp augmentation	90.7	83.8	92.1	93.4	89.7	32.2
+ Attention-guided Feature Refinement only	90.9	84.1	92.6	93.8	90.8	32.3
+ Multitask Loss + Cascade-based refinements only	90.9	84.0	92.8	93.7	91.0	32.2
+ Attention-guided Feature Refinement + Multitask Loss + Cascade-based refinements (both)	91.0	84.3	93.0	94.0	91.3	32.3
+ Separate decoders	**91.2**	**84.5**	**93.2**	**94.1**	**91.8**	**32.4**

The baseline shared-decoder model achieves an mAP@50 of 83.5%, Recall of 91.2%, drivable terrain mIoU of 92.6%, trail Accuracy of 87.4%, and Boundary IoU of 32.0%, with an inference speed of 90.0 FPS. These results provide the reference point for assessing the contribution of the subsequent architectural and optimization modifications.

Introducing backbone fine-tuning produces a consistent improvement across all perception tasks. The mAP@50 increases from 83.5% to 83.7%, while Recall improves from 91.2% to 91.5%. The drivable terrain mIoU also increases from 92.6% to 93.0%, accompanied by an increase in trail Accuracy from 87.4% to 88.6%. These changes indicate that adapting the feature extractor to the characteristics of TVD improves the representation of defence-oriented visual content.

The subsequent incorporation of Mosaic and MixUp augmentation provides additional gains. The resulting configuration achieves 83.8% mAP@50 and 92.1% Recall, while the drivable terrain mIoU and trail Accuracy increase to 93.4% and 89.7%, respectively. Boundary IoU also improves slightly to 32.2%. These results demonstrate that enhanced variation during training can improve the model's ability to handle changes in object appearance, scale, viewpoint, and surrounding terrain.

The individual contribution of attention-guided feature refinement was then examined using the same baseline setting. When this component is introduced independently, the drivable terrain mIoU increases to 93.8% and trail Accuracy reaches 90.8%. Compared with the baseline, these correspond to improvements of 1.2 and 3.4 percentage points, respectively. The results indicate that attention-based spatial refinement is particularly beneficial for preserving relevant terrain and trail structures while reducing the influence of irrelevant background regions.

A separate experiment was conducted to evaluate the effect of the multi-task loss formulation together with cascade-based refinements. This configuration increases Recall from 91.2% to 92.8%, representing an improvement of 1.6 percentage points over the baseline. Improvements are also observed in drivable terrain mIoU and trail Accuracy, which reach 93.7% and 91.0%, respectively. The stronger improvement in Recall indicates

that adaptive task optimization and cascade-based refinement assist the network in learning difficult target instances, including partially visible and occluded objects.

The combined configuration incorporating both attention-guided feature refinement and multi-task loss with cascade-based refinements produces further improvements. The model achieves 84.3% mAP@50, 93.0% Recall, 94.0% drivable terrain mIoU, 91.3% trail Accuracy, and 32.3% Boundary IoU. Relative to the baseline, these results correspond to gains of 0.8, 1.8, 1.4, and 3.9 percentage points in mAP@50, Recall, drivable mIoU, and trail Accuracy, respectively. This improvement across multiple tasks demonstrates the complementary nature of spatial feature refinement and task-aware optimization.

Finally, separate task-specific decoders are introduced to the combined configuration. This produces the best overall results, with mAP@50 of 84.5%, Recall of 93.2%, drivable terrain mIoU of 94.1%, trail Accuracy of 91.8%, and Boundary IoU of 32.4%. The inference speed also increases to 91.2 FPS. The improvement obtained after separating the decoders indicates that independent task-specific processing can alleviate competition among the different prediction objectives while allowing each branch to retain features most relevant to its respective task.

The ablation results therefore demonstrate that the performance gain of DefenceVision-Net is not attributable to a single modification. Backbone adaptation and data augmentation improve the learned visual representation, attention-guided refinement enhances spatial feature preservation, and the combined multi-task loss with cascade-based refinement improves task optimization. The final use of separate decoders further strengthens task-specific learning and produces the highest performance across all evaluated metrics. The gradual and consistent improvement observed throughout the ablation sequence supports the effectiveness of the proposed design for multi-task perception in complex and unstructured defence environments.

Failure Case Analysis

A systematic analysis of unsuccessful predictions was performed to identify the principal factors responsible for detection errors in DefenceVisionNet. Figure 5.4 illustrates representative failure cases observed during evaluation. One notable error occurs when a soldier is located at a considerable distance from the camera and the visual characteristics of the individual, such as uniform patterns and other distinguishing features, are insufficiently resolved. Changes in illumination, viewing angle, and human pose can further reduce the discriminative information available to the network. Under such conditions, the model may incorrectly assign the target to the more general Person category instead of identifying it as a Soldier.

The analysis of these failure cases reveals several factors that can limit the reliability of defence-oriented perception. First, the discriminative visual characteristics required for fine-grained classification may be lost when targets are small, distant, partially occluded, or captured at low resolution. Second, the prediction can be influenced more strongly by local appearance than by the broader context of the scene, limiting the ability of the network to exploit contextual relationships between objects and their surroundings. Third, different categories may exhibit substantial visual similarity, making their separation difficult when appearance varies across locations, operational conditions, viewpoints, and image sources.

Therefore, the major failure modes identified in the analysis are associated with inadequate fine-grained visual information, limited contextual reasoning, and inter-class similarity under domain variation. Addressing these factors through improved multi-scale

representation, stronger contextual modeling, and complementary sensing modalities represents an important direction for enhancing defence perception reliability.

Limitations and Future Scope

Although DefenceVisionNet achieves strong performance across the evaluated perception tasks, several limitations remain. The current framework relies primarily on RGB imagery, which can significantly degrade in situations where visible-light information is limited. Nighttime operations, fog, smoke, and dust can reduce object visibility and introduce ambiguity in both detection and segmentation. A potential approach for improving robustness under such conditions is the integration of complementary sensing modalities, such as thermal cameras, LiDAR, and radar. Multimodal perception could provide additional information when RGB observations become unreliable and thereby improve the consistency of autonomous perception in adverse environments.

Another limitation concerns the computational evaluation of the proposed framework. Although DefenceVisionNet demonstrates real-time inference performance, the reported experiments were conducted using an NVIDIA Tesla V100 GPU. Consequently, the current results do not fully characterize the behaviour of the model on resource-constrained embedded computing platforms typically considered for autonomous vehicle deployment. Future work should therefore include systematic benchmarking on embedded hardware, with particular emphasis on memory consumption, computational latency, power requirements, and real-time throughput.

The adoption of separate decoders for individual perception tasks also introduces additional computational overhead compared with a fully shared decoding structure. While task-specific decoders improve specialization and reduce interference between heterogeneous prediction objectives, their additional computational requirements may become important for edge deployment. Future investigations will therefore explore adaptive decoder-sharing strategies in which parameters are selectively shared according to task similarity. Task-routing mechanisms can also be investigated to dynamically direct features to the most relevant processing branches. In addition, lightweight transformer-based decoder architectures may provide an alternative means of achieving task-specific feature modeling while controlling computational complexity.

Overall, future development of DefenceVisionNet will focus on three complementary directions: multimodal sensing for improved operation under adverse visibility, efficient implementation on resource-constrained embedded platforms, and adaptive task-specific decoding for reducing computational overhead. These developments can further improve the robustness, scalability, and deployment readiness of unified perception systems for autonomous defence vehicles.

5.3.3 Summary of the Important Results

This chapter presented the development and experimental evaluation of **DefenceVisionNet**, a unified multi-task deep learning framework for autonomous defence vehicle perception in challenging and unstructured environments. The proposed framework jointly performs defence–civilian entity detection, drivable terrain segmentation, and trail/boundary detection within a single network. To support the development and evaluation of the proposed framework, the **Tactical Vehicle Dataset (TVD)** was developed by integrating images from multiple publicly available sources and standardizing their annotations and class

definitions. The major findings obtained in this chapter are summarized as follows:

1. The **Tactical Vehicle Dataset (TVD)** was developed to provide a diverse benchmark for defence-oriented visual perception. The final dataset contains **20,758 images across 31 defence–civilian categories**, covering different vehicle types, personnel, aircraft, and other entities encountered in defence environments. The dataset incorporates variations in object scale, viewpoint, illumination, background complexity, occlusion, and environmental conditions.

2. The proposed **DefenceVisionNet** integrates complementary perception tasks within a unified architecture, including defence–civilian entity detection encompassing defence-related vehicles, military personnel, and civilian persons in challenging scenes, along with drivable-terrain segmentation and trail/boundary detection. The shared encoder and feature-fusion network minimize redundant computation, while dedicated task-specific decoders enable specialized feature processing and help mitigate interference among heterogeneous prediction objectives.

3. The proposed architecture incorporates an **E-ELAN-inspired feature extraction backbone, C3k2, SPPF, C2PSA, and FPN-based multi-scale feature fusion** to improve representation learning across different object scales and scene structures. Attention-guided feature refinement further enhances spatial feature discrimination, particularly in the presence of clutter, camouflage, partial occlusion, uneven terrain, and weak scene boundaries.

4. For **defence–civilian entity detection**, DefenceVisionNet achieves **84.5% mAP@50** and **93.2% recall**. Compared with the evaluated baseline detectors, the proposed framework provides improved recognition of visually similar and partially occluded defence entities. In particular, the model demonstrates improved robustness for challenging categories where conventional appearance-based detectors are affected by camouflage, clutter, and illumination variations.

5. For **drivable terrain segmentation**, the proposed framework achieves a **94.1% mIoU**. This result demonstrates the ability of DefenceVisionNet to accurately identify traversable terrain under unstructured conditions. The attention-guided feature refinement mechanism and multi-scale feature fusion contribute to improved preservation of spatial details and more accurate terrain boundaries in scenes containing uneven surfaces, vegetation, waterlogged regions, shadows, and weakly defined road boundaries.

6. For **trail and boundary detection**, DefenceVisionNet achieves an overall **91.8% accuracy** and **32.4% Boundary IoU**. The dedicated trail/boundary decoder, deeper semantic features, and attention-guided refinement improve the representation of thin and elongated structures. The results indicate improved localization of navigable trails and terrain boundaries in off-road environments where explicit route markings are often unavailable.

7. The **statistical validation** conducted using three independent training runs demonstrates the consistency of the proposed framework. DefenceVisionNet obtains **84.5 ± 0.2% mAP@50**, **93.2 ± 0.3% recall**, **94.1 ± 0.2% drivable-terrain mIoU**, **91.8**

± **0.2% trail accuracy**, and **32.4 ± 0.12% Boundary IoU**. The corresponding inference speed of **91.2 ± 0.2 FPS** further indicates stable computational performance across independent runs.

8. The **ablation study** demonstrates the progressive contribution of the proposed architectural and training components. Attention-guided feature refinement improves both terrain segmentation and trail detection, while the multi-task loss and cascade-based refinements improve detection recall. Combining these components provides further gains across all perception tasks. The final configuration with **separate task-specific decoders** achieves the best overall performance, reaching **84.5% mAP@50**, **93.2% recall**, **94.1% drivable-terrain mIoU**, **91.8% trail accuracy**, and **32.4% Boundary IoU**.

9. The final DefenceVisionNet configuration contains approximately **41.2 million parameters** and achieves an inference speed of **91.2 FPS** at an input resolution of 640×640. Despite the additional capacity introduced by attention mechanisms and task-specific decoders, the framework maintains real-time processing capability, demonstrating its potential for deployment in autonomous defence vehicle perception systems.

10. Comparative evaluation with existing multi-task and single-task perception approaches, including **YOLOP, HybridNets, YOLOPv2, YOLOv11s, MultiNet, DLT-Net, and Faster R-CNN**, demonstrates the effectiveness of the proposed framework. DefenceVisionNet achieves the highest reported performance across the evaluated defence entity detection, drivable terrain segmentation, and trail/boundary detection tasks while maintaining real-time inference capability.

Overall, the experimental results establish that **DefenceVisionNet** provides an effective unified solution for multi-task perception in autonomous defence vehicles operating in unstructured environments. The combination of shared feature extraction, multi-scale feature fusion, attention-guided refinement, task-specific decoders, and cascade-based multi-task optimization enables the framework to jointly address object recognition, terrain understanding, and trail/boundary perception. The proposed framework achieves **84.5% mAP@50** for defence–civilian entity detection, **94.1% mIoU** for drivable terrain segmentation, and **91.8% accuracy** with **32.4% Boundary IoU** for trail/boundary detection, while maintaining a reported inference speed of **91.2 FPS**. The developed TVD dataset further provides a diverse foundation for evaluating defence-oriented autonomous perception under visually complex and operationally challenging conditions.

CHAPTER 6

Secure Model Deployment and Embedded Autonomous Perception Systems

---

This chapter presents the final phase of the research, focusing on the secure deployment and real-time embedded implementation of autonomous perception systems. Building upon the deep learning-based perception frameworks developed in the preceding chapters, this phase extends the research toward two complementary aspects: secure deployment of deep learning models and embedded realization of autonomous driver assistance systems.

The first part investigates post-quantum security mechanisms for protecting the authenticity and integrity of deep learning models in defence and intelligent transportation applications. The proposed QR-DefenceDrive framework employs ML-DSA-based digital signatures for cryptographic model verification in defence-oriented autonomous perception systems, while Q-DriveShieldNet integrates model-integrity protection with a unified multi-task perception architecture for intelligent transportation and autonomous driving environments.

The second part addresses the practical realization of autonomous perception on resource-constrained hardware through the development of a low-cost embedded Advanced Driver Assistance System (ADAS) prototype using a Raspberry Pi platform. The prototype combines vision-based perception, LiDAR-based obstacle detection, and an interactive Human–Machine Interface (HMI) to support real-time environmental awareness. This implementation demonstrates the feasibility of deploying autonomous perception capabilities on a low-cost embedded platform for real-time environmental monitoring and driver assistance.

Thus, the research presented in this chapter addresses the transition from deep learning-based perception development toward secure and embedded realization of autonomous systems. The work encompasses two complementary research directions: post-quantum protection of deployed deep learning models for defence and intelligent transportation applications, and real-time implementation of autonomous perception on low-cost embedded hardware.

6.1 Secure Mullti-task Deep learning Autonomous Perception Systems

Autonomous driving and Intelligent Transportation Systems (ITS) have advanced considerably with the development of deep learning-based perception and connected vehicular

technologies. Modern Advanced Driver Assistance Systems (ADAS) increasingly employ multi-task perception frameworks to perform several functions simultaneously. Sharing feature representations across multiple tasks can reduce computational redundancy and facilitate real-time operation. However, reliable perception remains challenging in unstructured environments characterized by damaged road surfaces, potholes, faded lane markings, adverse weather, low illumination, and dense traffic.

In addition to perception capability, the secure deployment of deep learning models has become an important consideration for connected and autonomous systems. Modern ITS increasingly rely on distributed computing, vehicle-to-network connectivity, cloud–edge infrastructures, and over-the-air (OTA) updates. These mechanisms provide flexibility in model distribution and system maintenance but also introduce potential security risks, including unauthorized model modification, replacement, and manipulation. Such threats are particularly significant when perception models are employed in applications where their outputs influence safety-critical decisions.

The security of deep learning models has been investigated through different approaches. Supply-chain attacks have demonstrated that malicious modifications can be introduced into machine learning models during their development or distribution [91]. Neural network watermarking has subsequently been explored for establishing ownership and identifying unauthorized use of trained models. Uchida *et al.* [92] proposed a method for embedding watermarks into deep neural networks, while the DeepMarks framework introduced a fingerprinting mechanism for model ownership verification [33]. Guo and Potkonjak [93] further investigated watermarking of deep neural networks for embedded systems. Although these approaches provide useful mechanisms for ownership protection and model identification, their primary objective differs from direct cryptographic verification of the authenticity and integrity of a deployed model artifact.

The emergence of quantum computing introduces an additional dimension to the security of connected autonomous systems. Shor's algorithm demonstrated that sufficiently powerful quantum computers could efficiently solve integer factorization and discrete logarithm problems, which form the basis of several widely used public-key cryptographic schemes [94]. Consequently, post-quantum cryptography (PQC) has emerged as a research direction for developing cryptographic mechanisms capable of resisting potential quantum attacks [95]. While PQC has been extensively investigated for secure communication and authentication, its application to the authentication and integrity protection of deep learning model artifacts is comparatively recent.

Recent work has begun to investigate this direction. Castiglione *et al.* [96] examined post-quantum digital signatures for enhancing trust in deep learning models and evaluated schemes including ML-DSA, Falcon, and SPHINCS+ in the context of distributed learning. Such studies demonstrate the feasibility of employing post-quantum signatures for protecting deep learning models; however, secure model deployment in autonomous perception introduces additional requirements related to model verification, computational overhead, and real-time operation.

These observations motivate the investigation of secure deep learning model deployment in autonomous and defence-oriented perception systems. In this research, the *QR-DefenceDrive* framework is developed for defence-oriented applications, employing the ML-DSA digital signature algorithm to provide cryptographic authentication and integrity verification of trained deep learning models. The framework enables verification of a model before deployment or execution and provides a mechanism for identifying unauthorized modifications.

The security perspective is also investigated in the context of intelligent transportation and autonomous driving through the development of *Q-DriveShieldNet*. Unlike the defence-oriented *QR-DefenceDrive* framework, *Q-DriveShieldNet* combines model-integrity protection with a unified multi-task perception architecture for driving environments. The framework is designed to jointly address comprehensive perception, multi-task optimization, computational efficiency, and secure model deployment.

Alongside secure model development, practical deployment of autonomous perception requires consideration of hardware and computational constraints. Deep learning models developed and evaluated on high-performance computing platforms cannot always be directly deployed on low-cost embedded devices. This motivates the development of a Raspberry Pi-based ADAS prototype incorporating vision-based perception, LiDAR-based obstacle detection, and an interactive Human–Machine Interface (HMI). The prototype provides a practical demonstration of real-time autonomous assistance using resource-constrained embedded hardware.

6.2 Embedded Realization of the Autonomous Driver Assistance System

The development of accurate perception models represents an important stage in an autonomous driver assistance system; however, practical deployment requires the models to operate reliably under the computational, memory, and power limitations of embedded platforms. High-performance computing systems can provide substantial processing capacity during model development, but their computational and energy requirements make them unsuitable for many low-cost or portable applications. Therefore, this work further investigates the embedded realization of autonomous perception through the development of a compact Advanced Driver Assistance System (ADAS) prototype.

The proposed embedded system [97], is designed primarily as an assistive platform for elderly and vulnerable drivers, for whom timely perception of the surrounding environment can be particularly important. Age-related changes in driving behaviour and takeover performance have been reported in partially automated driving scenarios [98]. Instead of attempting to reproduce the complete functionality of a production-level autonomous vehicle, the developed prototype focuses on a set of essential perception and assistance functions, namely lane detection, traffic sign recognition, traffic light recognition, and obstacle detection. These functions are implemented using a combination of vision-based processing and Time-of-Flight (ToF) ranging.

A Raspberry Pi 4B is employed as the principal embedded computing platform because of its compact form factor, relatively low power requirement, and ability to interface with cameras, displays, and external sensors. The platform executes the perception and decision logic while an Arduino Mega is used for low-level vehicle actuation. The resulting architecture separates high-level perception from low-level motor control, thereby simplifying the overall system and allowing individual modules to be modified independently.

Figure 6.1 illustrates the overall architecture of the developed embedded ADAS. The camera provides continuous visual input to the Raspberry Pi, where lane and traffic information are extracted. In parallel, the LiDAR sensors provide short-range distance measurements for obstacle detection. The processed information is subsequently used to generate assistance or vehicle-control commands. Relevant perception outputs are

presented through a touchscreen-based Human–Machine Interface (HMI), while low-level movement commands are transferred to the Arduino Mega for actuation.

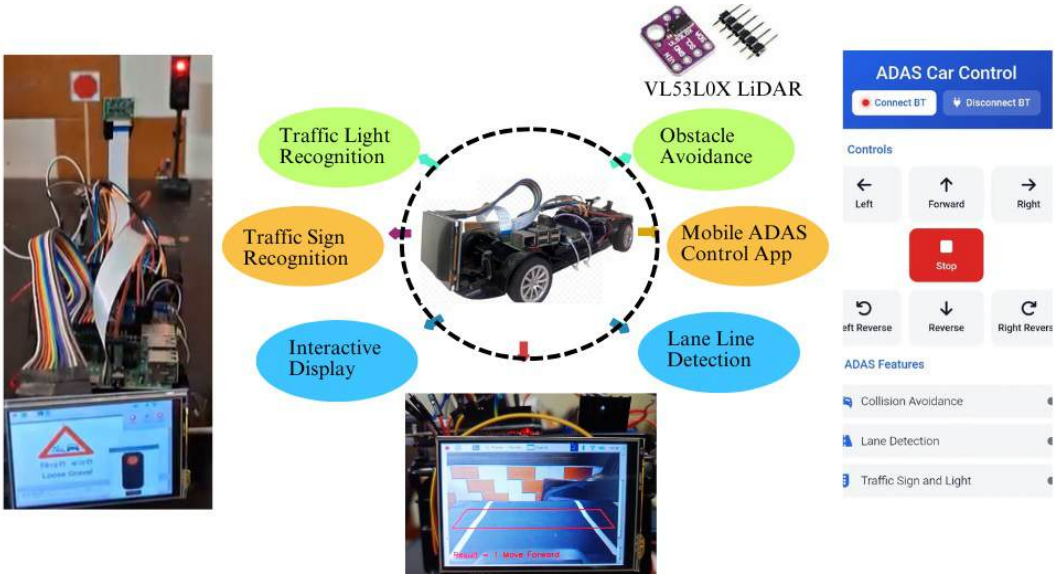

Figure 6.1: Overall architecture of the embedded autonomous driver assistance system integrating vision, LiDAR, embedded processing, vehicle control, and interactive feedback.

6.2.1 Hardware Architecture

The embedded prototype consists of a Raspberry Pi 4B, Pi Camera, VL53L0X ToF LiDAR sensors, Arduino Mega, motor driver, DC motors, steering servo, touchscreen display, and supporting power circuitry. The hardware arrangement is designed to provide a clear separation between perception, decision-making, and actuation.

The Raspberry Pi 4B serves as the central processing unit and executes the computer-vision and deep-learning components of the system. Its technical characteristics make it suitable for prototyping embedded vision applications where computational resources are more limited than those available on desktop or server-class GPUs [99]. A 5-megapixel Pi Camera is mounted at the front of the scaled vehicle platform and provides the visual stream required for lane and traffic-element analysis.

Short-range obstacle sensing is implemented using VL53L0X ToF sensors. Unlike conventional ultrasonic ranging, the VL53L0X determines distance using infrared time-of-flight measurements and provides digital distance information through the I^2C interface [100]. Multiple sensors are employed to increase the frontal sensing coverage. Their addresses are configured individually through the XSHUT control pins, allowing the sensors to operate simultaneously on the same I^2C bus.

The Arduino Mega performs the low-level control functions. Commands generated by the Raspberry Pi are transmitted to the microcontroller, which subsequently controls the drive motors and steering mechanism through the motor driver. This division of processing reduces the dependency of the low-level actuation layer on the computational workload of the vision-processing unit.

A 3.5-inch TFT touchscreen with a resolution of 320×480 pixels is integrated into the prototype to provide real-time visual feedback. The display presents the detected traffic information and lane status in a simplified form so that the driver can interpret the

information without requiring attention to a conventional computer interface. The use of visual cues and enlarged display elements is particularly appropriate for an assistive driving prototype intended to minimize unnecessary interaction with the system.

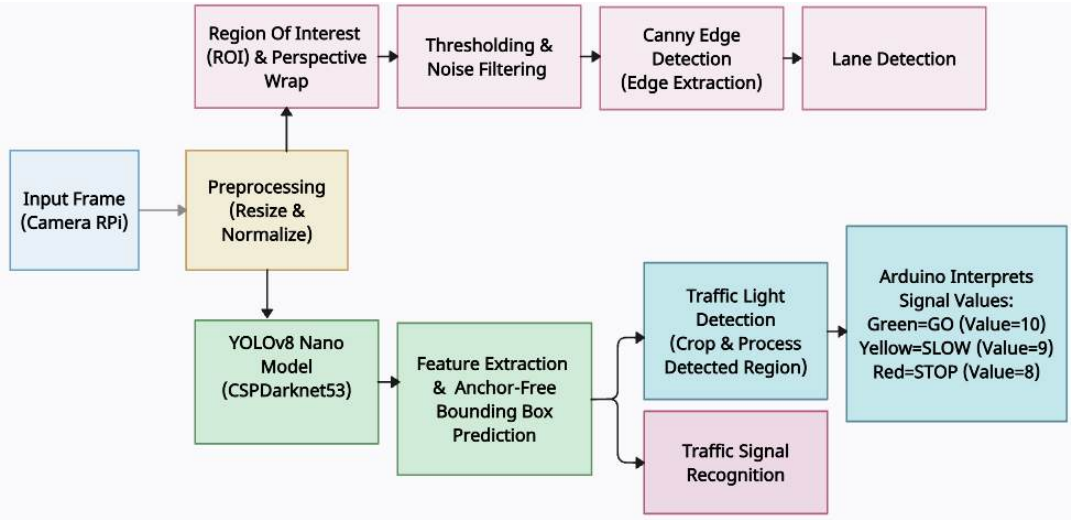

Figure 6.2: Processing flow of the vision-based perception module for lane detection, traffic light recognition, and traffic sign recognition.

6.2.2 Vision-Based Perception Pipeline

The vision subsystem as shown in Fig. 6.2 provides the principal source of environmental information for the embedded ADAS. A forward-facing camera continuously captures road scenes, which are processed to extract lane boundaries and identify relevant traffic signs and traffic lights. The perception pipeline is intentionally designed to combine conventional image processing with a lightweight deep learning detector. This combination reduces the computational burden while retaining the ability to recognize semantically meaningful traffic elements.

The complete visual processing pipeline is divided into two parallel functional branches. The first branch estimates the lane position using image-processing operations, whereas the second branch performs traffic sign and traffic light detection using a customized YOLOv8-nano model. The outputs from both branches are subsequently displayed on the embedded HMI and, where required, converted into control commands.

6.2.3 Lane Detection and Lane-Center Estimation

Lane estimation is used to determine the position of the vehicle with respect to the visible road boundaries. In the proposed implementation, a lightweight image-processing approach is adopted instead of a computationally intensive deep neural network. This design choice is motivated by the limited processing capability of the Raspberry Pi and the requirement for continuous frame-level operation.

The captured image is first restricted to a predefined Region of Interest (ROI) containing the road surface. ROI selection reduces the influence of irrelevant regions such as the sky, buildings, and roadside structures and consequently decreases the amount of image data requiring further processing. ROI-based approaches have also been investigated for improving the efficiency of lane detection [101].

Following ROI extraction, a perspective transformation is applied to obtain a bird's-eye representation of the road. This transformation reduces the apparent perspective distortion of lane markings and provides a more convenient representation for estimating their positions. Gaussian filtering is subsequently employed to suppress high-frequency image noise. Canny edge detection is then applied to identify strong intensity transitions corresponding to lane markings.

The resulting edge representation is converted into a binary form using a predefined threshold. A histogram of the processed region is used to estimate the dominant positions of the left and right lane boundaries. The midpoint between the estimated boundaries provides the lane center. The lateral displacement of the vehicle is obtained by comparing the estimated lane center with the geometric center of the camera frame.

The resulting displacement is mapped to a small set of steering commands. When the estimated vehicle position is approximately aligned with the lane center, a forward command is generated. Positive or negative lateral displacement results in corresponding left or right steering commands. This simplified control representation reduces the computational requirements of the embedded system while providing continuous lane-following capability. The overall lane-detection pipeline is shown in Fig.6.3

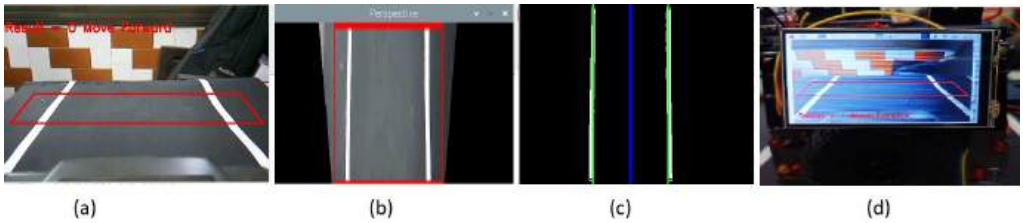

Figure 6.3: Embedded lane-detection pipeline showing the input road image, perspective transformation, lane-center estimation, and corresponding control output.

6.2.4 Traffic Sign and Traffic Light Recognition

Traffic signs and traffic lights provide important semantic information for driver assistance because their interpretation can directly influence the vehicle's intended movement. A compact YOLOv8-nano detector is therefore employed to perform traffic sign and traffic light recognition within a common detection pipeline. The use of a single detector avoids maintaining independent inference pipelines for different traffic elements and consequently simplifies embedded deployment.

The model receives resized camera frames and produces bounding boxes, class labels, and confidence values for the detected objects. The traffic-light classes correspond to the red, yellow, and green signal states, while the traffic-sign classes represent commonly encountered Indian road signs. The detector was trained using the Indian Traffic Sign dataset [49], comprising approximately 6,480 images and 7,056 traffic-sign instances across 87 categories. The data include variations in illumination, road type, weather, camera viewpoint, occlusion, and background complexity. Additional samples from the IRTSD [102] and Indian road traffic-light dataset [103] were incorporated to increase environmental diversity.

The complete dataset was divided into training, validation, and testing subsets using a 70:15:15 distribution. All images were converted to YOLO-compatible annotations and resized to 320×320 pixels for embedded inference. During training, augmentation

operations including horizontal flipping, brightness variation, simulated rain, glare, and partial occlusion were employed to improve the ability of the detector to handle variations encountered in practical road scenes.

The traffic-light output is further interpreted at the decision level. Detection of a red signal produces a stop condition, while yellow corresponds to a cautionary or slow-down state and green permits forward movement. These decisions are transmitted to the Arduino Mega when automatic control is enabled. Simultaneously, the detected class and associated visual information are presented on the touchscreen. Thus, the perception and HMI layers remain connected without requiring the driver to interpret raw camera frames. The overall results are shown in Fig.6.4

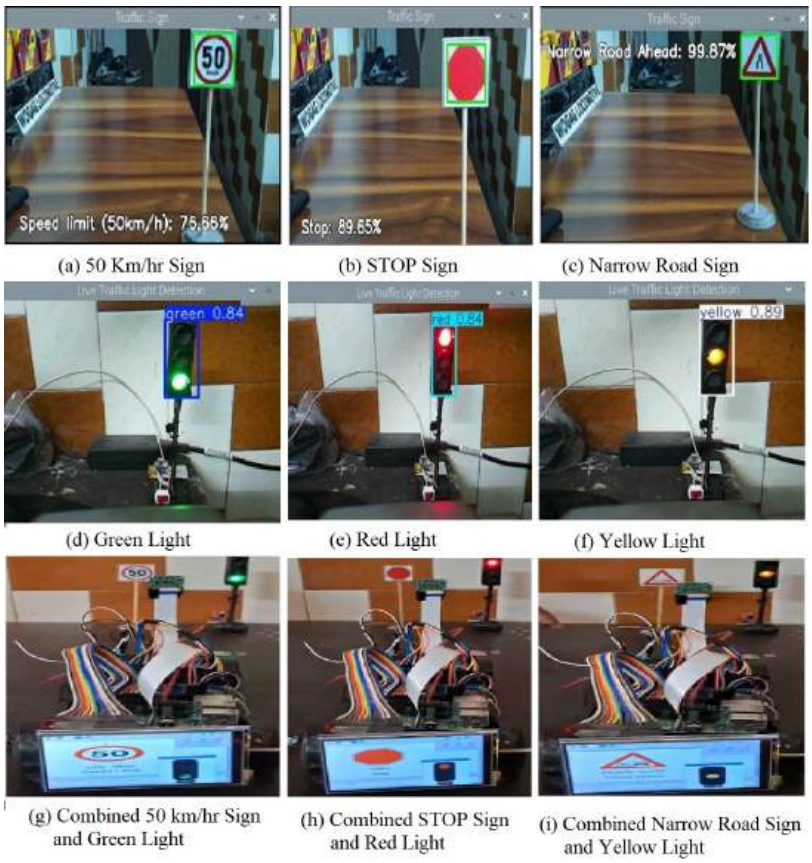

Figure 6.4: Real-time visualization of traffic sign and traffic light detections on the embedded touchscreen interface.

6.2.5 LiDAR-Based Obstacle Detection and Collision Avoidance

Camera-based perception provides semantic information about the road environment, but visual detection alone may not provide sufficiently reliable short-range distance estimation for collision avoidance on a resource-constrained prototype. To address this limitation, the system incorporates three VL53L0X ToF LiDAR sensors for direct distance measurement.

The sensors are positioned at the front of the prototype to provide coverage across different lateral directions. Each sensor is assigned an individual I^2C address using its XSHUT control line. The Arduino Mega continuously reads the distance measurements and compares them with a predefined safety threshold. When an object enters the specified

detection range, a collision-avoidance condition is generated and the corresponding halt command is issued.

For the experimental prototype, the effective obstacle-detection range is configured up to approximately 2 m. The use of multiple short-range distance sensors provides an additional sensing layer independent of the camera-based detector. This sensor redundancy is particularly useful in situations where visual information is affected by lighting, texture, or partial occlusion. ToF-based sensing has also been investigated in autonomous-vehicle perception and localization systems, supporting its use as a complementary sensing modality [104].

The obstacle-detection subsystem is deliberately separated from the deep-learning inference pipeline. Consequently, an obstacle can trigger a safety response even when the vision model is processing a complex frame or temporarily fails to identify the object. This architecture provides a simple but effective safety layer for the embedded prototype.

6.2.6 Embedded Model Optimization and Deployment

The trained traffic perception model is initially developed and validated using GPU-enabled computing resources. Direct execution of the complete training configuration on the Raspberry Pi is not practical because of the substantially lower computational capacity of the embedded platform. Therefore, the trained model is prepared for deployment through input-size reduction and numerical optimization.

The inference resolution is reduced to 320×320 pixels, which decreases the computational cost of processing each frame. In addition, INT8 quantization is applied to reduce the memory footprint and computational requirements of the deployed model. Quantization represents model parameters and intermediate computations using lower-precision numerical representations, thereby reducing the amount of data transferred and processed during inference.

The optimized model is subsequently executed on the Raspberry Pi using the embedded software stack. Camera acquisition is handled through the Pi Camera interface, while OpenCV-based operations are employed for image preprocessing and lane extraction. The deep-learning inference component is integrated with the remaining perception and control modules so that the system operates as a continuous processing loop.

The deployment architecture therefore separates the computationally intensive training stage from the resource-constrained inference stage. Training, model evaluation, and optimization are conducted using GPU-enabled resources, whereas the final inference model is executed locally on the Raspberry Pi. This approach avoids dependence on continuous cloud connectivity and demonstrates the feasibility of performing essential ADAS functions directly at the edge.

6.2.7 Integrated Operational Workflow

During operation, the embedded ADAS follows a continuous sensing, perception, decision, and feedback cycle. The Pi Camera first captures the forward road scene, while the LiDAR sensors simultaneously measure the distance to nearby obstacles. The captured image is processed by the lane-detection pipeline and the customized YOLOv8-nano detector. The former estimates the lane position, whereas the latter identifies traffic signs and traffic-light states.

The outputs of these perception modules are combined with the LiDAR measurements at the decision level. Lane deviation determines the corresponding steering direction, traffic-light recognition provides signal-dependent driving commands, and the LiDAR subsystem maintains the collision-avoidance condition. When an obstacle is detected within the predefined safety range, the obstacle-avoidance decision takes priority over normal movement commands.

The resulting information is simultaneously communicated to the actuation layer and the HMI. The Arduino Mega translates valid high-level commands into motor and steering actions, whereas the touchscreen provides the driver with the current perception status. This processing arrangement establishes a closed-loop embedded assistance system in which sensing, perception, decision-making, actuation, and driver feedback are performed continuously.

6.3 Experimental Evaluation of the Embedded ADAS

The developed prototype was evaluated with respect to the performance of the individual perception modules and the real-time behaviour of the integrated embedded system. The evaluation considered lane tracking, traffic sign and traffic light detection, obstacle response, inference speed, and computational resource utilization.

6.3.1 Lane Tracking Performance

The lane-following module maintained stable tracking for both straight and moderately curved road segments. The image-processing pipeline operated at approximately 20 frames per second on the Raspberry Pi 4B, demonstrating that continuous lane estimation can be achieved without using a computationally intensive deep-learning lane detector.

The measured lane-tracking precision was approximately 94.6%, with a normalized positional error of approximately 3.2%. The average command generation latency was approximately 115 ms per decision cycle, while memory consumption remained below 400 MB during continuous operation.

The obtained performance demonstrates the suitability of the lightweight lane-processing approach for low-cost embedded hardware. Although deep-learning approaches such as Ultra Fast Lane Detection [105] and PolyLaneNet [106] can achieve substantially higher frame rates on high-performance computing platforms, their computational requirements are considerably different from those of a Raspberry Pi-based system. The objective of the present implementation is therefore not to maximize benchmark throughput, but to obtain an appropriate balance between accuracy, latency, and embedded resource utilization.

6.3.2 Traffic Sign and Traffic Light Detection Performance

The customized YOLOv8-nano detector was evaluated using precision, recall, and mean Average Precision at an Intersection over Union (IoU) threshold of 0.50 (mAP@50). The model achieved an average precision of 94.0%, recall of 92.4%, and mAP@50 of 95.5% over the evaluated traffic-sign and traffic-light classes.

For a more detailed comparison, Table 6.1 presents the class-wise results obtained using the proposed YOLOv8-nano model and three lightweight baseline detectors. The results show that the customized YOLOv8-nano provides consistently higher performance across

the evaluated categories. The comparison indicates that the customized YOLOv8-nano

Table 6.1: Performance comparison of the customized YOLOv8-nano with lightweight object-detection models.

No.	Class	YOLOv8-n			YOLOv5s			MobileNetV2			EffDet-Lite0		
		P	R	mAP	P	R	mAP	P	R	mAP	P	R	mAP
1	Green L	0.957	0.967	0.976	0.930	0.940	0.952	0.901	0.887	0.910	0.892	0.874	0.903
2	Red L	0.961	0.974	0.988	0.936	0.949	0.965	0.913	0.892	0.918	0.900	0.881	0.910
3	Yellow L	0.923	0.955	0.986	0.907	0.925	0.940	0.882	0.860	0.895	0.870	0.845	0.885
4	Ped Cross	0.930	0.910	0.950	0.905	0.881	0.921	0.870	0.842	0.881	0.861	0.830	0.870
5	Hospital	0.950	0.940	0.970	0.926	0.910	0.948	0.890	0.865	0.900	0.878	0.855	0.892
6	Stop	0.950	0.920	0.960	0.922	0.896	0.940	0.888	0.860	0.892	0.875	0.848	0.884
7	Speed-20	0.940	0.890	0.940	0.910	0.882	0.917	0.862	0.835	0.870	0.850	0.821	0.863
8	Child-X	0.910	0.890	0.920	0.884	0.855	0.902	0.840	0.812	0.852	0.832	0.803	0.846
All	**Average**	**0.940**	**0.924**	**0.955**	**0.903**	**0.895**	**0.942**	**0.868**	**0.844**	**0.890**	**0.857**	**0.845**	**0.882**

provides an overall mAP@50 of 95.5%, compared with 94.2% for YOLOv5s, 89.0% for MobileNetV2, and 88.2% for EfficientDet-Lite0. The improvement is particularly evident for small traffic-light targets, where both localization accuracy and class discrimination are important. The results support the selection of YOLOv8-nano as the primary detection architecture for the embedded prototype.

The model also maintained an inference rate of approximately 20 fps on the Raspberry Pi 4B under the experimental configuration. This result is significant from an embedded-deployment perspective because the objective is to achieve continuous perception with limited hardware resources rather than maximize detection speed on a GPU platform.

6.3.3 Training Convergence and Generalization

The training behaviour of the customized YOLOv8-nano model is shown in Figure 6.5. The training and validation losses exhibit a progressive reduction over the training period, indicating stable optimization. Precision and recall increase during the initial stages and subsequently stabilize, while the mAP@50 metric reaches a value above 0.94.

The close behaviour of the training and validation curves indicates that the model does not exhibit a substantial divergence between the training and validation data. This behaviour is desirable for deployment because the detector must operate across road scenes that differ from the individual training samples in illumination, background, viewpoint, and object appearance.

6.3.4 Obstacle Detection Performance

The LiDAR-based obstacle-detection subsystem was evaluated using stationary objects and slowly moving objects placed within the configured sensing range. The system achieved an obstacle-detection accuracy of approximately 98.2% for static objects and 95.4% for slow-moving objects. The average response latency was approximately 110 ms.

The measured response time indicates that the sensing layer can provide a rapid safety response independently of the computationally heavier vision pipeline. The sensors provide approximately 90° of frontal coverage in the implemented arrangement, while the measured power consumption remains below 300 mW. These characteristics make the module suitable for integration into a low-power embedded assistance platform.

The results also demonstrate the benefit of combining complementary sensing modalities. Vision-based perception provides semantic information such as traffic signs, traffic

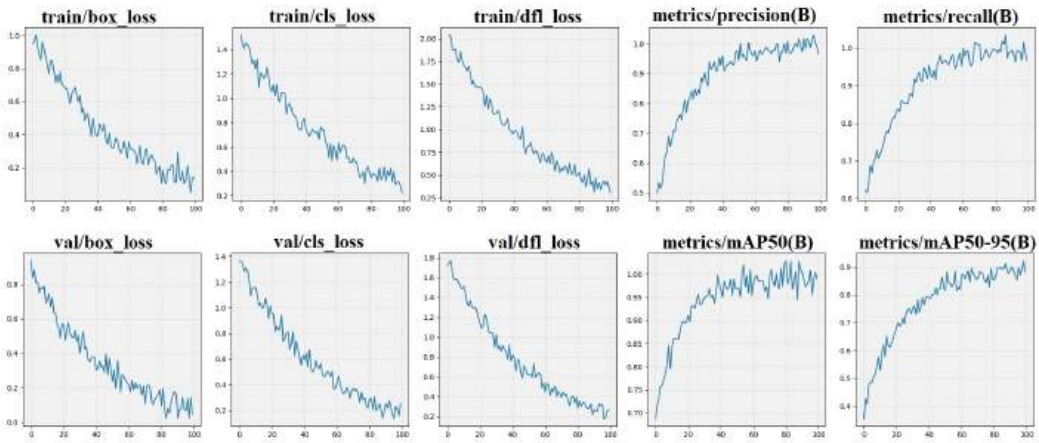

Figure 6.5: Training and validation performance curves of the customized YOLOv8-nano traffic perception model.

lights, and lane markings, whereas LiDAR provides direct short-range distance information. The combination therefore improves functional coverage without requiring a large number of computationally demanding perception models.

6.4 Discussion

The experimental evaluation demonstrates that essential ADAS functions can be implemented on a low-cost embedded platform when perception algorithms are selected according to the available computational resources. The proposed system does not depend on a single sensing modality. Instead, visual perception and short-range LiDAR sensing are assigned complementary roles. The camera provides semantic understanding of the road scene, whereas LiDAR provides an independent distance-based safety mechanism.

The results also demonstrate the importance of hardware-aware model selection. Although larger detection architectures can provide strong accuracy on GPU-based systems, their computational requirements can limit their applicability on low-cost embedded platforms. The customized YOLOv8-nano model provides a practical compromise between detection accuracy and inference speed, achieving an mAP@50 of 95.5% while maintaining approximately 20 fps on the Raspberry Pi 4B.

Another important aspect of the implementation is the modularity of the architecture. Perception, ranging, control, and HMI functions are implemented as separate modules connected through defined interfaces. Consequently, individual components can be replaced or upgraded without redesigning the complete system. For example, the lightweight traffic detector can subsequently be replaced by one of the unified multi-task perception frameworks developed in the preceding chapters. Similarly, additional sensors can be incorporated into the sensing layer without fundamentally changing the HMI or low-level actuation architecture.

From a practical perspective, the developed prototype demonstrates the transition from algorithmic development to embedded realization. The system provides a compact experimental platform for evaluating real-time perception, sensor integration, driver feedback, and hardware-constrained inference. This embedded realization therefore complements the secure model-deployment studies presented elsewhere in this chapter by

demonstrating how autonomous perception models can be translated into an operational edge-computing system.

6.4.1 Summary of the Important Results

This chapter presented the secure deployment and embedded realization of autonomous perception systems for defence and intelligent transportation applications. The chapter investigated two complementary aspects, namely post-quantum protection of deep learning models and practical implementation of autonomous driver assistance on resource-constrained embedded hardware. The major findings obtained from the embedded ADAS implementation are summarized as follows:

1. A **low-cost embedded ADAS prototype** was developed using a **Raspberry Pi 4B** as the primary processing platform. The prototype integrates camera-based visual perception, LiDAR-based obstacle sensing, vehicle control, and an interactive Human–Machine Interface (HMI) within a unified architecture. The implementation demonstrates the feasibility of realizing essential driver-assistance functions on resource-constrained hardware.

2. The embedded perception system combines three complementary functions, namely **lane detection, traffic sign and traffic light recognition, and obstacle detection**. The modular architecture enables these functions to operate independently while sharing the embedded processing platform. This design reduces dependence on a single computationally intensive perception model and facilitates subsequent replacement or enhancement of individual modules.

3. The **lane-detection module** provides reliable estimation of the lane centre and corresponding steering direction under the evaluated road conditions. The implemented pipeline based on region of interest selection, perspective transformation, image preprocessing, edge extraction, and lane-centre estimation achieves a **lane-tracking precision of 94.6%**. The normalized positional error is approximately **3.2%**, indicating accurate estimation of the vehicle's lateral position relative to the detected lane.

4. The lane-processing module maintains approximately **20 FPS** on the Raspberry Pi 4B, with a command response time of approximately **115 ms**. These results demonstrate that conventional image-processing operations can provide sufficiently responsive lane-following assistance without imposing substantial computational requirements on the embedded processor.

5. For **traffic sign and traffic light recognition**, a customized lightweight **YOLOv8-nano** model was trained using a combined traffic-sign and traffic-light dataset. The model achieves a **94.0% precision**, **92.4% recall**, and **95.5% mAP@50**. The results demonstrate that a compact detection architecture can provide accurate recognition of relevant traffic-control elements while remaining suitable for embedded deployment.

6. Comparative evaluation with lightweight object-detection architectures shows that the customized **YOLOv8-nano** provides the highest average detection performance among the evaluated models. It achieves an average **mAP@50 of 95.5%**, compared

with **94.2% for YOLOv5s**, **89.0% for MobileNetV2**, and **88.2% for EfficientDet-Lite0**. This indicates a favorable balance between detection accuracy and model complexity for the considered embedded ADAS application.

7. The traffic perception module was evaluated under different visual conditions using augmentation strategies including brightness variation, synthetic rain, glare, horizontal flipping, and partial occlusion. The resulting model demonstrated consistent recognition of traffic signs and traffic lights, including visually similar categories and different traffic-light states. The detected information is subsequently presented on the embedded touchscreen to provide immediate driver feedback.

8. Integration of **LiDAR and vision-based sensing** in the embedded ADAS prototype, achieving approximately **98.2%** accuracy for static obstacles and **95.4%** for slowly moving obstacles within 2 m, with a response latency of approximately **110 ms**. The complementary sensing improves environmental awareness while reducing the need for computationally intensive deep-learning processing.

9. The deep learning model was optimized for embedded execution through a lightweight network configuration and **INT8 quantization**. The deployed perception pipeline operates with a memory requirement of less than approximately **400 MB**, demonstrating that the proposed implementation can operate within the computational and memory limitations of the Raspberry Pi 4B.

10. An interactive **Human–Machine Interface (HMI)** was implemented using a touchscreen display to present detected traffic signs, traffic-light states, and relevant system information to the driver. The interface provides immediate visual feedback while maintaining a simple and compact presentation suitable for an embedded ADAS environment. A complementary mobile interface was also developed to support manual vehicle-control commands.

11. The complete prototype integrates **sensing, perception, decision-making, actuation, and driver feedback** within a single operational loop. Camera input is used for visual perception, LiDAR measurements provide obstacle information, the embedded processor performs the required computations, and the resulting decisions are communicated through the vehicle-control and HMI subsystems. The integrated implementation demonstrates the practical feasibility of the proposed perception-to-assistance pipeline.

12. Overall, the experimental evaluation demonstrates that the developed embedded ADAS can provide essential autonomous assistance functions while operating on a **low-cost Raspberry Pi 4B platform**. The combination of **94.6% lane-tracking precision**, **95.5% mAP@50 for traffic perception**, approximately **98.2% static-obstacle detection accuracy**, and low response latency demonstrates the effectiveness of the developed prototype for real-time assistive driving applications.

CHAPTER 7
Conclusion, Future Directions, and Social Impact

*In conclusion, this thesis presents a comprehensive investigation into **deep learning-based autonomous perception for driver assistance in unstructured environments**. The research addresses the limitations of conventional perception systems through the development of **unified, computationally efficient, and secure multi-task perception frameworks**, dedicated benchmark datasets, domain-specific perception models, post-quantum model authentication mechanisms, and an embedded ADAS prototype.*

*The developed perception frameworks address multiple critical tasks, including **traffic-object detection, traffic sign and traffic-light recognition, road-defect detection, lane detection, and drivable-area segmentation**, thereby providing comprehensive environmental understanding. The research progresses from **Indian traffic perception to road-defect detection, tactical vehicle perception, secure model deployment, and real-time embedded implementation**, demonstrating the applicability of the proposed approaches across both civilian and autonomous defence environments.*

*The experimental results demonstrate that the developed frameworks can achieve high perception accuracy while maintaining real-time inference capability. Furthermore, the incorporation of **post-quantum model authentication** strengthens the security and trustworthiness of model deployment, while the embedded implementation demonstrates the practical feasibility of the proposed approaches on resource-constrained platforms.*

The outcomes of this thesis provide a strong foundation for the development of accurate, robust, secure, computationally efficient, and practically deployable autonomous perception systems for next-generation intelligent transportation and autonomous defence applications.

7.1 Summary of the Work Done in the Thesis

The research work presented in this thesis was carried out through multiple phases, beginning with the investigation of existing multi-task driving perception systems and progressing towards the development and deployment of unified perception frameworks for unstructured environments. The major contributions of the research are summarized below.

The first phase focused on the development of a **unified multi-task perception framework for Indian road environments**. Existing autonomous driving perception approaches such as YOLOP, HybridNets, and YOLOPv2 provide simultaneous perception capabilities; however, their performance may degrade when applied to heterogeneous Indian traffic conditions. Indian roads are characterized by high traffic density, diverse

vehicle types, irregular traffic behaviour, poorly defined lane boundaries, road defects, occlusions, varying illumination, and adverse weather conditions. Therefore, a dedicated **Indian Traffic and Adverse Weather Dataset (ITAWD)** was developed to address the lack of a comprehensive dataset representing these conditions. The dataset contains **24,236 images belonging to 47 classes**, collected from approximately 30 km of driving sequences in Mumbai. The images were subsequently annotated and organized to support multi-task traffic perception.

To address the identified challenges, **SmartDriveNet** was developed as a unified camera-based multi-task perception framework. The proposed framework integrates enhanced feature extraction and attention mechanisms through **GELAN, Programmable Gradient Information (PGI), and Density Adaptive Attention Mechanism (DAAM)**. The developed framework enables simultaneous perception of multiple driving-environment components within a unified architecture. SmartDriveNet achieved **84.5% mAP@50 and 93.1% recall** for traffic-object detection, along with **93.9% drivable-area mIoU**. The results demonstrate the effectiveness of unified multi-task learning for comprehensive perception of complex Indian road environments.

The research was subsequently extended towards **highly unstructured and mission-critical environments**, with a particular focus on autonomous defence vehicles. Defence vehicles operate under conditions significantly different from conventional urban driving environments, including off-road terrain, camouflage, irregular surfaces, severe occlusions, varying illumination, and adverse environmental conditions. Existing road-oriented datasets and perception models are not specifically designed for such scenarios. Therefore, a dedicated **Tactical Vehicle Dataset (TVD)** was developed. The dataset contains **20,758 images representing 31 defence–civilian categories**. A subsequent extension incorporated an additional synthetic landmine category, resulting in a **32-class representation** for the extended dataset.

To address the perception requirements of autonomous defence vehicles, **DefenceVisionNet** was developed as a unified multi-task perception framework. The proposed framework extends the multi-task perception paradigm to tactical and unstructured environments and incorporates architectural improvements for robust feature representation. The framework jointly performs **defence–civilian entity detection, drivable terrain segmentation, and trail/boundary detection**. DefenceVisionNet achieved **84.5% mAP@50 and 93.2% recall** for defence–civilian entity detection, **94.1% mIoU** for drivable terrain segmentation, and **91.8% accuracy with 32.4% Boundary IoU** for trail/boundary detection. The model contains approximately **41.2 million parameters** and achieves an inference speed of approximately **91.2 FPS**, demonstrating its capability for real-time multi-task perception in challenging defence environments. These results demonstrate the feasibility of extending unified perception beyond conventional road environments to autonomous defence applications.

Another important research direction investigated in this thesis was the development of **Q-DriveShieldNet**, a quantum-resilient multi-task perception framework for intelligent transportation systems operating in unstructured environments. The framework addresses challenges associated with dense traffic, road defects, adverse weather, and the security of connected autonomous systems. It supports **traffic-object detection, traffic sign and traffic-light recognition, road anomaly detection, drivable-area segmentation, and lane detection**. Furthermore, **model fingerprinting and post-quantum digital signature authentication** were investigated to ensure model integrity, authenticity, and tamper detection prior to deployment. The framework was evaluated on challenging

driving datasets, demonstrating its potential for accurate, efficient, and secure multi-task perception.

The final phase of the research focused on the **real-time implementation of autonomous driver-assistance capabilities on an embedded platform**. A prototype ADAS was developed using a **Raspberry Pi 4B**, integrating lightweight vision-based perception algorithms with **LiDAR-based obstacle detection** and an interactive **human–machine interface**. This phase demonstrated the practical feasibility of transferring the developed perception methodologies from experimental environments to a resource-constrained embedded platform and provided a foundation for further development of low-cost real-time driver-assistance systems.

Overall, the research presented in this thesis establishes a **unified, multi-task, and multi-domain perception framework for autonomous driver assistance in unstructured environments**. The work progresses from dataset development and traffic perception to road-defect detection, tactical vehicle perception, and embedded deployment. The developed methodologies demonstrate that multiple perception tasks can be integrated into unified frameworks while maintaining high accuracy and real-time inference capability.

The major contributions of this thesis can therefore be summarized as follows:

- Development of the **SafeRide Helmet and License Plate Dataset (SHLPD)** for helmet-violation detection and automatic number plate recognition.
- Development of **ITAWD** dataset for heterogeneous Indian traffic and adverse weather conditions.
- Development of **SmartDriveNet** as a unified multi-task driving perception framework, combining shared feature extraction with task-specific prediction branches for efficient and robust perception.
- Development of the **Tactical Vehicle Dataset (TVD)** dataset for perception in unstructured defence environments.
- Development of **DefenceVisionNet** as a unified perception framework encompassing defence-related vehicles, military personnel, and civilian persons in challenging scenes, together with drivable-terrain and trail/boundary perception.
- Development of **Q-DriveShieldNet** and **QR-DefenceDrive** for secure deployment of deep learning-based perception systems in autonomous driving and defence applications, respectively.
- Development of a **real-time embedded ADAS prototype** using Raspberry Pi 4B, LiDAR, and an interactive human–machine interface.

7.2 Limitations of the Study

Although the proposed frameworks demonstrate promising performance for multi-task perception in civilian and defence environments, certain limitations remain. These limitations primarily arise from challenging environmental conditions, sensing requirements, computational resources, and the complexity of multi-task deployment.

One limitation is the dependence on RGB-based visual information for most of the perception tasks. The performance of camera-based systems can be affected by low-visibility conditions which may also result in ambiguous visual patterns, leading to reduced detection accuracy and increased false positives. In addition, non-standard traffic signal colours or damaged signals may further affect traffic-light recognition.

Another limitation is the limited representation of unusual and extreme conditions in the available training data. Although the developed datasets represent challenging civilian and defence scenarios, unusual traffic signal appearances, severely damaged road surfaces, extreme weather conditions, and rare environmental situations are difficult to capture comprehensively. Increasing the diversity of such samples and incorporating targeted data augmentation can improve the robustness of the developed models.

For autonomous defence applications, the reliance on RGB information can be particularly challenging in nighttime, foggy, smoky, and dusty missions. Incorporating complementary sensing modalities such as thermal imaging, radar, and LiDAR can provide additional information and improve perception reliability under degraded visibility. Although LiDAR was incorporated into the embedded ADAS prototype for obstacle detection, comprehensive multi-modal sensor fusion was not investigated across all perception tasks.

Another limitation concerns the computational requirements of multi-task perception models. Although the developed frameworks provide real-time inference capability, their evaluation and training involve high-performance GPU resources. Detailed benchmarking and optimization for highly resource-constrained embedded platforms remain areas requiring further investigation. Moreover, the use of independent task-specific processing can increase computational overhead when multiple perception tasks are executed simultaneously.

Finally, the present research needs to encompass the complete autonomous driving pipeline such as vehicle trajectory prediction, pedestrian intent recognition, decision-making, path planning, and vehicle control were beyond the primary scope of this thesis. Therefore, further integration of the developed perception frameworks with prediction, planning, and control modules, along with model security mechanisms, remains exploratory and is required to achieve comprehensive autonomous navigation.

7.3 Future Scope and Social Impact

As autonomous driving and intelligent transportation systems continue to evolve, the development of robust perception systems for unstructured environments remains an important research direction. The findings of this thesis provide several opportunities for further improvement and practical deployment. The future scope and potential social impact of the developed work are discussed below.

7.3.1 Future Scope/Directions

For the **multi-task traffic perception frameworks**, future work can focus on improving their robustness under more challenging environmental conditions, including severe weather, low illumination, heavy occlusion, and highly congested traffic. Additional categories of traffic objects and road elements can also be incorporated to provide more comprehensive environmental understanding.

Future research can investigate **multi-modal perception and sensor fusion** by integrating RGB cameras with LiDAR, radar, depth sensors, thermal cameras, GPS, and IMU information. Such fusion can provide complementary information and improve perception reliability under adverse visibility conditions.

For **road-defect detection**, future studies can include a wider range of defects such as rutting, surface deformation, damaged road markings, uneven surfaces, and other infrastructure abnormalities. The development of more comprehensive road-condition datasets can further improve the robustness of road-defect perception.

The developed models can further be optimized for **edge and embedded deployment** using model pruning, quantization, knowledge distillation, lightweight backbones, and hardware-aware architecture design. Such approaches can reduce memory consumption and computational requirements while maintaining acceptable perception accuracy.

For autonomous defence vehicles, future work can extend DefenceVisionNet towards **complete autonomous off-road navigation**. In addition to object detection and semantic understanding, future systems can incorporate terrain classification, traversability estimation, obstacle avoidance, landmine detection, target recognition, path planning, and autonomous navigation in complex terrain.

Future work can also focus on **explainable and uncertainty-aware AI**. Confidence estimation and uncertainty modelling can identify situations in which the perception model is uncertain, allowing the system to provide appropriate warnings or request human intervention. Explainable AI can further improve the transparency and trustworthiness of perception decisions in safety-critical applications.

The embedded ADAS prototype can be further enhanced into a **complete real-time driver-assistance system** by incorporating additional vehicle-state information, advanced warning mechanisms, event logging, GPS/IMU integration, improved human–machine interaction, and vehicle communication technologies.

Finally, the research can be extended beyond perception towards a **complete autonomous driving pipeline**, integrating perception with prediction, planning, decision-making, and vehicle control. Such integration would enable the developed perception frameworks to contribute directly to autonomous navigation.

7.3.2 Social Impact

The proposed research has considerable potential for social and technological impact because reliable environmental perception forms the foundation of safe and intelligent transportation systems. The major potential impacts are as follows:

- **Improved Road Safety:** The proposed systems can detect vehicles, pedestrians, traffic signs, traffic lights, lanes, drivable regions, potholes, cracks, and other road-related hazards. Timely identification of these elements can assist drivers in recognizing potentially dangerous situations and may contribute to accident prevention.

- **Driver Assistance in Unstructured Environments:** The primary focus of this thesis is perception under heterogeneous and unstructured road conditions. The developed systems can provide driver assistance in environments where conventional autonomous driving systems may struggle because of irregular traffic patterns, faded lane markings, road defects, occlusions, and adverse weather.

- **Road Infrastructure Monitoring:** Automatic pothole and crack detection can support the monitoring of road conditions. Such systems can potentially assist road authorities in identifying damaged road sections and prioritizing maintenance activities, thereby contributing to improved infrastructure management.

- **Support for Autonomous Defence Vehicles:** The developed tactical perception framework can support autonomous or semi-autonomous defence vehicles operating in hazardous and unstructured environments. Such systems have the potential to reduce human exposure to dangerous terrain and support safer operation in mission-critical scenarios.

- **Reduced Driver Cognitive Load:** A unified perception system capable of monitoring several environmental elements simultaneously can reduce the burden on human drivers, particularly in complex traffic conditions. Continuous monitoring of the surrounding environment can provide additional assistance when human attention is limited.

- **Contribution to Indigenous Autonomous Driving Research:** The development of datasets and perception frameworks specifically addressing Indian traffic and environmental conditions contributes towards the development of autonomous-driving technologies suited to local requirements. The research can provide a foundation for further development of indigenous intelligent transportation and driver-assistance systems.

- **Security and Trustworthy Autonomous Systems:** The increasing use of deep learning models in connected and autonomous vehicles introduces security risks associated with unauthorized model modification, tampering, and compromised over-the-air updates. Such attacks can alter perception outputs and potentially lead to incorrect decisions in safety- and mission-critical applications. The consequences can be particularly severe for **VIP transportation, ambulances, police vehicles, military logistics, and autonomous defence vehicles**, where perception errors may threaten human lives and critical operations. The investigation of model fingerprinting and post-quantum digital signature authentication in this thesis contributes towards improving model integrity and authenticity and supports the development of more secure and trustworthy autonomous systems in the presence of emerging quantum threats.

- **Socioeconomic Benefits:** Improved road safety, efficient road-condition monitoring, and affordable driver-assistance technologies can potentially reduce accident-related losses, improve transportation efficiency, and support better utilization of road infrastructure.

APPENDIX A
Density Adaptive Attention Mechanism (DAAM)

The Density Adaptive Attention Mechanism (DAAM) used in SmartDriveNet is summarized in Algorithm 1. The mechanism performs multi-scale feature aggregation followed by channel-wise and spatial attention refinement. The resulting feature representation is then projected and normalized before being passed to the subsequent network layers.

Algorithm 1: Density Adaptive Attention Mechanism (DAAM)

Input: Input feature map $F \in \mathbb{R}^{C \times H \times W}$
Output: Refined feature map F_{DAAM}
Step 1: Multi-scale feature extraction
$F_1 \leftarrow \text{Conv}_{1\times1}(F)$
$F_3 \leftarrow \text{Conv}_{3\times3}(F)$
$F_5 \leftarrow \text{Conv}_{5\times5}(F)$
$F_{\text{agg}} \leftarrow \text{Concat}(F_1, F_3, F_5)$
Step 2: Channel-wise attention
$W_c \leftarrow \sigma\big(\text{MLP}\left(\text{GlobalAvgPool}(F_{\text{agg}})\right)\big)$
Step 3: Spatial attention using shift convolution
$W_s \leftarrow \sigma\big(\text{ShiftConv}(F_{\text{agg}})\big)$
Step 4: Attention-based feature refinement
$F_{\text{attn}} \leftarrow W_c \odot F_{\text{agg}} + W_s \odot F_{\text{agg}}$
Step 5: Output projection and normalization
$F_{\text{DAAM}} \leftarrow \text{BN}(\text{Conv}_{1\times1}(F_{\text{attn}}))$
return F_{DAAM}

APPENDIX B
Multi-Task Gradient Balancing

The multi-task optimization strategy used in SmartDriveNet balances the contributions of detection and segmentation objectives during training. A GradNorm-based procedure is employed to dynamically regulate the task weights according to their relative gradient magnitudes. The procedure is summarized in Algorithm 2.

Algorithm 2: GradNorm-Based Dynamic Gradient Balancing for Multi-Task Learning

Input: Detection loss L_{det}, segmentation loss L_{seg}, initial weights α and β
Output: Balanced total loss L_{all}
Initialize:
$\alpha \leftarrow 1.0, \qquad \beta \leftarrow 0.8$

foreach *training step* **do**
Step 1: Compute task-specific gradients
$g_{\text{det}} \leftarrow \|\nabla_\theta L_{\text{det}}\|$
$g_{\text{seg}} \leftarrow \left\|\nabla_\theta L_{\text{seg}}\right\|$
Step 2: Compute normalized gradient magnitudes
$\bar{g} \leftarrow \frac{g_{\text{det}}+g_{\text{seg}}}{2}$
$r_{\text{det}} \leftarrow \frac{g_{\text{det}}}{\bar{g}}$
$r_{\text{seg}} \leftarrow \frac{g_{\text{seg}}}{\bar{g}}$
Step 3: Update task weights
$\alpha \leftarrow \alpha(r_{\text{det}})^\gamma$
$\beta \leftarrow \beta(r_{\text{seg}})^\gamma$
Normalize α and β such that
$\alpha + \beta = 1$
Step 4: Compute the balanced multi-task loss
$L_{\text{all}} \leftarrow \alpha L_{\text{det}} + \beta L_{\text{seg}}$
return L_{all}

REFERENCES

[1] Neumann, T., 2024. Analysis of Advanced Driver-Assistance Systems for Safe and Comfortable Driving of Motor Vehicles. *Sensors*, Vol. 24, No. 19, pp. 6223.

[2] Wu, D., Liao, M.-W., Zhang, W.-T., Wang, X.-G., Bai, X., Cheng, W.-Q. and Liu, W.-Y., 2022. YOLOP: You only look once for panoptic driving perception. *Machine Intelligence Research*, Vol. 19, pp. 550–562.

[3] Vayadande, K., Patil, R., Patni, A., Bhadane, P., Pawar, S., Ponnuru, R. and Ingale, V., 2024. Traffic Sign Recognition System Using YOLO: Societal System for Safe Driving. Techno-Societal 2022 (ICATSA 2022), pp. 157–166. Springer.

[4] Zhu, Y. and Yan, W. Q., 2022. Traffic Sign Recognition Based on Deep Learning. *Multimedia Tools and Applications*, Vol. 81, pp. 170461–170495.

[5] Kulkarni, R., Dhavalikar, S. and Bangar, S., 2018. Traffic light detection and recognition for self-driving cars using deep learning. Proceedings of the Fourth International Conference on Computing, Communication, Control and Automation (ICCUBEA), pp. 1–4.

[6] Garg, S., Baghel, A., Agarwal, A. and Toshniwal, D., 2024. Snowy Scenes, Clear Detections: A robust model for traffic light detection in adverse weather conditions. Proceedings of the KDD Workshop on Urban Computing (KDD-UC '24). Barcelona, Spain.

[7] Park, S.-S., Tran, V.-T. and Lee, D.-E., 2021. Application of various YOLO models for computer vision-based real-time pothole detection. *Applied Sciences*, Vol. 11, pp. 11229.

[8] Wang, Y., Qi, Q., Sun, L., Xian, W., Ma, T., Lu, C. and Zhang, J., 2024. An intelligent detection and classification model based on computer vision for pavement cracks in complicated scenarios. *Applied Sciences*, Vol. 14, pp. 2909.

[9] Yarram, S., Varma, G. and Jawahar, C. V., 2018. City-scale road audit system using deep learning. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 635–640, IEEE.

[10] Sultana, S., Ahmed, B., Paul, M., Islam, M. and Ahmad, S., 2023. Vision-based robust lane detection and tracking in challenging conditions. *IEEE Access*.

[11] Oguz, E., Kucukmanisa, A., Duvar, R. and Urhan, O., 2022. A deep learning-based fast lane detection approach. *Chaos, Solitons & Fractals*, Vol. 155, pp. 111722.

[12] Qiao, D. and Zulkernine, F., 2021. Drivable area detection using deep learning models for autonomous driving. 2021 IEEE International Conference on Big Data (Big Data), pp. 5233–5238.

[13] Li, S., Sultonov, F., Ye, Q., Bai, Y., Park, J.-H., Yang, C., Song, M., Koo, S. and Kang, J.-M., 2022. TA-Unet: Integrating Triplet Attention Module for Drivable Road Region Segmentation. *Sensors*, Vol. 22, pp. 4438.

[14] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems (NeurIPS), Vol. 28.

[15] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A. C., 2016. SSD: Single Shot MultiBox Detector. European Conference on Computer Vision (ECCV), pp. 21–37.

[16] Wang, C.-Y., Bochkovskiy, A. and Liao, H.-Y. M., 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7464–7475.

[17] Jocher, G., Chaurasia, A. and Qiu, J., 2023. Ultralytics YOLOv8. [Online] Available at: `https://github.com/ultralytics/ultralytics` [Accessed:18 August 2026].

[18] Wang, C.-Y., Yeh, I.-H. and Liao, H.-Y. M., 2024. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. Computer Vision – ECCV 2024, Lecture Notes in Computer Science, Springer.

[19] Jocher, G., Chaurasia, A., Qiu, J. and Team, U., 2024. YOLO11: Next-Generation Real-Time Object Detection. [Online] Available at: `https://docs.ultralytics.com/models/yolo11/` [Accessed: 18 August 2026].

[20] Teichmann, M., Weber, M., Zöllner, J. M., Cipolla, R. and Urtasun, R., 2018. MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving. 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1013–1020, IEEE.

[21] Qian, Y., Dolan, J. M. and Yang, M., 2020. DLT-Net: Joint detection of drivable areas, lane lines, and traffic objects. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 21, No. 11, pp. 4670–4679.

[22] Dat, V. T., Bao, N. V. H. and Hung, P. D., 2025. HybridNets: End-to-End Perception Network. *Pattern Recognition and Image Analysis*, Vol. 35, No. 2, pp. 106–118.

[23] Han, C., Zhao, Q., Zhang, S., Chen, Y., Zhang, Z. and Yuan, J., 2022. YOLOPv2: Better, Faster, Stronger for Panoptic Driving Perception. *arXiv:2208.11434*.

[24] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V. and Darrell, T., 2020. BDD100K: A diverse driving dataset for heterogeneous multitask learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2636–2645.

[25] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223.

[26] Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M. and Jawahar, C. V., 2019. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1743–1751.

[27] Arya, D., Maeda, H., Ghosh, S. K., Toshniwal, D. and Sekimoto, Y., 2022. RDD2022: A Multi-national Image Dataset for Automatic Road Damage Detection. *arXiv:2209.08538*.

[28] Lam, D., Kuzma, R., McGee, K., Dooley, S., Laielli, M., Klaric, M., Bulatov, Y. and McCord, B., 2018. xView: Objects in context in overhead imagery. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1–8.

[29] Li, K., Cheng, G., Bu, S. and Han, J., 2020. Object detection in optical remote sensing images: A survey and new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 159, pp. 296–307.

[30] Jafarzadeh, P., Zelioli, L., Farahnakian, F., Nevalainen, P., Heikkonen, J., Hemminki, P. and Andersson, C., 2023. Real-time military tank detection using YOLOv5 implemented on Raspberry Pi. Proceedings of the 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC), pp. 20–26.

[31] Gu, T., Dolan-Gavitt, B. and Garg, S., 2017. BadNets: Identifying vulnerabilities in the machine learning model supply chain. Proceedings of the IEEE Access Workshop on Machine Learning and Computer Security.

[32] Uchida, Y., Nagai, Y., Sakazawa, S. and Satoh, S., 2017. Embedding watermarks into deep neural networks. Proceedings of the ACM International Conference on Multimedia Retrieval, pp. 269–277. Bucharest, Romania.

[33] Rouhani, B. D., Chen, H. and Koushanfar, F., 2018. DeepMarks: A secure fingerprinting framework for deep neural networks. Proceedings of the ACM/IEEE International Conference on Computer-Aided Design (ICCAD), pp. 1–8.

[34] Castiglione, A., Esposito, J. G. and Loia, V., 2025. Enhancing trust of deep learning models with post-quantum digital signatures. *The Journal of Supercomputing*, Vol. 81, pp. 1191.

[35] Jain, S., Indu, S. and Goel, N., 2023. Comparative Analysis of YOLO Algorithms for Intelligent Traffic Monitoring. Proceedings of the International Conference on Data Analytics and Computing (ICDAC 2022), Lecture Notes on Data Engineering and Communications Technologies, Vol. 175, pp. 159–168. Singapore, Springer.

[36] Jain, S., Indu, S. and Goel, N., 2026. Helmet Violation Detection and Automatic Number Plate Recognition of Motorcyclists Using Deep Learning. 68th National Convention of the Indian Institution of Industrial Engineering (IIIE) and 6th International Conference on Industrial Engineering (ICIE-2026). Sardar Vallabhbhai National Institute of Technology, Surat, Gujarat, India.

[37] Jain, S., Indu, S. and Goel, N., 2025. SmartDriveNet: An Integrated Approach for Robust Driving Perception in Unstructured Environments and Adverse Weather Conditions. *Sādhanā*, Vol. 50, pp. 344.

[38] Jain, S., 2025. Indian Traffic and Adverse Weather Dataset (ITAWD), *Roboflow Universe*. [Online] Available at: `https://universe.roboflow.com/shilpa-mocpz/itawd` [Accessed: 03 April 2026].

[39] Jain, S., Indu, S. and Goel, N., 2026. A Unified Multitask Deep Learning Framework for Autonomous Defence Vehicle Perception: DefenceVisionNet. *Defence Science Journal*, Vol. 76, No. 5, pp. 681–692.

[40] Jain, S., 2026. Tactical Vehicle Dataset (TVD), *Roboflow Universe*. [Online] Available at: `https://universe.roboflow.com/shilpa-o3j8x/tactical-vehicle-dataset-tvd-94dmv` [Accessed: 03 April 2026].

[41] He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2961–2969.

[42] Jocher, G., 2020. Ultralytics YOLOv5. [Online] Available at: `https://github.com/ultralytics/yolov5` [Accessed:18 August 2026].

[43] Ji, A. and Ma, X., 2025. Vehicle detection and classification for traffic management and autonomous systems using YOLOv10. *Alexandria Engineering Journal*, Vol. 127, pp. 804–816.

[44] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A., 2010. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, Vol. 88, pp. 303–338.

[45] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollar, P., 2014. Microsoft COCO: Common Objects in Context. European Conference on Computer Vision, pp. 740–755.

[46] Yen, S. H., Shu, C. Y. and Hsu, H. H., 2020. Occluded Traffic Signs Recognition. Advances in Information and Communication. Advances in Intelligent Systems and Computing, vol. 1130, pp. 794–804. Springer.

[47] Qin, Z. and Yan, W., 2021. Traffic-Sign Recognition Using Deep Learning. Springer Series on Machine Learning and Data Mining in Pattern Recognition, pp. 13–25.

[48] Stallkamp, J., Schlipsing, M., Salmen, J. and Igel, C., 2011. The German Traffic Sign Recognition Benchmark: Multiple Object Classification with a Large-Scale Dataset. Proceedings of the International Joint Conference on Neural Networks (IJCNN), pp. 1453–1460.

[49] Megalingam, R. K., Thanigundala, K., Musani, S. R., Nidamanuru, H. and Gadde, L., 2023. Indian Traffic Sign Detection and Recognition Using Deep Learning. *International Journal of Transportation Science and Technology*, Vol. 12, No. 3, pp. 683–699.

[50] Wang, J.-G. and Zhou, L.-B., 2019. Traffic light recognition with high dynamic range imaging and deep learning. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20, pp. 1341–1352.

[51] Chuang, C.-H., Lee, C.-C., Lo, J.-H. and Fan, K.-C., 2023. Traffic light detection by integrating feature fusion and attention mechanism. *Electronics*, Vol. 12, pp. 3727.

[52] Behrendt, K., Novak, L. and Botros, R., 2017. A deep learning approach to traffic lights: Detection, tracking, and classification. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 1370–1377.

[53] Tiwari, G., Goel, R. and Bhalla, K. Road Safety in India: Status Report 2021. Tech. rep. Indian Institute of Technology Delhi, 2022.

[54] Rithish, H., Modhugu, R., Reddy, R., Saluja, R. and Jawahar, C. V., 2021. Evaluating Computer Vision Techniques for Urban Mobility on Large-Scale, Unconstrained Roads. arXiv preprint arXiv:2109.05226.

[55] Fan, R., Ozgunalp, U., Hosking, B., Liu, M. and Pitas, I., 2020. Pothole Detection Based on Disparity Transformation and Road Surface Modeling. *IEEE Transactions on Image Processing*, Vol. 29, pp. 897–908.

[56] Dhiman, A. and Klette, R., 2020. Pothole Detection Using Computer Vision and Learning. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 21, No. 8, pp. 3536–3550.

[57] Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X. and Ling, H., 2020. Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 21, No. 4, pp. 1525–1535.

[58] Salcedo, E., Jaber, M. and Carrión, J., 2022. A Novel Road Maintenance Prioritisation System Based on Computer Vision and Crowdsourced Reporting. *Journal of Sensor and Actuator Networks*, Vol. 11, No. 1, pp. 15.

[59] Ronneberger, O., Fischer, P. and Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241.

[60] Badrinarayanan, V., Kendall, A. and Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 12, pp. 2481–2495.

[61] Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J., 2017. Pyramid Scene Parsing Network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890.

[62] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S., 2017. Feature Pyramid Networks for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117–2125.

[63] Baheti, B., Innani, S., Gajre, S. and Talbar, S., 2020. Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1473–1481.

[64] Chen, L.-C., Papandreou, G., Schroff, F. and Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv preprint arXiv:1706.05587.

[65] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818.

[66] Zhu, J.-Y., Park, T., Isola, P. and Efros, A. A., 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2223–2232.

[67] Wang, Z., Ren, W. and Qiu, Q., 2018. LaneNet: Real-time lane detection networks for autonomous driving. *arXiv:1807.01726*.

[68] Hou, Y., Ma, Z., Liu, C. and Loy, C. C., 2019. Learning Lightweight Lane Detection CNNs by Self Attention Distillation. *arXiv:1908.00821*.

[69] Tabelini, L., Berriel, R., Paixão, T. M., Badue, C., De Souza, A. F. and Oliveira-Santos, T., 2021. Keep your Eyes on the Lane: Real-time Attention-guided Lane Detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 294–302.

[70] Qin, Z., Wang, H. and Li, X., 2020. Ultra Fast Structure-aware Deep Lane Detection. Computer Vision – ECCV 2020, pp. 276–291.

[71] Sultana, S., Ahmed, B., Paul, M., Islam, M. and Ahmad, S., 2023. Vision-Based Robust Lane Detection and Tracking in Challenging Conditions. *IEEE Access*, Vol. PP, pp. 1–1.

[72] Geiger, A., Lenz, P. and Urtasun, R., 2012. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3354–3361.

[73] Kenk, M. A. and Hassaballah, M., 2020. DAWN: Vehicle Detection in Adverse Weather Nature Dataset [Online]. Available at: `https://arxiv.org/abs/2008.05402` [Accessed: 03 April 2026]].

[74] Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X. and Ling, H., 2020. Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 21, No. 4, pp. 1525–1535.

[75] Fan, R., Ozgunalp, U., Hosking, B., Liu, M. and Pitas, I., 2020. Pothole Detection Based on Disparity Transformation and Road Surface Modeling. *IEEE Transactions on Image Processing*, Vol. 29, pp. 897–908.

[76] Pan, X., Shi, J., Luo, P., Wang, X. and Tang, X., 2018. Spatial as Deep: Spatial CNN for Traffic Scene Understanding. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, pp. 7276–7283.

[77] Feng, Z., Guo, S., Tan, X., Xu, K., Wang, M. and Ma, L., 2022. Rethinking Efficient Lane Detection via Curve Modeling. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17041–17049.

[78] U.S. Department of Defense, 2026. Defense Visual Information Distribution Service (DVIDS). [Online] Available at: `https://www.dvidshub.net` [Accessed:18 August 2026].

[79] Wikimedia Commons, 2026. Military Vehicles Media Repository. [Online] Available at: `https://commons.wikimedia.org/wiki/Category:Military_vehicles` [Accessed:18 August 2026].

[80] Skakov, M., Islamgozhayev, T. and Abdildayeva, A. A., 2025. An Image Dataset for Military Vehicle Detection and Classification (DMVDC). *Herald of the Kazakh-British Technical University*, Vol. 22, No. 1, pp. 74–83.

[81] DIUx, 2018. xView: A Dataset for Detecting Objects in Overhead Imagery. [Online] Available at: `https://xviewdataset.org/` [Accessed: 26 August 2026].

[82] Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M. and Zhang, L., 2018. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3974–3983.

[83] Wojke, N., Bewley, A. and Paulus, D., 2017. Simple Online and Realtime Tracking with a Deep Association Metric. 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649. Beijing, China, IEEE.

[84] Jain, S., 2026. SafeRide: Helmet & License Plate Dataset (SHLPD). Roboflow Universe [[Online]. Available at: `https://universe.roboflow.com/tlsor/saferide-helmet-license-plate-dataset-shlpd-d4gwy` [Accessed: 03 April 2026]].

[85] Hao, W. and Zhili, S., 2020. Improved Mosaic: Algorithms for more complex images. *Journal of Physics: Conference Series*, Vol. 1684, pp. 012094.

[86] Wang, H., Sui, C., Jiang, F., Li, S., Liu, H. and Wang, A., 2024. Value-guided adaptive data augmentation for imbalanced small object detection. *Electronics*, Vol. 13, pp. 1849.

[87] Chen, Z., Badrinarayanan, V., Lee, C.-Y. and Rabinovich, A., 2018. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. Proceedings of the 35th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, Vol. 80, pp. 794–803, PMLR.

[88] Ioannides, G., Chadha, A. and Elkins, A., 2024. Density Adaptive Attention is All You Need: Robust Parameter-Efficient Fine-Tuning Across Multiple Modalities. *arXiv preprint arXiv:2401.11143*.

[89] Jafarzadeh, P., Zelioli, L., Farahnakian, F., Nevalainen, P., Heikkonen, J., Hemminki, P. and Andersson, C., 2023. Real-time Military Tank Detection Using YOLOv5 Implemented on Raspberry Pi. Proceedings of the 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC), pp. 20–26.

[90] Suprayitno, S., Fauzi, W. A., Ain, K. and Yasin, M., 2023. Real-time military person detection and classification system using deep metric learning with electrostatic loss. *Bulletin of Electrical Engineering and Informatics*, Vol. 12, No. 1, pp. 1–8.

[91] Gu, T., Dolan-Gavitt, B. and Garg, S., 2017. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *arXiv preprint arXiv:1708.06733*.

[92] Uchida, Y., Nagai, Y., Sakazawa, S. and Satoh, S., 2017. Embedding Watermarks into Deep Neural Networks. Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 269–277. Bucharest, Romania, ACM.

[93] Guo, J. and Potkonjak, M., 2018. Watermarking Deep Neural Networks for Embedded Systems. Proceedings of the 2018 ACM/IEEE International Conference on Computer-Aided Design, pp. 1–8. San Diego, CA, USA, ACM.

[94] Shor, P. W., 1994. Algorithms for Quantum Computation: Discrete Logarithms and Factoring. Proceedings of the 35th Annual Symposium on Foundations of Computer Science, pp. 124–134. Santa Fe, NM, USA, IEEE.

[95] Bernstein, D. J. and Lange, T., 2017. Post-Quantum Cryptography. *Nature*, Vol. 549, pp. 188–194.

[96] Castiglione, A., Esposito, J. G., Loia, V., Nappi, M., Pero, C. and Polsinelli, M., 2025. Enhancing Trust of Deep Learning Models with Post-Quantum Digital Signatures. *The Journal of Supercomputing*, Vol. 81, No. 11, pp. 1191.

[97] Jain, S., Indu, S. and Goel, N., 2025. Real-time Embedded ADAS for Elderly Drivers with Vision-Based Interactive Display. International Conference on Emerging Electronics and Automation (E2A 2025).

[98] Pan, H., Payre, W., Xu, J. and Koppel, S., 2024. Age-related differences in takeover performance: A comparative analysis of older and younger drivers in prolonged partially automated driving. *Traffic Injury Prevention*, Vol. 25, No. 7, pp. 968–975.

[99] Raspberry Pi Ltd. 2023. Raspberry Pi 4 Model B: Technical Specifications. `https://www.raspberrypi.com/products/raspberry-pi-4-model-b/specifications/` [Accessed: 2026-09-02].

[100] STMicroelectronics, VL53L0X Time-of-Flight Ranging Sensor: Datasheet. https://www.st.com/resource/en/datasheet/vl53l0x.pdf [Accessed: 2026-09-02].

[101] An, X.-J., Li, J. and Shang, E.-K., 2011. Region-of-Interest Generation for Lane Detection Using High-Level Information. Proceedings of the 2011 International Conference on Machine Vision.

[102] Jackulin, J. M. A. and Aryan, 2024. (IRTSD-Datasetv1)-Indian Road Traffic Sign Detection Dataset. *IEEE DataPort*.

[103] Gautam, S. and Kumar, A., 2022. An Indian Roads Dataset for Supported and Suspended Traffic Lights Detection. *arXiv preprint arXiv:2209.04203*.

[104] Ma, Q., Jian, Q., Li, M. and Ullah, S., 2025. LiDAR-Visual Fusion SLAM for Autonomous Vehicle Location. *IEEE Internet of Things Journal*, Vol. 12, No. 13, pp. 25197–25210.

[105] Qin, Z., Wang, H. and Li, X., 2020. Ultra Fast Structure-Aware Deep Lane Detection. *arXiv preprint arXiv:2004.11757*.

[106] Tabelini, L., Berriel, R., Paixao, T. M., Badue, C., Souza, A. F. and Oliveira-Santos, T., 2020. PolyLaneNet: Lane Estimation via Deep Polynomial Regression. *arXiv preprint arXiv:2004.10924*.

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis Autonomous Driver Assistant System for Unstructured Environments

Total Pages 144 **Name of the Scholar** Shilpa Jain

Supervisor (s)

(1) Prof. S. Indu, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi

(2) Prof. Nidhi Goel, Department of Electronics and Communication Engineering, Indira Gandhi Delhi Technical University for Women, Delhi

Department Department of Electronics and Communication Engineering, Delhi Technological University, Delhi

This thesis is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin **Similarity Index:** 9% **Total Word Count:** 51174

Date: 07-09-2026

Candidate's Signature **Signature of Supervisor(s)**