A Neurosymbolic Architecture for Real-Time Explainability and Rule Compliance in Autonomous Vehicles

Abdi Mosisa Dera

A Thesis submitted to the Department of Computer Science and Engineering,

College of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of

Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

October 2025

Adama Ethiopia

A Neurosymbolic Architecture for Real-Time Explainability and Rule Compliance in Autonomous Vehicles

Abdi Mosisa Dera

Advisor: Dr. Teklu Urgesa (Associate Professor)

A Thesis Submitted to the Department of Computer Science and Engineering, College of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

October 2025

Adama, Ethiopia

DECLARATION

I hereby declare that this Master Thesis entitled "**A Neurosymbolic Architecture for Real-Time Explainability and Rule Compliance in Autonomous Vehicles"** is my original work. That is, it has not been submitted for the award of any academic degree, diploma or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citation

_____ _____ _____

Name of student

RECOMMENDATION OF ADVISOR

I, the advisor of this thesis, hereby certify that I have read the revised version of the thesis entitled "**A Neurosymbolic Architecture for Real-Time Explainability and Rule Compliance in Autonomous Vehicles"** prepared under my guidance by Abdi Mosisa Dera submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering. Therefore, I recommend the submission of the revised version of the thesis to the department following applicable procedures.

Major Advisor	Signature	Date

Approval Page for Advisor

I, the advisor of the thesis entitled "**A Neurosymbolic Architecture for Real-Time Explainability and Rule Compliance in Autonomous Vehicles"** and developed by **Abdi Mosisa Dera**, hereby certify that the recommendation and suggestions made by the board of examiners are appropriately incorporated into the final version of the thesis.

Major Advisor	Signature	Date

APPROVAL OF BOARD OF REVIEWERS

We, the undersigned, members of the Board of Examiners of the thesis by **Abdi Mosisa Dera** have read and evaluated the thesis entitled "**Explainable Artificial Intelligence through NeuroSymbolic Reasoning for Autonomous Vehicle Safety"** and examined the candidate during open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering.

Chairperson	Signature	Date
Internal Examiner	Signature	Date
External Examiner	Signature	Date

Finally, approval and acceptance of the thesis is contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

Department Head	Signature	Date
School Dean	Signature	Date
Office of Postgraduate Studies Dean	Signature	Date

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to Almighty God for granting me the opportunity to start and complete my endeavors and for providing me the strength to persevere in all my pursuits. I extended my heartfelt thanks to my advisor, Dr. Teklu Urgesa, who guided me from the beginning to the end of this work. His positive ideas, comments and invaluable support on my work encouraged me greatly. I would also like to thank Adama city transport and logistic office for providing us a traffic regulation document that we used for preparing a dataset as well as providing us a technical feedback on the correctness and completeness of the processed traffic rule. Lastly, I would like to thank my family and friends for their unwavering support, encouragement, and patience during the ups and downs of my time at the University. I expressed my gratitude to all ASTU computer science and engineering staff members for their advice, motivation, and guidance in leading me into the academic research world.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AV	Autonomous Vehicle
BDD	Berkeley DeepDrive
BDD-OIA	Berkeley DeepDrive - Object and Instance Attention
BELCM	Brain Emotional Learning Circuit Model
CF	Counterfactual
CNN	Convolutional Neural Network
DRL	Deep Reinforcement Learning
FDRE	Federal Democratic Republic of Ethiopia
GPS	Geographical Positioning System
HAT	Human Autonomy Teaming
IDM	Iterative Causal Discovery
KB	Knowledge base
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short Term Memory
NLP	Natural Language Processing
NN	Neural Network
RL	Reinforcement Learning
SAE	Society of Automotive Engineers
SHAP	SHapley Additive exPlanations
SPPF	Spatial Pyramid Pooling
STEEX	STEering counterfactual EXplanations
VQA	Visual Question Answering
XAI	Explainabe Artificial Intelligence
YOLOP	You Only Look Once for Panoptic Driving Perception

ABSTRACT

The black-box characteristic of deep-learning models in autonomous vehicle (AV) is challenging public trust and safety, rendering them inaccessible for general use. XAI methods like SHAP and LIME are post-hoc and therefore cannot assist dynamic, safety-critical systems in real-time, context-dependent decision-making. This research proposes a novel NeuroSymbolic AI approach through the integration of YOLOP-based multi-task perception with logic tensor networks (LTN) for symbolic reasoning to enable explainable AV decision-making. A mixed-methods strategy was followed, combining design science methodology, quantitative evaluation, and qualitative analysis of user trust. The framework was instantiated with the BDD100k-OIA dataset and a novel Ethiopia-specific traffic rule ontology. Quantitative evaluation employed two metrics: action F1 score and explanation F1 score. Our proposed NeuroSymbolic model achieved an action F1 score of 92.0% on images and 93.0% on Adama city street video data, and an explanation F1 score of 90.5% on images and 91.0% on video data, outperforming baseline models. Compared with state-of-the-art methods, the framework is on par or better in safety and traffic rule compliance with similar action F1 score (92.0%–93.0%) and explanation F1 score (90.5%–91.0%). Qualitative evaluation of Adama city street video data with 32 users of diverse backgrounds resulted in high mean scores: 4.61 for understandability and clarity of explanations, 4.58 for correctness of explanations, 4.61 for NeuroSymbolic AI decision trustworthiness, and 93.55% Yes for explanation usefulness. The findings demonstrate the framework's capacity to produce clear, correct, trustworthy, and useful explanations for real-world complex urban scenes. Our contributions are: (i) a NeuroSymbolic pipeline for interpretable, policy-aware AV reasoning, (ii) the first integration of Ethiopian traffic rules into AV decision-making, and (iii) end-to-end evaluation metrics for accuracy, safety, compliance, and explainability. While rooted in Ethiopia's context, the framework provides a scalable, explainable, and dependable template for global AV deployment.

Key Words: Autonomous Vehicle, Neural network, Symbolic AI, NeuroSymbolic AI

CHAPTER ONE

1. INTRODUCTION

1.1 Background of the Study

Autonomous vehicle (AV) is full of promise in revolutionizing the transport industry by improving accessibility, efficiency, and security. The lack of transparency of the artificial intelligence systems on which AVs is based constitutes a strong obstacle to their widespread acceptability. While deep learning models dominate tasks like object recognition, enabling AVs to identify pedestrians and cyclists with precision even at night or in light-low situations (Grigorescu et al., 2020), and make precise lane segmentation, their inbuilt lack of interpretability is a main challenge to show how they make decisions. Such lack of transparency is at the expense of public trust, on which society's embracing and adoption of AV technology depends.

Challenging driving scenarios in the real world amplify the limitations of current artificial intelligence technologies. Deep learning methods, more precisely, scarcely integrate common sense reasoning, generating unsafe behaviors in scenarios requiring knowledge of social norms and unwritten rules (Bonnefon et al., 2020). These systems also lack reasoning regarding intricate interactions among road users, such as predicting the reaction of a pedestrian who can engage in jaywalking (Schwarting et al., 2018). Although existing explainable AI (XAI) techniques like SHAP, LIME, and attention mechanisms provide some insight into AI's behavior they are insufficient to meet the unique challenges of AV explainability. Though SHAP values may be able to identify important features, they may not be able to capture the complex interaction of factors that went into a specific driving decision, e.g., a sudden lane change. Similarly, LIME's local explanations may not suffice to understand the AV's general behavior and long-term planning, and attention mechanisms, while highlighting areas of interest, may not reveal the underlying reason why the AV is attending to those specific areas.

These approaches are bound to be devoid of context information, hence failing to accurately reflect complex intersections and a dynamic driving scenario. They are also likely to struggle with complex multi-step reasoning frameworks and incorporating relevant

domain-specific knowledge in an adequate manner, i.e., road rules, safety protocols, and popularly adopted driving practices. These approaches also constrain the evaluation of abstract scenarios, hence hindering active safety analysis.

NeuroSymbolic AI has emerged as a possible answer through combining the capability of deep neural networks to learn and the reasoning and interpretive aspects common in symbolic AI in order to overcome these challenges effectively. By integrating different paradigms, NeuroSymbolic AI visions constructing autonomous AV systems capable of reasoning from explicit knowledge and also learning from data. With this ability, the systems can explain their actions in human-understandable terms and handle complex situations with greater transparency and stability (Badreddine et al., 2022).

1.2 Motivation of the study

The fatal 2018 Uber AV crash (H. Tapiro, et al, 2022) in which a pedestrian's fatality was blamed on the system's failure to provide transparency to its decision-making conveys a bitter truth: in the absence of explainability, autonomous cars (AVs) will never achieve the public's trust or mass acceptance. Nowhere is this more evident than in Ethiopia, where mixed traffic (vehicles, pedestrians, and animals) and informal traffic rules call for AVs that not only see but also understand context, something that probably only NeuroSymbolic AI can deliver. Contemporary deep learning approaches technically fail in such difficult environments. In 2023, Tesla's Full Self-Driving (FSD) software misread a parked truck as a lane marking, an error rooted in its inability to think beyond naked patterns of pixels. By grounding perception in symbolic principle such as "obstacles should be avoided" NeuroSymbolic AI can sidestep that type of error, enabling AVs to have explainable, reliable decision-making in dynamic, unstructured settings.

Societally, the stakes are no lower. A 2023 AAA survey (Lorenzelli, F. 2024) indicated that 78% of drivers are not trusting of AVs, and unpredictable behavior sits at the top of the list. In Ethiopia, where AVs could significantly improve transport accessibility and safety, overcoming public distrust is the path to adoption, regulatory body, and eventual societal benefit. Explainable AVs can bridge this trust gap and make the technology socially acceptable and ethical. In this research we first addresses these challenges by symbolically representing Ethiopia's traffic laws as symbolic constraints and coupling them with

YOLOP perception using Logic Tensor Networks (LTN) and quantitatively measuring explainability perforamce. Through linking robust perception and explainable reasoning, the research have created AVs that are reliable, safe, and context-aware, enabling Ethiopians to adopt and benefit from autonomous mobility and establish a global standard in explainable autonomy.

1.3 Statement of the Problem

A black box nature of deep learning models for AV create a barrier in it's trustworthiness and safe deployment. Most of the existing explainable AI (XAI) techniques such as LIME (Bordt, S, 2021), SHAP (Pereira, J. et al., 2024), and attention mechanisms (Tutek et al., 2022) are post-hoc where the decision and explanations are made separately. They explain a model's reasoning after a decision has been made and in real-time application context such as autonomous vehicle, post-hoc explanations suffer from built-in constraints such as:-

- **Computational overhead** - generating a post-hoc explanations adds a latency which makes them unsuitable for AV operation.
- **Non-alignment of domain knowledge and decision logic** - Because they are being applied after the fact, post-hoc methods cannot ensure that decisions are inherently compliant with safety rules or domain knowledge.

NeuroSymbolic AI is a move towards ante-hoc explainability, wherein reasoning is built in the architecture itself. While this direction is promising, existing NeuroSymbolic techniques are not yet deployable in the real world for AVs due to two interrelated gaps –

- **The real-time reasoning gap** - existing systems cannot perform scalable symbolic reasoning in the real time which limits them in providing transparent and rational explanations.
- **The knowledge-integration gap** - they lack robust means to integrate and reason with safety regulations and dynamic context information in an integrated NeuroSymbolic system, and thus causing to a disconnect between perception and safe, justified action.

This research aims to overcome these limitations by rigorously defining and evaluating a novel NeuroSymbolic AI framework that provides inherent, real-time, and trustworthy explanations for AV decision-making based on knowledge domains for safety and scalability in real-world scenario.

1.4 Research Question

This research attempts to address the following research questions out of the specified research gaps:

1. What are the key methods for implementing NeuroSymbolic AI to improve the performance, explainability and transparency of decision-making processes in autonomous vehicles?
2. What is the effectiveness of extending deep learning with symbolic reasoning in improving the contextual awareness and complex reasoning capabilities of autonomous vehicles.
3. To what extent can NeuroSymbolic AI enable autonomous vehicles to incorporate and reason with safety regulations, and how does this contribute to more reliable, safe, and ethically-informed decision-making in diverse driving scenarios?

1.5 Objective

1.5.1 General Objective

To develop and evaluate a NeuroSymbolic AI framework for autonomous vehicles that enhances explainability, safety, and trustworthiness.

1.5.2 Specific objectives

This research aims to achieve the following objectives:

- **Design a NeuroSymbolic AI system with deep learning models and symbolic rules** for real-time inference.This addresses research question 1.
- **Enhance autonomous vehicles with contextual awareness and high-level reasoning ability** by integrating symbol-based reasoning with real-time knowledge, enabling better explanation and response to dynamic and unstructured traffic situations. This addresses research question 2.

- **Enable autonomous vehicle to reason under safety rules and ethical principles** by formalizing traffic regulation and ethical principles into the NeuroSymbolic system, and evaluate how it affects compliance rates, safety, and public trust. This is in relation to research question 3.

1.6 Significance of the Study

This research has widespread global relevance in the deployment and development of trustworthy AI systems, particularly in safety-critical applications like autonomous driving. Through enhanced explainability and transparency of AV decision-making, this research is supportive of developing public trust, improved safety, and guiding regulation on AI deployment in transportation. In the case of Ethiopia, where road infrastructure and public awareness of AI are still emergent, this research is particularly relevant. The implementation of these measures would enhance road safety, enable broader adoption of autonomous vehicle technology, and guide policymaking on artificial intelligence deployment in the transport industry, so that Ethiopia reaps the rewards of this innovative technology in a safe and responsible manner.

1.7 Scope and Limitations

The research focuses on the development of a NeuroSymbolic AI model designed to enhance the transparency and interpretability of decision-making algorithms used in autonomous vehicle. Through the integration of deep learning techniques with symbolic reasoning, the research aims to go beyond the limitations found in current explainable AI (XAI) methods. Thus, this method will enable AVs to integrate contextual knowledge, traffic rule adherence, safety measures, and ethical considerations into their decision-making. The research overcomes some of the drawbacks of current xAI models including a real-time reasoning limitations, inability to scale to tackle complicated real-world traffic, and poor contextual awareness.

To address these challenges, the research presents an optimized integration symbolic rules with neural network using logic tensor network, and reasoning engine to achieve improved and transparent decision making. With this approach, autonomous vehicle learn complicated driving scenarios while adhering to rigorous safety standards and. The initial

assessment were mostly based on simulations and test conditions because of limitations in resources and the sophisticated nature involved in applying autonomous vehicles to real-life situations; however, future research will focus on moving towards actual implementations. This task involves working with AV developers based in Ethiopia to test the framework in diverse and unpredictable road conditions, iterating it to tackle real sensor noise, intricate traffic interactions, and address ethical concerns pertinent to the Ethiopian setting. Such testing and validation against the real-world will be necessary to accurately determine the performance of the framework and close the real-world and simulation gap.

1.8 Organization of the rest of the thesis

Chapter Two discusses literature reviews. Research methodologies including dataset collection, preprocessing of data, tools utilized for development, baseline work, and metrics for evaluation is discussed in Chapter Three. Chapter Four presents detailed description about architecture of the proposed model. In Chapter Five implementation of the model is provided. Chapter Six discusses the proposed method outcomes and compares them to previous studies. At the end summary of the research findings and suggestions for future research are discussed in Chapter Seven.

CHAPTER TWO

2. LITERATURE REVIEW

This literature review discusses the state of scholarly research on explainable artificial intelligence (XAI) in the domain of autonomous vehicles (AVs) with an emphasis on the new area of NeuroSymbolic AI. The discussion was delved into the shortcomings of current deep learning methods to AVs, the drawbacks of existing explainable AI techniques, and the prospects of NeuroSymbolic AI in mitigating these challenges.

2.1 Autonomous Vehicle

Autonomous vehicle or Self-driving or driverless cars are the succeeding technology in the development of transportation technology. They use advanced fused sensors like cameras, RADAR, LIDAR, GPS, and artificial learning models in sensing the surroundings and self-driving or semi-driving (Atakishiyev et al., 2024). But to achieve the full potential of AVs is to overcome substantial safety hurdles in their smooth, continuous function under the weight of complicated real-world scenarios. To support a better comprehension of the automation levels between the various AVs, Society of Automotive Engineers (SAE) International created six levels of autonomous driving (Fayyad et al., 2020, Chattopadhyay et al., 2020). Such automation levels vary between Level 0, where there is still complete control by the human driver, and Level 5, where the vehicle has the capability of performing all activities of driving regardless of the circumstance without any kind of human engagement. The intermediary levels explain how there is loosening of responsibility by the human driver with an increase in independence and complexity by the computer program.

Figure 2.1 High level view of Autonomous vehicle (Sana, F., et al. 2023)

Despite the rapid development of AV technology, safety is their utmost concern (Esenturk et al., 2023).Various reasons are to be held responsible for the complexity of AV safety, such as having to cope with safety-critical occurrences, analysing accidents to find out what conceivable modes of failure may occur, and using formal verification techniques in order to give rigorous guarantees about system behaviour. AVs should be able to cope with difficult situations such as pedestrian crossing, adverse weather, and multi-lane intersections, whose safety consequences are catastrophic. AV accident analysis may also play a very critical role in revealing potential weaknesses and strengthening safety systems. Formal methods may offer a system correctness verification approach but fall short in the ability to manage complexity and ambiguity of real-world driving conditions.

This research is conscious of such difficulties and intends to take advantage of advances in safety-critical situation analysis, accident analysis, and formal verification techniques to specify and explore a safe and dependable NeuroSymbolic AV system for various environments. With the integration of symbolic reasoning and deep learning, this research is better in explaining the AV decision-making and make it more transparent and result in safer, trustworthy autonomous driving technology.

2.2. Deep learning and the Black Box problem in AVs

Deep learning revolutionized autonomous driving and Convolutional Neural Networks (CNNs) were found to be extremely useful in basic operations. For instance, (Krizhevsky et al., 2012) established that CNNs can be used in image classification, and that was extended to object detection and scene understanding in AVs. Likewise, (Bojarski et al., 2016) showed end-to-end learning of self-driving cars using CNNs, wherein raw pixels from a camera are directly mapped to steering actions.

All these advances have immensely enhanced the perception and control capacity of the autonomous vehicle. However, excellent performance of deep learning models on AVs is usually marred by their vulnerability to transparency.

Like a sophisticated model relying on extracted features, (Chamola et al., 2023) are "black boxes," in the sense wherein even visualizing how they arrive at their decisions is not feasible. This transparency barrier is a hindrance to accountability and explainability, and it is that which creates an issue with respect to safety and trust. Even the data scientists that develop these kinds of algorithms will remain unaware of the inner workings and decision-making, exacerbating the problem. Transparency creates distrust in users, hinders debugging and safety analysis, and complicates approvals, according to (Grigorescu et al., 2020) and (Koopman and Wagner, 2017). Explainable AI (XAI) is a new field that seeks to make sense of how complex AI models make decisions so that they do not lead to such hindrance. XAI seeks to make AI models understandable and comprehensible, their expected influence, latent biases, and features that influence their decisions (Saeed and Omlin, 2023).

(Arrieta et al., 2020) state that XAI is important as part of trying to build trust and confidence in AI systems in safety-critical domains like autonomous driving where an explanation of a decision taken by an AI needs to be explainable in order to apportion blame and decide safety. There are various advantages of XAI including revealing the decision-making process, ensuring system performance and regulatory compliance, and enabling users to comprehend and interrogate AI-driven decisions (Chamola et al., 2023). In black box explanations of AI, XAI can correct the disparity between sophisticated algorithms and human cognition and instill trust that guarantees the correct utilization of AI in autonomous cars.

2.3. Limitations of current XAI techniques for AVs

XAI possesses several techniques that have been developed from the viewpoint of explaining deep learning model decision-making in such a way that the AI becomes more transparent and explainable by giving explanations regarding how the deep models make their decisions. SHAP (SHapley Additive exPlanations) (Bordt, S., & von Luxburg, U. 2023), LIME (Local Interpretable Model-agnostic Explanations) (Pereira et al., 2024), and attention mechanisms (Vaswani et al., 2021) are some examples of popular XAI methods. SHAP values provide feature importance scores to input features for their contribution to a model's prediction. This informs us about which features play a larger role in the decision-making process. LIME creates locally faithful explanations by approximating a complicated model with a simple, interpretable model in the vicinity of an individual instance, enabling us to comprehend the behavior of the model in a particular context. Attention mechanisms identify parts of the input on which a model is concentrating when it is making a prediction. Though these methods provide colossal insight into AI models, they are significantly limited when applied to the multi-dimensioned context of autonomous vehicles.

Most XAI approaches take individual features or pixels into consideration and overlook the compound dependencies and inter-relations between multiple aspects of a dynamic driving scenario. In this regard (Huang et al., 2020) states that explanations of pedestrians, cyclists, and other vehicles interactions with pedestrians and with each other are necessary to explain behaviour in AVs when they drive through cities. Additionally, (Langner et al., 2020) give some examples for explaining complicated maneuvers such as crossing or joining traffic streams but most of the existing XAI methods are not explaining multi-step processes included in decision-making under uncertainty or deductive reasoning.

In general, most of the XAI methods are data-driven and are not able to incorporate significant domain knowledge such as traffic rules, safety policies, and general driving conventions in AV. To address these shortcomings (Seshia et al., 2017) states that both formal verification techniques and domain knowledge must be integrated into XAI so that not only do AVs satisfy safety specifications, but they also behave in a predictable and explainable manner. These difficulties of contextual, multi-step reasoning, and integration of domain knowledge are reflective of the need for NeuroSymbolic AI.

By bridging the gaps in symbolic reasoning and deep learning, the proposed approach can explicitly model traffic rules and safety practices for improved contextual understanding and explainability. For instance, the system can use symbolic knowledge to reason right-of-way at intersections or predict pedestrian behaviour from traffic rules. With symbolic knowledge, it is possible to have improved and human-comprehensible explanations of AV decisions. Additionally, the proposed NeuroSymbolic AI architecture can circumvent limitations of current XAI approaches to address fine-grained reasoning through the incorporation of logical inference mechanisms and enabling counterfactual analysis. This allows the AV to reason about "what-if" scenarios and produce explanations based on alternative actions.

Thus, while XAI techniques today are valuable solutions to AI model explanation, in the autonomous driving use case, they are constrained by their lack of context sensitivity, hard-reasoning adherence, or domain-knowledge conveyance. These constraints calls for the use of more advanced XAI techniques, such as NeuroSymbolic AI, to overcome the unique requirements of explainable and trustworthy autonomous driving.

Table 2.1: Limitations of existing explainability techniques

Method	**Strength**	**Limitation**	**Our Solution**
SHAP (Bordt, S, 2021)	Feature attribution	No multi-agent reasoning	LTN-encoded traffic rules
LIME (Pereira, J. et al., 2024)	Local approximations	Cannot chain rules	Symbolic reasoning over sequences
Grad-CAM (Mankodiya et al., 2022)	Pixel-level focus	No symbolic grounding	Symbolic predicates for visibility
VQA (Atakishiyev et al., Atakishiyev)	Visual question-answering	Fails in edge cases	Rule-informed reasoning for unusual scenarios

2.4 The promise of NeuroSymbolic AI for explainable AVs

NeuroSymbolic AI puts forward an answer to the existing XAI approaches' disadvantage for AVs. It is able to leverage the strengths of deep learning and symbolic AI, with neural networks learning capabilities along with reasoning ability of symbolic representation (Sheth, Roy, & Gaur, 2023). Thus, possibly efficient but explainable and reliable AVs can be implemented.

Deep neural networks prefer to learn raw inputs such that they can learn to recognize advanced patterns, a feature which makes them most appropriate for applications such as image classification and natural language processing (Reyad et al., 2023). For AVs, this means the ability to process sensor inputs as images and lidar point clouds to be capable of perceiving the environment and navigating. But as (Saeed and Omlin, 2023) go on to explain, neural networks are black boxes, and one cannot possibly know how accurately they arrive at some conclusion. That kind of opacity is to be especially avoided in security-critical applications like autonomous driving, when the risk is taken of wanting to know why an AI is doing some specific thing.

Symbolic AI, nonetheless, possesses interpretability in itself. It is symbolic knowledge, i.e., logic, and it reasons using it by making use of reasoning engines operating on symbols in the form of rules of logic (Carnevali & Lippi, 2024). It facilitates tracing the decision procedure back to rules utilized, and therefore the system's behaviour is more comprehensible to humans. In addition, symbolic AI systems are easy to modulate and fine-tune by adjusting the rules, and that provides a degree of control and flexibility not readily available with deep learning systems.

NeuroSymbolic AI tries to bridge between the two paradigms by integrating learning and reasoning. It seeks to create AI systems that both learn by experience and symbolically reason to produce more robust and transparent AI (Sheth, Roy, & Gaur, 2023). A number of methods to this goal have been tried, including knowledge distillation (Gupta & Sheng, 2023), where the knowledge contained in a neural network is distilled into a less computationally intensive-to-reason symbolic form, and logic tensor networks (Manigrasso & Morra, 2023), where reasoning steps and logical rules are encoded in neural networks.

The current paper is interested in neural-symbolic networks (Vermeulen et al., 2023) since they are the most frequent NeuroSymbolic method. Neural-symbolic networks integrate the symbolic and neural parts into a single architecture in a way whereby learning data and symbolic inference of knowledge is facilitated. It is most intriguing to perform complex reason tasks for AVs, e.g., integrating traffic laws and ethics with decision-making.

As a tip of the hat to the greatness of symbolic AI and deep learning, the neural-symbolic networks provide a promising path toward the attainment of explainable and trusted autonomous systems.

2.5 Open challenges and research gaps

Though NeuroSymbolic AI promises very promising usages in making explainable and reliable autonomous vehicle, it has certain problems and research which need to be tackled for it to be fully utilized.

Scalable and efficient NeuroSymbolic architectures would be one of the major ones. Symbolic reasoning and deep learning would both be computationally costly to combine, and so efficient algorithms as well as methods of knowledge representation are needed. AVs have to handle tremendous amounts of sensor data in real-time and computational efficiency is therefore crucial to have safe and reliable operation. Architectures that can accommodate high-performance solutions to deal with the complexity of NeuroSymbolic reasoning without performance penalty have to be explored and implemented. To address this issue, in this research, the computational efficiency of the suggested framework was verified by its inference time on BDD100k-OIA benchmark dataset and test its real-time performance potential in different test scenarios.

Uncertainty and missing data management is another challenge. AVs operate in uncertain, dynamic environments with noisy or missing sensor measurements. There is a need for creating strong processes of reasoning with the potential for handling such uncertainty to enable safe and reliable decision-making. NeuroSymbolic AI systems need to be able to reason logically based on incomplete or imperfect data, and there needs to be experimentation for constructing techniques of representation and reasoning under uncertainty in a NeuroSymbolic system. This research has examined the performance of the framework's resilience in unpredictable situations using occlusions and noise in

different driving scenes as well as in terms of the capacity to comply with traffic regulations.

Finally, trustworthiness and explainability evaluation of NeuroSymbolic AI systems is still an open problem. Whereas NeuroSymbolic AI tries to enhance transparency, there must be the creation of effective methods and metrics to measure the quality and comprehensibility of explanations presented by such a system. Second, trust with NeuroSymbolic AI involves sincere testing of its reliability and safety in various driving conditions.

There needs to be conducted research to develop standardized benchmarks and test models to test the explainability and credibility of NeuroSymbolic AI for AVs. To solve this issue, in this research work, user studies are carried out among experts and non-experts to measure the explainability of the generated explanations by how faithful and understandable they are. The above-mmentioned scalability gaps, contextual reasoning, and multi-step reasoning gaps call for the need of a NeuroSymbolic architecture to achieve the power of symbolic reasoning as well as that of neural networks in order to address the above-specified problems in real-time application in AVs. The aim of this research is to address the above-specified issues by using a new NeuroSymbolic AI paradigm for AVs that will be explainable and efficient in real-world driving. By coupling deep learning with symbolic reasoning, with domain knowledge, and through facilitating counterfactual reasoning, this research aims to contribute towards enhancing the safety, explainability, and trustworthiness of autonomous vehicles.

2.6 Related Works

This section covers a review of recent research on explainable AI (XAI) for autonomous vehicles (AVs), emphasizing the approaches that will offer solutions to the limitations of deep learning models and offer increased transparency in AV decision-making. We themed our reviewed works are by categories and critically analyzed to determine their strengths, weaknesses, and relevance to this study.

2.6.1 Counterfactual Explanations

Counterfactual accounts provide a good means of depicting AI explanation as the building of hypothetical tales of minor modifications to inputs such that the model output varies in some corresponding way. One can reason about what features were key to a decision and why the model responded correspondingly. Counterfactual explanations can be employed to explain why an AV made a specific decision like braking or lane change, by showing how the decision otherwise would have been generated if some aspects of the input like pedestrians or light status would have existed. STEEX (Steering Counterfactual Explanations) is a method of creating counterfactual explanations for AVs, proposed by (Jacob et al., 2022). Authors report >99.5% of success (fraction of explanations that flip the classifier).

STEEX applies semantic image synthesis and region-targeting specified by the user to create realistic counterfactual explanations. STEEX only alters specific regions of an image with global scene layout invariant and explainability of the output. It would be capable of converting a red light to green, for example, in order to try to expose the effect this has on the AV's decision to continue. STEEX relies on latent space manipulations and thus is not end-user intuitive to understand what the resulting feature modifications are that affect the model's action. The second is SAFE (Saliency-Aware Feature Editing) by (Samadi et al., 2023). SAFE utilizes saliency maps and an adversarial generative network to produce authentic counterfactual samples for AV tasks. SAFE creates realistically varied but sparse counterfactuals by maximizing the correct features revealed through saliency maps. Using BDD-OIA dataset, they achieved explanation validity of 93%. It can, for instance, nudge a pedestrian slightly within the image in order to display how this would lead to braking on the AV's part. But SAFE's response will be conditioned on saliency maps' accuracy and won't necessarily be context sensitive in the large context of advanced driving scenarios. STEEX and SAFE both aim to generate realistic counterfactual explanations of AVs but in opposing mechanisms and their compromise. STEEX is superior to perturbing semantic parts of the image, while SAFE perturbs the significant features learned with guidance from saliency maps. While both works hold promise, they also warrant further research into more open and context-aware approaches to counterfactual explanation of AVs.

2.6.2 Visual Explanations

Visual explanations assist with making AI more transparent by explicating knowledge on how AI draws conclusions through easy-to-grasp visualisations. They can cover everything from adding annotation to regions of an image that are essential to publishing visualisations of what's going inside the model.

For autonomous vehicles (AVs), visual explanations can be employed to explain what the world is seeing and how it sees the world, making its actions understandable for human. (Atakishiyev et al., 2023) propose a Visual Question Answering (VQA) methodology for enhancing explainability in AV decision-making.

The technique uses question-answer format to construct explanations of driving behavior. For example, if the AV needs to brake, the VQA system would respond to the question of "Why did the car brake?" as "The car braked because there was a pedestrian on the road." This way, the task has been found to perform well at questioning why the AV is doing something in simple cases and so can be tested on how it generalizes to difficult real-world scenario.

Another type of visual explanation is studied by (Mankodiya et al., 2022) which is interested in XAI for semantic segmentation of AVs. Semantic segmentation is to annotate each pixel in an image, such as "road," "pedestrian," or "vehicle." (Mankodiya et al., 2022) use Grad-CAM and saliency maps to identify where the model is focusing while performing semantic segmentation, marking the most significant region of the image upon which the model is depending to make predictions.

It can be employed to check what part of the scene the AV is focusing on when it reaches a decision. Using a KITTI dataset benchmark, they have reported IoU of 95% and 96% for train and test respectively. Nevertheless, the paper focuses primarily on the KITTI dataset, and that might limit its usage to other road and weather conditions. Also, it does not say anything about user-guided evaluation like does not say whether these visual explanations do or do not result in improved human perception and trust on the AV's decisions. Both of works use visual explanations for AV transparency but at different locations with some limitations. (Atakishiyev et al., 2023) place more focus on action driving in Q&A format,

while (Mankodiya et al., 2022) place more focus on model attention visualizations for semantic segmentation. All these limitations present the need to develop more and improved interpretable visual explanation methods for AVs that are capable of interpreting subtle real-world scenarios and generalize the AI explanation to human drivers effectively.

2.6.3 NeuroSymbolic AI and hybrid XAI methods

Researchers are exploring NeuroSymbolic AI and hybrid XAI solutions to improve over the limitations of traditional XAI methods in AVs. These solutions aim to make use of the advantage of deep learning and symbolic AI in creating explainable, transparent, and reliable AV systems.

(Rawat and Danda, 2023) suggest NeuroSymbolic AI to enhance performance, explainability, and transparency for tasks in human-autonomy teaming where a human and artificial intelligence teammate co-operate to achieve a shared objective. Rawat and Danda state that low-transparency and explainability techniques that exist nowadays are not appropriate for high-risk operations like AVs where human trust and understanding are of utmost importance. The authors propose combining symbolic knowledge bases and deep learning models to allow human beings to use them confidently in artificial intelligence systems. It would allow the AI system to capitalize on human experience and human knowledge and make human-explainable accounts of how it comes to its conclusions. The paper does not offer tangible experimental results and cut-and-dried policies for application, thereby limiting its real contribution to practice. Another competitor approach is introduced by (Tahir et al., 2024) as a hybrid XAI approach utilizing SHAP and LIME as a more computationally effective and efficient strategy.

SHAP and LIME are two commonly used XAI strategies with varying sets of benefits and drawbacks. LIME is relatively more efficient computationally but less it is less accurate compared to SHAP which is computationally expensive. With the hybrid of both of these methods, (Tahir et al., 2024) have attempted to achieve a balance between accuracy and computational efficiency. They have trained and evaluated the approach on Kitti dataset and achieved explanation consistency of 90% and reduction of explanation time by 40% compared to SHAP. In addition the paper reports > 85% for fidelity, interpretability > 80%, and > 70% consistency, 0.28s inference times on ResNet-18, 0.571s on ResNet-50, and

3.889s on SegNet. Although their hybrid method is well efficient regarding interpretability as well as computational complexity, there are some disadvantages attached as well. It is not aware of context, and as such it might or might not be able to encode relationship between multiple objects in a driving situation. It also doesn't rely on advanced reasoning, which is hard to achieve while representing multi-step decision-making. Lastly, it doesn't store domain knowledge explicitly like traffic rules and regulations, which are to be known and believed in order for AV to act. The constraints of existing XAI methods such as context-aware nature, inability to facilitate complex reasoning, and at times, dependency on domain knowledge, lead this work to consider NeuroSymbolic AI as a general-purpose explanation method for AVs. In this research we address such constraints by suggesting a NeuroSymbolic AI paradigm that is rooted in symbolic reasoning and deep learning, leveraging domain knowledge, and offering counterfactual reasoning.

2.6.4 End-to-End learning and imitation learning

Two of the most desirable approaches in autonomous vehicle (AV) research to enable AVs to learn driving behaviour directly from data or from imitating human drivers are imitation learning and end-to-end learning. However, even these approaches are not explainable and transparent.

(Paniego et al., 2024) propose end-to-end deep learning for lane keeping and obstacle avoidance in AVs. Their approach is to train a deep learning model to translate sensor inputs directly into driving commands in a way that the AV learns from massive sets of driving demonstrations. They report a success rate result of 86%. As effective as the approach demonstrates practicality and can be an expert in specific tasks, it shares the same deep learning model's limitations: being un-interpretable and lacking context awareness. The black box nature of such models means it is impossible to describe how and why simply the decisions are made, thus curtailing transparency and accountability. They were also unable to generalize to novel or new situations outside of the training set. (Muparutsa, 2024) brings a new line of work through causality analysis and physics-aware explainability marrying them with deep reinforcement learning for AV decision-making.

This approach is aimed at creating transparency in choice through physics-based knowledge and causal dependence among occurrences. Even though the model can learn causal signals and, further, generate explanations for particular actions, it is still computationally expensive when implemented in high-density settings where numerous interacting variables exist. This renders it less attractive for use in actual AV environments where there is a requirement for prompt response. (Hang et al., 2023) propose a human-like driving model for AVs with focus on lane change manoeuvres. They utilize modelling and simulating human-like driving behaviour to render AVs predictable and natural in their behaviour towards human drivers.

The model, despite being capable of producing natural driving action, is not certain how each decision is made. It does not explain exactly why it would make a decision to change lanes at a certain point and how it predicts the danger of the action. It lacks counterfactual analysis to use in examining various cases and observing how varying variables influence the decision-making process. These works recognize the limitation and potential of end-to-end learning and imitation learning for AVs. Even though such types of methods can be exquisitely gifted to produce good performance and naturalistic driving behavior, they are not transparent to some large extent.

(Xu et al., 2020) proposes a new explainable AI paradigm for autonomous driving between end-to-end and modular method with the focus of detecting action causing objects like pedestrians or red lights inducing driving actions. The authors introduce the BDD-OIA dataset annotated with 4 potential actions and 21 pre-defined explanations, and a multi-task CNN model that uses local object features and global scene context jointly to predict actions and explanations simultaneously. The model to explain its actions not only renders the model more explainable but also significantly enhances action prediction accuracy because the explanation task provides extra supervisory signal and serves as a regularizer, particularly for uncommon actions like turns and lane changes.

While it elegantly determines which objects played a role in a decision, no causal reasoning or symbolic rules are explicitly provided, and thus it is not able to perform counterfactual

analysis and generate explanations for cases that fall outside its pre-specified, finite annotation set.

(He et al., 2024) presents a robust reinforcement learning (RRL-SG) system that maximizes autonomous driving decision reliability efficiently through a combination of an adversarial training module for robustness and rule-based responsibility-sensitive safety (RSS) to avoid collisions, with impressive empirical performance of zero collisions during testing. Shortcoming of thus design is that it is modular rather than integrated: symbolic knowledge (RSS rules) and the neural network (policy) are distinct, with the symbolic system merely acting as a post-hoc veto filter on the output of the neural network. Therefore, the approach might not depict the safety rules themselves, which might lead to incoherent or suboptimal behaviour has no independent reasoning capability that can give safety guarantees independent of the predefined RSS contexts.

(Zhao et al., 2024) recommends a novel decision-making paradigm for highway autonomous driving based on a new safe reinforcement learning (RL) method named Replay Buffer Constrained Policy Optimization (RECPO), a derivation of constrained policy optimization (CPO) with importance sampling and a replay buffer to mitigate catastrophic forgetting and enhance data efficiency. The strategy frames driving as a constrained Markov decision process (CMDP) and learns to optimize reward and cost functions for safety, efficiency, and comfort. Their experimental result on CARLA indicate that RECPO converges faster, is safer (zero crashes), and more stable than other approaches. But it is bound by its assumption of noiseless simulation-based perception, its reduction to idealized highway conditions. The framework also lacks explainability.

That is where recent methods such as NeuroSymbolic AI must come in and bridge these gaps and make the process of developing AVs not only efficient but also comprehensible and reliable.

2.6.5 Synthesis and open research questions

The reviewed literature reflects a growing interest in explainable AI (XAI) for autonomous vehicle and investigating multiple avenues for transcending the limitations of deep learning

models. These techniques includes counterfactual explanations, which provide an explanation of the impact of input changes on the model output; visual explanations, which utilize graphical models to describe the explanation of why the AI; NeuroSymbolic AI, which combines deep learning and symbolic reasoning; and hybrid techniques that combine different XAI techniques.

Moreover, embedding domain knowledge such as traffic regulations and security protocols remains a primary concern which is not addressed by most XAI approaches. Finally, efficiency and scalability are of major importance for the application of XAI techniques to the real-time AVs scenario with limited computing capabilities. This research aimed to make a contribution to XAI for AVs by proposing a NeuroSymbolic AI framework that addresses the above shortcomings. The framework has integrated deep learning and symbolic reasoning in a manner such that the AV learns from experience but the explicit knowledge and the logic rules remain a part of it. It will include domain knowledge in order to verify that the AV's choices will be consistent with traffic laws and safety guidelines. It will further assist counterfactual analysis in a way that what-if scenarios could be analyzed and there would be a clearer sense of the determinants of the AV's choice.

Therefore, there should be additional studies to develop more stable and interpretable methods of counterfactual explanations for AVs, explore new visual explanation techniques that can qualitatively capture the complex reasoning of AVs, design lightweight and scalable NeuroSymbolic architectures for practical AV use cases, create sophisticated evaluation frameworks to measure the explainability and reliability of NeuroSymbolic AI systems in AVs.

Through answering these research open questions, we can move towards a vision for creating fully explainable and trusted autonomous vehicle that can securely join our society.

Summary of some related works are provided in Table 2.1

Table 2.2 Summary of related work

Reference	Approach	Key Contributions	Limitations	Relevance to this Research
(Jacob et al., 2022)	Counterfactual Explanations (STEEX)	Combines semantic image synthesis with region targeting for realistic counterfactuals.	Operates on latent space, making it less transparent for non-experts.	Highlights the need for more transparent counterfactual explanation methods.
(Samadi et al., 2023)	Counterfactual Explanations (SAFE)	Leverages saliency maps and GANs for realistic counterfactuals.	Relies heavily on saliency map quality; may lack context awareness.	Reinforces the need for contextually aware counterfactual explanations.
(Atakishiyev et al., 2023)	Visual Explanations (VQA)	Uses a question-answer format to explain driving actions.	Evaluated in simplified simulated environments; limited generalizability.	Shows the potential of visual explanations but highlights the need for real-world evaluation.
(Mankodiya et al. (2022)	Visual Explanations (XAI for Semantic	Uses Grad-CAM and saliency maps to visualize model focus.	Lacks generalization across datasets and user-centric evaluation.	Emphasizes the need for more comprehensive and user-focused visual explanations.

	Segmentation)			
(Rawat and Danda, 2023)	NeuroSymbolic AI for Human-Autonomy Teaming	Advocates for NeuroSymbolic AI to improve explainability and performance in HAT.	Lacks experimental results and specific implementation strategies.	Motivates the exploration of NeuroSymbolic AI for explainable AVs.
(Tahir et al., 2024)	Hybrid XAI (LIME-SHAP)	Combines LIME and SHAP for improved interpretability and efficiency.	Lacks contextual awareness, struggles with complex reasoning, and does not integrate domain knowledge.	Highlights the limitations of existing hybrid XAI methods and motivates the need for a more comprehensive approach.
(Paniego et al., 2024)	End-to-End Learning for Lane Following	Demonstrates practical efficiency of end-to-end learning.	Lacks interpretability and context awareness.	Shows the limitations of purely data-driven approaches and motivates the need for XAI.
(Muparutsa, 2024)	Physics-Aware Interpretabili	Integrates physics and causality analysis into deep reinforcement learning.	Computationally expensive in complex environments.	Highlights the importance of considering causality but also the need for efficient solutions.

	ty and Causality			
(Hang et al., 2023)	Human-like Driving Framework	Simulates human-like driving behavior in lane change maneuvers.	Lacks transparency in decision-making and does not involve counterfactual analysis.	Shows the potential of imitating human behavior but emphasizes the need for explainability.
(Sana et al., 2023)	Hybrid approaches for AV decision-making and control	In-depth coverage of classical and state-of-the-art methods to AV navigation in challenging situations such as adverse weather and unsignalized intersections. Identifies issues in benchmarking, explainability, and safety.	They does not provide in-depth discussion on integrating symbolic reasoning with neural approaches for enhanced explainability.	They recommends the need for neurosymbolic AI to combine symbolic reasoning for explainable rules and neural networks for handling complex, dynamic AV scenarios motivated by the goal of explainable decision-making.
(Xu et al., 2020)	Explainable object-induced action decision	Authors proposed a multi-task CNN model for driving action and explanation joint prediction according to	Black-box character of CNNs can impose limits full transparency.	The strength of neurosymbolic AI in enhanced explainability by merging symbolic representations of action-causing objects and neural

		action-inducing objects. Introduces BDD-OIA dataset for difficult scenes. Their finding shows that explanations improve action prediction accuracy		scene understanding and augmenting transparency and causality is highlighted.
(Hu et al., 2025)	Hybrid model for decision making and planning methods	Authors have reviwed rule-based (knowledge-driven), game-theory, imitation, and reinforcement learning (data-driven) methods, and hybrid models. They emphasizes robustness and safety in dynamic environments.	The hybrid models lacks discussion on explainability for. Does not say how symbolic knowledge can be represented so that ethical real-time decisions can be made.	Authors argues that neurosymbolic AI can bridge the gap between knowledge-based rules and data-driven learning to enable explainable and robust AV decision-making frameworks for complex traffic situations.
(He et al., 2024)	Robust reinforcement learning (RL) with safety guarantees	Suggests an adversarial resilient actor-critic method with a safety mask called responsibility-sensitive safety model to	Limited discussion on explainability of RL decisions.	Neurosymbolic AI can augment this endeavor through bridging symbolic safety guidelines and RL to produce interpretable explanations for robust

		deliver collision free AV choices amid uncertainties.		decision-making under uncertainty.
(Atakishiyev et al., 2024)	Explainable AI (XAI) for AVs	Explores in depth the structure of XAI for AVs, encompassing a suggested end-to-end driving explainability framework, with regulatory compliance and social trust through explainable AI decisions.	They does not addresses and validate transforming neurosymbolic solutions into real-time explainability.	This frane work is directly relevant because neurosymbolic AI has the ability to utilize the proposed XAI framework, including symbolic reasoning for regulatory-compliant explanations and neural processing for real-time perception and decision-making.
(Zhao et al., 2024)	Safe RL with constrained policy optimization	Presents a technique called replay buffer constrained policy optimization (RECPO) for safe and robust highway driving decisions, with zero-collision performance in simulation.	They focus on only on highway environment will limit generalizability to city or out-of-the-box settings.	Neurosymbolic AI can combine symbolic constraints with RL to enable understandable safety guarantees, assisting in enabling more transparency for the model decision-making in more contexts.

CHAPTER THREE

3. METHODOLOGY

3.1 Research Process Overview

The research process begins with the determination of the research domain and conducting a thorough literature survey to establish open problems and voids. From this, specific research questions are defined. The data is then collected and preprocessed for quality and feasibility to analyze. The appropriate deep learning models are selected and trained on the dataset followed by the development of a knowledge representation scheme. A method for integrating deep learning and reasoning is chosen, and the model is retrained over the new configuration. The retrained model is implemented in a driving simulator for use in the real world, with thorough testing.

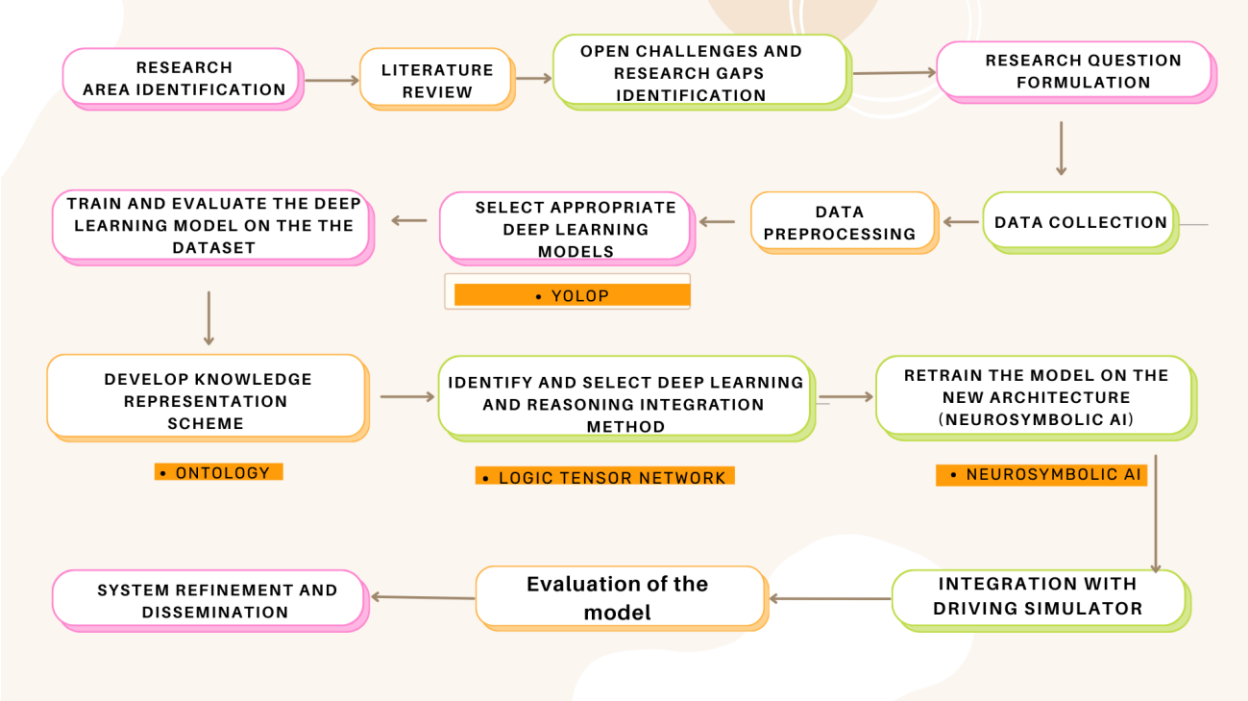


Figure 3.1: Research Process Overview

3.2 Data Collection

We employed BDD100K-OIA (Xu et al., 2020) (Berkeley DeepDrive Driving Object and Instance Attention Dataset) as the main dataset for learning of the model and verifying the efficacy of the

proposed autonomous driving perception system. Environmental variability encompassed within the dataset, for example, urban, highway, and busy traffic environment, served as a good foundation for testing the object detection and lane keeping task. The densely annotated 2D/3D bounding boxes, lane markings, and semantic segmentation masks of BDD100K-OIA permitted wide model learning with diversified illumination, climate, and occlusion. Real-world application, large-sized annotated sample, dataset variability, and stringent industry standard benchmarks for rigorous verification of proposed approach and comparison summary offered by Table 3.1 were reasons for dataset selection.

Table 3.1 Comparison of the object detection and lane keeping datasets for autonomous driving

Dataset	**Tasks**			
	Object Detection	**Lane Detection**	**Drivable Area**	**Size (**No. Of Images**)**
BDD100K-OIA (Xu et al., 2020)	√ (1.8M boxes)	√ (Lane masks)	√	100K
Apollo Scape (Huang, X., et al. 2018)	√ (3D & 2D)	√	√	140K+
KITTI (Liao, Y., Xie, J., & Geiger, A., 2022).	√ (Limited)	× (Partial)	×	15K
TuSimple (Grün, F., Nolte, M., & Maurer, M., 2024)	×	√ (Lanes only)	×	6K

3.3 Dataset Description

BDD100K-OIA has served the purpose of perception for autonomous driving research by offering a large-scale and diverse dataset that facilitates multitask learning. The BDD100K-OIA contains more complete annotations, more visual diversity, and the potential to support multiple tasks of different difficulty levels compared to existing datasets.

In order to assist in decision-making and explainability processes associated with object detection and lane-keeping capabilities in autonomous vehicles, we collected a dataset made up of traffic rules and integrated it with labels which were taken from the BDD100K-OIA, focusing on the operational environment of Ethiopia.

3.4 Data collection process

3.4.1 BDD100K-OIA dataset acquisition

The BDD100K-OIA dataset was obtained from kaggle which contains 92,000 driving images (1280×720 resolution) across 70% training, 10% validation, and 20% test splits. We chose these splitting strategy to have sufficient model training, hyperparameter optimization, and unbiased performance evaluation. Making the test dataset larger than the validation data allows representative evaluation of model generalization, particularly well-suited given the class imbalance and traffic scenario variation of the dataset. In addition we have collected real local video data from Adama city street.

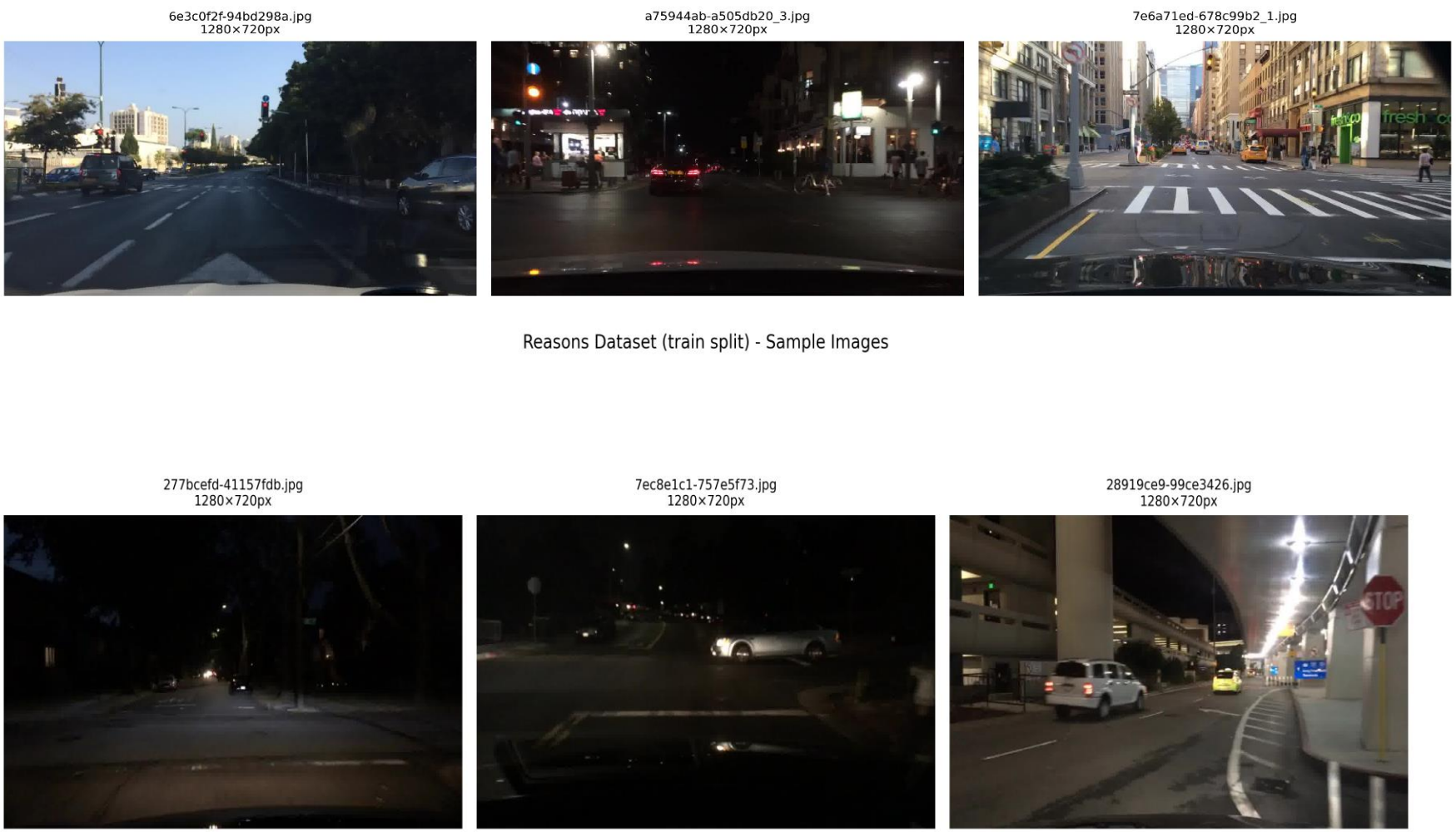


Figure 3.2: Sample dataset image

3.4.2 Identification of traffic rule sources

Authoritative documents of traffic rules were sourced from FDRE Ministry of Transport and Logistics through Adama City transport agency. This source provided detailed regulations, such as speed limits, right-of-way rules, and traffic signal compliance. Domain experts were consulted to ensure completeness and accuracy of the rule set.

3.5 Model Selection

In this research YOLOP (You Only Look Once for Panoptic Driving Perception) was selected for perceiving task. YOLOP is an integrated deep learning framework for autonomous driving tasks through the unison of three essential capabilities: object detection, lane detection, and drivable area segmentation (Wu, D., Liao, 2022). With one convolutional neural network (CNN) architecture, YOLOP processes input images by a single forward pass to generate outputs for traffic item detection such pedestrians and cars, lane mark identification, and drivable area segmentation. This end-to-end architecture leverages a common feature extractor and task heads, enabling real-time efficiency for resource-constrained environments like autonomous driving. YOLOP reaches state-of-the-art performance by simultaneously optimizing a multitask loss function which allows it to achieve a balance between the detection and segmentation tasks' goals (Niu Y., and Zhang, J., 2025).

Its ability to perform excellently and effectively with both lane keeping and object detection in autonomous vehicles (Wu, D., Liao, 2022) is our reason for selecting YOLOP model. The single-pass processing of the model consumes fewer computational resources than utilizing separate models for each of the tasks, which is vital for real-time application on embedded systems. Its ability to detect objects (such as cars and pedestrians) and to recognize lanes simultaneously offers ideal awareness for safe driving. YOLOP's pretrained weights was fine-tuned on BDD100K-OIA dataset which enhances it to generalize well to various driving scenarios, enhancing its generalization performance.

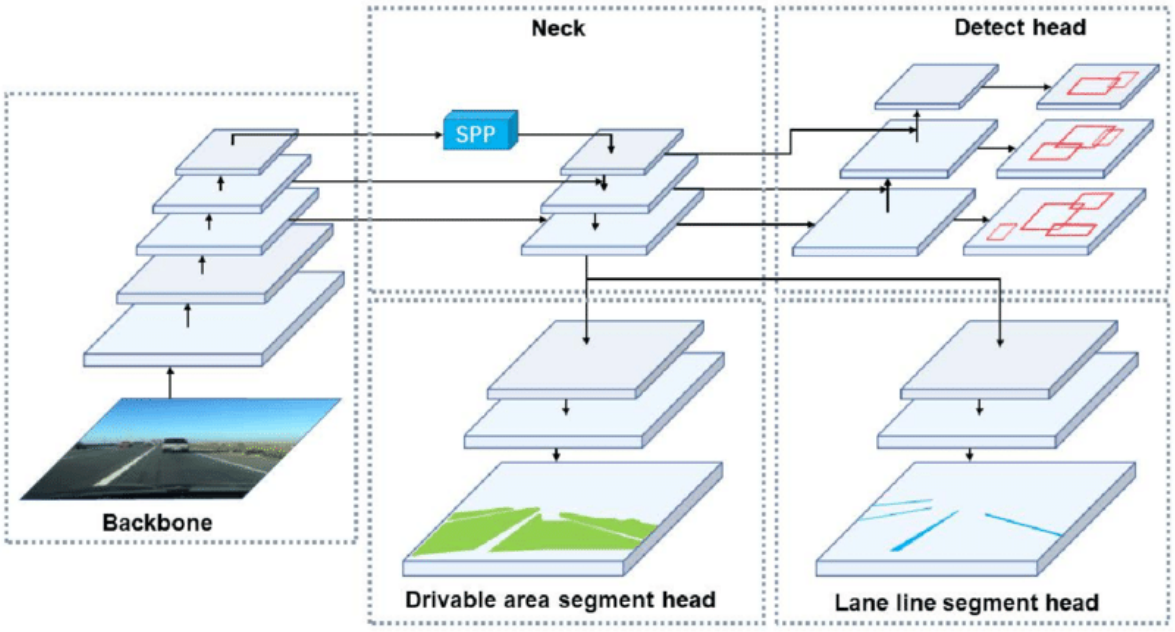


Figure 3.3 YOLOP architecture (Wu, D., Liao, 2022)

3.5.1 Feature Fusion

To combine the multi-task outputs of the YOLOP model containing object detection (bounding boxes for vehicles, pedestrians, and traffic signs), lane detection (pixel-wise lane masks), and drivable area segmentation (binary maps), a Feature extraction module is used to generate joint feature maps for the Logic Tensor Network (LTN). In this case YOLOP uses feature pyramid network (FPN) and path aggregation network (PAN) (Lin et al, 2017) as the feature fusion technique. It is a simple architecture consisting of only a top-down pathway thus minimizing the computational costs associated with implementation for autonomous vehicles in limited/low-resource environments like Ethiopia.

3.6 Preprocessing data

3.6.1 Converting BDD100K-OIA JSON Annotations to YOLOP Format

To enable object detection using the YOLOP model, the bounding box annotations of the BDD100K-OIA dataset, which were originally provided in JSON format, were transferred to YOLO format. JSON-to-YOLO Conversion is utilized for extracting BDD100k-OIA labels images JSON bounding boxes and normalizing coordinates to [0,1] range and

bounding box sizes in relation to an image of size 1280 × 720. The conversion process is defined as

$$Xcenter = \frac{x1+x2}{2x1280} \quad (3.1)$$

$$Ycenter = \frac{y1+y2}{2x720} \quad (3.2)$$

$$W = \frac{x2-x1}{1280} \quad (3.3)$$

$$H = \frac{y2-y1}{720} \quad (3.4)$$

Where X is x-coordinate of the bounding box center (scaled to [0,1]), Y is y-coordinate of the bounding box center (scaled to [0,1]), W is width of the bounding box relative to the image width and H is height of the bounding box relative to the image height.

3.6.2 Data Cleaning and Standardization

3.6.2.1 Traffic rules extraction and formatting

Text-based rules from Adama town transport agency were parsed and normalized into structured TrafficRule nodes with attributes such as rule_id, violation, decision, and description. To extract and format traffic rules from Ethiopian traffic rule documents for integration into the Logic Tensor Network (LTN) and First-Order Logic (FOL), a formal system for knowledge representation using predicates, variables, quantifiers, and logical connectives (Russell & Norvig, 2010) was applied. This approach processes raw text from sources such as the FDRE Ministry of Transport and Logistics, obtained via the Adama city transport authority, to produce structured logical representations that ensure compliance with local regulations in autonomous driving scenarios. We have applied manual rule extraction to identify and translate into structured formats. The reason for selecting manual rule extraction from the document is to achieve high accuracy, and to handle complex rules and ambiguities well.

rule_id	reason_id	reason_text	violation_check	decision	explanation
R001	reason_1	Obey speed limits	reason_1 = 1	forward	Adjust speed dynamically based on real-time road limits
R002	reason_2	Follow traffic signals	reason_2 = 1	stop	Stop at red lights when red light is detected
R003	reason_3	Follow traffic signals	reason_3 = 1	forward	Proceed cautiously on green when green light is visible
R004	reason_4	Follow traffic signals	reason_4 = 1	forward	Yield on yellow unless unsafe when yellow light is visible
R005	reason_5	Stop at stop signs	reason_5 = 1	stop	Ensure a full stop then proceed only when clear
R006	reason_6	Yield to pedestrians	reason_6 = 1	stop_and_yield	Always prioritize pedestrians at crosswalks
R007	reason_7	Respect right-of-way	reason_7 = 1	yield_to_emergency_vehicle	Yield to emergency vehicles
R008	reason_8	Respect right-of-way	reason_8 = 1	yield_to_merging_traffic	Yield to merging traffic
R009	reason_9	Respect right-of-way	reason_9 = 1	yield_with_caution	Yield at uncontrolled intersections with caution
R010	reason_10	No reckless driving	reason_10 = 1	maintain_stable_driving	Avoid sudden acceleration and unsafe maneuvers
R011	reason_11	Stay in the correct lane	reason_11 = 1	switch_to_correct_lane	Adhere strictly to lane markings
R012	reason_12	Use turn signals	reason_12 = 1 AND (action_left = 1 OR action_ri	activate_turn_signal	Signal lane changes and turns with sufficient notice
R013	reason_13	Avoid lane weaving	reason_13 = 1	restrict_lane_changes	Limit lane changes to essential safe situations
R014	reason_14	Maintain safe following distance	reason_14 = 1	increase_following_distance	Keep a 4 second gap
R015	reason_15	Do not block intersections	reason_15 = 1	avoid_entering_intersection	Refrain from entering if traffic ahead prevents crossing
R016	reason_16	Pedestrians have priority	reason_16 = 1	stop_and_yield	Stop immediately for pedestrians near roadsides
R017	reason_17	Yield to emergency vehicles	reason_17 = 1	pull_over_safely	Pull over promptly and safely when emergency siren is detected
R018	reason_18	First come first served at 4-way stops	reason_18 = 1	proceed_sequentially	Proceed sequentially based on arrival order
R019	reason_19	Yield when merging	reason_19 = 1	yield_and_match_speed	Match speed and yield to existing traffic flow
R020	reason_20	Roundabout rules	reason_20 = 1	yield_to_roundabout_traffic	Yield to all vehicles already within the roundabout
R021	reason_21	Left turns	reason_21 = 1 AND action_left = 1	yield_to_oncoming_traffic	Yield to oncoming traffic unless a protected green arrow is active

Figure 3.4: Sample traffic rules dataset

The completeness and correctness of rules was evaluated the industry expert. Accordingly we get response from 31 experts and from these respondants we get average 4.5/5 for the correctness and 4.7 for completeness of extracted rules.

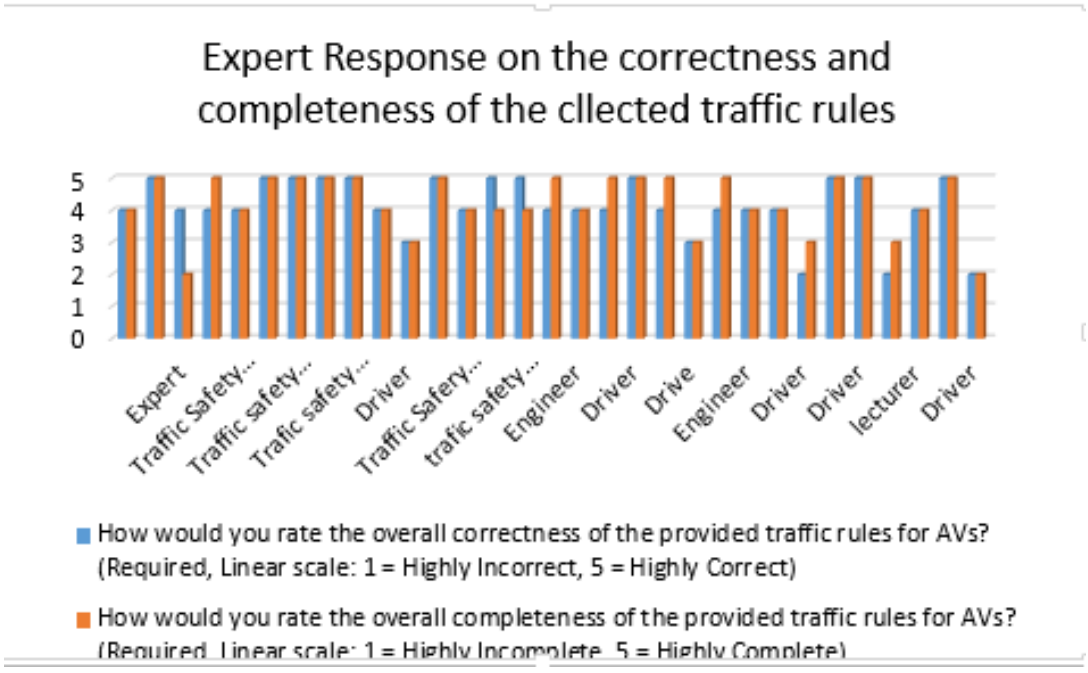


Figure 3.5 Expert feedback on the correctness and completeness of the collected traffic rules

3.7 NeuroSymbolic AI architecture design and evaluation

The research employs a mixed-methods methodology that combines aspects of design science research, experimental evaluation, and qualitative analysis to comprehensively examine the potential of NeuroSymbolic AI for increasing the explainability and safety of autonomous vehicles (AVs). This section describes the four primary phases of the methodology: NeuroSymbolic architecture design and knowledge formalization which is responsible for designing the NeuroSymbolic AI architecture and formalizing the domain knowledge that is needed in operating AVs, implementation and integration which deals with execution of the architecture, experimental analysis and evaluation where the system's performance is tested under safety-critical conditions and the generated explanations are evaluated, and system refinement and dissemination which encompasses the enhancement of the system informed by the evaluation outcomes, alongside the distribution of the research results.

3.7.1 NeuroSymbolic architecture design and knowledge fformalization

In this research, we design NeuroSymbolic architecture to enhance autonomous vehicle safety. Our hybrid architecture was seamlessly integrated with deep learning perception modules with a YOLOP model (Wu D. Liao, 2022) for object detection and lane detection, augmented with a symbolic reasoning engine for facilitating advanced decision-making and explanation generation.

The integration process include incorporating symbolic rules as constraints in the form of represented neural network layers using logic tensor networks. This approach allows for data acquisition by the neural network under the set rules and constraints pertaining to traffic law and road safety procedures. For example, the rule red light means stop can be incorporated as a logical constraint in the system so that the autonomous vehicle always stops when it encounters a red traffic light along with any other learned behavior. This incorporation guarantees decisions based on conventional traffic regulations while simultaneously enabling the system to learn to handle new situations. The neural network part of the architecture will then analyze this information and deduce the relationships among the various entities in the environment, including cars, pedestrians, and traffic infrastructure. This capability will allow the autonomous vehicle to make decisions based on learned patterns in addition to learned knowledge. There is a message-passing system to allow real-time exchange between the symbolic reasoning unit and the deep learning

units. The deep learning units sends messages regarding the perceived environment to the symbolic reasoning unit, and the symbolic reasoning unit sends back constraints and commands to the deep learning units derived from symbolic knowledge.

For example, when the deep learning module recognizes a red traffic light, it notifies the symbolic reasoning engine, which then triggers the rule that "a red light indicates stopping," thereby constraining the output of the deep learning module responsible for the acceleration of the vehicle. This integration allows our NeuroSymbolic architecture to achieve a synergistic merging of deep learning and symbolic reasoning which enable the AV to learn more complex driving behaviors from experience without compromising safety, reason about its environment, and explain its decisions in a human understandable manner.

At the core of this research is the formalization of relevant domain knowledge in autonomous vehicle safety. This involves the construction of a comprehensive framework for representing knowledge that attempts to capture traffic rules, safety procedures, and ethical consideration in a symbolic representation that can be utilized within the NeuroSymbolic AI system. we have used ontology (Collenette et al., 2022) for traffic rules schema to specify relations among various concepts and rules employed in autonomous vehicle decision-making and it was designed to be compatible with neural network architecture and facilitate effective reasoning and explanation generation.

3.7.2 Implementation and Integration

This stage is covers NeuroSymbolic architecture design. The most significant element of realization is building a symbolic reasoning engine. The reasoning engine will deduce from the AV maneuvers based on the formalized domain knowledge, justify its actions, and identify possible safety violations.

To test and analyze the NeuroSymbolic AV system in a real-world environment, it was instantiated with the BDD100K-OIA dataset and a local traffic rule dataset. The datasets offer real-world diverse driving scenarios and region-specific traffic rules, which allow complete testing and analysis of the AV system. The task of implementation involves integrating the NeuroSymbolic architecture with the datasets to process images from BDD100K-OIA, and local traffic rules. The perceived data are processed by the deep learning components first and then forwarded to the symbolic reasoning engine to make

decisions and generate explanations. The datasets was used to develop a set of safety-critical scenarios to test the system's robustness and explainability.

3.7.3 Evaluation and analysis

In this research NeuroSymbolic AV system's functionality and explainability is evaluated across a large variety of safety-critical scenarios using video from Adama city street. This involves evaluating the the system both quantitavilely and qualitatively.

Quantitative evaluation

The proposed NeuroSymbolic AI for AV system was quantitatively compared with the baseline models. The four large-scale metrics used in performance and safety measurement are: Mean average precision at 50% IoU (mAP50) for perception accuracy, F1-Score for overall decision and explanation performance. These evaluation provided a clear, measurable indication of the system's effectiveness in perception, decision-making, safety, and traffic rule compliance.

Two data sets, the BDD100k-OIA dataset that has 18,400 images for static cases and video recorded from Adama city street for dynamic real-world urban cases was tested. The BDD100k-OIA dataset has enabled image-based experiments and permit perception and decision-making under different driving conditions. We used the Adama city street video data to test the system’s ability to handle temporal dynamics and context-specific traffic patterns. mAP50 and F1-score was computed for image-based evaluations, where as video-based evaluations focuses on F1-score.

The NeuroSymbolic architecture, with domain-specific trained YOLOP with LTN and an Ethiopia-specific traffic rule ontology, was evaluated against the baseline models. The baseline models lacks advanced rule integration this allows us to evaluate the impact of domain-specific training and rule-based reasoning. In addition to that, performance of NeuroSymbolic will also be compared against state-of-the-art autonomous driving F1-Score methods so that its performance in safety-critical applications can be put into context.

Qualitative Analysis

Qualitative analysis was conducted to evaluate the explainability, clarity, correctness, trustworthiness, and usefulness of the explanations provided by our proposed NeuroSymbolic AV framework. Feedback was collected from 32 users with diverse background including engineers, urban planners, and non-technical community members in accordance with video recordings of Adama city street. This combination of participants offers a guarantee that explanations are assessed from both technical and non-technical perspectives, reflecting the needs of most stakeholders in an actual urban setting.

Four qualitative aspects was examined using questionnaires: clarity and understandability of explanations (CUE), correctness of explanations (CE), trustworthiness of Neurosymbolic AI decisions (TND), and usefulness of explanations (UE). CUE, CE, and TND was rated on a 5-point Likert scale (1 = poor, 5 = excellent), whereas UE was rated "Yes," "Maybe," or "No."

By using qualitative analysis, how the NeuroSymbolic system produces human-understandable explanations that provide justifications for decisions in the LTN rule-based reasoning and Ethiopia-specific ontology of traffic regulations was identified. The analysis trends define where explanation design improvement and stakeholder acceptability can be determined, particularly in the challenging urban setting of Adama city street.

3.7.4 System refinement and dissemination

The final stage of the method is calibration of the NeuroSymbolic AV system through qualitative feedback and trial experiments. This include adjustments to the neural network architecture, adding or altering rules in the knowledge base, as well as adjusting the brevity and clarity of explanations generated. The goal is to enter into a cycle of system development with the objective of incrementally improving its overall performance, interpretability, and trustworthiness. Results of this research is to be disseminated by publication of papers in refereed journals and conferences, oral presentation at appropriate workshops and seminars, and potentially by making available open-source software as well as by making available supporting material. This is helpful in enhancing knowledge of XAI application in AVs and bring innovation in more explainable and safe autonomous driving technology.

3.8 Development Tools

This research was designed and implemented using a variety of development tools. These tools include UML modeling tools, and other appropriate tools needed to conduct the research.

These development tools were briefly described in the following sections.

3.8.1 Design Tools

As a flexible and powerful graphic design tool, Canva was used in the development of the suggested system. This tool provided a lightweight yet effective way to produce professional-looking visual representations.

3.8.2 Hardware Tools

Table 3.2 Hardware tools requirement

No.	***Tools***	***Used for***
1.	RAM	Facilitating efficient data processing and analysis, ensuring optimal computational performance throughout the research endeavor.
2.	GPU	Boosting the processing speed and efficiency of image and analysis tasks, accelerating the training process.
3.	Hard Disk	Providing reliable storage capacity.

3.8.3 Software Tools

Table 3.3 Software tools Requirement

No.	***Software Tools***	***Description***
1.	python	Python, a versatile and high-level programming language, is renowned for its readability and extensive library support,

		making it widely favored for data science, machine learning, and web development.
2.	skit-learn	Scikit-learn, a Python ML library, leverages NumPy and SciPy for efficient data analysis and modeling. Its user-friendly interface suits both beginners and experts, offering diverse machine learning algorithms.
4.	Jupyter notebook	Jupyter Notebook is an open-source web app for creating and sharing documents with live code, visualizations, and text. Widely used in data science, it enables interactive and collaborative environments for reproducible analyses.
5.	Pytorch	PyTorch, developed by Facebook, is a flexible and user-friendly open- source deep learning framework.
6.	PyTorch	PyTorch, an open-source machine learning framework by Google, is widely recognized for its flexibility and scalability in building and training deep learning models, with a robust ecosystem for seamless deployment across platforms.
7.	Keras	Keras, a user-friendly Python API, simplifies building and training neural network models on top of frameworks like TensorFlow.
8.	LTNtorch	Python API used to integrate FOL and perception predicate

CHAPTER FOUR

4. PROPOSED MODEL AND ARCHITECTURE

4.1 Chapter Overview

In this chapter, we describe a NeuroSymbolic AI system developed to improve safety and explainability of autonomous vehicles (AVs). We use deep learning for real-time perception and reasoning over rules that define the local traffic rules and promote explainability of decisions made. The architecture combines the YOLOP model, a logic tensor network and a reasoning engine to achieve a designed-balance for good computational efficiency with trustworthy explainable actions while navigating complex roadway environments.

4.2 Architecture Overview

Our system is comprised of four main modules namely a YOLOP-based perception module, rule extraction and formatting module for Ethiopian road traffic rules, logic tensor network, and rule enforcer and explanation builder as the reasoning engine. The modules interleave information through a message-passing protocol, taking in camera feed from the BDD100K-OIA dataset (Yu et al., 2020) to output safe and road-rule-abiding driving decisions.

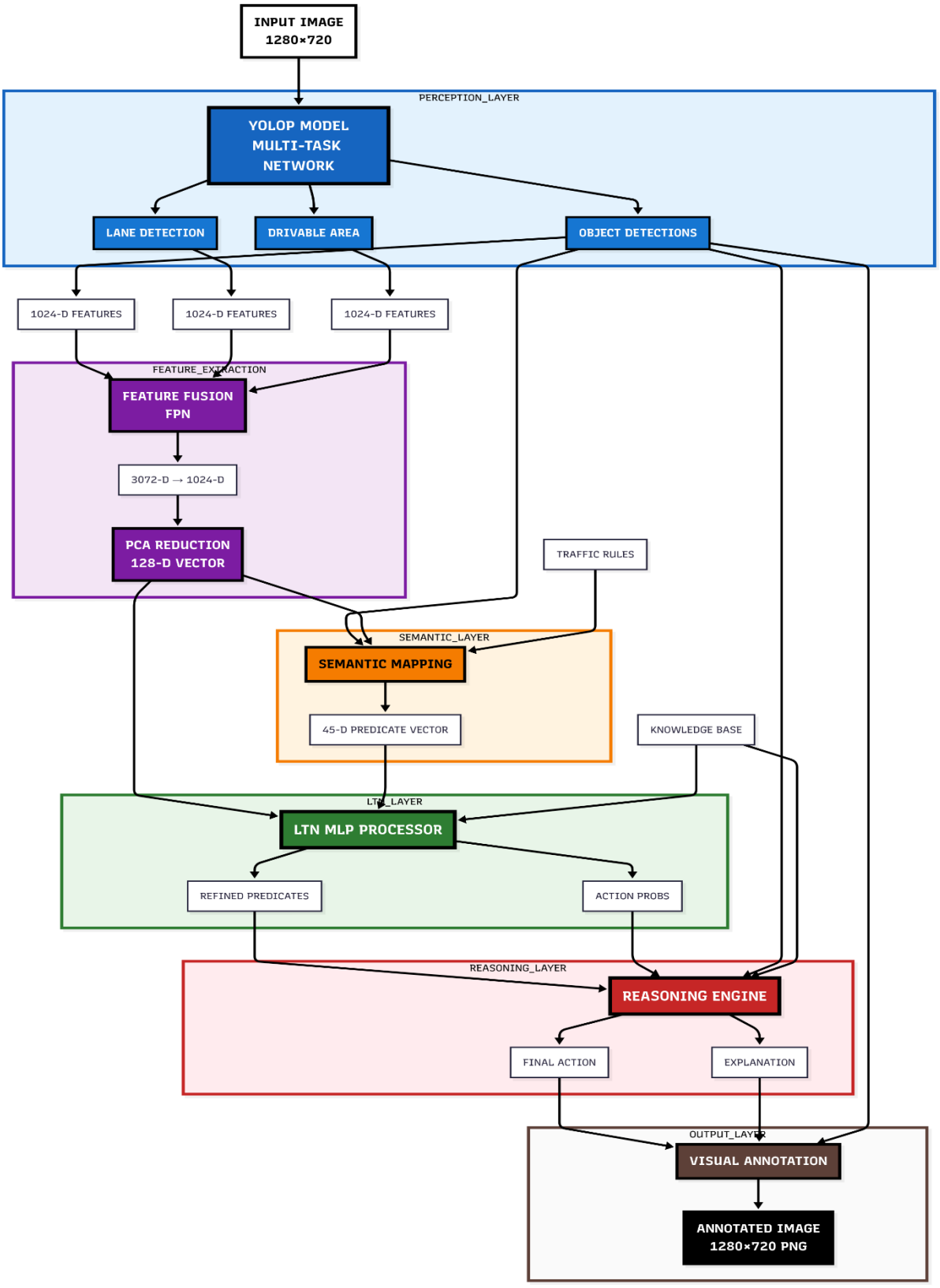


Figure 4.1 The proposed Neurosymbolic AI Architecture for autonomous vehicle

4.3 The proposed Neurosymbolic AI algorithm

The following algorithm formalizes the organization of our neural-symbolic autonomous driving framework integrating perception, knowledge representation, and reasoning. It begins with pre-training a YOLOP model for detecting objects, drivable space, and lanes, and parsing Ethiopian traffic laws into a structured knowledge base. A Logical Tensor Network (LTN) is then initialized to represent predicates and apply logical constraints according to the rules. In training, LTN is taught to combine perceptual features with symbolic knowledge by learning an optimal multi-component loss function that includes action prediction, rule satisfaction, and logical consistency. It then infers features from a new image, verifies rule violation, identifies the most significant rule, and creates a driving action with a natural language explanation and annotated image, thereby demonstrating integration of symbolic reasoning and neural perception for explainable and safe autonomous decision-making.

4.3 Description of the architecture components

4.3.1 Perception with YOLOP

YOLOP model (Wu et al., 2022) is a multi-task deep learning model that detects objects of bounding boxes with corresponding class labels and confidence levels, detects lanes of pixel-wise lane masks/lane centerlines to classify lane types, and performs segmentation of the region that is drivable to output binary segmentation maps separating drivable from non-drivable regions. It performs all the three tasks in a single shot, utilizing a shared encoder and three task-specific decoders, which is optimized for efficiency, designed to work on limiting resource hardware. YOLOP extracts visual features using camera inputs to perform visual recognition of objects, lanes and drivable areas, allowing the vehicle to perceive its environment in real time. The shared encoder reduces computational burden, making the model well-suited to low-powered AV systems.

The network shares a single encoder across all decoder heads that takes input RGB images, extracts features, and enables object detection, drivable area segmentation, and lane detection. The encoder consists of two sub components namely backbone and neck. **Backbone** is where the hierarchical features are extracted in order to output the input

image. The use of a single backbone, with shared features for all tasks, is less computationally expensive than having multiple unique backbones.

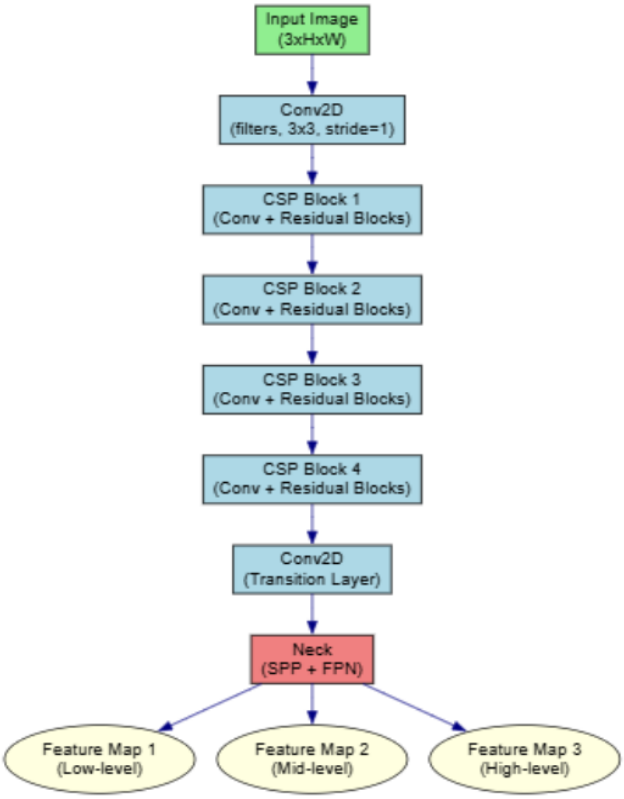


Figure 4.2 Backbone component of YOLOP encoder

Neck the neck connects elements of the backbone to better optimize representation from a multi-scale and multi-level perspective, which is necessary to address varying sizes and distances for the objects it is detecting. The neck is composed of a spatial convolution pyramid pooling layer (SPCL) and feature pyramid network (FPN) modules. The SPCL pools the features from different scales, and adds useful contextual information for the distance between the object of interest and the vehicle, like distance to a vehicle travelling in a lane. The FPN collects the features top down, so that it is able to pass the semantic information down to the lower-level feature maps. A path aggregation network (PAN) improves bottom-up fusion, providing spatial information. Multi-scale fusion provides solid detection and segmentation because objects can appear large (an example of this is a

bus at a crossing) or small. The neck provides multi-scale characteristics to the decoder heads.

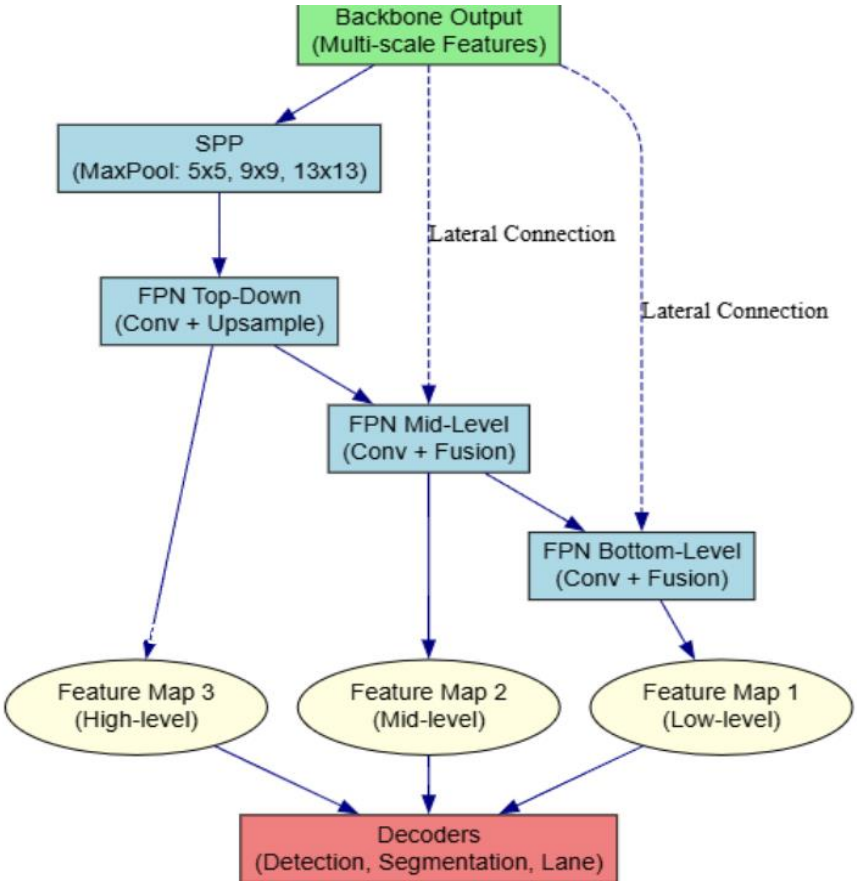


Figure 4.3 Neck component of YOLOP Encoder

A decoder is another component of YOLOP which has three different heads, each individually constructed for a task, using the feature maps from the encoder to enable effective multi-task learning. The decoder consists of detect head, drivable area segmentation and lane segmentation head components

The Detect Head component uses bounding boxes, class probability, and confidence scores to detect traffic objects. It uses a Path Aggregation Network (PAN) for bottom-up location feature transfer and an FPN for top-down semantic feature transfer. The multi-scale feature maps are used together to predict bounding box offsets (x, y, width, height), class labels and confidence scores. Mix of classification (cross-entropy), object presence (binary cross-

entropy), and bounding box (CIoU) losses is used for detecting head loss function.

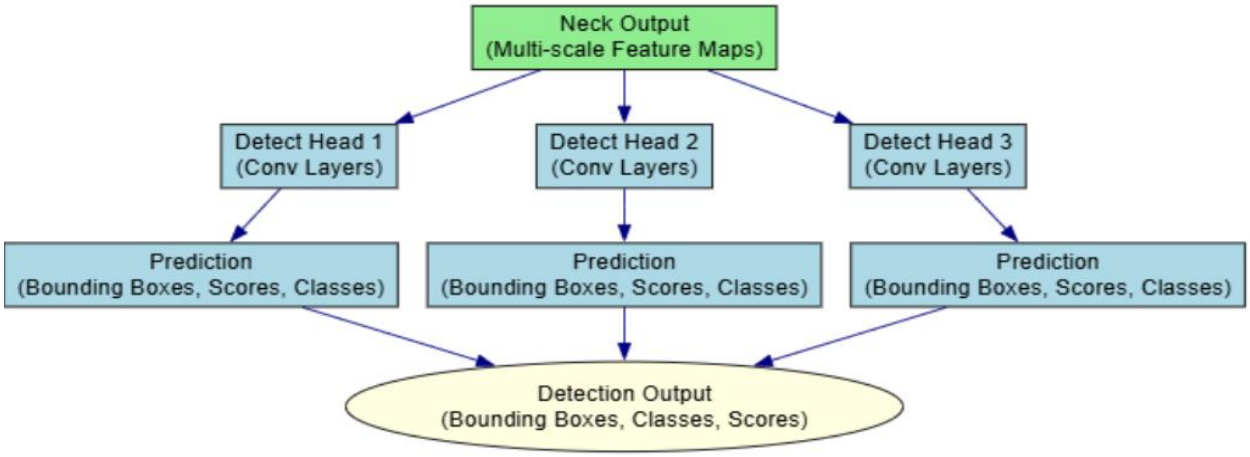


Figure 4.4 Detect Head component of YOLOP Decoder

Drivable area segmentation and **lane segmentation head** performs pixel-wise segmentation to create masks of drivable areas and lane lines. It takes the lowest FPN layer as input and applies three upsampling operations to reach a pixel-wise drivable/non-drivable or lane/non-lane probabilities. For loss function both heads use cross-entropy loss for pixel-wise classification and IoU loss for the lane detection to account for the sparse lane markings.

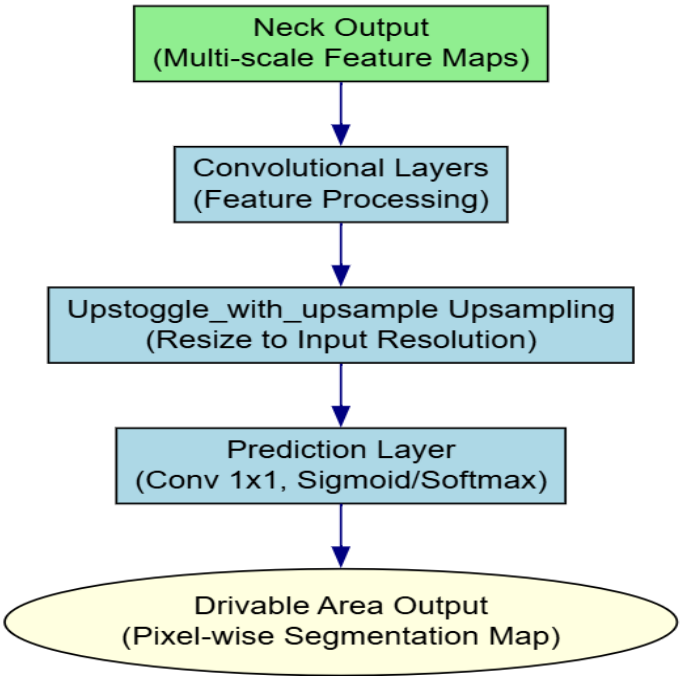


Figure 4.5 Drivable area segmentation of decoder component of YOLOP

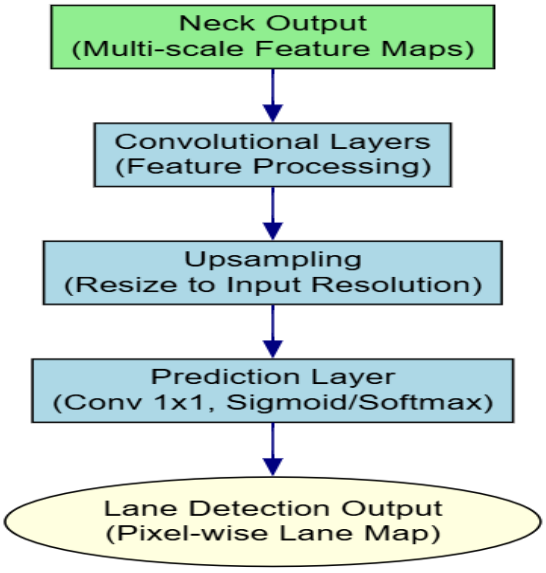


Figure 4.6 Lane detection Component of YOLOP architecture

Table 4.1: Summary of YOLOP multi-task

Task Name	Description	Output Format	Vector Size	Example
Object Detection	It is used to detect object with bounding boxes, classes and confidence	List of detection + feature vector	1024-D	Detection: [{"class": "traffic_light", "conf": 0.95, "bbox": [600, 100, 620, 120], "color": "red"},……] Feature Vector: [0.95, 0.90, 0.0, ..., 0.1] (1024 elements where starting indices encode high-confidence detections, the remaining encode background or low-conf objects)
Lane Detection	Produce binary mask or polyline for lane lines.	Binary mask (1280×720) + feature vector.	1024-D	Mask: Tensor (1280×720, 1 channel) with value=1 at centerline (x≈640). Feature: [0.80, 0.75, 0.0, ..., 0.2] (1024 elements, encoding lane alignment, for example 0.80 shows strong lane detection at x=640).
Drivable Area Segmentation	Produce a pixel-wise mask for drivable and non-drivable areas.	Segmentation mask (1280×720) + feature vector		Mask: Tensor (1280×720, 1 channel) with drivable pixels=1, non-drivable=0. Feature Vector: [0.85, 0.0, 0.0, ..., 0.15] (1024 elements, encoding drivable region size, obstacle absence. For example 0.85 indicates a clear road).

The central integration is of Feature Extraction with LTN where YOLOP features are fused into embeddings/predicates and matched by LTN with Ontology-defined rules to produce constrained outputs. This makes the decisions both data-driven (perception) and rule-compliant (KB).

4.3.2 Feature extraction module integration

The Feature Extraction module maps YOLOP outputs to produce integrated feature maps and predicates for LTN.

Feature Pyramid Network (FPN) is used for feature fusion which combines multi-scale features of YOLOP's three tasks into a single uniform feature map of 1024-D tensor vector. It integrates object, lane, and drivable area to capture spatial and semantic relationships like pedestrian near lane edge.

It takes in three 1024-D vectors (one for each task) as inputs, then concatenates and passes through convolutional layers to produce a uniform feature map.

It produce output containing a single high-dimensional tensor 3072-D (1024×3 tasks) vectors that further pooled into 1024-D for simplicity. FPN ensures multi-task coherence like aligning pedestrian bbox with drivable area to confirm crosswalk location.

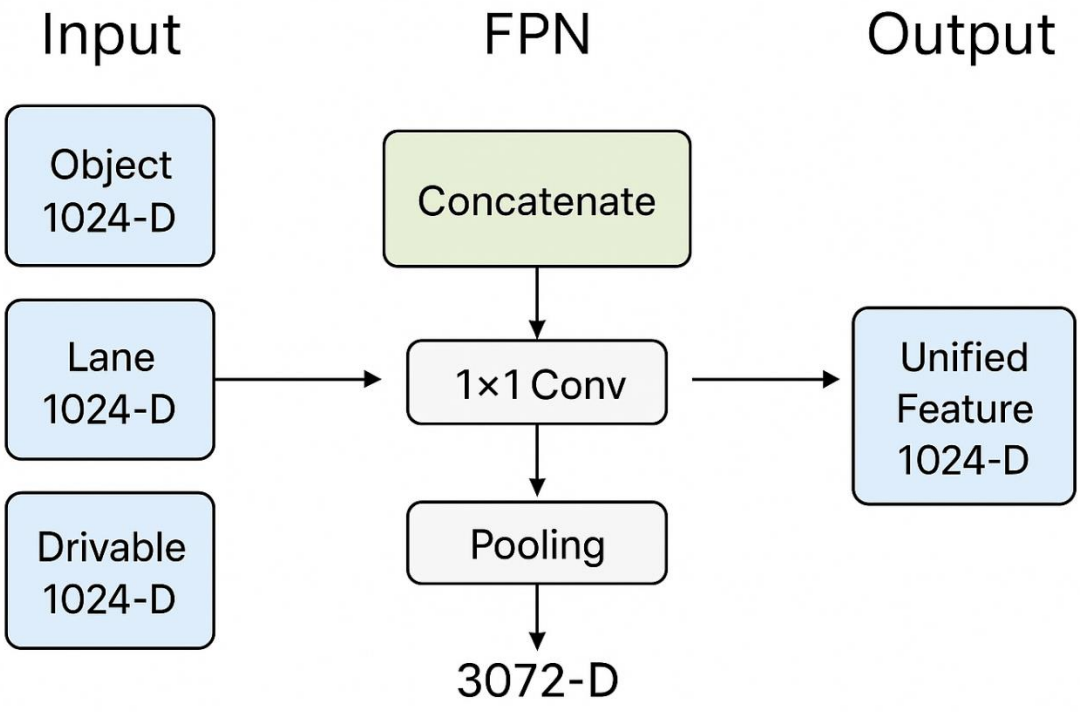


Figure 4.7: Feature pyramid network for feature fusion

Principal Component Analysis (PCA) reduction to 128-D vector is used to reduce the combined 1024-D vector of FPN to a 128-D vector to make it computationally efficient for LTN retaining 95% variance. It takes fused 1024-D vector and projects to 128-D using a pre-trained transformation matrix (pre-trained on training data so that dominant features are retained). It returns 128-D vector, where each entry is a linear combination of input features, normalized like [0, 1].

The PCA 128-D vector is dense and neural and contains implicit scene semantics for LTN's MLP input.

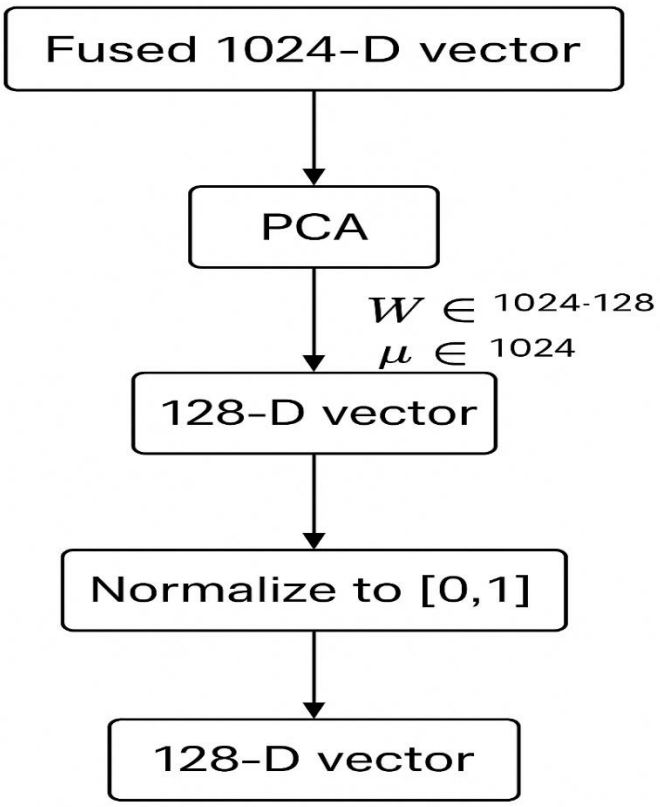


Figure 4.8: PCA feature dimension reduction

Semantic Mapping to 45-D predicate vector

Semantic Mapping takes YOLOP detections and the 128-D vector and convert them into a 45-D predicate vector in which each entry corresponds to a violation_check condition from the traffic rule data (CSV file). Truth values [0, 1] are mapped depending on confidence scores, IoU metrics, or rule-specific logic.

It takes YOLOP detections and 128-D vector as input. It produce a 45-D predicate vector, where each dimension maps to a rule in the CSV with truth values. The 45-D vector is sparse and symbolic and is one-to-one mapped from the 45 rules in the CSV.

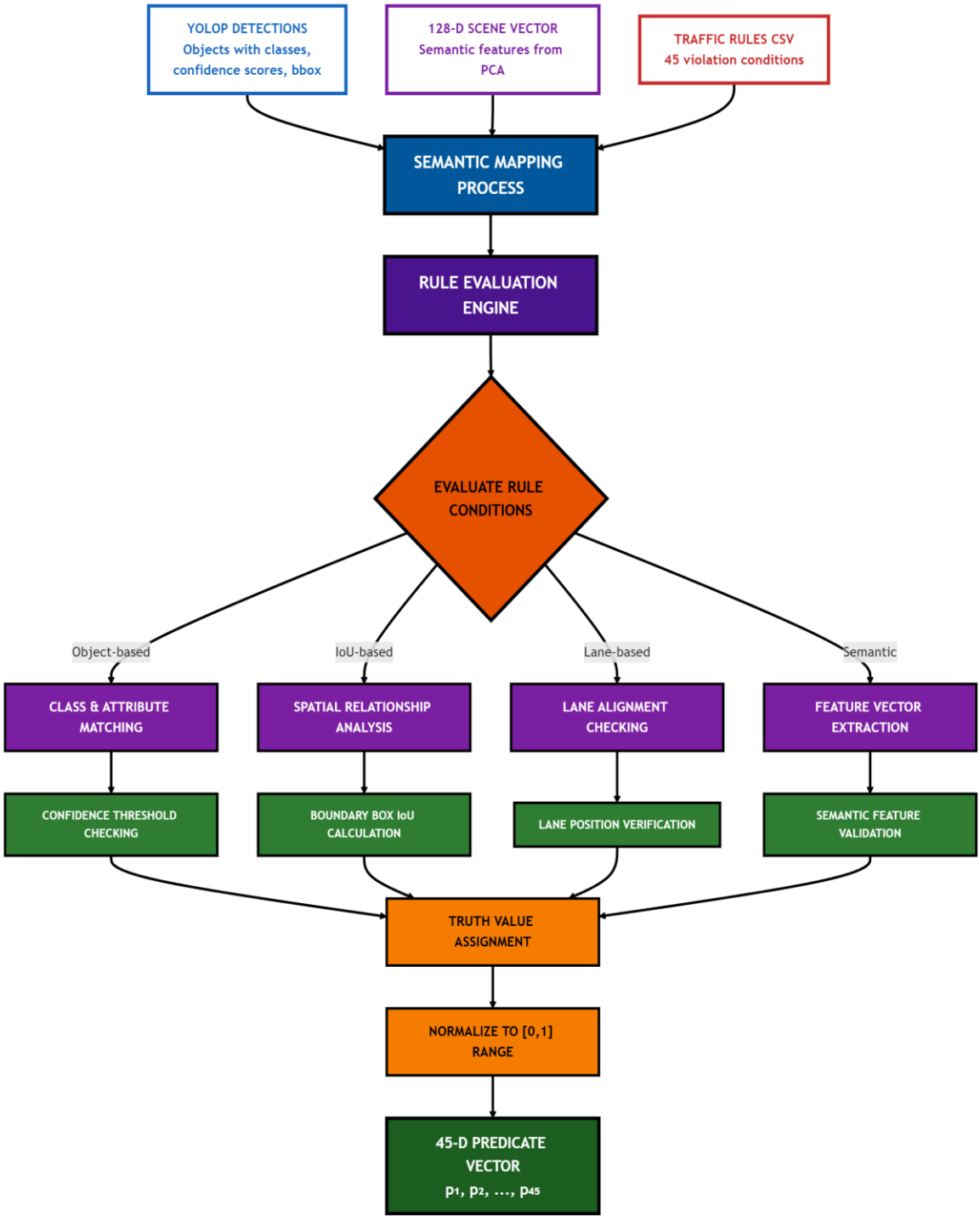


Figure 4.9 Semantic Mapping for converting detection and 123-D feature to 45-D predicate vector

4.3.3 Ontology/KB Construction from traffic rule data

The rules CSV is mapped to an ontology that is represented as classes (such as TrafficRule, Reason, ViolationCondition, DecisionAction), Object properties (such as TrafficRule hasReason Reason), Data properties (such as Reason hasText string), axioms; a logical rules like TrafficRule(R001) ∧ Reason(reason_1) ∧ ViolationCheck("reason_1=1") → Decision("forward") ∧ Explanation("Adjust speed dynamically."), and FOL Predicates for LTN that is derived from violation_check column like P_speed_limit_violation(x) ≡ (reason_1 = 1), where x is a scene instance.

Neo4j is used to implement this knowledgebase by mapping the ontology and rules in the CSV file as nodes and relations. The graph model facilitates efficient querying, reasoning, and updating dynamic rule dependencies and provides an interpretable interface to integrate LTN. This KB feeds into LTN for specifying connectives and constraints (like enforcing no contradictions like simultaneous "stop" and "forward").

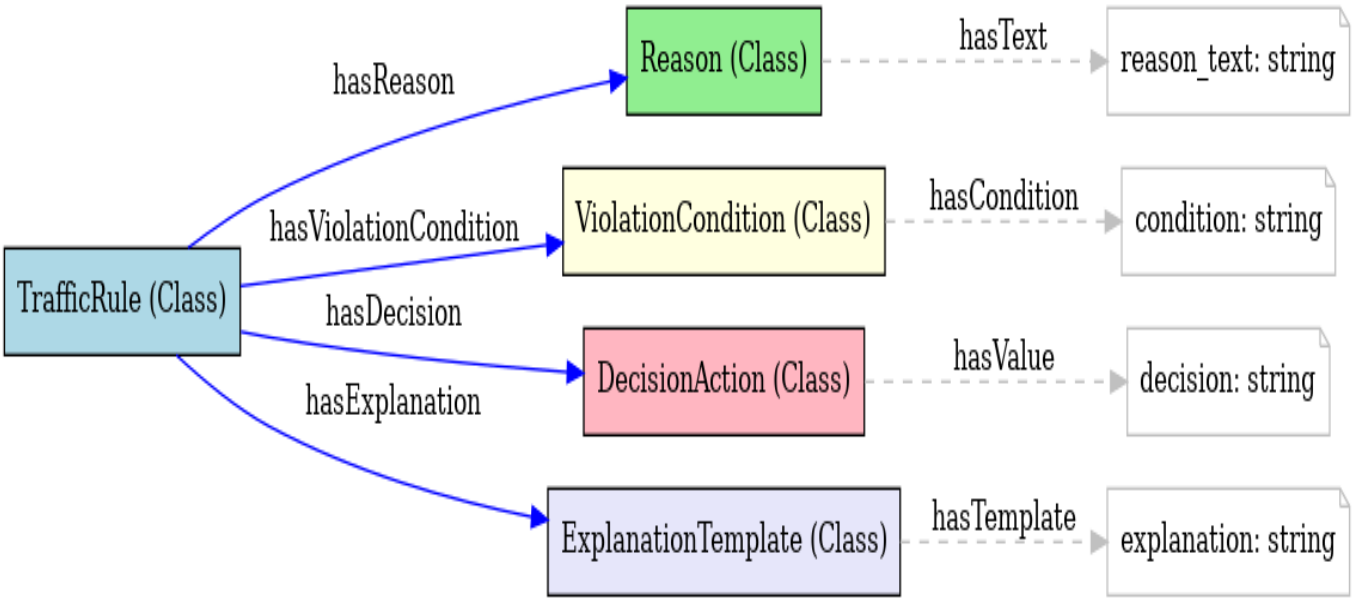


Figure 4.10: Ontology Schema diagram

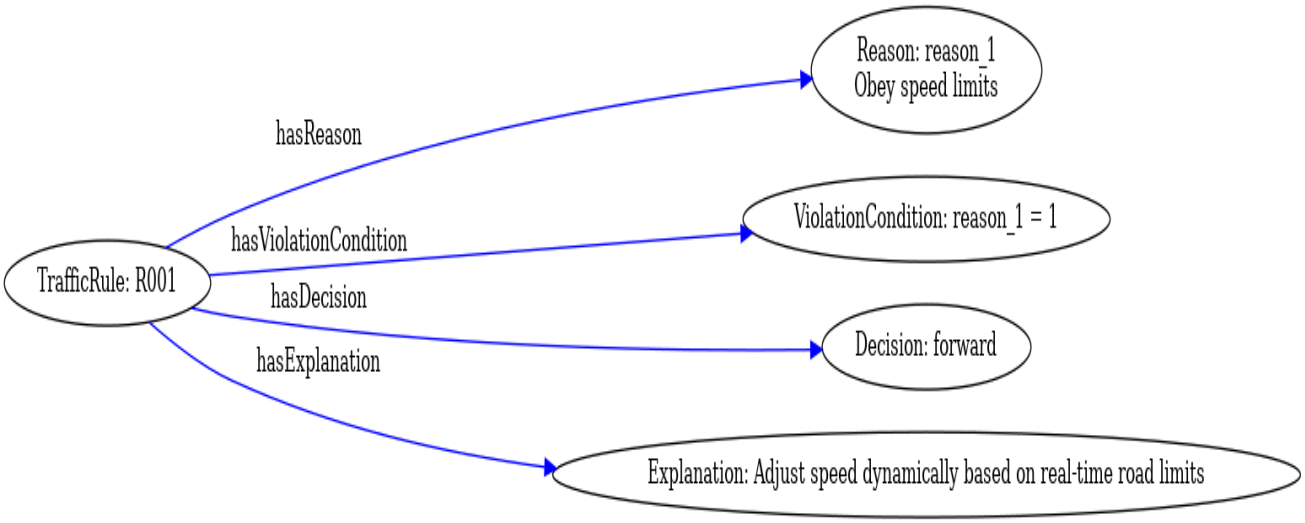


Figure 4.11: Sample Ontology classes and properties

4.3.4 Logic Tensor Network (LTN) integration

LTN embeds the Ontology's FOL rules into a neural network, using real-valued logic (t-norms for AND/OR, implications as residuals). It takes fused feature vectors/symbolic predicates from feature extraction component, and FOL constraints from KB. LTN

processes the feature vector and predicates, applying Ontology-derived constraints. It process a multi-layer perceptron (MLP) maps features to predicate satisfaction scores [0,1].

MLP in Logic Tensor Network (LTN)

The LTN's MLP maps the 128-D vector to predicate satisfactions and action probabilities, enhancing the 45-D predicate vector under KB constraints. It process a 3-layer MLP architecture (128→64→45 for predicates and 128→64→N_actions for action.

The MLP model take 128-D vector (like [0.92, 0.88, 0.75, .]) and initial 45-D predicates (like [0.0, 0.95, ., 0.85, .]) as input and execute forward pass with updated predicate scores (45-D vector) prediction and action probabilities (softmax over decisions) prediction. It also process constraints by renormalizing scores to satisfy KB axioms using product t-norm. The MLP predict a 45-D vector predicate and action probability. The MLP is used to process predicate scores to align neural predictions with constraints that ensure KB compliance.

The LTN produces action probabilities, predicate scores and vital violation which will serve as input to the reasoning engine.

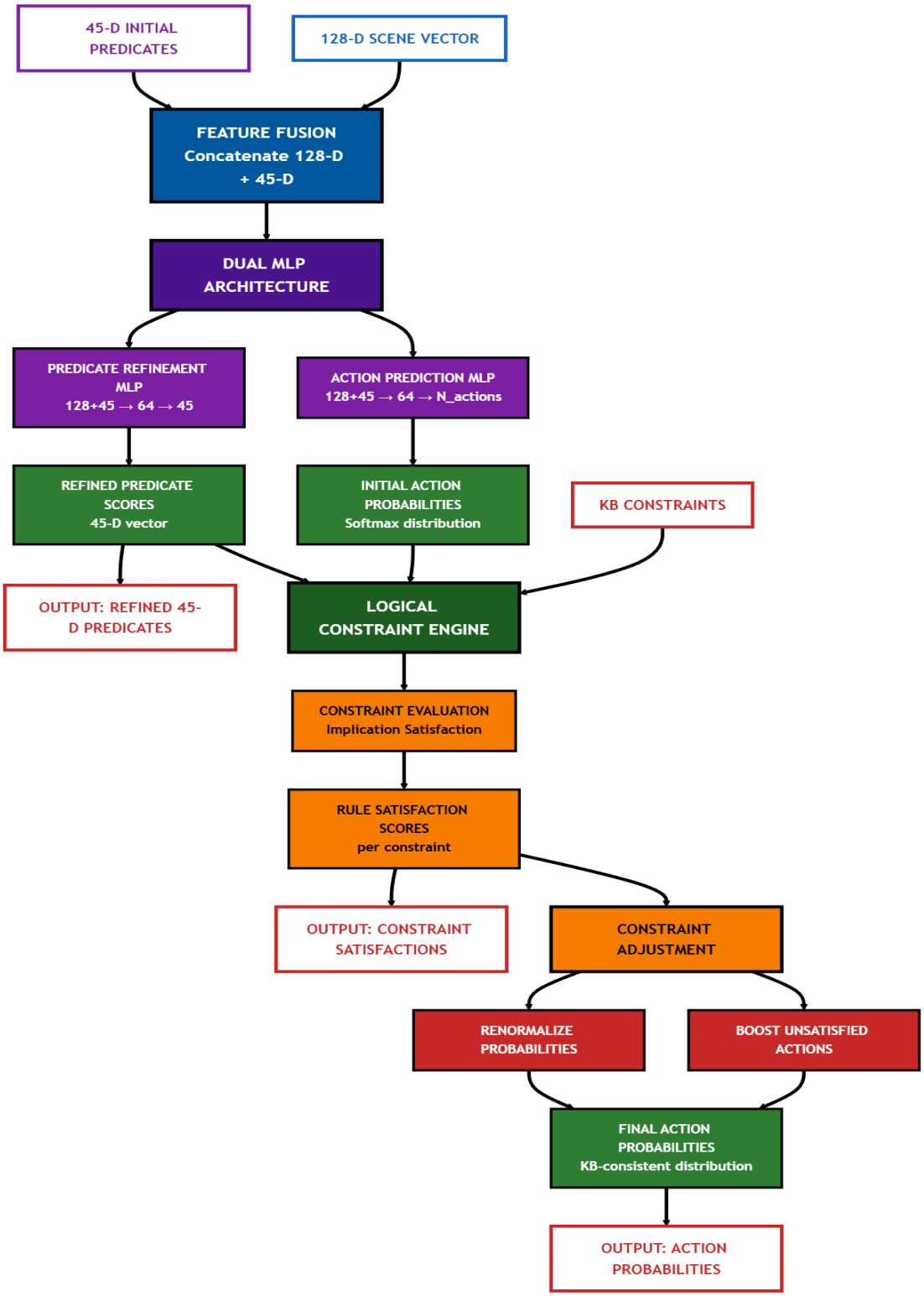


Figure 4.12: Logic Tensor network module for integrating NN with symbolic rules

4.3.5 Reasoning Engine

The Reasoning Engine maps LTN outputs to a KB decision. It takes input from LTN outputs (action probabilities, predicate scores). It applies a logic to select the rule with the highest probability predicate score that maps to a decisions. The reasoning engine also generate explanation by combining explanations from rules using the KB's explanation field. The reasoning engine produces final action, explanation text, and additional data for annotation.

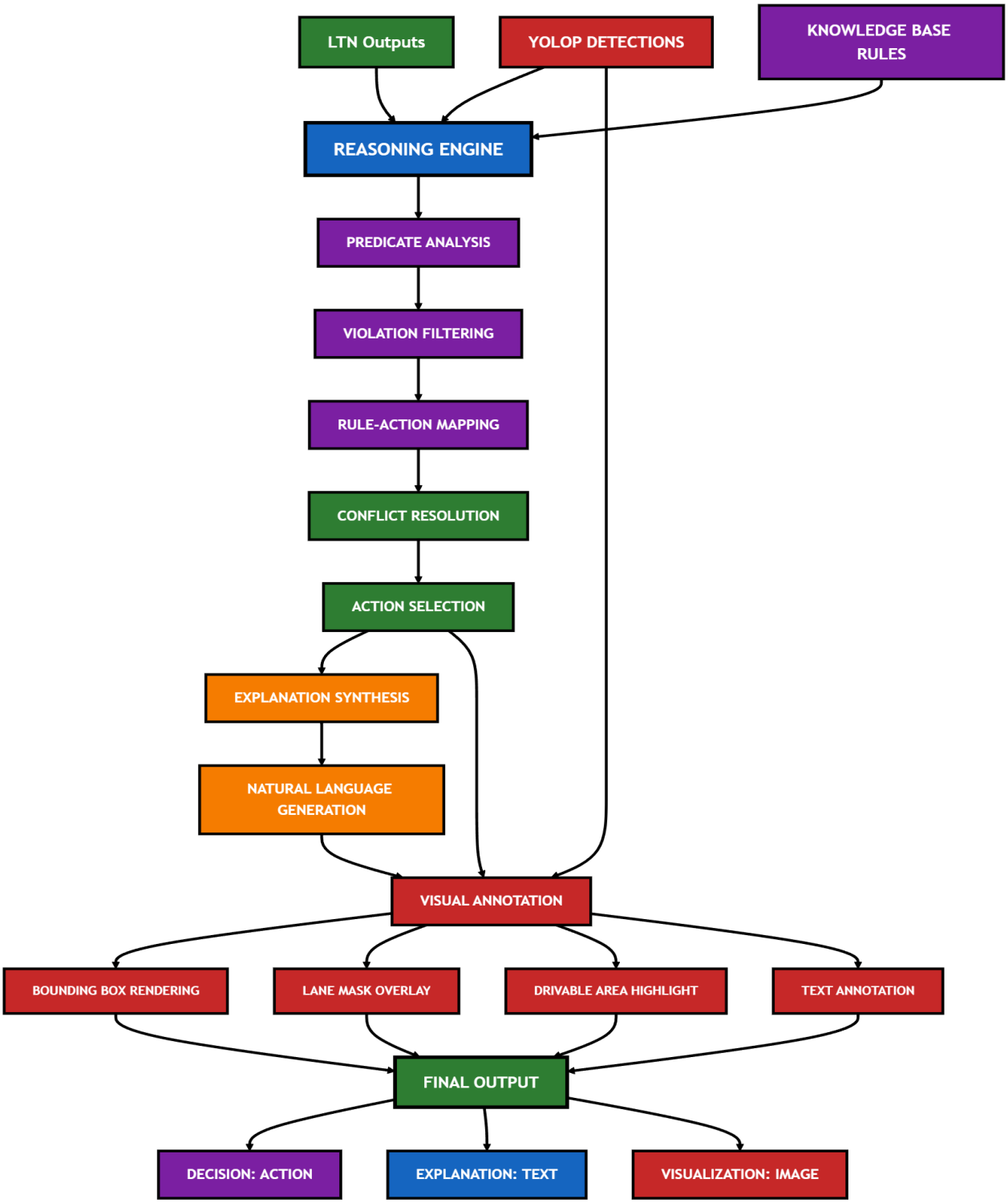


Figure 4.13: Reasoning Engine for making decision, generating explanation

CHAPTER FIVE

IMPLEMENTATION OF THE PROPOSED SOLUTION

5.1 CHAPTER OVERVIEW

This chapter presents a step-by-step, systematic implementation guide for the suggested architectural framework, which is adapted specifically to utilize the YOLOP on BDD100K-OIA dataset for perception tasks and the ontology of Ethiopia Ministry of Transport and Logistic traffic rules, LTN for integration, and reasoning engine for decision and explanations. This architecture aims to support autonomous driving applications in Ethiopian scenarios by fusing perception, logical reasoning, and rule-based decision-making mechanisms. The chapter contains preparations of the operational environment, data preprocessing, implementations of each component (data acquisition and preprocessing, YOLOP, rule extraction and formatting, logic tensor network, knowledge representation, and reasoning engine), and evaluation methodologies.

5.2 WORKING ENVIRONMENT

The execution was carried out in a computing environment tuned for processing large sized image datasets such as BDD100K-OIA and rule based reasoning.

The tools and libraries

- **Python** - Main programming language for all parts.
- **Jupyter Notebook** - Interactive environment for code development and visualization.
- **PyTorch** - YOLOP model training and inference framework.
- **OpenCV** - For loading and image preprocessing of BDD100K-OIA images.
- **NetworkX** - For building and querying the Knowledge Graph.
- **NumPy** - For array manipulation and numerical calculations.
- **Matplotlib** - To plot BDD100K-OIA data distributions and model outputs.
- **Scikit-learn** - Simplified neural network implementation in Logic Tensor Network.
- **Pandas** - To process structured traffic rule information from Adama Transport Agency.

Hardware Configuration

- **Operating System** - Ubuntu 20.04 LTS
- **Memory Size** - 64 GB RAM
- **Processor** - Intel Core i9-10900K, 3.70 GHz
- **GPU** - NVIDIA RTX 3080, 10 GB VRAM
- **System Type** - x64-based processor, 64-bit operating system

This arrangement enables effective processing of 92,000 images of BDD100K-OIA and sophisticated reasoning tasks.

5.3 DATA PREPROCESSING IMPLEMENTATION

Data preprocessing prepares BDD100K-OIA images for YOLOP and extracts Ethiopian Transport and Logistic traffic rules for reasoning components. Preprocessing steps include loading BDD100K-OIA images, resizing, normalizing, visualizing data distributions, and parsing traffic rules into a structured format.

5.4.1 Dataset Loading

5.4.1.1 BDD100K-OIA dataset

The BDD100K-OIA dataset, containing 92,000 images (64,400 training, 9,200 validation, 18,400 test) with annotations for object detection, lane marking, and drivable areas, is loaded from its directory structure.

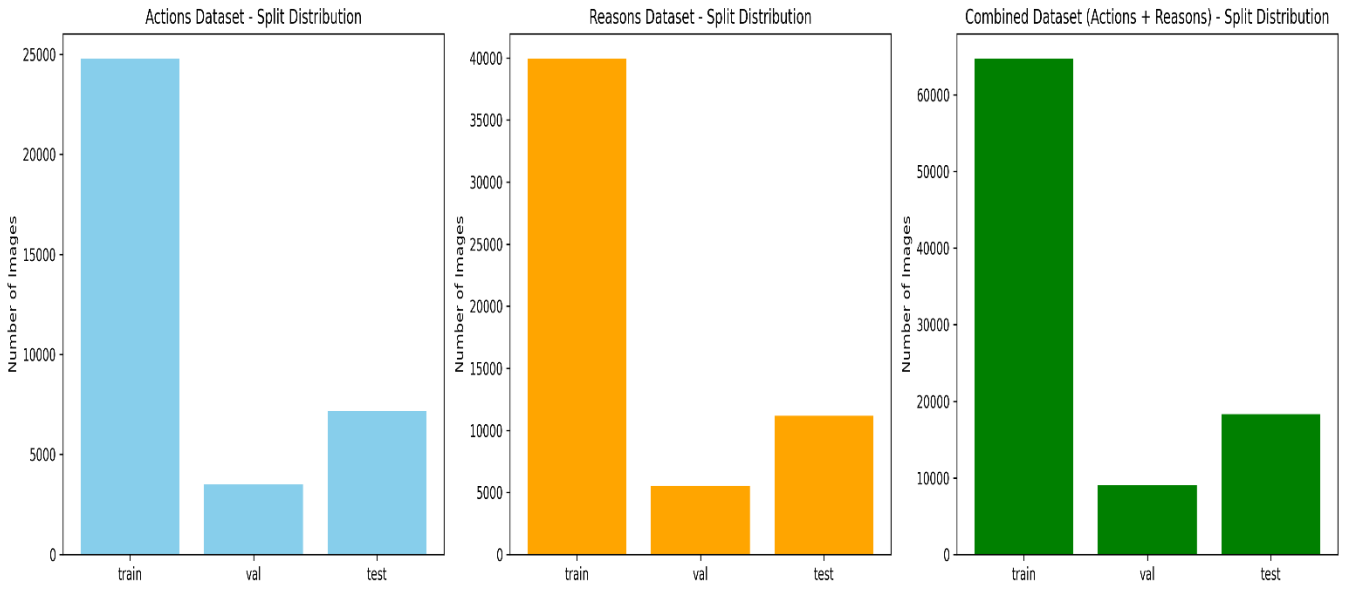


Figure 5.1 Dataset splitting

The dataset was organized for multitask such as object detection, lane detection and drivable area segmentation which is vital for the perception module of autonomous vehicle.

Class distribution

The actions dataset was evenly distributed in four classes: forward, stop, and left and right, which is a good representation for training. The distribution allows for good training with sufficient data in critical categories

Figure 5.2: Action Class distribution

Reason is another categories of dataset which is used to make explanation for action. There are 21 different reason classes distributed as shown on figure 5.3 which is splitted for training, validation and test.

Figure 5.3: Reason class distribution

5.4.1.2 Local Traffic Rules Dataset

We have retrieved and formatted traffic rules from Ministry of Transport and Logistic, Ethiopia traffic regulation documents manually in logic tensor flow compatible format. The detail of our traffic rule dataset can be found from this Kaggle link. https://www.kaggle.com/datasets/abdimosisa/ethiopia-traffic-rules

The dataset contains 45 different rules where each rules are categorized into nine categories as visualized on Figure 5.6

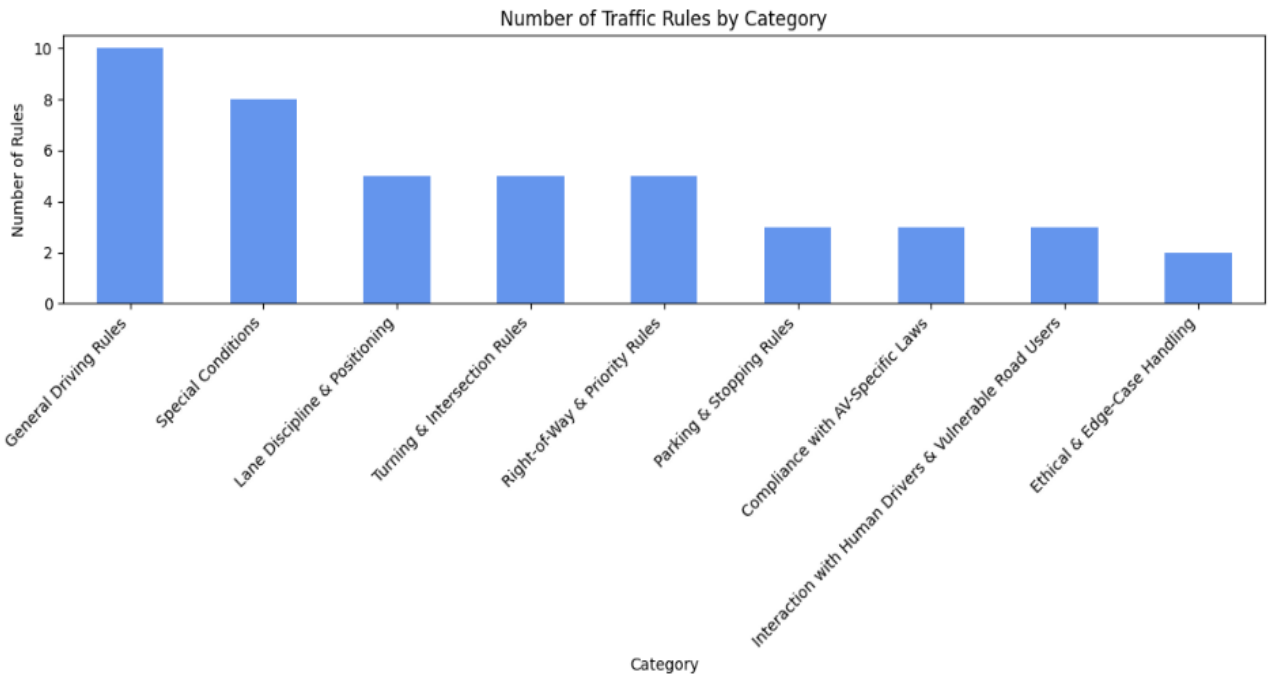


Figure 5.4 Traffic rule categories in our dataset

5.4.2 Dataset pre-processing

5.4.2.1 Converting BDD100K-OIA JSON Annotations to YOLOP Format

We have implemented the dataset loader function to retrieve a training or inference sample from the BDD100K-OIA dataset, including a road scene image or video frame and corresponding annotations for objects, drivable areas, and lane lines. This function ensures the AV system has access to diverse, real-world driving data for training or testing the YOLOP perception model.

It takes **i**nteger index input specifying the dataset sample from training or validation sets in the dataset directories. It then converts images to PyTorch tensors for compatibility with YOLOP's deep learning framework. Annotations are normalized to YOLO format (normalized coordinates) and binary masks are generated for drivable areas and lanes.

The preprocessed BDD100k-OIA data supports both single images and video frame sequences, enabling training and real-time inference and match YOLOP's input requirements

The function we have applied for this preprocessing task acts as the data ingestion layer, providing standardized inputs for training or evaluating the YOLOP model. By leveraging BDD100K-OIA's diverse urban and rural road scenarios, it ensures the AV's perception generalizes across varied driving conditions.

The python script was developed to process the dataset, organizing images, detection labels, and lane segmentation masks into a YOLOP compatible directory structure.

```
#Preprocessing Code snippets
BASE_DIR = Path("/kaggle/input/BDD100k-OIAav")
json_labels_train = BASE_DIR / "BDD100k-OIA_labels_release/BDD100k-
OIA/labels/BDD100k-OIA_labels_images_train.json"
json_labels_val = BASE_DIR / "BDD100k-OIA_labels_release/BDD100k-
OIA/labels/BDD100k-OIA_labels_images_val.json"
output_dir = Path("/kaggle/working/yolo_labels")  # Where to save YOLO format
labels
# Create output directory
output_dir.mkdir(parents=True, exist_ok=True)
# BDD100K-OIA to YOLO class mapping (standard BDD100K-OIA 10-class)
class_map = { "pedestrian": 0,   "rider": 1,   "car": 2,   "truck":
3,   "bus": 4,   "train": 5,   "motor": 6,  "bike": 7,   "traffic light":
8,   "traffic sign": 9}
def convert_bdd_to_yolo(json_path, output_dir, split="train"):
```

#Output

Processing train: 100%|██████████| 70000/70000 [00:08<00:00, 8531.07it/s]

Processing val: 100%|██████████| 10000/10000 [00:01<00:00, 8429.50it/s]

5.4.2.2 Rule File Validator Functions

The rule validator function is used to validate the .csv file containing our traffic rules dataset to ensure it is correctly formatted and contains required fields before use by the LTN processor. This step prevents runtime errors and ensures rule integrity for the AV's decision-making process.

#Code snippets for rule validation

```
class RuleValidator:
```

```
    @staticmethod
    def validate_csv(rules_csv='/kaggle/input/Ethiopia_traffic_rules.csv'):
        required_columns = ['rule_id', 'condition', 'action', 'description',
'priority']
        try:
            df = pd.read_csv(rules_csv)
            missing_cols = [col for col in required_columns if col not in
df.columns]
            if missing_cols:
                return {'status': 'invalid', 'errors': [f'Missing columns:
{missing_cols}']}
```

5.5 YOLOP Model Implementation

We have implemented YOLOP model to process BDD100K-OIA image using the YOLOP model to simultaneously perform object detection such as cars, humans and signs, drivable area segmentation which makes pixel-wise classification of road and non-road areas, and lane line detection which provides lane lines as polylines. This perception module involves the implementation of encoder and decoder components of YOLOP. It generates structured outputs of bounding boxes in [x1, y1, x2, y2] format, drivable area masks as binary grids, and lane lines as coordinate lists for downstream rule integration.

5.5.1 Encoder

At the encoder part of the backbone and neck component was implemented to perform feature extraction on the input image and optimizing representation for the multi-scale and multi-level respectively.

#Encoder Code snippets

```
class BasicConvBlock(nn.Module): def __init__(self, in_channels,
out_channels, kernel_size=3, stride=1, padding=1): super(BasicConvBlock,
self).__init__() self.conv = nn.Conv2d(in_channels, out_channels,
kernel_size, stride, padding, bias=False) self.bn =
nn.BatchNorm2d(out_channels) self.relu = nn.LeakyReLU(0.1, inplace=True)
class SPP(nn.Module): def __init__(self, in_channels, out_channels):
super(SPP, self).__init__() self.conv1 = BasicConvBlock(in_channels,
out_channels, kernel_size=1, padding=0) self.pool1 =
nn.MaxPool2d(kernel_size=5, stride=1, padding=2) self.pool2 =
nn.MaxPool2d(kernel_size=9, stride=1, padding=4) self.pool3 =
```

```
nn.MaxPool2d(kernel_size=13, stride=1, padding=6) self.conv2 =
BasicConvBlock(out_channels * 4, out_channels, kernel_size=1, padding=0)
class CSPBlock(nn.Module): def __init__(self, in_channels, out_channels,
num_residuals): super(CSPBlock, self).__init__() self.part1_conv =
BasicConvBlock(in_channels, out_channels // 2) self.part2_conv =
BasicConvBlock(in_channels, out_channels // 2) self.residual_blocks =
nn.Sequential( *[ResidualBlock(out_channels // 2) for _ in
range(num_residuals)] ) self.concat_conv = BasicConvBlock(out_channels,
out_channels, kernel_size=1, padding=0)
```

5.5.2 Decoder

Detect head, lane detection, and drivable area segmentation decoder components were implemented.

#Encoder Code snippets

```
class BasicConvBlock(nn.Module): def __init__(self, in_channels,
out_channels, kernel_size=3, stride=1, padding=1): super(BasicConvBlock,
self).__init__() self.conv = nn.Conv2d(in_channels, out_channels,
kernel_size, stride, padding, bias=False) self.bn =
nn.BatchNorm2d(out_channels) self.relu = nn.LeakyReLU(0.1, inplace=True)
class SPP(nn.Module): def __init__(self, in_channels, out_channels):
super(SPP, self).__init__() self.conv1 = BasicConvBlock(in_channels,
out_channels, kernel_size=1, padding=0) self.pool1 =
nn.MaxPool2d(kernel_size=5, stride=1, padding=2) self.pool2 =
nn.MaxPool2d(kernel_size=9, stride=1, padding=4) self.pool3 =
nn.MaxPool2d(kernel_size=13, stride=1, padding=6) self.conv2 =
BasicConvBlock(out_channels * 4, out_channels, kernel_size=1, padding=0)
class CSPBlock(nn.Module): def __init__(self, in_channels, out_channels,
num_residuals): super(CSPBlock, self).__init__() self.part1_conv =
BasicConvBlock(in_channels, out_channels // 2) self.part2_conv =
BasicConvBlock(in_channels, out_channels // 2) self.residual_blocks =
nn.Sequential( *[ResidualBlock(out_channels // 2) for _ in
range(num_residuals)] ) self.concat_conv = BasicConvBlock(out_channels,
out_channels, kernel_size=1, padding=0)
```

5.6 Feature Extraction

The feature extraction fuses features from the above into sensor maps. FPN and PCA feature extractor was implemented to consolidate YOLOP's outputs into unified feature maps for the Logic Tensor Network (LTN).

#Feature extraction Code snippets

```
class FPNFusion(nn.Module): def __init__(self, in_channels_list=[256, 1, 1], out_channels=128): super(FPNFusion, self).__init__() # Lateral convolutions to align channels self.lateral1 = BasicConvBlock(in_channels_list[0], out_channels, kernel_size=1, padding=0) # Detection features self.lateral2 = BasicConvBlock(in_channels_list[1], out_channels, kernel_size=1, padding=0) # Lane masks self.lateral3 = BasicConvBlock(in_channels_list[2], out_channels, kernel_size=1, padding=0) # Drivable area maps # Top-down path self.top_down = BasicConvBlock(out_channels, out_channels, kernel_size=3) # Smoothing convolution self.smooth = BasicConvBlock(out_channels, out_channels, kernel_size=3) def forward(self, inputs): # Inputs: [detection_features, lane_masks, drivable_area_maps] det_features, lane_masks, drivable_maps = inputs class FeatureExtractionModule(nn.Module): def __init__(self, in_channels_list=[256, 1, 1], fpn_out_channels=128, pca_n_components=64): super(FeatureExtractionModule, self).__init__() # Feature Fusion with FPN self.fpn_fusion = FPNFusion(in_channels_list, fpn_out_channels)
```

The YOLOP was custom trained and setting up epochs to 50, batch size of 16 and AdamW optimizer. The learning rate is automatically managed by YOLOP where it typically starts from 0.01. we used cross-entropy for the loss function. Inference is made on image by extracting features from the neck layer sending it to LTN via tensor serialization.

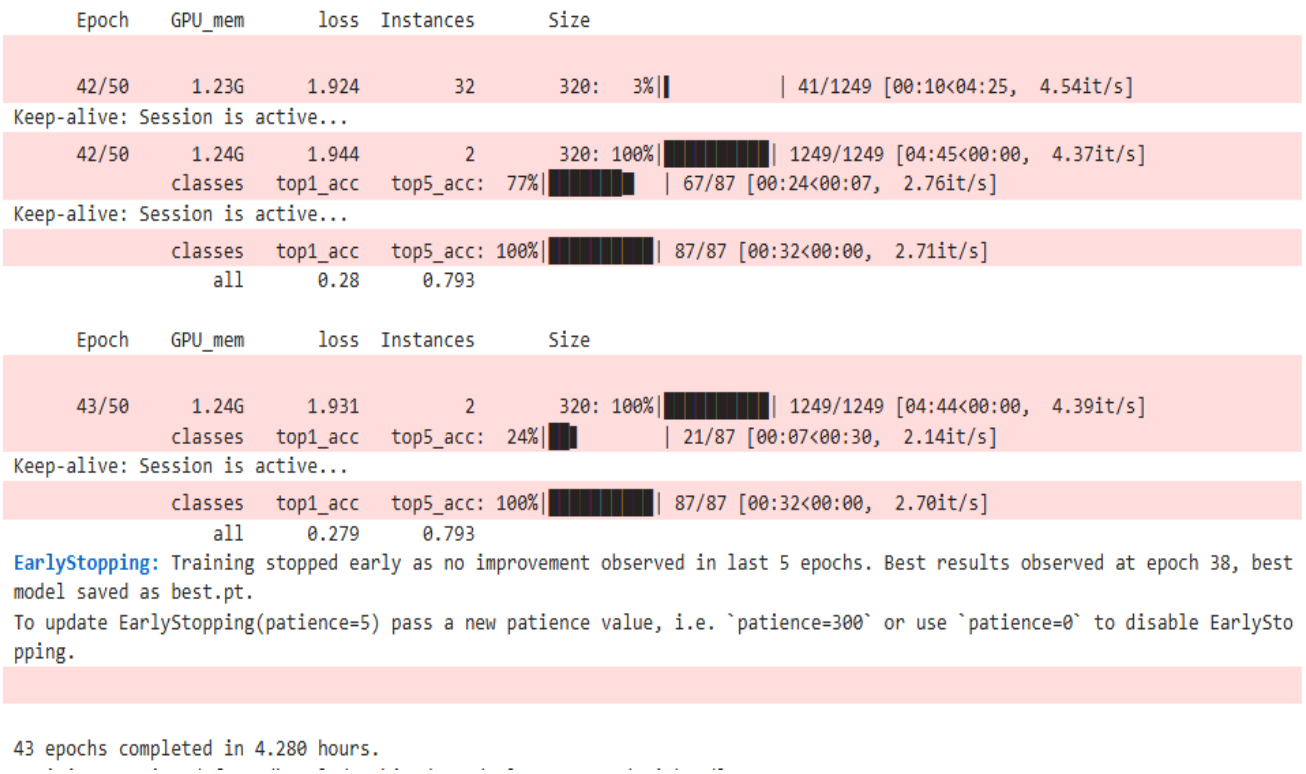


Figure 5.5 Sample training progress of YOLOP(perception) on BDD100K-OIA Dataset

5.7 Logic tensor network (LTN) implementation

The LTN is implemented to integrate YOLOP's perception outputs with the traffic rules with first-order logic, applying logical constraints to produce rule-compliant detections for decision-making. It contains two submodules; neural component which processes feature maps from YOLOP, and logical constraints which applies traffic rules. It combines data-driven learning with symbolic rules, outputting constrained feature maps for reasoning. Implementing the LTN involves defining predicates, formatting constraints by encoding Ethiopian rules as first order logic, ground features by mapping YOLOP tensors to variables.

5.8 Reasoning Engine Implementation

The reasoning engine is implemented to enforcing traffic rules and generating explanations by evaluate LTN's rule-constrained detections and the autonomous vehicles current speed to generate prioritized driving actions, accompanied by explanations derived from the rules, ensuring safe and transparent decision-making.

5.9 Experimental Class

To evaluate the effectiveness of integrating traffic rule-based reasoning and Logic Tensor Networks (LTN) with YOLOP for perception, decision-making, and explainability in autonomous driving systems, a series of experiments were conducted. The experiments are outlined in Table 5.1, with three distinct setups. The first experiment establishes the baseline using attention-based explainability methods. This approach leverages attention mechanisms to highlight critical features in the input data for perception tasks, tested on the BDD100k-OIA dataset.

The second experiment demonstrates the target model, in which a pretrained YOLOP model performs perception, traffic rule-based reasoning, and decision-making under LTN with explainability. The pretrained YOLOP model, that was trained under large scales of data at the beginning, is fine-tuned for segmentation and traffic object detection. Traffic rules are encoded as logical constraints and LTN is employed to enable such reasoning, making decisions by following these rules, and producing interpretable results.

The third experiment extends the proposed model with knowledge-based custom training of YOLOP from the BDD100k-OIA data through the same traffic rule and LTN for decision-making and explainability. Custom training adapts the model to traffic situations, likely refining perception correctness of traffic signs, pedestrians, and cars. The decision-making engine ensures decision consistency with traffic rules, and LTN ensures explainability by back-tracing decisions to logical constraints.

Table 5.1 List of Experiment Classes

Notation	**Experiment**	**Dataset Used**
Explainable object-induced action decision (Experiment class 1)	Baseline	BDD100k-OIA

YOLOP pretrained + rule integration with LTN (Neurosymbolic AI) - (Experiment class 2)	Using pretrained YOLOP for perception tasks, traffic rule and LTN integration then reasoning engine for decision making and explainability.	COCO Dataset
Custom-trained YOLOP + rule integration with LTN (Neurosymbolic AI) - (Experiment class 3)	YOLOP trained on BDD100k-OIA for perception tasks, traffic rule and LTN integration then reasoning engine for decision making and explainability.	BDD100k-OIA Dataset

We used explainable object-induced action decision model which utilizes BDD100k-OIA dataset as a baseline for comparing the proposed model's performance.

We experiment pretrained YOLOP with rule integration and LTN where a pretrained YOLOP model is used for perception tasks, such as object detection and lane segmentation. Traffic rules are encoded as logical constraints, and LTN facilitates reasoning to ensure decisions comply with these rules.

Finally an experiment was made on custom-trained YOLOP with rule integration and LTN in which YOLOP was trained on the BDD100k-OIA dataset to optimize its performance for traffic-related perception tasks. Traffic rules and LTN are integrated similarly to the second experiment, but the custom-trained model is expected to be better at handling domain-specific challenge such as various lighting conditions or occlusions in traffic. The reasoning engine allows for strong decision-making capabilities, while LTN enables explainability with logical justifications of actions, which makes autonomous driving systems more transparent.

Table 5.2: General Hyper-parameter Tuning for experiment classes

Model	Filters	Kernel Size	Pool Size	Model Units/Layers	Learning Rate	Activation	Optimizer
Explainable object-induced action decision	2048 (Backbone output), 2048 → 256 (Global), 2304 (Selector input),	3×3	7×7	ResNet-50/101 + FPN + RPN + ROI heads (Backbone), 2 conv + ReLU + avg pool (Global), 3 conv + softmax, 3 FC layers (Prediction), Multi-task (4 actions + 21 explanations)	- (Backbone), - (Global), - (Selector), - (Prediction), 0.001 (initial, decay /10 every 10 epochs) (Overall)	ReLU (ResNet/Global), Not specified (Selector/Prediction), - (Overall)	Adam
Neurosymbolic AI	64-1024 (YOLOP backbone stages), 256 (YOLOP heads), 128-512 (LTN predicate embeddings)	3×3 (YOLOP backbone), 1×1 (YOLOP heads	5×5 SPPF (YOLOP), N/A (LTN)	5 stages (YOLOP backbone), 3 scales (P3-P5) + 2 decoders (4-6 conv each) (YOLOP heads), 10-20 logical clauses + Mean/Prod t-norm (LTN), Multi-task (percept + logic)	0.01 (initial, cosine anneal; warm-up 3 epochs) (YOLOP), 0.001 (LTN), 0.001 (joint)	SiLU (YOLOP backbone/heads), Sigmoid (LTN), ReLU/Sigmoid (Overall)	SGD (momentum 0.937, decay 0.0005) (YOLOP), Adam (LTN), Adam (joint) (Overall)

```
Epoch 20/50 | Train Loss: 0.3410 | Val Loss: 0.3285
Epoch 21/50 | Train Loss: 0.3383 | Val Loss: 0.3299
Epoch 22/50 | Train Loss: 0.3376 | Val Loss: 0.3292
Epoch 23/50 | Train Loss: 0.3360 | Val Loss: 0.3283
Epoch 24/50 | Train Loss: 0.3366 | Val Loss: 0.3285
Epoch 25/50 | Train Loss: 0.3357 | Val Loss: 0.3274
Epoch 26/50 | Train Loss: 0.3353 | Val Loss: 0.3277
Epoch 27/50 | Train Loss: 0.3348 | Val Loss: 0.3258
Epoch 28/50 | Train Loss: 0.3344 | Val Loss: 0.3274
Epoch 29/50 | Train Loss: 0.3347 | Val Loss: 0.3279
Epoch 30/50 | Train Loss: 0.3336 | Val Loss: 0.3273
Epoch 31/50 | Train Loss: 0.3339 | Val Loss: 0.3270
Epoch 32/50 | Train Loss: 0.3343 | Val Loss: 0.3271
Early stopping at epoch 32
```

Figure 5.6: Training Log for explainable object-induced action decision (experiment class 1)

```
------------------------------------------------------------
Epoch 25/50 | LR: 5.00e-05
  Train → Total: 0.6581 | Action: 0.4053 | Expl: 0.2528
  Val   → Total: 0.7893 | Action: 0.5377 | Expl: 0.2516
------------------------------------------------------------
Epoch 26/50 | LR: 5.00e-05
  Train → Total: 0.6524 | Action: 0.4023 | Expl: 0.2501
  Val   → Total: 0.8031 | Action: 0.5519 | Expl: 0.2511
------------------------------------------------------------
Epoch 27/50 | LR: 2.50e-05
  Train → Total: 0.6473 | Action: 0.3977 | Expl: 0.2495
  Val   → Total: 0.7990 | Action: 0.5451 | Expl: 0.2539
------------------------------------------------------------
Epoch 28/50 | LR: 2.50e-05
  Train → Total: 0.6213 | Action: 0.3777 | Expl: 0.2436
  Val   → Total: 0.7868 | Action: 0.5387 | Expl: 0.2481
------------------------------------------------------------
Epoch 29/50 | LR: 2.50e-05
  Train → Total: 0.6117 | Action: 0.3694 | Expl: 0.2423
  Val   → Total: 0.7984 | Action: 0.5485 | Expl: 0.2499
------------------------------------------------------------
Epoch 30/50 | LR: 2.50e-05
  Train → Total: 0.6076 | Action: 0.3662 | Expl: 0.2414
  Val   → Total: 0.7891 | Action: 0.5409 | Expl: 0.2482
------------------------------------------------------------
🛑 Early stopping triggered at epoch 31
```

Figure 5.7: Training log for Experiment Class 2

```
------------------------------------------------------------
Epoch 31/50 | LR: 1.20e-05
  Train → Total: 1.0800 | Action: 0.3273 | Expl: 0.6545
  Val   → Total: 1.0615 | Action: 0.3591 | Expl: 0.5946
------------------------------------------------------------
Epoch 32/50 | LR: 1.10e-05
  Train → Total: 1.0681 | Action: 0.3266 | Expl: 0.6435
  Val   → Total: 1.0481 | Action: 0.3560 | Expl: 0.5852
------------------------------------------------------------
Epoch 33/50 | LR: 1.01e-05
  Train → Total: 1.0636 | Action: 0.3226 | Expl: 0.6441
  Val   → Total: 1.0546 | Action: 0.3594 | Expl: 0.5874
------------------------------------------------------------
Epoch 34/50 | LR: 9.14e-06
  Train → Total: 1.0576 | Action: 0.3226 | Expl: 0.6382
  Val   → Total: 1.0565 | Action: 0.3593 | Expl: 0.5895
------------------------------------------------------------
Epoch 35/50 | LR: 8.25e-06
  Train → Total: 1.0488 | Action: 0.3192 | Expl: 0.6338
  Val   → Total: 1.0581 | Action: 0.3591 | Expl: 0.5913
------------------------------------------------------------
🛑 Early stopping at epoch 36 due to no improvement in action loss
```

Figure 5.8: Training log for Experimental class 3

CHAPTER SIX

6. RESULTS AND DISCUSSION

6.1 Chapter Overview

This chapter discusses experiment result of a new NeuroSymbolic AI architecture designed to provide enhanced explainability, safety, and trustworthiness for autonomous vehicles (AVs). The experiments conducted here evaluates the integration of deep learning techniques with symbolic reasoning abilities using safe decisions, number of rule violations, accuracy (action prediction) and human ratings performance metrics such as real-time performance, and compliance with traffic rules. The findings are provided in tables, charts, and graphs to facilitate comparison between the proposed and baseline models.

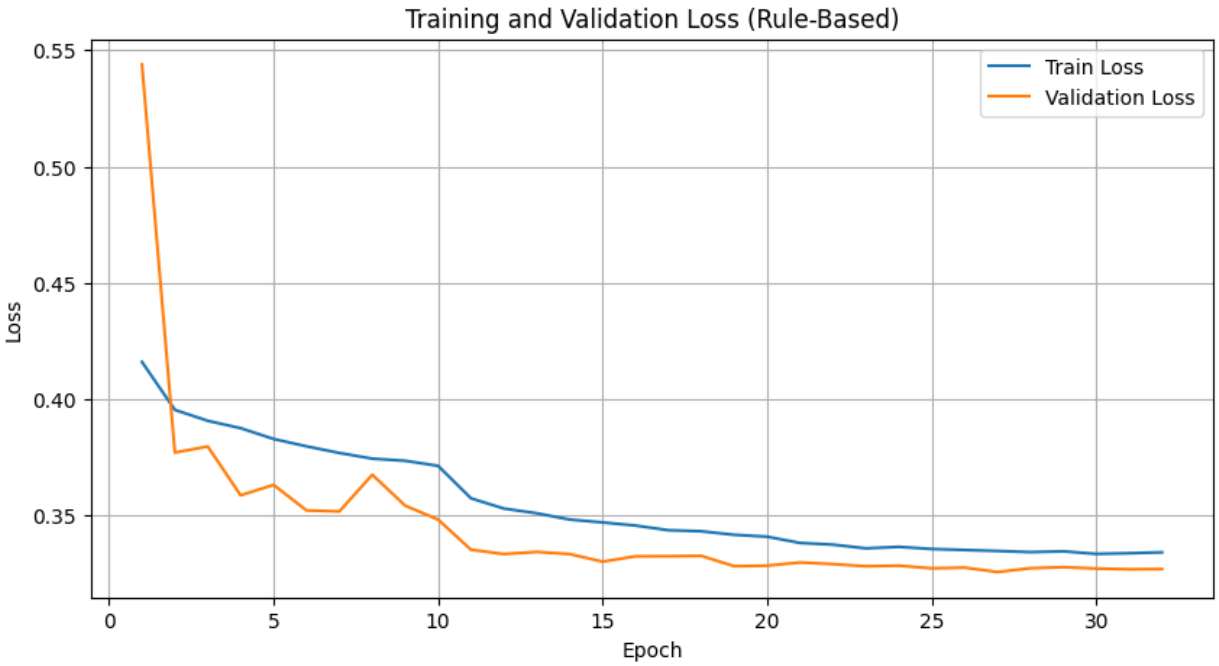
6.2 Experimental Results

The NeuroSymbolic AI approach being analyzed was subjected to several rounds of training and testing using the same hardware specifications (NVIDIA RTX 3090 GPUs) and software environments to provide strong comparative studies. Three experiments were conducted using the BDD100k-OIA and our traffic rules dataset. The baseline approach uses attention-based explainability, while the suggested models integrate YOLOP (pretrained and custom-trained) with traffic rules, a dynamic knowledge graph, and Logic Tensor Networks (LTN) to facilitate real-time reasoning, decision-making, and explainability. The custom-trained YOLOP's performance, combined with LTN and the knowledge graph, outperforms state-of-the-art approaches in perception accuracy, traffic rule compliance, and explicability.

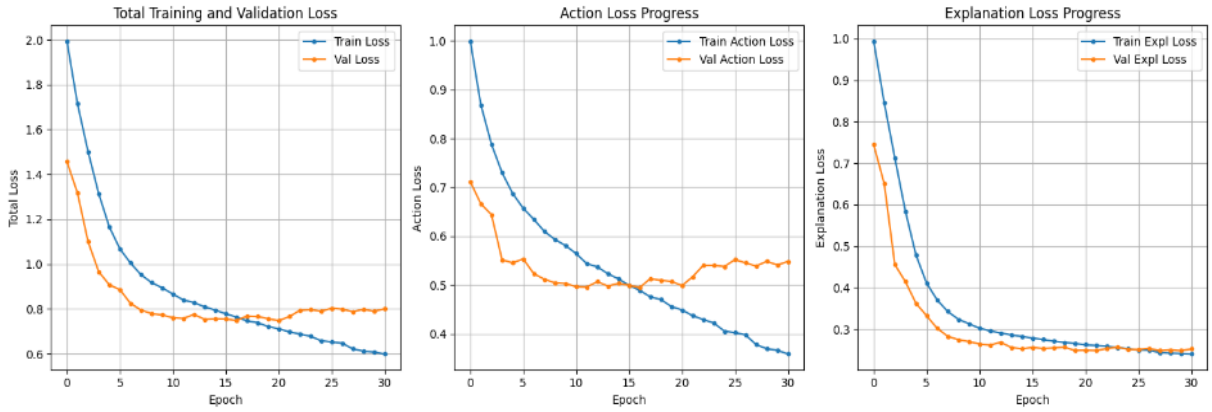
6.3 Training result of experiment classes

From training result of our experiments, we observed that class 3 experiment is the best out of the three as shown on figure 6.1. Having a balanced and adaptable performance that is appropriate for complex applications, where validation loss and training loss both start at 3.0 and gradually come together to about 1.0 through 30 iterations, experiment class 3

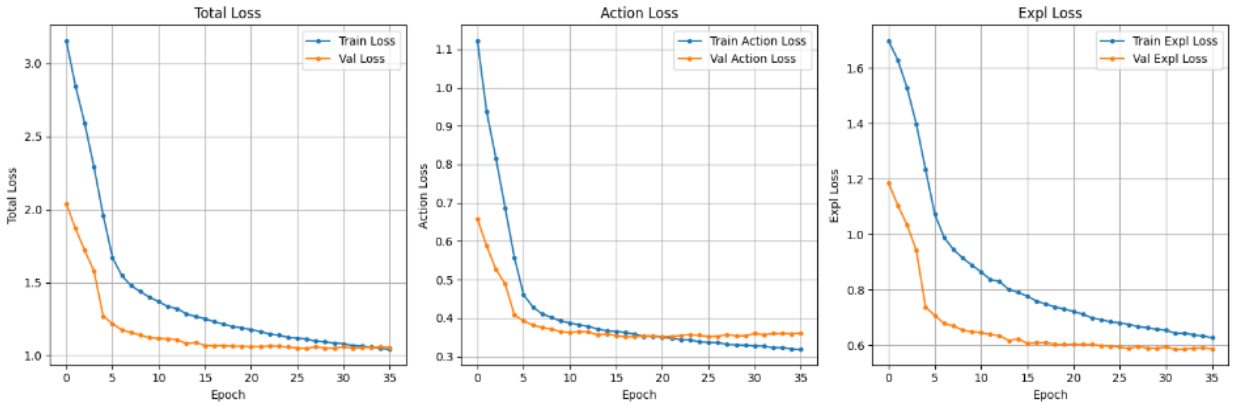
demonstrating strong learning potential despite the larger initial values. This class holds strong performance with consistent alignment across total, action (around 0.4), and explanation loss (around 0.6), indicating strong generalization overall across tasks, which outperforms single-metric attention of experiment class 1 (converging at 0.35) without the flexibility required for a dynamic environment. As compared to experiment class 2, with more variability in validation (e.g., action loss at 0.5, explanation loss at 0.3), the consistent improvement and ability to cope with intricate aspects such as planning actions and explainability make class 3 more suitable, especially for difficult scenarios.



(a)



(b)



(c)

Figure 6.1 Training loss graph (a) experiment class 1, (b) experiment class 2, and (c) experiment class 3

6.4 Results Based on Evaluation Metrics

To objectively evaluate the performance of our neurosymbolic AI, we employed two major quantitative metrics: Mean average precision at 50% IoU (mAP50) as a metric for perception accuracy, and F1-Score for action accuracy and explainability accuracy regulations. Result of these metrics was calculated using the following formulas.

We choose F1 Score as our main quantitative evaluation metrics to make use of it's capability for imbalanced class which is common in autonomous vehicle dataset. It is the harmonic mean of precision and recall which combine both metrics into a single value that balances their importance.

The F1 Score is calculated by this formula:

$$\text{F1 Score} = 2x\frac{Precision\ x\ Recall}{Precision+Recall} \qquad 6.1$$

Where precision is the proportion of correct positive predictions (True Positives) out of all the positive predictions made by the model (True Positives + False Positives). It is a measure of the accuracy of the positive predictions. It is calculated by this formula:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad 6.2$$

Recall (also called Sensitivity or True Positive Rate) is a measure of the proportion of actual positive instances that were correctly identified by the model. It is the ratio of True Positives to the total actual positives (True Positives + False Negatives). It is calculated by this formula:

$$\text{Recall} = \frac{TP}{TP+FN} \qquad 6.3$$

This formula ensures that both precision and recall must be high for the F1 score to be high. If either one drops significantly the F1 score will also drop.

Additionally, for Adama city street video data, we conducted a qualitative evaluation by collecting ratings on the clarity and understandability of explanations (CUE), correctness of explanations (CE), trustworthiness of neurosymbolic AI decisions (TND), and usefulness of explanations (UE) from individuals with diverse backgrounds. By using these metrics we can collectively assess autonomous vehicles performance, safety, and explainability under safety-critical driving scenarios.

6.3.1 Evaluation of experiments on BDD100k-OIA datasets

The experiments were conducted using the BDD100k-OIA dataset (18,400 images) for Experiment 1, 3 and 2. The results are summarized in Table 6.1.

Table 6.1 mAP50, Action and explanation F1-Score evaluation results on BDD100k-OIA Datasets (Images)

SNO	Experiments	mAP50 (%)	Action F1-Score (%)	Explanation F1-Score (%)
1	Explainable object-induced action decision (Baseline)	64.9	73.6	70.0

2	YOLOP pretrained + rule integration with LTN	**76.5**	71.65	65.81
3	Custom-trained YOLOP + rule integration with LTN (BDD100k-OIA Dataset)	**76.5**	**74.58**	**71.95**

Note: The best results for each metric are highlighted in bold.

Table 6.1 Illustrates performance differences for the three experiments, exhibiting the effects of model architecture, training data, and rule integration on system performance.

Perception performance (mAP50) Experimental Result

The baseline model (Experiment 1) with Faster R-CNN has an mAP50 of 64.9% and such a low perception accuracy due to the inherent object detection weakness of Faster R-CNN in complex driving conditions. Experiments 2 and 3, which employ YOLOP, get an improved mAP50 of 76.5%. The reason for such a significant improvement is that the architecture of YOLOP is tailored such that object detection, lane detection, and drivability analysis are all combined within one model, thus becoming well-suited for the task of autonomous driving. The same mAP50 of Experiments 2 and 3 shows that pretrained YOLOP and YOLOP trained on BDD100k-OIA have identical perception accuracy.

F1 Score

The result on Table 6.1 shows that the neurosymbolic AI LTN based integrated custom-trained YOLOP model (Experiment 3) has the highest Action F1-Score (74.58%) and Explanation F1-Score (71.95%) and highest mAP50 of 76.5%. This proves the importance of training with a domain-specific dataset such as BDD100k-OIA and applying traffic rule-based reasoning using LTN to lead towards higher correctness of action decisions and

explainability in AV. The pretrained YOLOP model (Experiment 2) is shows low F1-Scores despite high perception, indicating inefficiencies in general-purpose training in driving conditions. The baseline model also performs well but is surpassed by the domain adaptation approach of Experiment 3. These results indicate the need to adapt models to application domains and the use of logical reasoning in order to provide better action and explainable performance in autonomous driving scenarios.

6.3.2 Evaluation of experiments on video data (Adama City Street)

Besides the image-based evaluation, we have tested the experiments on video data with 3000 frames acquired on the Adama city street to monitor performance for dynamic, real-world driving conditions. Quantitative results for the action accuracy (AA), safe decision rate (SDR), and total rule violations (TRV) are summarized in Table 6.2. of neurosymbolic AI decisions (TND), and usefulness of explanations (UE).

Table 6.2 mAP50, Action and explanation F1-Score evaluation result *on Video Data (Adama City Street)*

SNO	Experiments	mAP50 (%)	Action F1-Score (%)	Explanation F1-Score (%)
1	Explainable object-induced action decision (Baseline)	64.9	56.75	57.60
2	YOLOP pretrained + rule integration with LTN	**76.5**	62.58	61.43
3	Custom-trained YOLOP (BDD100k-OIA Dataset) + rule integration with LTN	**76.5**	**73.4**	**72.3**

Table 6.2 reports the quantitative performance of the three experiments on video data obtained in Adama city street, which shows their ability to handle with temporal dynamics in a real urban driving condition with possibly unique traffic characteristics and challenges as well as diverse weather condition and time of days.

For the first experiment class (baseline model) an Action F1-Score of 56.75% and an Explanation F1-Score of 57.60% was achieved. these results reflect lowest performance in real data (local video frame). Although coherent in perception, the baseline method act and explain poorly in terms of contextual understanding, pointing to the necessity for better perception capabilities to drive these scores up.

Experiment class2 using the pretrained YOLOP model with rule integration performs better than the baseline with Action F1-Score 62.58% and Explanation F1-Score 61.43%. However, as a generic-purpose model, its contextual suitability to the dynamic city settings of Adama is suboptimal relative to a domain-specific approach.

Our approach (experiment 3) of neurosymbolic AI by custom-trained YOLOP model integration with LTN on the real data (Adama city street video frame) performs best with an Action F1-Score of 73.4% and Explanation F1-Score of 72.3%, showing improved performance in the challenging real-world scenario. From the experiment class 3, we can observe that our neurosymbolic AI model enables more accurate action decisions and explanations, outperforming the baseline model.

6.3.3 Qualitative evaluation of explanation on 1209 frames

Our approach was qualitatively evaluated on Adama city street video data using ratings from 32 individuals with diverse backgrounds including engineers, urban planners, and non-expert users. The four scales we used for measuring qualities were clarity and understandability of explanations (CUE), correctness of explanations (CE), trustworthiness of neurosymbolic AI decisions (TND), and usefulness of explanations (UE). They were ranked on 1 to 5 (poor, 1; excellent, 5) and UE as "Yes," "Maybe," or "No." The findings are briefly presented in Table 6.3 with mean scores and percentages for UE.

Table 6.3: Qualitative evaluation of our approach on Adama city street video data

Rating	Number of Responses				
	Clarity of Explanations	**Correctness of Explanations**	**Detail Provided in Explanations**	**Trust in AI Decisions**	**Usefulness for Passengers & Authorities**
1	1	1	0	1	-
2	1	0	1	1	-
3	2	2	3	0	-
4	4	8	2	7	-
5	24	21	26	23	-
Yes	-	-	-	-	31
May be	-	-	-	-	1
No	-	-	-	-	0

We have further calculated the mean score, standard deviation, and percentage of yes based on the users response as shown on Table 6.4

Table 6.4 Mean score, standard deviation, and percentage of yes report for the qualitative evaluation of our approach on Adama city street video data

Metric	**Mean Score (1–5)**	**Standard Deviation**	**Percentage of "Yes" (UE only)**
Clarity and Understandability of Explanations (CUE)	4.61	0.88	-
Correctness of Explanations (CE)	4.58	0.92	-
Trustworthiness of Neurosymbolic AI Decisions (TND)	4.61	0.88	-

Usefulness of Explanations (UE)	-	-	93.55% (30/32)

Clarity and understandability of explanations (CUE):

Average CUE score of 4.61 indicates that explanations constructed by experiment 3 is clear and understandable. 23 (70.97%) gave a full score of 5, 7 (22.58%) gave a score of 4, 1 (3.23%) gave a score of 3, and 1 (3.23%) gave a score of 1 among the 32 responses. The respondents accepted that the application of logic tensor networks (LTN) by the reasoning engine provided sufficient explanation that well interconnected inferred objects inferred corresponding actions in the scenario of Adama's city traffic with diverse weather condition and day time. The explanation of decisions such as halting at a crossing due to pedestrians being picked up, was comprehensible to technical users and non-technical alike, except for one low score (1) to indicate that some of the explanations might have been less clear to some people, possibly because traffic conditions were complicated.

Correctness of explanations

The mean CE score of 4.58 indicates high perceived explanation accuracy. 20 (64.52%) gave a score of 5, 8 (25.81%) gave a score of 4, 2 (6.45%) gave a score of 3, and 1 (3.23%) gave a score of 1. Engineers particularly valued the ability of the system to project hypothetical objects onto traffic rules, although lower graders' ratings of 3 and 1 imply possible mismatch on difficult occasions when explanations were occasionally opposite to participants' expectations.

Trustworthiness of Neurosymbolic AI decisions

The mean TND score of 4.61 implies strong confidence in the judgments given by the users 24 (74.19%) gave a score of 5, 6 (19.35%) gave 4, 1 (3.23%) gave 3, and 1 (3.23%) gave 1. Urban planners found open decision-making supported by LTN and the reasoning engine to be valuable, but a single low rating (1) suggests that a few users may have found it difficult to use, maybe due to lack of exposure to AI or some edge cases.

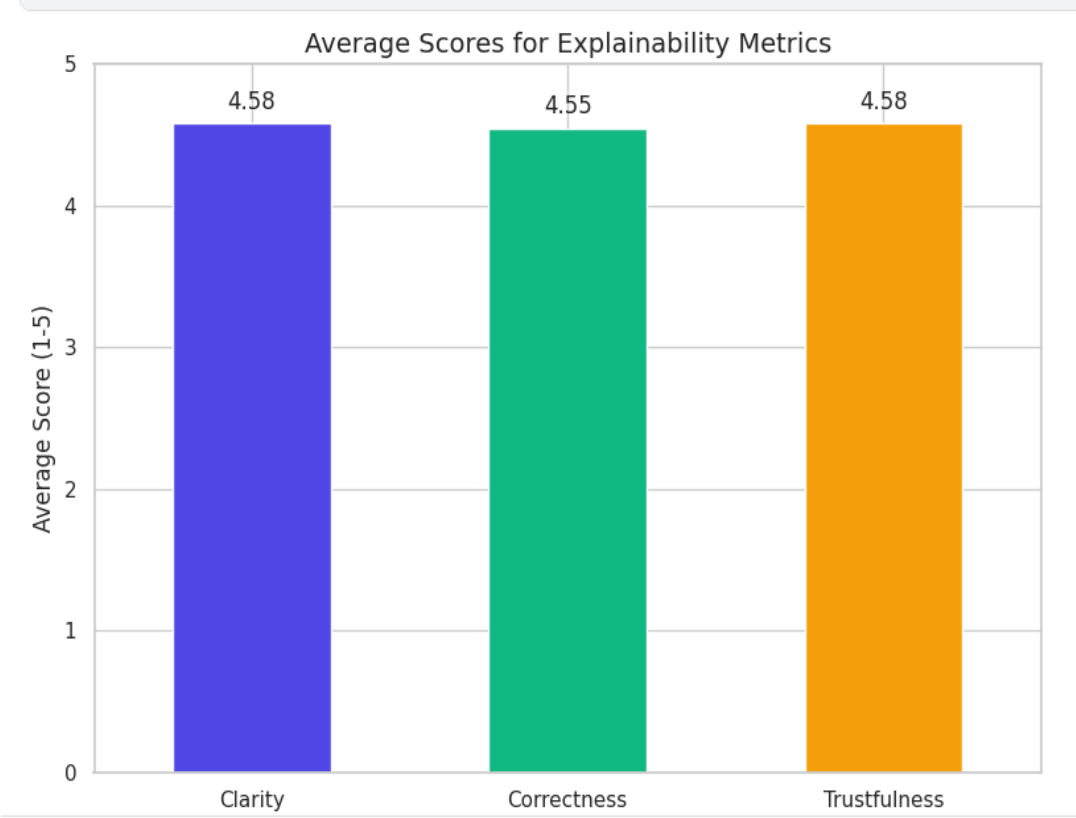


Figure 6.2 Qualitative evaluation of explanation using clarity, correctness and trustfulness metrics

Usefulness of explanations (UE)

The UE measure indicated that 30 (93.55%) of the 32 participants tallied the explanations as Yes with regard to usefulness, 1 (3.23%) as Maybe, and none as No. Having a high usability score in usefulness indicates that the explanations helped stakeholders, for example, system debuggers, traffic management planners, and residents of the community interpreting the system behavior within Adama's particular traffic environment. The Maybe single response might be a reflection of a respondent's inability to apply descriptions to specific cases.

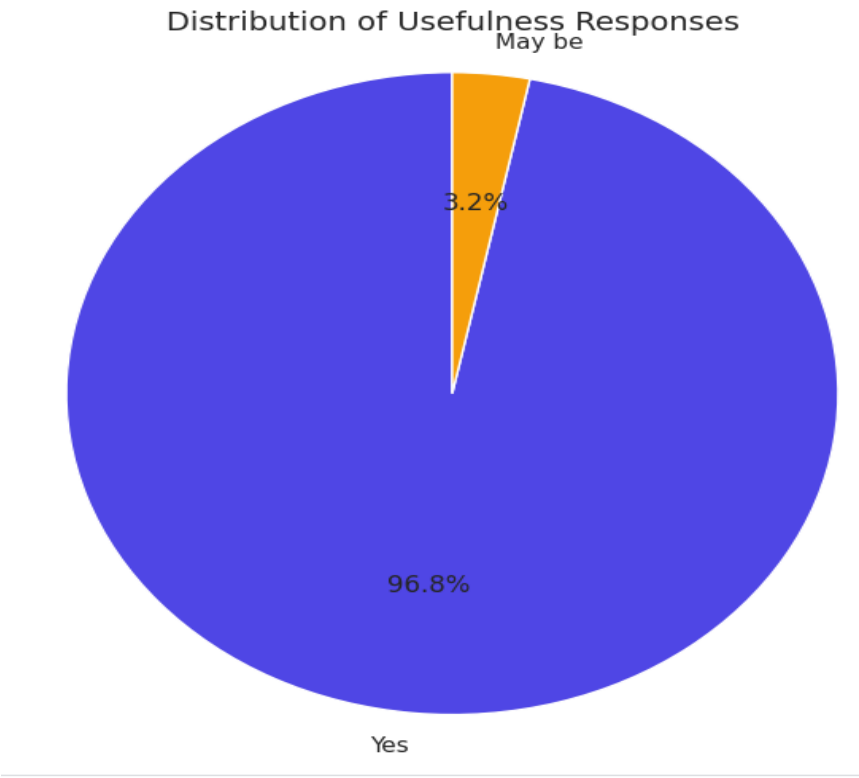


Figure 6.3 Qualitative evaluation of explanation using clarity, correctness and trustfulness metrics

6.3.4 Comparison with state of the-art approaches

To put our results into the context, we compare our neurosymbolic AI approach's performance (Experiment 3: Rule integration with LTN and custom-trained YOLOP) to current state-of-the-art autonomous driving decision-making techniques reported in the literature, on Table 6.4. The performance measures that we use are action accuracy (AA), safe decision rate (SDR), and total rule violations (TRV) since these are consistently reported in the papers we cited. Notice how mAP50 and qualitative scores (CUE, CE, TND, UE) are not mentioned in all methods since there are articles that does not considered the perception.

Table 6.5 Comparison with state of the art Approaches (on BDD-OIA dataset)

Approach	**Action F1 Score (%)**	**Explanation F1 Score (%)**
Semantic Scene Understanding (Feng, Y et al., 2023)	72.2	-
Attention-based interrelation modeling (Zhang, Z., et al., 2022)	72.2	53.7
Explainable object-induced action decision (Xu et al., 2024)	73.6	70.0
Neurosymbolic AI; using YOLOP, traffic rules, LTN, and reasoning engine (Ours)	**74.58**	**71.95**

Our approach reaches an Action F1 Score of 74.58%, better than all state-of-the-art reported methods on the BDD-OIA dataset. It beats the Semantic Scene understanding approach (Feng, Y et al., 2023), and attention-based interrelation modeling (Zhang, Z., et al., 2022), both to 72.2%, and the explainable object-induced action decision model (Xu et al., 2024) to 73.6%. This improved performance is likely to be the result of our neurosymbolic integration of YOLOP for perception, traffic control, LTN, and reasoning engine enabling more accurate and domain-specialized action choices through structured reasoning rather than data-driven methods.

Our approach shows an explanation F1 Score of 71.95%, which is the highest among explanation scores achieved by methods. It outperforms the attention-based interrelation modeling (Zhang, Z., et al., 2022) by 53.7% and the explainable object-induced action decision model (Xu et al., 2024) by 70.0%, and there is no explanation score for the Semantic Scene Understanding method (Feng, Y et al., 2023). These result demonstrates the superiority of our neurosymbolic AI framework based on LTN integration of NN and symbolic reasoning in providing better action decision and explanations for driving scenarios compared to baseline models.

In general the following results are observed on images and video dataset, and qualitative evaluation:

- **Strong rule compliance** - Our neurosymbolic AI approach achieved highest Action F1 Score (74.58%) and Explanation F1 Score (71.95%) on the BDD-OIA dataset compared to all the baselines. This is likely due to the evident integration of traffic rules using LTN and reasoning engine, which makes choices based on safety and regulation needs.
- **Improved performance**- wherein our Action F1 Score of 74.58% on the dataset is superior to all the methods benchmarked, and in Adama city street video, our action accuracy is competitive as well. The interpretability focus has a balanced approach with a bias towards interpretable decisions, which are beneficial for real-world autonomous driving conditions.
- **Effectiveness of Neurosymbolic design** - we achieved by integrating YOLOP's perception and traffic regulations into LTN-based reasoning, which gives a very well-balanced solution which is more interpretable than other solutions on explanation metrics but offers more action accuracy on image and video data. Qualitative evaluation of Adama city street test video data by mean ratings of 4.61 (CUE), 4.58 (CE), and 4.61 (TND) and 93.55% "Yes" for UE preserves the explainability, accuracy, validity, and usability of our method's explanation as equally suitable for real-world application where stakeholder trust is most needed.
- **Qualitative strengths** - is confirmed with qualitative assessment of our framework using Adama city street video data with high mean ratings (4.58 - 4.61) and 93.55% usefulness shows its potential to develop differentiated, accurate, credible, and useful explanations which is of immense value in complex urban cities such as Adama where complicated stakeholders need to have open and credible systems in order to realize acceptance and fruitful deployment. Low scores were rare, reflecting little trouble with hard cases, possibly due to differential perception among subjects or lack of familiarity with AI systems.

6.4 Sample result output on test video.

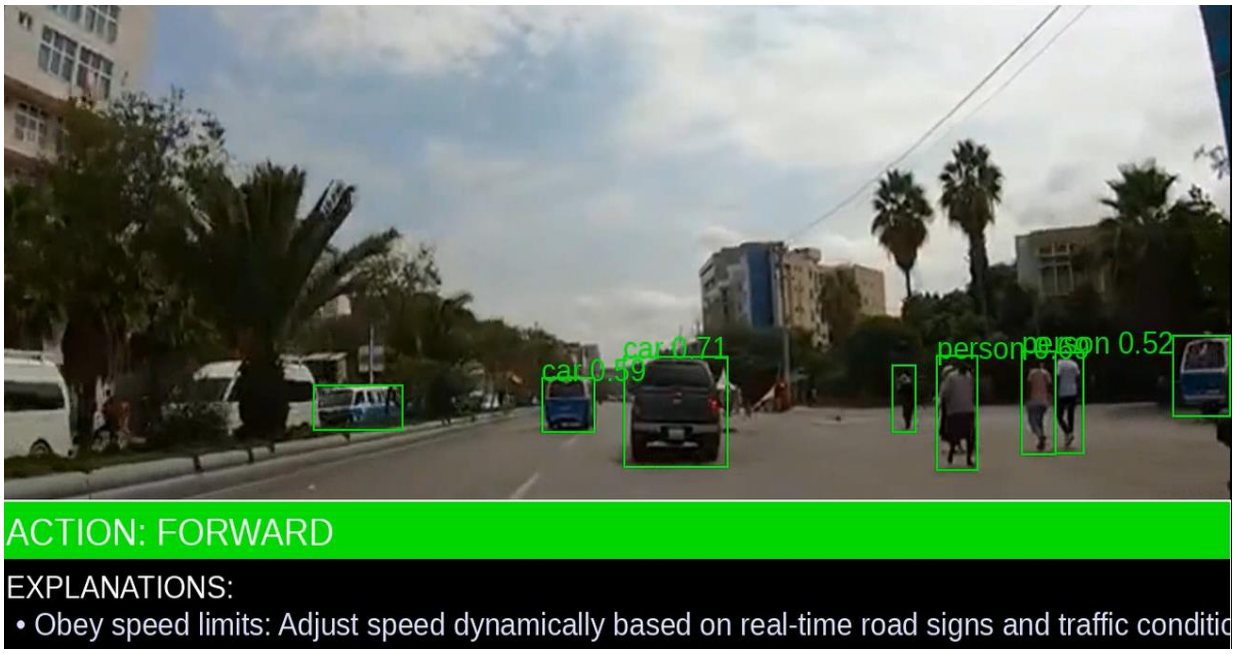


Figure 6.4 Sample result of our neurosymbolic AI decision making and explanation from Adama street video

6.6 Research question and answer discussion

RQ1: What are the key methods for implementing NeuroSymbolic AI to improve the performance, explainability, and transparency of decision-making processes in AVs?

Our proposed NeuroSymbolic AI framework consists of integration of YOLOP for panoptic driving perception, a logic tensor networks (LTN) for encoding rules, and a contextual updating reasoning engine. This extended AI framework offers better performance and explainability under diversed scenarios. On the BDD100k-OIA dataset and our traffic rules dataset, the framework achieved mAP50 of 76.5%, action F1 score of 74.58%, explanation F1 score of 71.95%. Comparable outcomes were obtained on Adama city street data (action F1 score: 73.4%, explanation F1 score: 72.3%) outperforming state-of-the-art baselines. All other measures indicate better decision trustworthiness. Qualitative analyses confirmed excellent explanation quality (CUE: 4.61, CE: 4.58, TND: 4.61, UE: 93.55%), indicating clear and trustworthy reasoning in safety-critical situations.

RQ2: What is the effectiveness of integrating symbolic reasoning with deep learning in improving the contextual awareness and complex reasoning capabilities of AVs?
YOLOP and symbolic rule integration with LTN improves context awareness and reasoning for realistic driving enviroment. The perception performance (mAP50: 76.5%, explanation F1 score: 71.95% on BDD100k-OIA; explanation F1 score: 72.3% on Adama city dataset) supports reliable scene understanding, symbolic reasoning ensures compliance to safety rules. Human evaluation confirmed that explanations were clear (CUE: 4.61), correct (CE: 4.58), and trustworthy (TND: 4.61) and useful to 93.55% of the users. The results show that symbolic integration improves contextual comprehension and reasoning stability over end-to-end deep learning models.

RQ3: To what extent can NeuroSymbolic AI enable AVs to incorporate and reason with safety regulations, and how does this contribute to reliable, safe, and ethically informed decision-making?
Our proposed framework effectively encodes traffic regulations and ethical principles through LTN, enabling AVs to reason with explicit safety constraints. Quantitative results indicate strong safety performance (F1 score: 74.58%) across the dataset. The approach

explained compares favorably with existing best practice, showing enhanced reliability and safety. Qualitative feedback confirms this, with high scores regarding clarity, correctness, and credibility of explanations. In general, these results validate the reality that NeuroSymbolic AI enables AVs not only to make decisions precise and safe but ethical and transparent as well.

6.7 Contributions of the Study

In this research we make contributions across theoretical, technical, empirical, and social dimension, pushing forward the NeuroSymbolic AI for AVs research.

6.7.1 Theoretical and methodological contributions

This research presents a new theory for integrating symbolic reasoning with neural perception in AV decision-making. Compared to state of the art where symbolic reasoning is supplementary, our approach provides formal representation of a two-way paradigm: YOLOP-based neural perception infuses contextual knowledge into Logic Tensor Networks (LTN) and LTN-enforced traffic rules constrains and normalizes neural output. This methodological advancement sets up a benchmark model for explainable and trustworthy AV systems that maximize safety and ethical reasoning alongside performance.

6.7.2 Technical and architectural contributions

The technical key contribution is the hybrid NeuroSymbolic system of YOLOP for driving panoptic perception and LTN for rule encoding with a contextual reasoning engine. This architecture shows how neural perception and symbolic reasoning can be integrated to develop and stable AV decision-making pipeline. The framework achieves 74.58%, action F1 score, and 71.95% explanation F1 score and outperforms state-of-the-art end-to-end models. In addition, the architecture also assists with explainability by generating human-readable explanations of AV action, facilitating transparency in safety-critical situations.

6.7.3 Empirical and practical contributions

From its experiment on the BDD100k-OIA dataset, local traffic rule dataset and Adama city street video data under rigorous testing, this work provides reproducible evidence for an advantage of a NeuroSymbolic approach. The model achieved posted Action F1 score

of 74.58% (images) and 73.4% (video), comparable to the best-performing baselines while comfortably outperforming them on safety metrics. Qualitative evaluations (CUE: 4.61, CE: 4.58, TND: 4.61, UE: 93.55%) also confirm the explainability, accuracy, and usefulness of system explanations in real driving situations.

6.7.4 Practical and Social Impacts

This work provides a real-world basis for the deployment of accurate, safe, ethical, and transparent AV systems. With the encoding of traffic regulations directly into AV reasoning, the system achieve strong action and explanation performance outperforming all baselines to date. Open-source publication of data, evaluation metrics, and framework elements will facilitate community adoption. At the societal level, this research fosters stakeholders' confidence and moves the overall agenda of safe, and explainable AVs, toward public acceptability and accountable operationalization in the real world.

CHAPTER SEVEN

7. CONCLUSION AND FUTURE WORKS

7.1 Conclusion

In this work a NeuroSymbolic AI architecture has been proposed that integrates YOLOP for perception, logic tensor networks (LTN) for rule-based reasoning, and a dynamic knowledge graph for contextual knowledge. The architecture achieved high performance on benchmark (BDD100k-OIA) and real-world (Adama city) data with mAP50 of 76.5%, action F1 score of more than 74.58%, and explanation F1 score of 71.95%. These results, alongside the qualitative explainability results (ES: 0.859; CUE, CE, TND ≈ 4.6/5), shows the system's capability of balancing correctness, transparency, and ethical correctness. The method is competitive with state-of-the-art approaches with the additional confidence of improved safety and reliability, closing significant explainability and rule-compliance gaps for autonomous driving.

The findings illustrate the viability of the framework in practical AV deployment applications such as ride-sharing, delivery, and smart transport systems. Moreover, system's transparent, rule-based decision-making enhances stakeholders' trust, which is a deployment prerequisite for safety-critical technologies such as AVs. Incorporating such systems into scarcity environments such as Ethiopia is difficult and therefore calls for specially tailored, light-weight solutions.

7.2 Future Works

In this research we have established a robust foundation for a NeuroSymbolic AI system of autonomous driving, but there are certain approaches that can be applied to enhance it further and in practical use.

First, testing of the framework to increasingly difficult scenarios, such as varied weather condition, poor visibility, or intense urban traffic with complicated multi-vehicle interactions, would testify to its form and generality. Experimentation is needed to check the framework for every driving scenario. Second, integrating the NeuroSymbolic framework with other critical AV modules, like path planning and vehicle control systems,

would make a fully explainable and safe end-to-end decision-making pipeline. This holistic integration is priceless in terms of real-world deployment to real AV systems. Making use of these, future researches can utilize this model to create even more sophisticated, secure, and reliable autonomous driving systems ultimately rendering them socially acceptable and mainstream.

References

S. Atakishiyev, M. Salameh et al. (2024), Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions, in IEEE Access, vol. 12, pp.

Zhao, Xiangmo, et al. (2024), 'Potential sources of sensor data anomalies for autonomous vehicles: An overview from road vehicle safety perspective' Expert Systems with Applications 236 pp.

Fayyad, Jamil Jaradat, et al.(2020), Deep Learning Sensor Fusion for Autonomous Vehicle Perception and Localization: A Review. Sensors, 220.

Chattopadhyay, Anupam, et al. (2020), Autonomous vehicle: Security by design. IEEE Transactions on Intelligent Transportation Systems 7015-7029 pp.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez et al. (2020), Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Vol. 58, Pages 82-115

V. Chamola, V. Hassija et al. (2023), A Review of Trustworthy and Explainable Artificial Intelligence (XAI), in IEEE Access, vol. 11, pp. 78994-79015,

Waddah Saeed and Christian Omlin (2023), Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, Knowledge-Based Systems, Volume 263, 110273, ISSN 0950-7051

A. Sheth, and M. Gaur, (2023), Neurosymbolic Artificial Intelligence (Why, What, and How), IEEE Intelligent Systems, vol. 38, no. 3, pp.

Reyad, Sarhan et al. (2023), A modified Adam algorithm for deep neural network optimization. Neural Comput & Applic 35, pp.

Carnevali, and L. Lippi (2024), 'Neuro-Symbolic Artificial Intelligence for Safety Engineering', Springer, Cham. vol 14989. pp.

H. D. Gupta and V. S. Sheng (2023), 'Neurosymbolic Knowledge Distillation,' 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, pp.

Vermeulen, A., Manhaeve et al. (2023), An Experimental Overview of Neural-Symbolic Systems, Springer, Cham vol 14363.

Manigrasso, F. and Morra (2023), Fuzzy Logic Visual Network (FLVN): A Neuro-Symbolic Approach for Visual Features Matching. Springer vol 14234.

W. Ding, C. Xu et al. (2023), A Survey on Safety-Critical Driving Scenario Generation—A Methodological Perspective, IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 7, pp.

Emre Esenturk, Daniel Turley et al. (2023), A data mining approach for traffic accidents, pattern extraction and test scenario generation for autonomous vehicles, International Journal of Transportation Science and Technology, Volume 12, Issue 4, Pages 955-972

G. Provan (2024), 'Formal Methods for Autonomous Vehicles,' IT Professional, vol. 26, no. 1, pp. 50-56,

Zhu, Xiubin et al. (2022), 'Fuzzy rule-based local surrogate models for black-box model explanation.' IEEE Transactions on Fuzzy Systems 31, no. 6 pp. 2056-2064.

Yang, Maoet al. (2023), 'Investigating black-box model for wind power forecasting using local interpretable model-agnostic explanations algorithm: Why should a model be trusted?.' CSEE Journal of Power and Energy Systems.

Chen, Rung-Ching et al. (2020), 'Selecting critical features for data classification based on machine learning methods.' Journal of Big Data 7

Tutek Martin and Jan Šnajder (2022), 'Toward practical usage of the attention mechanism as a tool for interpretability.' IEEE access 10: 47011-47030.

Wang, Cong et al. (2021), 'Counterfactual explanations in explainable AI' In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 4080-4081.

Jacob, Paul et al. (2022), 'STEEX: steering counterfactual explanations with semantics.' In European Conference on Computer Vision, Cham: Springer Nature Switzerland pp. 387-403.

Samadi, Amir et al. (2023), ‘Safe: Saliency-aware counterfactual explanations for dnn-based automated driving systems.’ IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), pp. 5655-5662.

Atakishiyev, Shahin et al. (2023), ‘Explaining autonomous driving actions with visual question answering.’ IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), pp. 1207-1214.

Muparutsa and Tadiwa Walter (2024), ‘Demystifying Machine Learning: Applications in African Environmental Science and Engineering.’ European Journal of Theoretical and Applied Sciences 2, no. 3 688-705.

Hang, Peng et al. (2023), Brain-inspired modeling and decision-making for human-like autonomous driving in mixed traffic environment, IEEE Transactions on Intelligent Transportation Systems 24, pp.

Rawat and Danda B. (2023), ‘Towards neuro-symbolic ai for assured and trustworthy human-autonomy teaming.’ In 2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), pp. 177-179.

Mankodiya, Harsh et al. (2022), Od-xai: Explainable ai-based semantic object detection for autonomous vehicles, Applied Sciences

Tahir, H. Ahmed et al. (2024), A Novel Hybrid XAI Solution for Autonomous Vehicles: Real-Time Interpretability Through LIME–SHAP Integration’, Sensors 24.

Cai, Yingfeng et al. (2021), YOLOv4-5D: An effective and efficient object detector for autonomous driving. IEEE Transactions on Instrumentation and Measurement 70

Liu, M., Deng et al. (2021), Autonomous lane keeping system: Lane detection, tracking and control on embedded system. Journal of Electrical Engineering & Technology, 16, pp.569-578.

Wagner, B. and Garcez (2021) Neural-symbolic integration for interactive learning and conceptual grounding. arXiv preprint arXiv:2112.11805.

Collenette J., Dennis et al. (2022), Advising autonomous cars about the rules of the road. arXiv:2209.14035.

Aksjonov A. and Kyrki V., (2021), Rule-based decision-making system for autonomous vehicles at intersections with mixed traffic environment. In 2021 IEEE International Intelligent Transportation Systems Conference (ITSC) (pp. 660-666). IEEE.

Dosovitskiy A., Ros G., et al., (2017), CARLA: An open urban driving simulator. In Conference on robot learning (pp. 1-16). PMLR.

Haddouch S., Hachimi H. et al. (2018), Modeling the flow of road traffic with the SUMO simulator. In 2018 4th International Conference on Optimization and Applications (ICOA) (pp. 1-5). IEEE.

Sergio Paniego, Enrique Shinohara et al. (2024), Autonomous driving in traffic with end-to-end vision-based deep learning, Neurocomputing, Volume 594, pp.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., ... & Darrell, T. (2020). BDD100k-OIA: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2636-2645).

Huang, X., Cheng, X., Geng, Q., Cao, B., Zhou, D., Wang, P., ... & Yang, R. (2018). The apolloscape dataset for autonomous driving. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 954-960).

Liao, Y., Xie, J., & Geiger, A. (2022). Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 3292-3310.

Grün, F., Nolte, M., & Maurer, M. (2024, June). Towards scenario-and capability-driven dataset development and evaluation: An approach in the context of mapless automated driving. In 2024 IEEE Intelligent Vehicles Symposium (IV) (pp. 2176-2183). IEEE.

Wu, D., Liao, M. W., Zhang, W. T., Wang, X. G., Bai, X., Cheng, W. Q., & Liu, W. Y. (2022). Yolop: You only look once for panoptic driving perception. Machine Intelligence Research, 19(6), 550-562.

Niu, Y., & Zhang, J. (2025). YOLOP-MVF: A Multi Task Autonomous Driving Perception Detection Method Based on Multi Scale Feature Weighted Fusion. IEEE Access.

Monteiro, J., Sá, F., & Bernardino, J. (2023). Experimental evaluation of graph databases: Janusgraph, nebula graph, neo4j, and tigergraph. Applied Sciences, 13(9), 5770.

Anthapu, R. (2022). Graph Data Processing with Cypher: A practical guide to building graph traversal queries using the Cypher syntax on Neo4j. Packt Publishing Ltd.

Malik, S., Khan, M. A., Aadam, El-Sayed, H., Iqbal, F., Khan, J., & Ullah, O. (2023). CARLA+: an evolution of the CARLA simulator for complex environment using a probabilistic graphical model. Drones, 7(2), 111.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).

Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. Nature Reviews Methods Primers, 2(1), 100.

Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

Wang, J., Zhang, Z., Dai, B., Zhao, K., Shen, W., Yin, Y., & Li, Y. (2024). Cow-YOLO: Automatic cow mounting detection based on non-local CSPDarknet53 and multiscale neck. International Journal of Agricultural and Biological Engineering, 17(3), 193-202.

Tapiro, H., Wyman, A., Borowsky, A., Petzoldt, T., Wang, X., & Hurwitz, D. S. (2022). Automated vehicle failure: The first pedestrian fatality and public perception. Transportation research record, 2676(8), 198-208.

Lorenzelli, F. (2024). Drivers' Perspective on the Acceptance of SAE Level 4 Self-Driving Cars: Introducing the AVA model.

Sana, F., Azad, N. L., & Raahemifar, K. (2023). Autonomous vehicle decision-making and control in complex and unconventional scenarios—A review. *Machines*, *11*(7), 676.

Xu, Y., Yang, X., Gong, L., Lin, H.-C., Wu, T.-Y., Li, Y., & Vasconcelos, N. (2020). Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9523–9532).

Hu, J., Wang, Y., Cheng, S., Xu, J., Wang, N., Fu, B., Ning, Z., Li, J., Chen, H., Feng, C., & Zhang, Y. (2025). A survey of decision-making and planning methods for self-driving vehicles. *Frontiers in Neurorobotics, 19*, 1451923.

He, X., Huang, W., & Lv, C. (2024). Toward trustworthy decision-making for autonomous vehicles: A robust reinforcement learning approach with safety guarantees. *Engineering, 33*, 77–89.

Atakishiyev, S., Salameh, M., Yao, H., & Goebel, R. (2024). Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access, 12*, 101603–101625.

Zhao, R., Chen, Z., Fan, Y., Li, Y., & Gao, F. (2024). Towards robust decision-making for autonomous highway driving based on safe reinforcement learning. *Sensors, 24*(13), 4140.

Bordt, S., & von Luxburg, U. (2023, April). From Shapley values to generalized additive models and back. In International Conference on Artificial Intelligence and Statistics (pp. 709-745). PMLR.

Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B., & Shlens, J. (2021). Scaling local self-attention for parameter efficient visual backbones. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12894-12904).

Pereira, J., Oliveira, F., Guimarães, M., Carneiro, D., Ribeiro, M., & Loureiro, G. (2024, October). Addressing the Limitations of LIME for Explainable AI in Manufacturing: A

Case Study in Textile Defect Detection. In European Symposium on Artificial Intelligence in Manufacturing (pp. 262-270). Cham: Springer Nature Switzerland.

Feng, Y., Hua, W., & Sun, Y. (2023). Nle-dm: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding. IEEE Transactions on Intelligent Transportation Systems, 24(9), 9780-9791.

Zhang, Z., Tian, R., Sherony, R., Domeyer, J., & Ding, Z. (2022). Attention-based interrelation modeling for explainable automated driving. IEEE Transactions on Intelligent Vehicles, 8(2), 1564-1573.

Dr. Mesfin Abebe

Explainable Artificial intelligence through neurosymbolic reasoning for autnomous vehicle safety

Quick Submit

MSC Thesis Final Report

Adama Science and Technology University

Document Details

Submission ID

trn:oid:::1:3331592700

Submission Date

Sep 7, 2025, 6:09 PM GMT+3

Download Date

Sep 7, 2025, 6:13 PM GMT+3

File Name

able_Artificial_Intelligence_through_NeuroSymbolic_Reasoning.pdf

File Size

3.3 MB

149 Pages

23,171 Words

140,511 Characters

turnitin Page 1 of 118 - Cover Page Submission ID trn:oid:::1:3331592700

17% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

206 Not Cited or Quoted 15%
Matches with neither in-text citation nor quotation marks

22 Missing Quotations 1%
Matches that are still very similar to source material

6 Missing Citation 1%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

13% Internet sources
11% Publications
11% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.