PEDESTRIAN INTENTION PREDICTION FOR AUTONOMOUS VEHICLES

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

DOCTOR OF PHILOSOPHY

in
Electronics and Communication Engineering
by

NEHA SHARMA
(Enrollment No.: 2K20/PHDEC/507)

Under the supervision of

PROF. INDU SREEDEVI
Professor
Department of Electronics and Communication Engineering

DR. CHHAVI DHIMAN
Assistant Professor
Department of Electronics and Communication Engineering

To the

Department of Electronics and Communication Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

September, 2025

ACKNOWLEDGMENTS

I owe tremendous debt and would like to express deep feelings of gratitude for the support and guidance of several people who have helped me to accomplish the research program with the support and direction of several persons. This challenging and rewarding experience has definitely helped me grow in character as well as academically. It gives me a great pleasure to now have the opportunity to express my gratitude towards them.

First and foremost, thanks to the Almighty for giving me strength and inspiration to carry out this research work. I owe a deep sense of gratitude to all his comprehensive soul whose divine light has enlightened my path throughout the journey of my research.

I take the opportunity to humbly submit my sincere and heartfelt thanks to my gurus (research supervisors), **Prof. S. Indu** and **Dr. Chhavi Dhiman** from the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, for their invaluable guidance, enthusiastic encouragement, and persistent support. I am truly grateful from the core of my heart for their meticulous approach, wonderful assistance of their perspective, and fruitful discussions on my research topic. Their immense contribution and rare dedication in providing the much-needed guidance, is worth of highest honor. Their careful supervision and personal attention have given me a lot of confidence and enthusiasm, during the different stages of my doctoral investigations. They are academic giants under whose watch am molded to a seasoned research scholar. I invariably fall short of words to express my sincere gratitude for their patience and motivation.

I am extremely thankful to the **Head of the Department** of Electronics and Communication Engineering, Delhi Technological University, Delhi, and other faculty members for their endless support and cooperation throughout this dissertation. I am thankful to all staff members of the department of Electronics and Communication Engineering for their kind help and support during the entire period of my research.

I would like to extend my heartfelt gratitude to my dear friends at Delhi Technological University, Delhi, for filling my journey with constant laughter, shared lunches, unwavering support, and timely assistance. I am especially thankful to **Mr. Aman Jolly, Ms. Aashania Antil, Mr. Gaurav Kumar** and **Ms. Hemanshi Chugh** for being an integral part of this experience.

I also wish to express my sincere appreciation to my former colleagues and mentors from KIET Group of Institutions, Ghaziabad—**Dr. Shruti Pandey, Dr. Ramesh Singh**, and other fellow colleagues—whose guidance, encouragement, and belief in my potential not only supported me during my tenure there but also inspired me to pursue my research interests beyond the professional realm. Their motivation gave me the strength to take a significant leap of faith: to leave a secure, well-paying job and commit fully to the pursuit of academia for the next 4–5 years with undeterred focus.

This thesis is dedicated to my family and teachers, whose boundless love, support, encouragement, and blessings have sustained me through 31 years of life. I am profoundly grateful to my parents, **Mr. Alok Anil** and **Mrs. Kiran Anil**, whose unwavering encouragement during times of doubt and discouragement, steadfast belief in my capabilities since childhood, and constant presence have been my greatest source of strength. I am equally thankful for the love and support of my extended family, particularly my maternal uncle, **Mr. V.P. Sharma,** who has been a continuous pillar of support throughout this academic journey.

Lastly, I am grateful to everyone, supporters and skeptics alike, whose actions, whether encouraging or discouraging, have shaped this journey into a deeply enriching and transformative experience. This thesis stands as a testament to all who have touched my academic and personal life.

NEHA SHARMA

DELHI TECHNOLOGICAL UNIVERSITY

Formerly Delhi College of Engineering

Shahbad Daulatpur, Main Bawana Road, Delhi –42

CANDIDATE'S DECLARATION

I **Neha Sharma** hereby certify that the work which is being presented in the thesis entitled **Pedestrian Intention Prediction for Autonomous Vehicles** in partial fulfillment of the requirements for the award of the Degree of Doctor in Philosophy, submitted in the **Department of Electronics and Communication Engineering**, Delhi Technological University is an authentic record of my own work carried out during the period from **January 2021** to **September 2025** under the supervision of **Prof. S. Indu** and **Dr. Chhavi Dhiman**.

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

Candidate's Signature

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

Prof. Sreedevi Indu	**Dr. Chhavi Dhiman**	**Prof. Sumantra Dutta Roy**
Supervisor	Co-Supervisor	External Examiner
Professor, ECE Dept.	Assistant Professor, ECE Dept.	Professor, EE Dept.
DTU, Delhi	DTU, Delhi	IIT Delhi

DELHI TECHNOLOGICAL UNIVERSITY

Formerly Delhi College of Engineering

Shahbad Daulatpur, Main Bawana Road, Delhi –42

CERTIFICATE BY THE SUPERVISOR(S)

Certified that **Neha Sharma** (Enrollment No.: 2K20/PHDEC/507) has carried out her research work presented in this thesis entitled "**Pedestrian Intention Prediction for Autonomous Vehicles**", for the award of **Doctor of Philosophy** from the Department of Electronics and Communication Engineering, Delhi Technological University, under our guidance and supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Prof. S. Indu
Supervisor
Department of ECE
Delhi Technological University,
Delhi –110042, India

Dr. Chhavi Dhiman
Co-Supervisor
Department of ECE
Delhi Technological University,
Delhi –110042, India

Place and Date

ABSTRACT

According to the Global Status Report on Road Safety 2023, vehicle crashes cause numerous annual deaths, particularly impacting vulnerable road users. Pedestrians, lacking protective gear, face high vulnerability and substantial injury risk in collisions. Consequently, the growing advancement of Autonomous Vehicle (AV) technology is being explored to enhance road safety and convenience for all users. AV technology can reduce accidents attributed to human errors like fatigue, misperception, and inattention. Leading automotive manufacturers and tech giants like BMW, Tesla, and Google are actively advancing AV technology in this pursuit.

Predicting pedestrians' road-crossing decisions is pivotal for achieving a reliable driverless experience through AVs. Initial studies emphasised pedestrian dynamics to anticipate crossing intent. Yet, analysing merely the trajectory proves inadequate for understanding underlying intentions. Beyond trajectory, various factors impact pedestrian road-crossing decisions. These factors fall into three primary modalities: pedestrian-specific (encompassing pose, appearance, etc.), context-specific (involving scene infrastructure and social interaction with co-pedestrians), and hybrid modality encompassing comprehensive human cognitive aspects while observing a pedestrian on the road. Nonetheless, dealing with such diverse modalities necessitates an efficient multimodal fusion framework that can capture adequate discriminatory features for classification. Moreover, interpreting pedestrian interactions with the surrounding environment is highly challenging in a dynamic ego-centric setting.

With the rise of deep learning, researchers started using deep neural networks (DNNs) to analyse large amounts of data and automatically learn features indicative of pedestrian intention. These models are trained on large datasets of pedestrian behaviour and show improved accuracy over traditional rule-based methods. This has led to the development of end-to-end models involving convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants that process raw sensory data, such as camera images or lidar point clouds, to make predictions. These

approaches are seen as more robust and capable of handling complex scenarios where single-modality approaches may fail, as they can learn the relationships between different modalities and make predictions in a more integrated manner.

This thesis explores deep learning-based approaches for predicting pedestrian intentions in autonomous vehicles. Pedestrian intention prediction is a multi-stage process comprising input acquisition, feature extraction and encoding, spatiotemporal modelling, multimodal fusion, and final decoding or classification. Each stage plays a crucial role in ensuring accurate predictions, with variations in approach depending on the specific output required, such as pedestrian crossing intent classification or trajectory anticipation.

The first stage of the process involves acquiring input data in the form of video frames and trajectory coordinates spanning a specific time window. These inputs can be sourced from real-time surveillance systems or pre-recorded video sequences captured from multiple camera angles. This data undergoes pre-processing to extract spatial and temporal features aligned with model requirements. Convolutional Neural Networks (CNNs), such as EfficientNet, are used to derive spatial representations from RGB sequences and segmentation maps, capturing posture, orientation, and environmental cues. To model temporal dependencies, Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks process historical trajectory data, enabling the inference of motion trends for accurate behaviour prediction.

Following feature extraction, the system proceeds to spatiotemporal modelling, which aims to capture the evolving interactions between pedestrians and their surrounding environment over time. This thesis investigates two distinct approaches for this task: Graph Convolutional Networks (GCNs) and Co-Learning Transformers. The GCN-based approach, incorporating a multi-head adjacency matrix, structures pedestrian trajectory data as a graph, enabling the model to learn relational dependencies among individuals. In contrast, the Co-Learning Transformer approach focuses on temporal modelling, capturing long-range dependencies and refining motion features through attention mechanisms.

Given that pedestrian intention prediction depends on multiple input modalities, an effective fusion strategy is critical for integrating these diverse sources of information. This thesis employs several advanced fusion mechanisms to address this challenge. Adaptive Fusion dynamically adjusts the importance of features based on contextual cues, allowing the model to prioritize relevant information. Co-Learning Architectures enable different modalities to contribute distinct and informative perspectives, enhancing the overall representation. The Multi-Head Shared Weights Mechanism promotes feature consistency across modalities by sharing parameters, thereby reducing redundancy and improving generalization. Finally, the Progressive Denoising Attention Mechanism incrementally filters out irrelevant noise while emphasizing salient patterns, leading to more refined and robust feature representations.

The final stage of the process involves decoding the fused feature representations to generate meaningful predictions about pedestrian behaviour. This thesis explores two primary decoding approaches. Pedestrian Intention Classification employs a classifier, such as a SoftMax layer to infer whether a pedestrian intends to cross the street, based on their observed behaviour and contextual cues. Trajectory Prediction, on the other hand, utilizes generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) to forecast future trajectories by learning from historical motion patterns.

The performance of each proposed pedestrian intention prediction approach is tested with various publicly available datasets and compared with earlier state-of-the-art algorithms. Finally, the research work is concluded followed by future research direction as well as possible future applications which are highlighted and discussed in detail.

LIST OF PUBLICATIONS

- **N. Sharma**, C. Dhiman, and S. Indu, "Pedestrian Intention Prediction for Autonomous Vehicles: A Comprehensive Survey," ***Neurocomputing***, vol. 508, pp. 120–152, 2022, doi: https://doi.org/10.1016/j.neucom.2022.07.085. **Impact Factor: 6.5.**

- **N. Sharma**, C. Dhiman, and S. Indu, "Visual–Motion–Interaction-Guided Pedestrian Intention Prediction Framework," ***IEEE Sensors Journal***, vol. 23, no. 22, pp. 27540–27548, 2023**,** doi: 10.1109/JSEN.2023.3317426. **Impact Factor: 4.5**

- **N. Sharma**, C. Dhiman, and S. Indu, "Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration," ***IEEE Transactions on Intelligent Vehicles***, vol. 9, no. 11, pp. 6960-6970, 2024, doi: 10.1109/TIV.2024.3391898. **Impact Factor: 14.3**

- **N. Sharma**, C. Dhiman, and S. Indu, "Predicting Pedestrian Intentions with Multimodal IntentFormer: A Co-learning Approach," ***Pattern Recognition***, vol. 161, p. 111205, 2025**,** doi: https://doi.org/10.1016/j.patcog.2024.111205. **Impact Factor: 7.6**

- **N. Sharma**, C. Dhiman, and S. Indu, "Cross-Modal Pedestrian Behavior Prediction: A Dual Task Approach with Progressive Denoising Attention and CVAE". ***IEEE Transactions on Intelligent Transportation Systems,*** doi: 10.1109/TITS.2025.3578023. **Impact Factor: 8.4**

- **N. Sharma**, C. Dhiman, and S. Indu, "A Deep Unified Pedestrian Detection Framework", in *IEEE Delhi Section Conference (DELCON),* pp. 1-6, 2022, doi: 10.1109/DELCON54057.2022.9753544.

- **N. Sharma**, C. Dhiman, and S. Indu, "Intelligent Pedestrian Intention Prediction Framework" In *IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, pp. 1-5, 2022, doi: 10.1109/SOLI57430.2022.10295014.

- N. Sharma, C. Dhiman, and S. Indu, "LLM-Guided Visual Reasoning for Scene-Aware Pedestrian Intention Prediction" In *International Conference on Pattern Recognition and Machine Intelligence (PReMI)*, 2025 (**Accepted**).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The autonomous vehicles market size is forecast to increase by USD 624 billion at a compound annual growth rate (CAGR) of 39.3% between 2024 and 2029[1]. This rapid expansion is driven by the significant economic advantages of autonomous vehicle technology, including lower driving costs, enhanced fuel efficiency, and broader societal benefits. Integrating autonomous vehicles into the transportation system is expected to significantly enhance safety and efficiency by eliminating the reliance on human drivers. Human error is a primary contributor to road accidents, and the implementation of autonomous technology has the potential to mitigate this risk, thereby improving overall traffic safety for both motorists and pedestrians. Furthermore, self-driving technology redefines the travel experience by providing a seamless and error-free journey. With no need to concentrate on road conditions, passengers can allocate their travel time to work, leisure, or other productive activities, ultimately increasing convenience and societal productivity [2].

Despite the highly promising future of AVs and its booming economic ventures, creating a fully autonomously working car remains an unfulfilled desire of many tech giants even after garnering huge success now and then in Advanced Driving Assistance Systems (ADAS) by the research community. According to The Global Status Report on Road Safety published by the World Health Organization (WHO) [3], [4], the number of deaths on roads globally has reached an unprecedented high of 1.35 million annually. Nearly half of these road accidents are victims of vulnerable road users (VRU). Huge challenges persist when developing appropriate infrastructure and proper safety traffic regulations to facilitate the harmonious co-existence of AVs and VRUs in urban traffic scenarios. One of the most challenging issues autonomous vehicles face is mimicking humans' perceptions and understanding many social cues in everyday traffic scenarios to avoid fatal vehicle-to-VRU collisions [5]. This is to prevent severe injury to the latter as they don’t have any special protective equipment.

Additionally, it creates a secure and more congenial atmosphere for every road user agent. Hence, early anticipation of VRU's intention is desired so that AVs can design their manoeuvres accordingly [6].

There are a variety of terminologies, like action prediction, behaviour analysis, and intention estimation, which are employed to delineate what exactly a pedestrian is about to do or what trajectory he/she will take in a particular traffic scenario. Action refers to physical movement, whether walking, waving hands, etc. Behaviour is a set of observable events seen as a generalized response that one undertakes in response to a stimulus. Hence, on one hand, action or behaviour is an observable event with ground-truth availability. In contrast, intention, on the other hand, is the intrinsic state of mind that can't be discerned just by looking but requires meticulous inference from behaviour or past actions. In other words, intention involves a deeper semantic comprehension of a human's physical or mental activities [7]. Most AVs resort to conservative driving to circumvent challenges associated with understanding VRU's intention to predict its forthcoming action. Conservative driving involves driving very slowly, avoiding complex interactions, choosing a less complicated path regarding scene understanding and VRU's footfall, and often stopping to avoid road mishaps. Such an approach ensures the safety of VRUs, but this can adversely impact the usual traffic flow, leading to high fuel wastage and decreased inefficiency. Action prediction approaches find their implementation in areas where estimation of future frames or prediction of the motion of pedestrians is required [8]-[11].

Various approaches are employed for this challenging task of intention prediction, including interpreting the forthcoming actions of vulnerable road users, particularly pedestrians, as they exhibit higher degrees of freedom and complexity in their movements. They are very agile, can execute any trajectory, might not follow designated lanes for crossing, abruptly change motion, be occluded in the presence of scenic obstacles, engage implicitly on the road through eye gaze or hand wave, or be diverted while talking over a phone or with fellow pedestrians. Their conduct on the road is more or less affected by several factors, like demographics, gait, traffic density, whether walking in a group or alone, road width, road structure, and many more. All

these factors form contextual data for pedestrian intention detection involving scene dynamics, pedestrian kinematics and social behaviour with other co-pedestrians. Several studies have shown the relationship between one or two factors and the behaviour of pedestrians so that AVs can make calculated decisions beforehand to prevent any mishap [12].

Therefore, high precision and accuracy are imperative in pedestrian intention prediction, as they directly impact human safety and cannot be compromised for technological advancements. This thesis explores various approaches to pedestrian intention prediction, aiming to anticipate short- and long-term actions. This chapter introduces the fundamental concepts of pedestrian intention prediction for autonomous vehicles, discusses the associated challenges, and highlights the significance and motivation behind this study, formulating the problem statement. The final section presents the major research contributions of this thesis, including theoretical formulation and experimental validation, followed by an outline of the thesis organization.

1.1 Pedestrian Intention Prediction

A combination of visual, dynamic, and motion cues is exhibited by pedestrians when they intend to cross the road, offering valuable clues to their crossing behaviour [13]-[14]. For instance, a pedestrian may cross the road if he/she is approaching the crosswalk and looking at the incoming vehicle to ask for a passage. On the other hand, a person standing still at the curb, showing no signs of motion or visual gait towards the crossing action, is less likely to cross the street in a short while. Hence, the pedestrian's positive crossing intent refers to observable behaviour and cues exhibited by a pedestrian, indicating a deliberate intention to cross a road or street. This intent is manifested through various actions, such as standing or approaching marked crosswalks, waiting at traffic lights designated for pedestrians, making eye contact with drivers, standing at or approaching zebra crossings, and raising a hand or arm as a signalling gesture to drivers. This kind of behaviour signifies a conscious decision by the pedestrian to engage in the act of crossing, contributing to overall road safety

awareness. Contextual factors, including co-pedestrians' behaviour and traffic signals or signs, may further influence the perception of positive crossing intent. Crossing intention confidence is a numeric score estimated from human reference data [15]-[16].

In the context of pedestrian crossing intention detection, a few fundamental temporal parameters shape the foundation of predictive systems: short-term intention, long-term intention, observation length and time-to-event (TTE). These parameters intricately influence the accuracy and responsiveness of intention predictions by determining the historical context and temporal proximity to the crossing event. Understanding their roles is pivotal for designing efficient and contextually aware systems that enhance pedestrian safety and optimize interactions with autonomous technologies.

Short-term intention prediction predicts the immediate behaviour or response of VRUs (Vulnerable Road Users) over the next few seconds (2-3 seconds), focusing on actions such as walking, stopping, crossing, or waiting [5], [17]-[20]. Whereas ***long-term intention prediction*** estimates the trajectory or final destination of VRUs by incorporating contextual and scene infrastructure details to improve trajectory accuracy beyond 3 seconds [21]-[27].

Observation Length: Observation length refers to the number of consecutive time steps for which historical pedestrian data is considered during the training process of a pedestrian crossing intention detection system. In other words, the duration of past behaviour and cues are considered for predicting a pedestrian's intention to cross the road.

***Time-to-Event (TTE)*:** Time-to-event (TTE) is the temporal difference between the last time step of the observation length and the occurrence of the actual crossing event. It quantifies the interval from when the system last observes the pedestrian's behaviour to when the pedestrian starts crossing the road.

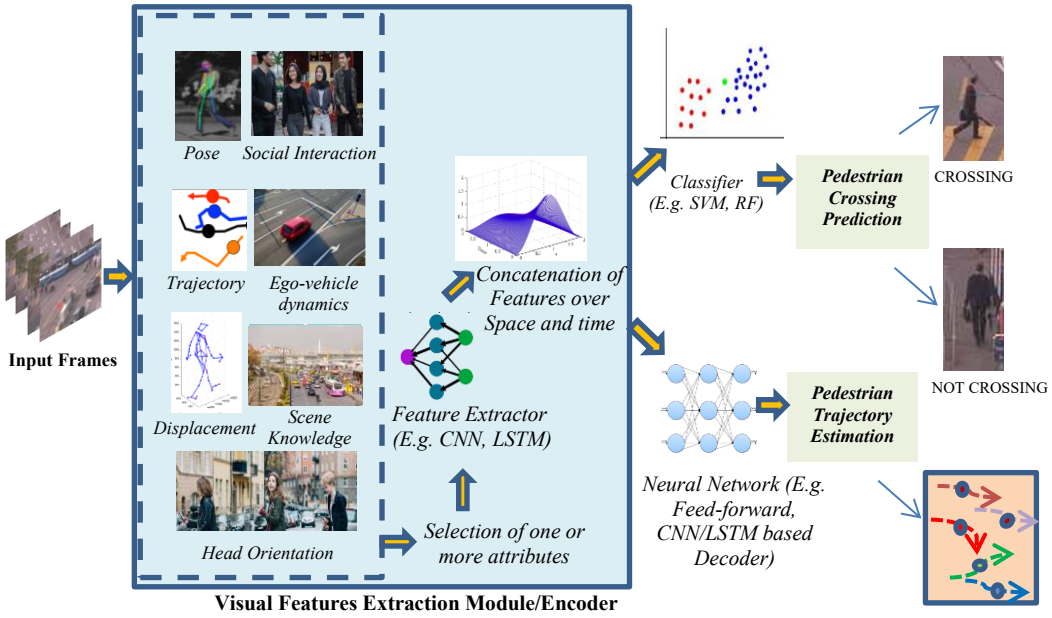


Fig 1.1: A generalised framework for Pedestrian Intention Prediction

The observation length and TTE are interconnected parameters crucial for designing effective pedestrian crossing intention detection systems. Striking the right balance between these factors is essential to ensure timely, accurate, and context-aware predictions, contributing to enhanced pedestrian safety and smoother interactions between pedestrians and autonomous systems [17]-[18].

Fig 1.1 describes the pedestrian intention prediction process into three primary stages: input, feature extraction and encoding, and decoding or classification, which varies based on the desired output. The input stage consists of frames from real-time or pre-recorded video sequences captured by various camera systems from multiple angles. These frames undergo a pre-processing phase, during which relevant attributes are extracted to align with the specific requirements of the proposed algorithm. Various feature extractors can encode features across spatial and temporal dimensions. The final stage incorporates a classifier or a neural network-based decoder to facilitate pedestrian crossing predictions and trajectory anticipations, respectively.

A comprehensive classification of pedestrian intention estimation approaches is presented in Fig 1.2, which encompasses a wide range of techniques explored in the

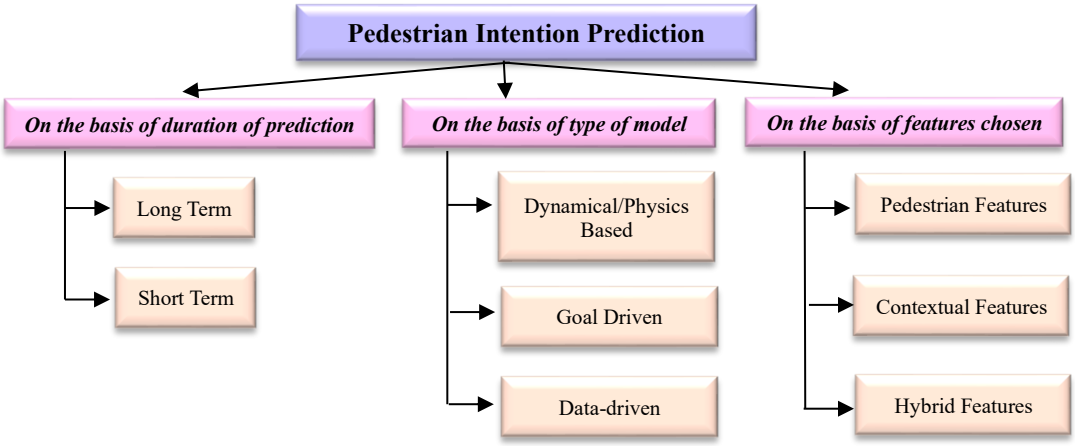


Fig 1.2: Taxonomy of Pedestrian Intention Prediction

literature. This classification is structured around three key parameters: duration of prediction, type of model, and choice of input features.

- Duration-based classification divides prediction techniques into long-term and short-term approaches, depending on the temporal window to anticipate pedestrian actions [28]-[31].
- Model-based classification categorizes approaches into dynamical/physics-based [32]-[35], goal-driven [36]-[41], and data-driven models [9], [26]-[28], [36], [42]-[53], leveraging different methodologies to interpret pedestrian behaviour.
- Feature-based classification distinguishes between pedestrian-specific[5], [32], [43], [50], [54]-[63], contextual[9], [64]-[73], and hybrid features[74]-[77], highlighting the input data types contributing to intention estimation.

1.2 Challenges in Pedestrian Intention Prediction

Pedestrian intention prediction is critical to ensuring safety in urban environments, particularly in human-vehicle interaction scenarios. Accurately forecasting whether a pedestrian will cross a street is essential for autonomous vehicles, driver-assistance systems, and intelligent transportation infrastructure. However, predicting pedestrian behaviour remains a highly complex task due to variability in human motion, environmental uncertainties, and limitations in sensor data. Unlike vehicles, which follow predefined traffic rules, pedestrians exhibit

unstructured and often unpredictable movements, making it difficult to develop a universally reliable prediction model.

One of the primary challenges in this domain is pedestrians' inconsistent and erratic movement patterns, especially in crowded urban settings. Pedestrians frequently change direction, pause, or accelerate unexpectedly, often influenced by distractions, urgency, or social interactions. Conventional motion models struggle to capture these non-linear behavioural variations, reducing prediction accuracy in real-world scenarios. Moreover, the challenge intensifies at busy intersections and crosswalks, where multiple pedestrians interact with each other and external elements, further complicating trajectory prediction [78]-[79].

Another significant issue is the reliance on multimodal data, which includes visual inputs, trajectory coordinates, and environmental context. While multimodal approaches enhance prediction accuracy, they introduce missing, noisy, or unreliable data vulnerabilities. Sensor failures, occlusions caused by vehicles or street objects, and adverse weather conditions can disrupt data acquisition, resulting in incomplete or erroneous inputs. Existing models often lack robust mechanisms to handle missing modalities, making them unreliable in dynamic real-world settings. Furthermore, current predictive architectures face challenges in contextual reasoning, particularly in associating pedestrian behaviour with environmental cues. Traffic signals, approaching vehicles, road infrastructure, and pedestrian flow patterns all play crucial roles in determining crossing intentions. However, many models fail to establish a cohesive relationship between these contextual elements and pedestrian dynamics, leading to suboptimal performance in complex traffic conditions [20]-[80].

Finally, computational efficiency and real-time feasibility pose additional challenges in pedestrian intention prediction. Many state-of-the-art models prioritize accuracy but overlook memory and processing constraints, making them impractical for real-time deployment in autonomous vehicles and edge-computing systems. High computational overhead can lead to delayed predictions, reducing the effectiveness of pedestrian detection in fast-moving traffic scenarios. Optimizing prediction models to

balance accuracy, speed, and efficiency is crucial for enabling their widespread implementation in intelligent traffic management systems [81]-[82].

1.3 Role of Deep Learning in Pedestrian Intention Prediction

In the early stages of research, researchers employed random models, including the Gaussian mixture regression model[78] and the hidden Markov model[79] to simulate pedestrian motion patterns based on either precise dynamical modelling or knowledge of prior end goals, limiting their ability to reasonably predict future interactions and their applicability to complex motion scenes. Fig. 1.3 illustrates an example of pseudo-goal candidates generated by matching test input with expert trajectories [80]. These candidates are then encoded, refined through a social attention network (Social ATTN), and utilized to produce final trajectory predictions.

Nonetheless, the recent surge of deep learning algorithms has outperformed these traditional approaches in handling complex scenarios without showing reliance on any dynamic motion modelling or prior knowledge of end goals. Several trajectory-based techniques [22], [23], [81], [82] that rely on past trajectories of the pedestrian to predict its forthcoming action of whether crossing or not crossing the road also fail to

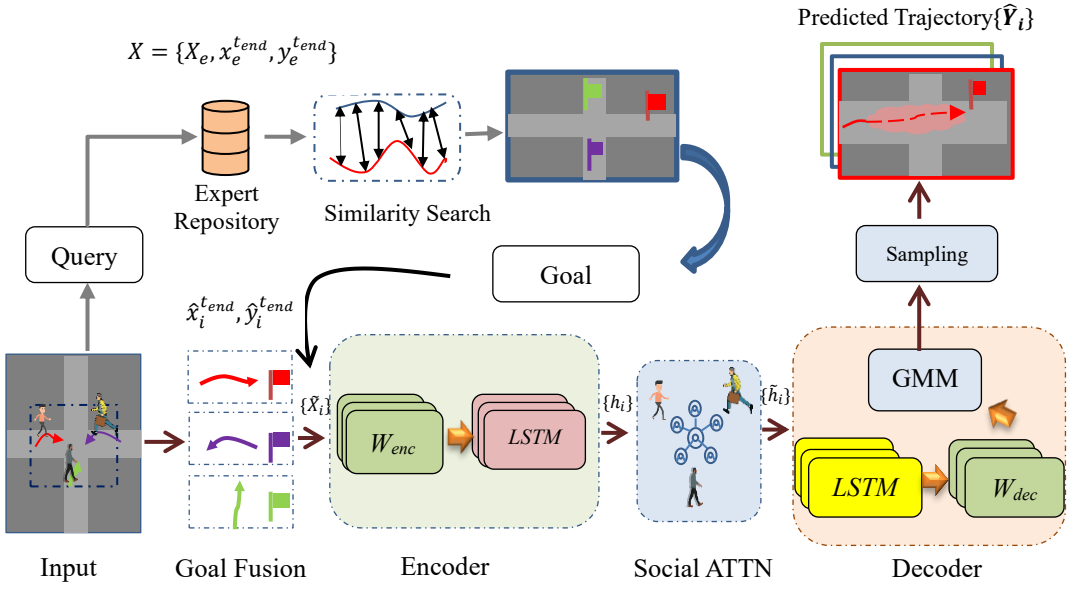


Fig. 1.3: Example of a Goal-driven approach for trajectory prediction [80]

anticipate the covert intention of the pedestrian at times. Intention or behavioural attributes of pedestrians might not necessarily be reflected in his past trajectories. Instead, a holistic view and comprehension of the context, scene, pedestrian behavioural attributes, and interaction with fellow pedestrians are vital for visual perception of future pedestrian actions. For instance, trajectory-based techniques might falsely predict a pedestrian checking for a bus to be a crossing event on the road. Hence, comprehending the cardinal cause or intention of the pedestrian behind any action event will help to anticipate its future action accurately. Furthermore, designing such systems that can predict the underlying intention of the pedestrian not only helps in anticipating their goal but also sheds the burden on AVs by shifting the focus on only those pedestrians who intend to cross the road [2], [11], [83].

1.4 Research Motivation

Pedestrian intention prediction is a critical area of research, driven by the need to enhance road safety and improve the integration of autonomous vehicles into urban environments. As vulnerable road users, traffic incidents disproportionately affect pedestrians, underscoring the importance of accurately anticipating their movements. In the European Union, pedestrians account for approximately 22% of all road fatalities, with 69% occurring within urban areas. This highlights the heightened risk pedestrians face in city settings, where vehicle interactions are frequent. Similarly, pedestrians constitute about 30% of all road-related deaths in Japan, emphasizing a global concern for pedestrian safety [4].

The increasing prevalence of larger vehicles like SUVs has further exacerbated pedestrian dangers. These vehicles often have design features that reduce driver visibility and increase the severity of collisions. In Australia, the popularity of such vehicles has been linked to a rise in road fatalities and serious injuries, particularly among pedestrians, cyclists, and motorcyclists. Despite efforts to improve pedestrian safety, recent data indicates that challenges persist. For instance, in Nashville, Tennessee, pedestrian fatalities decreased by 30% in the first half of 2024 compared to the same period in 2023. However, the total number of deaths remains significantly

higher than a decade ago, with 33 pedestrian fatalities reported in 2024, up from 18 in 2014 [83]. These statistics underscore the need for advanced systems to predict pedestrian intentions to prevent accidents.

Moreover, developing reliable pedestrian intention prediction models is crucial for adopting autonomous vehicles. Ensuring that these vehicles can effectively anticipate and respond to pedestrian behaviour is vital for public trust and the successful integration of autonomous technology into daily transportation systems. The key motivation behind studying pedestrian intention prediction lies in the pressing need to reduce pedestrian fatalities and injuries, adapt to evolving vehicle trends, and support the safe deployment of autonomous vehicles in complex urban landscapes. This thesis contributes to the field by addressing key challenges in pedestrian intention prediction, including handling multimodal data, mitigating the impact of noisy or missing information, and improving computational efficiency for real-time applications. The proposed approaches enhance the accuracy and robustness of predictive models in complex urban environments by integrating advanced deep-learning architectures, attention mechanisms, and context-aware modelling techniques. Through these advancements, the research refines existing methodologies and establishes a foundation for future innovations in AI-driven perception and human behaviour modelling.

1.5 Problem Formulation

The central problem addressed in this thesis revolves around the accurate and efficient prediction of pedestrian intentions in dynamic and uncertain urban environments, a critical requirement for the safe operation of autonomous vehicles. Pedestrian behaviour is inherently complex and unpredictable, often characterized by abrupt changes such as sudden stops, accelerations, and shifts in direction, influenced by a variety of contextual factors including distractions, urgency, and social interactions. Capturing these nuanced motion patterns demands models capable of understanding fine-grained spatiotemporal dependencies.

Furthermore, the integration of multimodal data comprising visual cues, trajectory information, and environmental context introduces additional challenges due to potential data corruption or loss caused by sensor failures, occlusions, and adverse weather conditions. These issues compromise model reliability, particularly in real-world deployments. Addressing this, there is a pressing need for robust learning strategies that can reason causally about pedestrian behaviour even in the presence of incomplete or noisy inputs.

Another key challenge lies in interpreting the complex and often subtle interactions between pedestrians and dynamic environmental elements, such as traffic signals and nearby vehicles. Traditional models struggle to maintain contextual awareness in such settings, leading to limited accuracy in intention prediction. Furthermore, ensuring the computational efficiency of such models remains difficult, as high memory consumption and processing overhead hinder their real-time applicability in autonomous driving systems.

This thesis formulates the problem as a two-fold task: short-term pedestrian intention prediction, focused on determining whether a pedestrian is likely to cross the street (crossing intention); and long-term pedestrian behaviour prediction, which involves forecasting the future trajectory of the pedestrian over a longer time horizon. Both tasks require the design of a robust, context-aware, and computationally efficient framework capable of learning from multimodal data and accurately modelling the complex dynamics of pedestrian behaviour in real-time.

1.6 Research Objectives

The principal objective of this thesis is to address the challenges inherent in predicting pedestrian intentions for autonomous vehicles, such as the dynamic nature of urban traffic, the randomness of pedestrian decisions and actions, and the necessity of interpreting these actions within diverse contextual frameworks. Furthermore, to enhance the understanding of scene context and adaptively respond to the variability in pedestrian dynamics, this research aims to transition from relying on single

modalities to robustly integrating multiple modalities for improved prediction accuracy. To this end, the following research objectives have been proposed:

RO.1 To review the different pedestrian intention prediction techniques for Autonomous Vehicles (AVs).

RO.2 To propose an efficient pedestrian intention prediction model by utilising various sources of contextual information of road scenes.

RO.3 To develop a multimodal architectural design for robust pedestrian intention prediction by AVs.

RO.4 To design a pedestrian intention prediction model in long term with scene semantic understanding.

RO.5 To build a joint framework for pedestrian intention prediction in both short term and long term.

1.7 Research Contributions

This thesis presents a set of research contributions aimed at advancing pedestrian intention prediction and trajectory forecasting under complex real-world conditions. Each contribution addresses a specific challenge in multimodal modelling, data robustness, contextual understanding, or computational efficiency.

One of the primary challenges in pedestrian intention prediction arises from the inherently inconsistent and unpredictable nature of pedestrian movement, which often includes sudden stops, accelerations, and abrupt changes in direction. These behaviours are typically influenced by a range of contextual factors such as distractions, urgency, and social interactions. To effectively model these dynamics, this work introduces several methodologies, such as Interaction Encoder constructed using Graph Convolutional Networks (GCNs) and a Progressive Denoising Attention Mechanism, enabling a more nuanced understanding of spatiotemporal motion patterns.

The dependence on multimodal data comprising visual inputs, trajectory coordinates, and environmental context introduces additional complexity due to the potential for sensor failures, occlusions, and adverse weather conditions. To enhance robustness under such conditions, a Counterfactual Training Approach is employed. This method improves the model's causal reasoning capabilities by explicitly modelling the relationships between observed behaviour and contextual features, thereby enhancing reliability in real-world deployment scenarios.

Understanding the intricate relationships between pedestrian behaviour and environmental elements, such as traffic signals and oncoming vehicles, is critical for accurate intention prediction. This thesis addresses this by integrating a Co-Learning Transformer Architecture, the aforementioned GCN-based Interaction Encoder, and a Context-Aware Feature Fusion Module (CAFFM) in the proposed works. These components enhance the model's contextual awareness and enable more precise prediction of pedestrian intentions in complex urban environments.

Finally, achieving real-time performance remains a significant concern due to the high computational and memory demands of deep learning-based models. To address this, this thesis incorporates Multi-Head Shared Weight Mechanisms (MHSWM), Shared MLP Heads, and a Progressive Encoder-Decoder Architecture, all of which contribute to reducing model complexity and computational overhead while preserving predictive accuracy.

1.8 Outline of the Thesis

The thesis entitled, '**Pedestrian Intention Prediction for Autonomous Vehicles**' is structured into six chapters, followed by a comprehensive bibliography. The organization of the thesis is as follows:

Chapter 1: Introduction presents the research motivation, outlines the challenges of pedestrian intention prediction, and discusses the role of deep learning in addressing them. It includes the problem formulation, research objectives, key contributions, and an overview of the thesis structure.

Chapter 2: Literature Review offers a detailed review of state-of-the-art methodologies, assessing their strengths and limitations regarding prediction duration, input feature types, and model architectures, thereby identifying research gaps and defining the objectives addressed in the thesis.

Chapter 3: Short-term Intention Prediction details two innovative approaches for short-term crossing intention prediction. The first utilises appearance, context, motion dynamics, and social interactions, integrating a Multi-Head Attention-based Graph Convolutional Network ($MHA - AdjMat$) to capture complex pedestrian behaviours and improve predictive accuracy. The second approach introduces a three-stage transformer encoder structure driven by a Co-learning module and Multi-Head Shared Weight Attention for efficient multimodal data fusion, enhanced by a Co-learning Adaptive Composite (CAC) loss to optimise training and feature representation.

Chapter 4: Long term Intention Prediction presents a GAN-based methodology for long-term trajectory prediction, addressing pedestrian movement's complexity and stochastic nature through adaptive learning strategies and contextual attention mechanisms. This chapter discusses using a Dynamic Progressive Generator and an Adaptive Fuzzified Discriminator to boost prediction accuracy, reduce mean squared error, and enhance model generalisation, particularly in ambiguous scenarios.

Chapter 5: Unified Short-term and Long-term Intention Prediction introduces a unified framework for concurrent short- and long-term pedestrian intention prediction, utilising a three-phase counterfactual training method and Progressive Denoising Attention (PDA) for effective cross-modal feature integration. The approach incorporates a Conditional Variational Autoencoder (CVAE) refined with a Context-Aware Feature Fusion Module (CAFFM) to optimise trajectory prediction accuracy.

Chapter 6: Conclusion, Future Scope and Social Impact presents a concise summary of the key ideas, findings, and contributions corresponding to each research objective addressed in the thesis. It also outlines potential directions for future research and discusses the broader social implications of the proposed methods and their applications.

CHAPTER 2
LITERATURE REVIEW

This chapter reviews state-of-the-art approaches for pedestrian intention prediction, categorized into short-term and long-term forecasting methods. Short-term prediction focuses on anticipating immediate pedestrian actions, crucial for real-time applications like autonomous navigation and intelligent surveillance [17]-[20]. At the same time, long-term forecasting aims to predict movement patterns over an extended period, benefiting urban planning and traffic management [21]-[27]. Recent advancements in deep learning have significantly improved these models by leveraging multimodal features such as visual data, trajectory coordinates, and environmental context [9], [15]-[16]. Various fusion strategies, including early, late, and adaptive fusion, enhance predictive accuracy, while social interaction modelling through Graph Convolutional Networks (GCNs) and attention mechanisms further refine behavioural understanding.

Short-term approaches often utilize recurrent neural networks (RNNs), transformers, and hybrid models to process motion cues and contextual information for immediate decision-making [18], [85]-[86]. In contrast, long-term trajectory forecasting relies on transformer-based architectures and generative models like GANs and VAEs to capture uncertainty in pedestrian motion [22]-[23]. These methods improve safety in autonomous systems by enabling proactive decision-making in dynamic environments. By systematically evaluating multimodal architectures, fusion techniques, and interaction-aware modelling, this chapter highlights key advancements and identifies open challenges in pedestrian intention prediction research as follows:

2.1 Short-term intention prediction

This section explores the critical role of multimodal feature representation, learning architectures, fusion strategies, and spatiotemporal modelling of pedestrian interactions in short-term pedestrian intention prediction. Integrating diverse input

modalities, such as trajectory, pose, and visual context, is crucial in capturing pedestrian intent, particularly in complex or ambiguous scenarios. Appropriate learning architectures, including RNNs, CNNs, GNNs, and Transformers, enable the extraction of meaningful spatial and temporal dependencies, improving predictive accuracy. Fusion strategies, ranging from early feature concatenation to advanced self-attention mechanisms, facilitate the effective integration of multimodal information, enhancing model generalization. Furthermore, spatiotemporal modelling of pedestrian interactions provides deeper insights into motion patterns and environmental influences, refining intent prediction in dynamic traffic settings. These elements collectively contribute to developing more robust and adaptive pedestrian intention prediction frameworks.

2.1.1 Multimodal Feature Representations

Within existing literature, various features have been employed to alleviate the cognitive load of employed intelligent frameworks. The predominant feature for predicting pedestrian intent has been trajectory or historical motion data, evident in numerous studies [9]. Nonetheless, relying solely on trajectory proves inadequate when no historical data exists, or the trajectory is abrupt [12]. Combining pose key-point information with trajectory has shown promise in advancing intention prediction [54]. Visual appearance features also offer significant cues regarding pedestrian intent and future actions. Recent pioneering research highlights the critical role of visual context features in understanding a pedestrian's traffic environment, as these features provide essential cues for predicting pedestrian behaviour. Additionally, in dynamic scenes, integrating ego-vehicle motion information enhances the assessment of a pedestrian's relative movement concerning onboard cameras, thereby improving situational awareness and predictive accuracy [17]-[18]. However, existing approaches have often overlooked the incorporation of richer contextual information, which is crucial for robust and generalizable pedestrian intention prediction. The Biped model [84] attempted to address this limitation by independently and jointly encoding multiple modalities, offering a more comprehensive understanding of pedestrian behaviour. Nevertheless, its heavy reliance on semantic scene parsing constrained its

adaptability, making it less effective in diverse and unstructured environments where contextual variations are significant.

2.1.2 Multimodal Learning Architectures

Several seminal architectures [17], [18], [84] are proposed hitherto that endeavour to fuse multi-source inputs optimally for efficient and accurate pedestrian crossing prediction. SF-GRU [17] fused local context, appearance, bounding box, pose and ego-vehicle speed hierarchically using GRU as the encoder. Along similar lines, Yang et al. [18] proposed the fusion of two different channels for visual: local context and are proposed hitherto that endeavour to fuse multi-source inputs optimally for efficient and accurate pedestrian crossing prediction. Nonetheless, these approaches were restricted since they do not consider the impact of human social conduct and interactions with the surrounding environment, which are inevitable in assessing a pedestrian's short-term intention. Moreover, these works lack rich feature representations of distinct pedestrian modalities and efficient integration of these modalities for enhancement. Subsequent seminal works [80], [81], [85] demonstrated the potential of attention mechanisms and transformers to fuse spatiotemporal features. In a recent work, Bai et al. [86] introduced a progressive feature fusion module with a self-attention mechanism to extract relevant multimodal features selectively. Pedestrian Graph+.[87] infers spatiotemporal relationships autonomously through network learning. In another seminal work, Yao et al. [88] designed a human visual learning-inspired Attention Relation Network for deeper traffic scene comprehension. However, such evolved multimodal architectures that seamlessly integrate diverse modalities, enhancing both the learning efficacy of the model and the intention prediction performance, remain limited.

2.1.3 Fusion Strategies

In deep learning architectures, fusion techniques are pivotal for integrating information across multiple modalities to enhance prediction accuracy, thereby influencing the effectiveness of intention prediction tasks. Early feature fusion

methods, such as straightforward or weighted concatenation of features before the final classification network, are employed in notable works [89], [90]. However, these approaches may not fully capture the complex intermodal relationships essential for optimal performance, potentially limiting the integration of diverse modal information. Several pioneering works [17], [18] employ Multi-Stream Architecture, processing each modality separately within network branches and combining their outputs later. While this allows for learning modality-specific representations and weighting each modality's importance in predictions, it may hinder capturing critical intermodal dependencies and interactions.

In contrast, advanced fusion techniques like self-attention mechanisms, as seen in noteworthy works [86], [91], [92], enhance pedestrian intention prediction by emphasizing relevant factors and dynamically selecting multimodal features. For instance, Bai et al. [86] introduce a progressive feature fusion module using a self-attention mechanism to select useful multimodal features from global to local perspectives for pedestrian crossing prediction. Sharma et al. [91] propose an adaptive fusion module to dynamically weigh all the visual, motion and interaction features, enhancing performance. Additionally, cross-modal Transformer architectures, as explored in another notable study [93], capture dependencies between data types and model interactions between pedestrians and traffic agents, considering both pedestrian and ego-vehicle dynamics. Despite recent advancements, current methodologies often face challenges in effectively interpreting correlations across different modalities, limiting their generalizability to unseen cases.

2.1.4 Spatiotemporal Modelling of Pedestrian interactions

Modelling subtle nuances of interactions among pedestrians in a dynamic traffic scene, influencing their crossing intention, is pivotal in mimicking human-like subconscious decision-making in AVs and ADAS systems. The inherent randomness and dynamic nature of these interactions in space and time pose challenges for learning models. Recently, spatiotemporal modelling has been widely used in pedestrian intention prediction, particularly with the development of deep learning models

capable of handling spatial and temporal information. Spatio-temporal modelling is a crucial aspect of pedestrian intention prediction as it allows for modelling both pedestrian behaviour's spatial and temporal dimensions. Spatial modelling refers to the modelling of the physical space in which the pedestrian is operating, including the location and orientation of the pedestrian in the environment. This information is vital for understanding the pedestrian's surroundings, potential obstacles, and interactions with other objects. On the other hand, temporal modelling refers to the time aspect of pedestrian behaviour. This information is essential for identifying sudden behavioural changes that may indicate an intention to cross the street [81], [94].

Leveraging the unprecedented success of RNNs and CNNs in several computer vision applications, the last decade has witnessed an increase in their usage in modelling sequential behaviour of pedestrians over time. RNNs help capture their motion patterns by allowing the network to maintain information about the pedestrian's motion over time. CNNs learn to identify significant features, such as the shape and movement of the pedestrian and the fully connected layers, and then use these features to make a prediction. Hamed et al. [89] employed a combination of CNN and Time-Distributed Layers (TDL) to visually represent pedestrians, with the LSTM layer learning the temporal context. Rasouli et al. [17] introduced an RNN encoder-decoder architecture that captures a visual representation of the image surrounding pedestrians concatenated with pedestrian dynamics. Inspired by this, Yao et al. [18] utilized an encoder-decoder architecture and a novel Attention Relation Network (ARN) to induce a spatiotemporal understanding for anticipating pedestrian crossing intentions. Other groundbreaking works [95], [96] integrated a hybrid combination of CNNs and RNNs for spatiotemporal encoding. However, RNNs and CNNs are challenging to train when there is sparse data, which could be the case in most pedestrian datasets. Furthermore, the vanishing gradient issue in RNNs for longer sequences and inefficiency in capturing the global relationship of the pedestrian with scene objects by CNNs make the overall performance of the CNN-RNN-based architectures suffer in the long run [97].

Several approaches [29], [82], [87], [98], [99], [100], [101] have also explored Graph Neural Networks (GNNs) to capture the interactions between pedestrians and their environment. These approaches depict each pedestrian as a node in the graph, and edges are added between nodes to model pedestrian relationships. Liu et al. [81] utilised graph convolution to understand the intricate spatiotemporal relationships in a scene, incorporating both pedestrian-centric and location-centric perspectives. Similarly, Naik et al. [99] analysed the relationship between pedestrians and the scene using a Scene Spatio-temporal Graph Convolution Network. Chen et al. [29] advanced this concept further by employing graph autoencoders to comprehend the impact of the surroundings on pedestrian crossing decisions. Zhang et al. [100] integrated Graph Attention Networks (GAT) into Graph Convolutional Networks(GCNs) to strengthen further the ability to model complex social interactions. In another interesting work, Riaz et al. [82] proposed a GNN-GRU-based architecture PedGNN that takes a sequence of pedestrian skeletons as input to predict crossing intentions. Ling et al. [101] utilised GCN(Graph Convolutional Network) with spatial, temporal and channel attention to strengthen feature extraction for more accurate and fast prediction. However, GNNs can struggle to generalize to unseen graphs, as they depend heavily on the graph structure and node features. This can be a limitation for anticipating pedestrian intention where the graph structure is subject to change over time [102].

To the extent of our knowledge, the examination of Transformers in pedestrian intention prediction is a novel and under-researched area, with only a handful of works that have addressed it [85], [103], [104], [105]. Achaji et al. [103] proposed a Transformer model with bounding boxes as the only required input. However, it relies solely on bounding box information, which fails to capture the road context and may misinterpret movements similar to crossing behaviour. The PIT framework [104] incorporated a sophisticated integration of a temporal fusion block and a self-attention mechanism, enabling the modelling of the dynamic relationships between the pedestrian, ego-vehicle, and environment. This progressive processing of temporal information enables the capture of dynamic interactions between elements in a manner that is more congruent with human-like behaviour. Additionally, Osman et al. [85] introduced a novel adaptive mechanism that dynamically assigns weights to the

significance of current and previous frames, utilizing an attention mask within the Transformer, thereby promoting dynamic spatiotemporal modelling. In another seminal work, Zhang et al. [105] capture temporal correlations within pedestrian video sequences using a Transformer module and address the uncertainty of complex pedestrian crossing scenes.

2.2 Long-term Intention Prediction

This section analyses long-term pedestrian intention prediction, focusing on the challenges of forecasting movement over extended time horizons, including non-linear motion patterns, abrupt directional changes, and multimodal trajectory distributions. An overview is provided of recurrent and Transformer-based models, assessing their effectiveness in capturing temporal dependencies and spatial interactions. The discussion then extends to advanced generative models, such as GANs and CVAEs, which have been developed to model the inherent uncertainty of pedestrian motion by generating diverse trajectory distributions. Finally, key limitations are highlighted, including integrating environmental context, multimodal fusion, and stability in generative learning. These remain critical for improving the accuracy and generalizability of long-term intention prediction models.

2.2.1 Recurrent and Transformer-based Trajectory Prediction

Trajectory prediction methodologies have significantly progressed, particularly with integrating recurrent and transformer-based models. Xue et al. [106] presented a novel method featuring dual temporal attention mechanisms and an embedded location-velocity attention layer within a specialized tweak module. Yu et al. [107] leveraged transformative mechanisms to adeptly model intra-graph crowd interactions and inter-graph temporal dependencies to capture intricate spatial-temporal dynamics. Taking a distinctive approach, Tao et al. [108] integrated rich information into Long Short-Term Memory (LSTM), effectively addressing dynamic interactions, long-trajectory correlations, and semantic scene layouts. In parallel, Wong et al. [109] estimated continuous key points and defined spectrum interpolation

sub-networks for trajectory modelling at both key points and interaction levels. These advancements [106], [107] showcase a nuanced understanding of advanced scene dynamics and pedestrian interactions within dynamic environments. However, inherent challenges, including abrupt directional changes and irregular non-linear motion patterns, such as sudden stops or velocity fluctuations, contribute to systematic errors in trajectory predictions [12], [23], [94], [106], [110].

2.2.2 Deep Generative Models for Trajectory Prediction

In contrast to preceding recurrent and transformer-based methodologies, generative models like GANs and CVAEs demonstrate unparalleled adaptability to abrupt changes and irregular motion patterns. Furthermore, the recent evolution in deep generative models has marked a transformative shift from predicting a single optimal trajectory to generating a distribution of potential future trajectories. Ivanovic et al. [111] adopted a Gaussian Mixture Model (GMM) for target trajectory assumption, presenting the Trajectron network to predict GMM parameters through a spatio-temporal graph. Trajectron++ [110] extended this approach to accommodate dynamics and heterogeneous input data. BiTrap [22] and SGNet [23] both leverage Conditional Variational Autoencoders (CVAEs) to handle the multimodality and uncertainty of human movements. BiTrap [22] enhances prediction accuracy through a goal-conditioned bidirectional approach that considers past and future contexts. However, its emphasis on a single endpoint may limit its capacity to model the full range of possible trajectories. SGNet [23] incorporates multi-temporal goal estimation, improving long-term accuracy and adding granularity to predictions. However, both models could benefit from better integration of environmental context to refine their predictions.

Mangalam et al. [24] addressed human trajectory prediction by modelling intermediate stochastic goals known as endpoints. Recognizing the inherent stochasticity in future human motion patterns, Y-Net [25] learned goal and path multimodalities by leveraging scene semantics. Su et al. [112] designed SIT to learn the spatiotemporal correlation of pedestrian trajectories via attention mechanisms.

Wang et al. [23] acknowledged the temporal variance in the goal of a moving agent by estimating goals at multiple temporal scales for more accurate trajectory prediction. Gao et al. [113] enhanced the model's awareness of diverse social interaction patterns through Social-DualCVAE, conditioned on past trajectories and unsupervised classification of interaction patterns. Recently, Yue et al. [114] integrated neural social physics to model pedestrian stochastic motion patterns followed by a CVAE to generate predictions. While GANs are known for generating realistic outputs, they face challenges like mode collapse and training instability. In contrast, Variational Autoencoders (VAEs) offer a more stable and reliable approach by learning latent space representations that encapsulate the underlying structure of trajectory data, making them better suited for precise and diverse predictions [115].

2.3 Research Gaps

Through an analysis of prior state-of-the-art methods for pedestrian intention prediction for autonomous vehicles, several research gaps have been identified as follows:

- Limited studies [14], [16] address the unpredictable movements of pedestrians in busy urban areas.
- Existing models [18], [87] fail to predict crossing intention effectively at TTEs greater than 1 second.
- There is a notable lack of research [14], [116] focused on addressing noisy or missing modality data in multimodal pedestrian crossing intention models.
- Prior multimodal architectures [106], [117] lack efficient integration of contextual understanding, such as the relationship between environmental factors and the dynamic behaviour of pedestrians.
- Although recent advancements have focused on improving prediction performance, limited attention has been given to optimizing memory footprint, resulting in architectures that are computationally demanding [15], [72].

2.4 Conclusion and Future Scope

This chapter has provided an in-depth review of state-of-the-art approaches for pedestrian intention prediction, distinguishing between short-term and long-term forecasting techniques. It highlighted the evolution of multimodal learning, advanced fusion strategies, and spatiotemporal modelling methods that have significantly enhanced predictive performance in dynamic environments. Short-term models benefit from real-time multimodal integration, whereas long-term models emphasize trajectory uncertainty and contextual reasoning through generative frameworks. Despite these advances, challenges remain in dealing with noisy or missing modalities, modelling social and contextual interactions effectively, and ensuring computational efficiency.

Addressing these gaps serves as the principal motivation for the present thesis. The subsequent chapters will introduce four novel methodologies, each specifically designed to tackle the limitations identified in short-term and long-term intention prediction, respectively. These contributions aim to advance the development of robust, context-aware, and computationally efficient predictive frameworks that are better aligned with the demands of real-world autonomous and assistive systems.

CHAPTER 3

SHORT-TERM INTENTION PREDICTION

In complex urban environments, accurately anticipating pedestrian behaviour is essential for ensuring the safety and reliability of autonomous vehicles and intelligent transportation systems. Among the various facets of human motion forecasting, short-term pedestrian intention prediction, focused on forecasting immediate actions such as initiating a crossing, plays a critical role in enabling timely decision-making in real-world traffic scenarios. Its applications are particularly significant in autonomous driving, where vehicles must respond rapidly to sudden pedestrian movements to avoid potential collisions.

Building upon this premise, this chapter introduces two short-term pedestrian intention prediction models aimed at overcoming key challenges in multimodal scenarios, particularly contextual integration and computational efficiency. A major limitation of existing approaches is their inability to effectively model the relationship between environmental factors and pedestrian behaviour, resulting in suboptimal contextual understanding. Furthermore, the high computational complexity of current models necessitates memory optimization to enhance efficiency without compromising predictive accuracy. To address these issues, the proposed frameworks undergo systematic experimental evaluation and a comprehensive analysis of results, discussions, and a comparative assessment against state-of-the-art methods.

3.1 Visual-Motion-Interaction Guided Pedestrian Intention Prediction Framework

The capability to comprehend the intentions of pedestrians on the road is one of the most crucial skills that the current autonomous vehicles (AVs) are striving for to become fully autonomous. In recent years, multimodal methods have gained traction by employing trajectory, appearance, context, etc., to predict pedestrian crossing intention. However, most existing research works still lag rich feature representational ability in a multimodal scenario, restricting their performance. Moreover, less

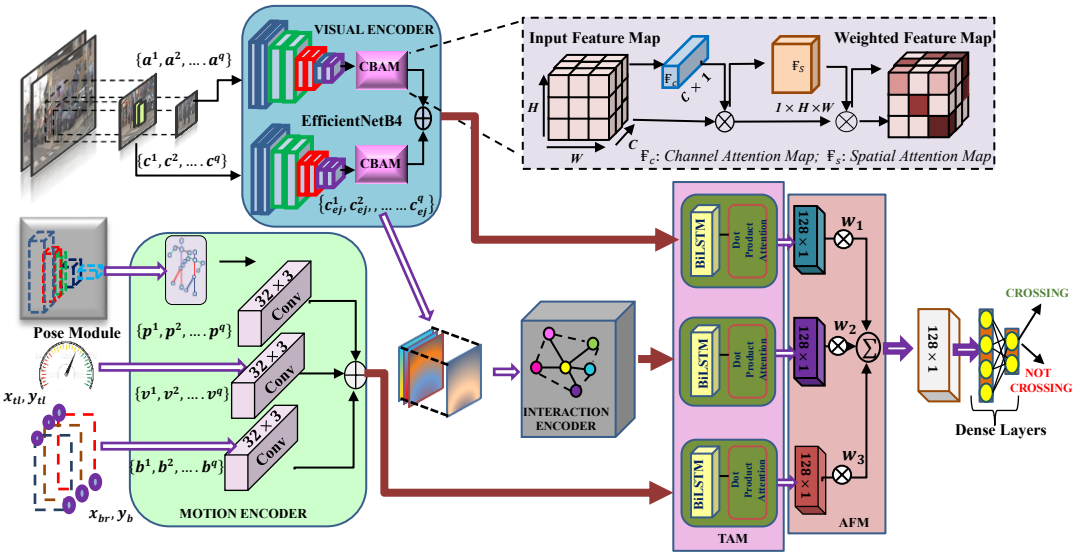


Fig 3.1: Illustration of the Visual-Motion-Interaction-Guided (V-M-I) framework

emphasis is put on pedestrian interactions with the surroundings for predicting short-term pedestrian intention in a challenging ego-centric vision. An efficient Visual-Motion-Interaction-guided (VMI) intention prediction framework has been proposed to address these challenges. This framework comprises a Visual Encoder (VE), Motion Encoder (ME) and Interaction Encoder (IE) to capture rich multimodal features of the pedestrian and its interactions with the surroundings, followed by temporal attention and adaptive fusion module to integrate these multimodal features efficiently. The proposed framework outperforms several SOTAs on benchmark datasets: PIE/JAAD with Accuracy, AUC, F1-score, Precision and Recall as 0.92/0.89, 0.91/0.90, 0.87/0.81, 0.86/0.79, 0.88/0.83 respectively. Furthermore, extensive experiments are carried out to investigate different fusion architectures and design parameters of all encoders. The proposed VMI framework predicts pedestrian crossing intention 2.5 sec ahead of the crossing event.

3.1.1 Proposed Methodology

The traffic scene environment is dynamically varying, and with onboard cameras, not just the traffic scene but the relative size and distance of the objects/pedestrians to the ego-vehicle in motion is also changing continuously. Unlike human drivers, who can decipher non-verbal cues of the surrounding traffic

environment and make decisions accordingly, AVs lack this inherent capability. This complex scenario motivates us to utilise multimodal information involving visual, motion and contextual features from the dynamic traffic scene to make AVs mimic human's cognitive ability to anticipate pedestrians' intentions on the road. In this work, the pedestrian crossing intention prediction can be formulated as a binary classification task wherein the motive is to find the probability of a pedestrian 'j' intention to cross or not, בּ$_j \in (0,1)$, provided the past observations of visual, dynamic and interaction information of the pedestrian and the ego-vehicle speed for $'q'$ time steps. The proposed implementation of the Visual-Motion-Interaction-Guided (V-M-I) framework for intention prediction of pedestrians is described in Fig 3.1. This architecture employs multimodal features extracted from a traffic scene, involving target pedestrian visual appearance and non-visual dynamic features. Furthermore, interaction features are extracted from the surrounding context, since pedestrian interactions with co-pedestrians on the road also play a pivotal role in influencing crossing behaviour. This architecture is comprised of the following essential components.

3.1.1.1 Visual Encoder (VE)

The VE encodes the visual features of the pedestrian and its surroundings as described below:

Appearance – The sequence of RGB images of a traffic scene captures the variations of pedestrians' appearance temporally [18], [119]. The visual appearance features $A_j = \{a_j^1, a_j^2, a_j^3, \dots \dots a_j^q\}$ of the pedestrian $'j' \in (1, m)$ for past observed $'q'$ time steps are captured using image portions of the size of the bounding box consisting of the pedestrian in a scene.

Context – The local contextual information of the target pedestrian depicts its relationship with the dynamic traffic scene elements in its surrounding [12] [29]. The surrounding local environment features $C_j = \{c_j^1, c_j^2, c_j^3, \dots \dots c_j^q\}$ are extracted using a larger image portion size to include the immediate contextual details. This is achieved by extending the dimension by at least twice the size of the pedestrian $'j'$ bounding

box and masking it to include only the surrounding details.

In generating image representations, transfer learning has emerged as a formidable approach, leveraging pre-trained models to extract meaningful features. Inspired by the insights derived from such transfer learning paradigms [120]-[122], the proposed framework endeavours to balance the trade-off between the model's performance and complexity in terms of size and number of parameters of the model by utilising the EfficientNetB4 [123] model. This model has 82.9% Top-1 Accuracy on ImageNet [124], while being $17.47x$ smaller and having $18x$ fewer parameters than best existing ConvNet [125] so far. Both the appearance and the local context features are processed in parallel through EfficientNetB4 pre-trained on ImageNet. This is followed by Convolutional Block Attention Module (CBAM) [126] to emphasize relevant features across both channel and spatial dimensions as depicted in Fig. 3.1. Let the appearance and context features after processing via EfficientNetB4 $A_e = \{a_{ej}^1, a_{ej}^2, a_{ej}^3, \dots \dots a_{ej}^q\}$ and $C_e = \{c_{ej}^1, c_{ej}^2, c_{ej}^3, \dots \dots c_{ej}^q\}$ of dimension $C \times H \times W$ where C, H and W denotes the number of channels, height and width dimension of the feature space. These processed features are then passed through an average pooling mechanism to reduce feature dimensions. The parallel branches are finally concatenated to give a modified visual feature representation as shown in Eqns. (1)- (3) as follows:

$$A_e' = \mathbb{F}_c(A_e) \otimes A_e; A_e'' = \mathbb{F}_s(A_e') \otimes A_e' \tag{3.1}$$

$$C_e' = \mathbb{F}_c(C_e) \otimes C_e : C_e'' = \mathbb{F}_s(C_e') \otimes C_e' \tag{3.2}$$

$$A_e" \oplus C_e" = \{A_e"C_e"_j^1, A_e"C_e"_j^2, \dots \dots A_e"C_e"_j^q\} \; \forall \; j(1,n) \tag{3.3}$$

where $\mathbb{F}_c \in \mathbb{R}^{C\times1\times1}$ and $\mathbb{F}_s \in \mathbb{R}^{1\times H\times W}$ corresponds to 1-D channel attention and 2-D spatial attention map, respectively. The procedure of mapping attention both spatially and channel-wise via CBAM is represented in Fig. 3.1, the darker the colour, the higher the weight assigned to the feature. The CBAM output then undergoes global average pooling to reduce the dimensions of the feature vector for further computations.

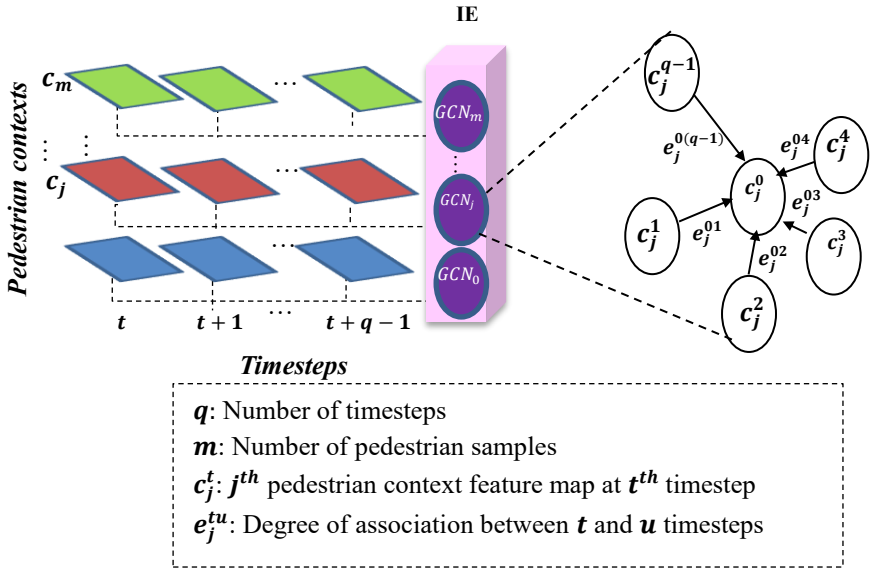


Fig. 3.2: Visualisation of Interaction Encoder

3.1.1.2 Motion Encoder (ME)

The representation of these features is described as follows:

Pose: In this work, HRNet [127] generates a set of 17 pose key points per pedestrian sample using a bounding box sequence. This network has achieved ~76% mAP on the MS COCO [128] dataset while maintaining higher resolution representation throughout the estimation process. This network surpasses the performance of several SOTA pose estimation modules [129] utilised in quite a few existing works[17], [18]. The key points are obtained in the form of x and y coordinates resulting in a vector of 34 values per pedestrian sample. These values undergo normalization and then concatenation into a feature vector for further processing for $'q'$ past observations and are represented as $P_j = \{p_j^1, p_j^2, p_j^3, \dots \dots p_j^q\}$

Trajectory: The location of a pedestrian $'j'$ in a 2D coordinate space is provided with top-left $\{x_{tl}, y_{tl}\}$ and bottom-right $\{x_{br}, y_{br}\}$ coordinate points represented as $B_j = \{b_j^1, b_j^2, b_j^3, \dots \dots b_j^q\}$.

Speed: This consists of speed value measurements of the ego-vehicle in km/h given as - $S_j = \{s_j^1, s_j^2, s_j^3, \dots \dots s_j^q\}$

Motivated by the scientific findings [130] about the low computational

complexity and faster training time of Conv 1-D over Conv 2-D for one-dimensional data, the ME processes pose, trajectory and speed through a Conv 1-D parallelly with 32 filters of kernel size, 3 and stride, 1. These transformed features are then combined and passed to the TAM.

3.1.1.3 Interaction Encoder (IE)

Pedestrians' interactions with traffic scene elements dynamically impact the road crossing intention [12], [87]. Inspired by some of the seminal works [29], [87], the proposed approach leverages graph convolutional networks (GCNs) [131] to model the temporal relationship of pedestrian interactions across consecutive frames. In the proposed architecture, the nodes of the graph $G: \{N_j, E_j\}$ represents $'t^{th}'$time-step encoded feature map of the target pedestrian $'j'$ context and edges E_j represents the associations existing among these feature maps corresponding to different time steps in a traffic scene as shown in Fig. 3.2. These inter- associations within a graph are reflected via an adjacency matrix $R_{q \times q}$ computed as shown in Algorithm 3.1, where 'q' is the total time steps considered per sample. For each j^{th} pedestrian sample, received context feature map is applied with linear transformation followed by a reshape operation to extract Query, Key and Value matrices, as described in Step 4, Algorithm 3.1. Step 5 computes multi-head attention $H_{q \times b \times d}$ by calculating individual attention heads h^i and then concatenating all $'b'$ attention heads. Following this, $H_{q \times b \times d}$ is then mapped with number of timesteps again as $H_{q \times r}$. Hence, the Multi-Head Attention-based Adjacency Matrix [132], $R_{q \times q}$, derived at Step 6, ensures extraction of the implicit contextual details of the dynamic interactions of the target pedestrian at consecutive timesteps. Further, GCN is applied to the graph representation derived above with nodes $'N_j'$ and multi-head attention-based adjacency matrix $'R_j'$ depicting the relationship between nodes. The adjacency matrix requires a normalisation step to curb issues of vanishing or exploding gradients as network training may be sensitive to the range of scale of values. The normalization step includes generating the Laplacian matrix as represented below:

$$\check{R} = D^{-\frac{1}{2}}(R + I)D^{-\frac{1}{2}} \tag{3.4}$$

wherein self-loops are also considered to incorporate the target node's features in the propagation, by adding Identity matrix I to the adjacency matrix R and D is the diagonal degree matrix containing row-wise summation of $(D + I)$ matrix. Here, the spectral propagation rule for the convolution of the graph features is employed [133]. Where, GCN layer is modelled by applying a linear transformation operation as a scalar product of the adjacency matrix and the hidden feature, followed by a Gaussian Error Linear Unit (GeLU) [134] activation function for the next layer $'l'$ as shown in Eqn. (3.5)

$$G(l + 1) = GeLU\left(\left(D^{-\frac{1}{2}}(R + I)D^{-\frac{1}{2}}\right)X_l W_l\right) \tag{3.5}$$

where $X(l)$ is the previously hidden layer output of the GCN convolution layer. For $l = 0$, $X(0) = C_e{}''$.

3.1.1.4. Temporal Attention Module (TAM)

In this module, the Bidirectional long short-term memory (BiLSTM) layer provides enhanced temporal representations of the input sequence by leveraging learning through bi-directional layers. The following attention module weighs the most relevant parts of the feature map $'\Upsilon'$along the temporal dimension. The attention weight vector $\widehat{\varkappa}_k$ for k^{th} branch of the architecture is given in Eqn. (6) and (7) as follows:

$$s_k = softmax(\tanh(W\Upsilon_k + bias)) \tag{3.6}$$

$$\widehat{\varkappa}_k = \sum_d \alpha_k \Upsilon_k \tag{3.7}$$

Where d is the dimension along which attention vector computation is carried out. In this case, it is the output vector length of the BiLSTM layer. The number of hidden units of the BiLSTM layer employed is 64.

Table 3.1: Training specifications of the proposed framework

Training Parameters	JAAD	PIE
Optimizer	ADAM	ADAM
Learning Rate	2×10^{-5}	5×10^{-5}
# Epochs	60	70
L2 Regularization	0.0001	0.0001
Loss Function	Binary Cross Entropy	Binary Cross Entropy
Batch Size	8	16

3.1.1.5 Adaptive Fusion Module (AFM)

The generated features from the VE, IE and ME, followed by the TAM, have varying impacts on the pedestrian intention prediction, therefore, an adaptive fusion is introduced to accordingly weigh all the encoded features. These hidden representations of the proposed fusion have been accumulated in a vector representation, $'E'$ as shown in the Eqn. (3.8) as follows:

$$E = \sum_{k=1}^{3} w_k \hat{\pi}_k \tag{3.8}$$

where $\hat{\pi}_k$ represents the encoded hidden states of the prior TAM and w_k are the trainable weights with HeNormal initialisation [135]. This is followed by dense layers of 64 and 8 units and a final activation function to give *'crossing'* or *'not crossing'* predictions.

3.1.2 Experimental Work and Results

This section presents the experimental evaluation of the proposed pedestrian intention prediction model. The implementation details, including architectural configurations, training procedures, and computational setup, are outlined, followed by a description of the datasets used for evaluation. A comparative analysis with state-of-the-art methods is then conducted to assess the effectiveness of the proposed models. Finally, an ablation study is performed to examine the contribution of individual components, providing insights into their impact on overall model performance.

3.1.2.1 Implementation Details

The experimental settings for implementation are outlined in Table 3.1. The computation of the pose key points, appearance and contextual features is done before training. Data augmentation involves horizonal image flipping for balancing crossing/non-crossing samples to mitigate prediction bias.

3.1.2.2 Datasets

The proposed framework is evaluated using two publicly available benchmark datasets, namely JAAD[136] and PIE[6]. The JAAD dataset is having 346 video clips with a total duration of 240 hours recorded at 30 frames per sec(fps). Each clip ranges from $5 - 15\ sec$ with a resolution of 1920×1080 and 1280×720. The bounding boxes and tracking ids are provided for each pedestrian. The driver's action is implicitly encoded as vehicle speed for training the model. The PIE dataset consists of 1842 pedestrian tracks with longer sequences and increased pedestrian samples with annotations compared to JAAD. The dataset configuration follows the training/validation/test split as recommended in [137] for JAAD [6] for PIE.

3.1.2.3 Comparison with State-of-the-art methods

The performance of the proposed framework has been compared against the following SOTA methods:

- *Pie_traj* [6]: employs an RNN-based encoder-decoder to extract vital information regarding pedestrian appearance and surrounding context and pedestrian dynamics.
- *Stacked Fusion GRU* (SF-GRU) [17]: hierarchically stacks GRU encoder that fuses features of high relevance like appearance at the beginning and others like ego-speed at the last.
- *Feature Fusion and Spatio-temporal Attention* (FFSTA) [18]: employs a hybrid architecture to fuse features from visual and non-visual branches. The local and global context are combined in the visual branch while other features like pose,

Table 3.2: Comparison of existing SOTAs with the proposed method on the PIE and JAAD dataset

Methods	Year	PIE/JAAD				
		Acc	**AUC**	**F1**	**Prec**	**Rec**
PIE_traj [6]	2019	0.79	-	0.87	-	-
SF-GRU [17]	2020	0.87/0.84	0.85/0.80	0.78/0.62	0.74/0.54	0.64/0.73
FFSTA [18]	2022	0.85/0.83	0.83/0.82	0.71/0.63	0.69/0.51	0.72/0.81
BiPed [84]	2020	0.91	0.90	0.85	0.82	0.88
MMA [119]	2020	0.89/0.89	0.88/0.88	0.81/0.81	0.77/0.77	0.85/0.85
IA [88]	2021	0.84/0.87	0.90/0.70	0.88/0.92	0.96 /0.66	0.81
PG+[87]	2022	0.89/0.86	0.90/0.88	0.81/0.65	0.83/0.58	0.79/0.75
Ours	**-**	**0.92/0.89**	**0.91/0.90**	**0.87/0.81**	**0.86/0.79**	**0.88/0.83**

bounding box and vehicle speed are hierarchically fused in the non-visual branch.

- *BiPed* [84]: proposes a bifold encoding approach for the individual and shared representation encompassing trajectory, grid locations, ego speed and global context.
- *Multi-model Atrous*(MMA) [119]: processes diverse input modalities through the visual encoding and dynamics encoding branch, followed by corresponding attention modules and subsequent fusion for joint representation.
- *Intent and Action* (IA) [88]: presents a human visual learning-inspired Attention Relation Network ensuring a deeper understanding of the scene semantics and other pedestrian-specific features.
- *Pedestrian Graph+*(PG+) [87]: employs a fully convolutional graph-based neural network that inputs context, human pose key points, and ego-speed.

The performance of the proposed approach in comparison to the state of the arts is represented in Table 3.2. It is evident from the table that the proposed VMI-guided framework performs better in terms of all evaluation metrics against SOTA models, except for the F1 score and precision on the PIE dataset and the F1 score and recall on the JAAD dataset. The intent and action [88] perform a little better (∼1%) in the F1 score and also have a significant leap in precision approximately by 10 %. This drop-in precision in our proposed work is compensated with an overall

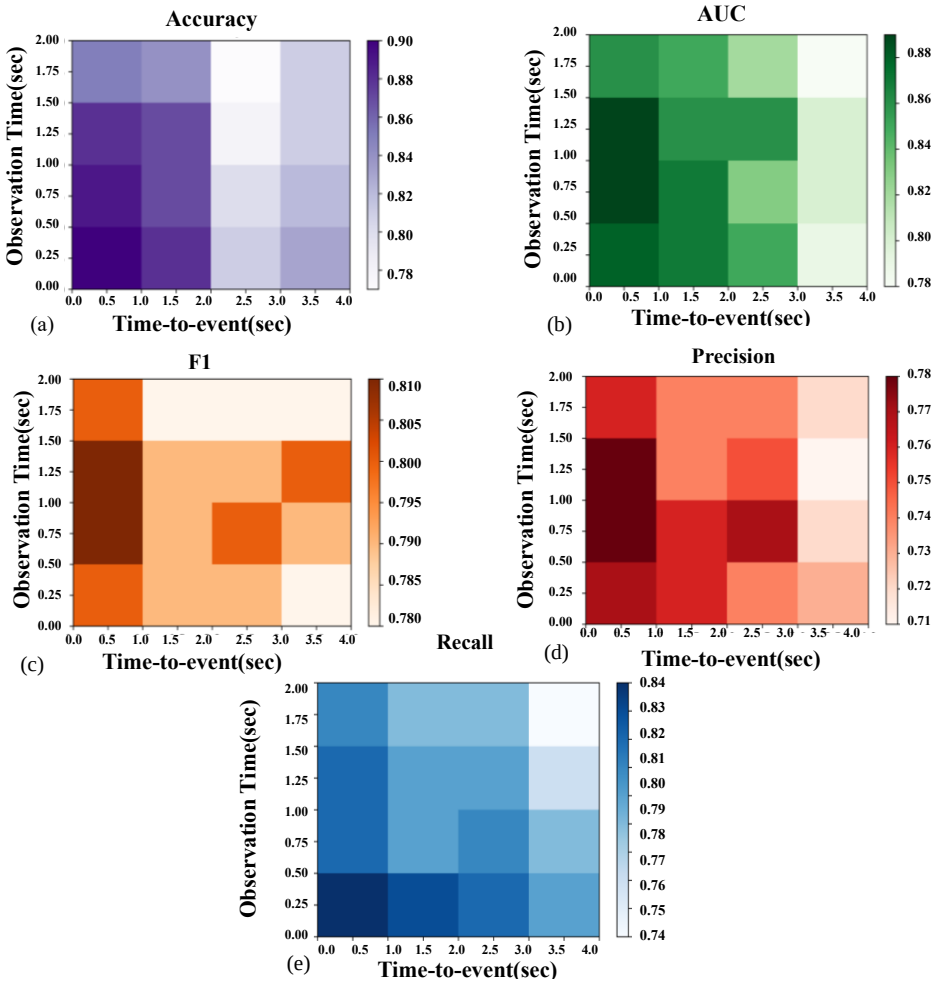


Fig. 3.3: Impact of Time to Event (TTE) and Observation Sequence Lengths (OSL) on Crossing Intention Prediction, evaluated in terms of (a) Accuracy, (b) AUC, (c) F1 Score, (d) Precision, and (e) Recall metrics.

improvement in accuracy (~ 8.7%), AUC score (~1%) and recall (~8.7%) on the PIE dataset. Similarly, the considerable jump in performance metrics like accuracy (~8.64%), AUC(~2.3%) and precision(~28.6%) overcomes the decline in F1 score (~12%) in the JAAD dataset against [88] by our proposed approach. The optimal observation length of 0.5 sec with 2.5 sec time-to-event has been found empirically which will be discussed later in the ablation study.

3.1.2.4 Ablation Study

This section presents an ablation study to assess the impact of key design choices in the proposed framework. The effects of varying Time to Event (TTE) and Observation Sequence Lengths (OSL) are examined, followed by an analysis of different fusion strategies. Furthermore, the contributions of the Motion Encoder,

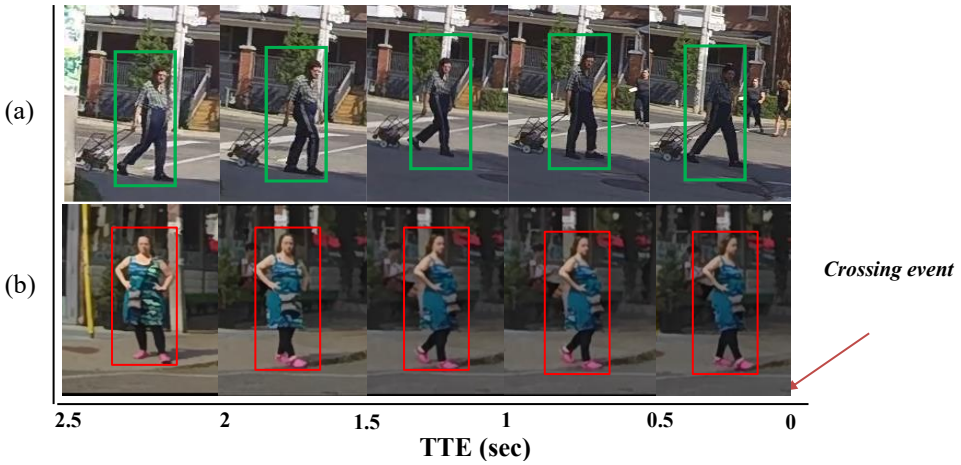


Fig. 3.4: Qualitative samples of pedestrian short-term intention prediction: (a) Correctly predicted intention. (b) Failure case where ***Green*** indicates crossing, ***Red*** denotes non-crossing based on TTE.

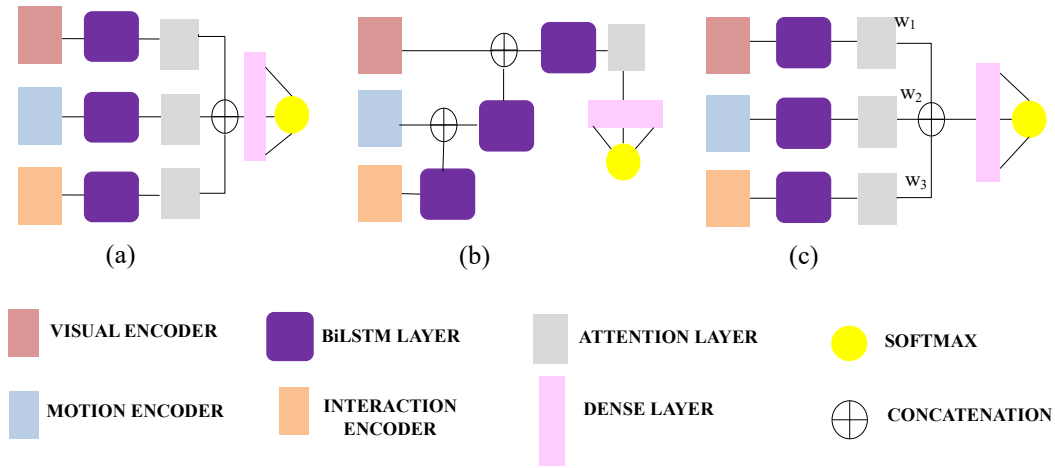


Fig. 3.5: Overview of Pedestrian Intention Prediction Fusion Architectures. (a) Parallel Fusion (PF), (b) Hierarchical Fusion (HF), and (c) Adaptive Fusion (AF) approaches.

Interaction Encoder, and Visual Encoder are evaluated to understand their relevance. providing a stronger rationale for the architectural and methodological choices. The analysis are as follows:

i. **Impact of choosing different Time to event (TTE) and Observation Sequence Lengths (OSL)**: The high variability in traffic scene dynamics with every passing second has a considerable impact on the crossing intention prediction of the pedestrian. OSL also impacts the prediction process. The longer the OSL, the greater the information acquired. However, it may also add some insignificant details that may result in erroneous prediction results. In this section, different TTE points are considered on the timeline of the crossing/not crossing event ranging

Table 3.3: Ablation study on different fusion architectures

Fusion	Encoder	PIE/JAAD				
		Acc	AUC	F1	Prec	Rec
HF	GRU	0.83/0.84	0.78/0.81	0.73/0.71	0.76/0.71	0.71/0.71
	LSTM	0.85/0.84	0.80/0.83	0.74/0.74	0.78/0.72	0.70/0.76
	BiLSTM	0.85/0.85	0.80/0.85	0.77/0.75	0.78/0.74	0.75/0.77
	BiLSTM+A	0.86/0.85	0.85/0.86	0.77/0.76	0.79/0.74	0.76/0.78
PF	GRU	0.85/0.83	0.83/0.82	0.78/0.72	0.77/0.69	0.78/0.75
	LSTM	0.86/0.83	0.83/0.82	0.78/0.74	0.77/0.72	0.79/0.77
	BiLSTM	0.86/0.85	0.81/0.84	0.78/0.76	0.80/0.74	0.77/0.79
	BiLSTM+A	0.88/0.86	0.84/0.87	0.80/0.77	0.81/0.75	0.79/0.80
AF	GRU	0.87/0.85	0.85/0.81	0.80/0.77	0.79/0.75	0.81/0.79
	LSTM	0.87/0.85	0.84/0.84	0.82/0.78	0.84/0.76	0.80/0.81
	BiLSTM	0.90/0.87	0.90/0.86	0.84/0.79	0.85/0.77	0.83/0.82
	BiLSTM+A	**0.92/0.90**	**0.91/0.89**	**0.87/0.81**	**0.85/0.78**	**0.88/0.83**

A : Attention

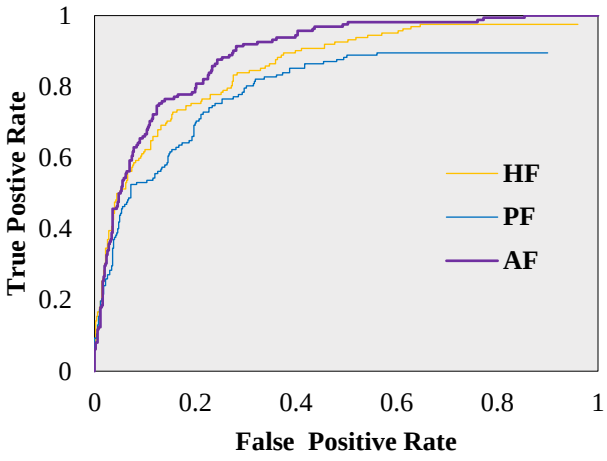


Fig. 3.6: ROC Curves illustrate the Performance of Parallel Fusion (PF), Hierarchical Fusion (HF), and Proposed Adaptive Fusion (AF) architectures

from 0-4 sec, sampled at every 0.2 sec and 2 secs long OSL is sampled at every 0.2 sec as shown in Fig. 3.3. It is observed that there is a gradual decline in the overall performance of the proposed approach as the TTE increases and vice versa as the TTE approaches close to 0 since the intention of the pedestrians becomes evident by that time. There is also a slight gain in accuracy, AUC and precision up to 1.5 sec of OSL but at the expense of a decrease in recall. Moreover, a more balanced metric F1 score shows no huge variations ($\sim$3%) with the increase in OSL. Notably, the proposed approach works satisfactorily with all parameters $\sim$(>

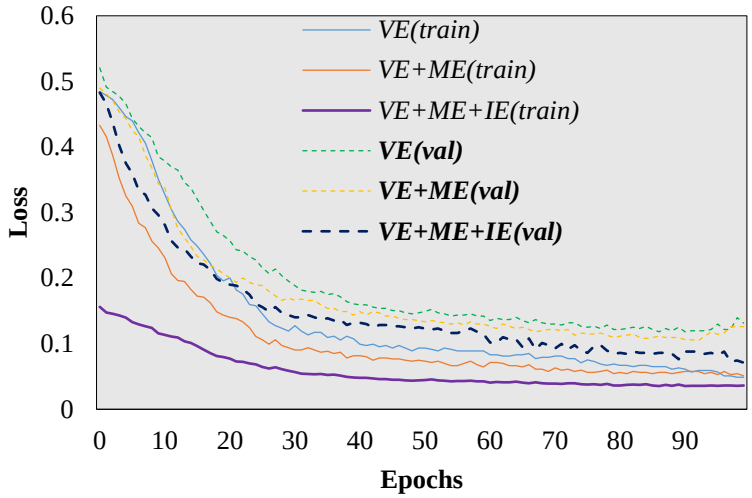


Fig. 3.7: Training and validation loss analysis for encoder combinations (VE, VE+ME, VE+ME+IE), with validation losses (---) and training losses (−).

80%) even 3 $secs$ to TTE as opposed to SF- GRU [17] where F1 score, precision and recall drop to $\sim 60-65\%$ at 3 sec to TTE. Furthermore, the overall performance of the proposed approach does not drop below $12-13\%$ with respect to the highest attained metric value even at the end TTE and OSL values. Nonetheless, in SF-GRU[17], the TTE and OSL variations cause a drop as high as 33.3%. This proves the robustness of our approach against varying OSL and TTE. Fig. 3.4 (a) illustrates a qualitative sample prediction on the PIE dataset 2.5 sec ahead of the crossing event, while Fig. 3.4 (b) depicts a failure case by a sudden pedestrian direction change.

ii. **Fusion Strategies*:*** Different fusion architectures for multimodal features have been employed in this ablation study inspired by the works [17], [18] as depicted in Fig. 3.5. Table 3.3 shows the performance of different fusion architectures on the PIE and JAAD datasets. It is observed that the proposed adaptive fusion with the BiLSTM + Attention layer achieves the best performance overall. The ROC Curve visualisations for different fusion approaches on the PIE dataset are shown in Fig. 3.6. This study also explores four encoder variations (GRU, LSTM, BiLSTM, BiLSTM + Attention), highlighting the apparent increasing performance trend of the BiLSTM followed by the attention layer across different fusion architectures.

—

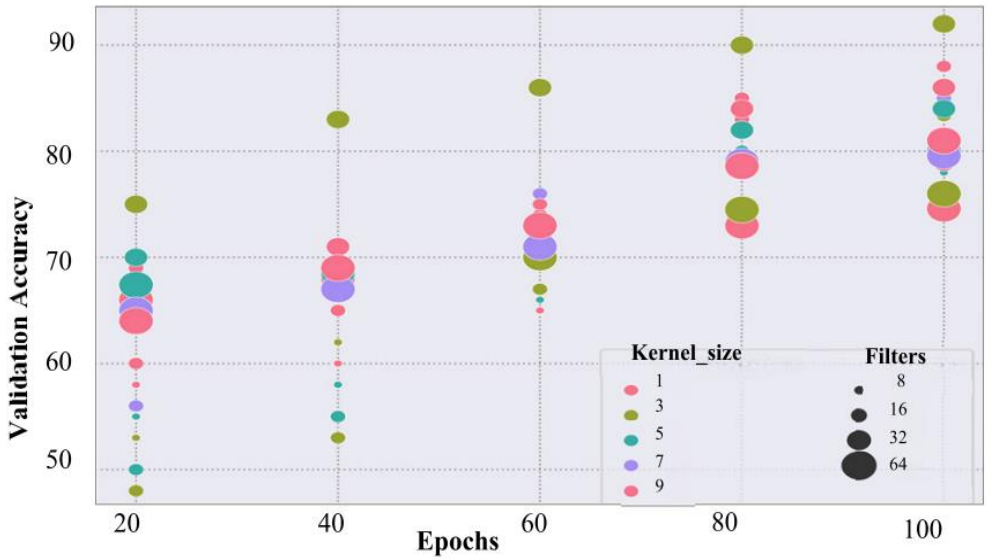


Fig. 3.8: Hyperparameter analysis of the convolutional layer in the Motion Encoder (ME)

iii. **Relevance of different encoders*:*** This section studies the training performance of VE, ME and IE both individually and jointly. The analysis of loss curves in Fig. 3.7 indicates effective training for all combinations, with a steeper validation loss curve observed when utilising all three encoders VE+ME+IE, showcasing the model's enhanced generalisation capability achieved through integration of multiple modalities.

iv. **Motion Encoder**: The ME utilises the Conv-1D layer to extract significant motion features from the pedestrian's pose, trajectory and ego speed. The impact of the number of filters and the kernel size of the Conv-1D layer on the training performance is shown in Fig. 3.8, where kernel size of 3 and 32 filters yields the highest validation accuracy of ~92 % on the PIE dataset after 100 epochs. This is attributed to the smaller kernel size's ability to generate fine-grained features, though it lacks neighbouring context. Conversely, larger kernel sizes like 5,7 or 9 may overlook intricate details. The plot indicates that more filters are required for the ME to capture complex and stochastic pedestrian motion. A significant drop of ~ 10-20% in accuracy is observed if the number of filters is increased or decreased from 32 during training for a fixed kernel size. However, the number of parameters also grows as we increase the kernel size or the number of filters. Therefore, a Conv-1D layer with 32 filters and a kernel size of 3 is employed in this work to minimise the trade-off between accuracy and the number of parameters.

Table 3.4: Ablation study on Interaction Encoder Components

Graph Node Features	Adjacency Matrix	PIE/JAAD				
		Acc	**AUC**	**F1**	**Prec**	**Rec**
Pose	Uni	0.81/0.81	0.84/0.82	0.78/0.75	0.77/0.72	0.78/0.77
	DI	0.82/0.82	0.83/0.80	0.79/0.74	0.78/0.71	0.79/0.76
	SHA	0.84/0.84	0.83/0.82	0.80/0.76	0.80/0.73	0.79/0.78
	MHA	0.84/0.83	0.82/0.81	0.78/0.73	0.79/0.71	0.77/0.75
Trajectory	Uni	0.82/0.83	0.83/0.84	0.79/0.74	0.79/0.71	0.79/0.78
	DI	0.81/0.81	0.83/0.81	0.79/0.72	0.80/0.69	0.77/0.76
	SHA	0.82/0.85	0.87/0.86	0.80/0.76	0.81/0.73	0.78/0.79
	MHA	0.86/0.83	0.83/0.82	0.82/0.75	0.83/0.72	0.80/0.77
Appearance	Uni	0.83/0.86	0.85/0.84	0.79/0.76	0.81/0.73	0.77/0.78
	DI	0.83/0.85	0.84/0.83	0.80/0.75	0.82/0.74	0.78/0.76
	SHA	0.85/0.86	0.85/0.85	0.81/0.77	0.81/0.74	0.80/0.79
	MHA	0.87/0.85	0.87/0.83	0.79/0.77	0.79/0.76	0.79/0.77
Context	Uni	0.88/0.86	0.85/0.83	0.84/0.76	0.82/0.74	0.86/0.78
	DI	0.87/0.85	0.87/0.83	0.83/0.74	0.80/0.72	0.85/0.76
	SHA	0.91/0.87	0.88/0.85	0.84/0.78	0.81/0.75	0.87/0.81
	MHA	**0.92/0.89**	**0.90/0.89**	**0.87/0.81**	**0.85/0.79**	**0.88/0.84**

v. **Interaction Encoder*:*** The graph node features and the adjacency matrix significantly enhance the graph's efficiency in modeling subtle interactions of the target pedestrians within a traffic scene. Table 3.4 shows the impact of diverse node features on performance metrics. It is observed that the context features secure the highest metric values, followed by appearance. The ablation study also explores distinct adjacency matrix computation methods: *Uniform*(Uni) which involves random initialisation of the adjacency matrix with values ranging from 0 to 1, *Distance-Inverse*(DI) that employs inverse Euclidean distance between pedestrians [138], *Single-head Attention*(SHA) that utilises a self-attention module inspired by the work [139] and lastly the adjacency computation using *Multi-head attention* (MHA) described in this work [132]. Notably, the MHA method outperforms other adjacency matrix computation methods in terms of classification performance. This is attributed to SHA's limited human behavioural learning range and Uni's inability to adequately represent intrinsic human interaction randomness. The DI method too restricts the interaction to sheer separation based limiting its performance.

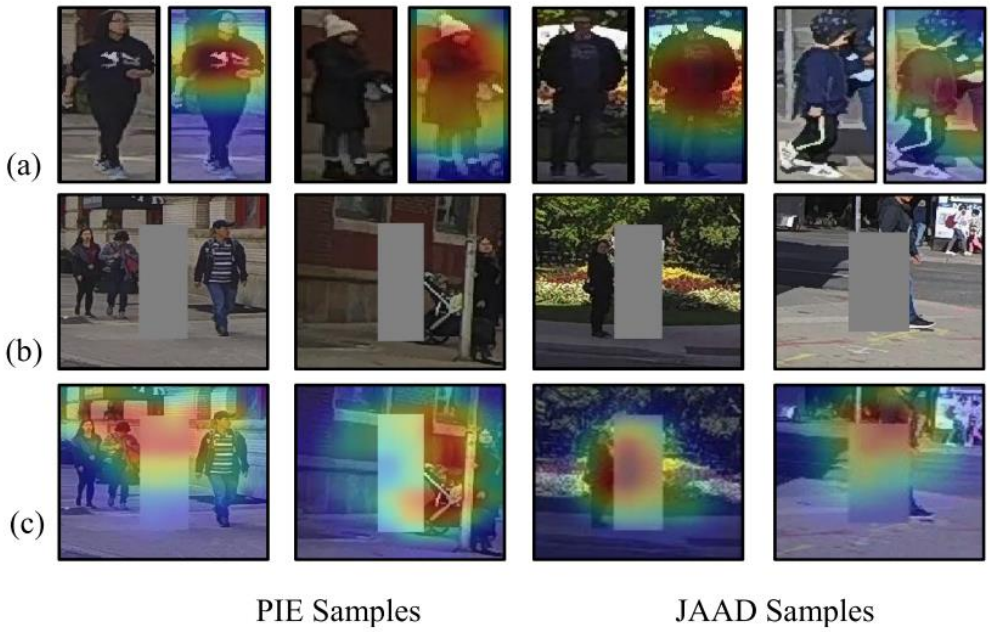


Fig. 3.9: GradCAM [140] visualizations showing key focus areas: (a) pedestrian ROI, (b) pedestrian image with surrounding context and (c) contextual cues influencing intent prediction.

vi. **Visual Encoder*:*** The GradCAM [140] visualisations for the CBAM feature maps in the VE are obtained by superimposing on the real cropped images enhancing model explainability as depicted in Fig. 3.9. It is inferred from both the PIE and JAAD dataset images that the torso region is given higher weights in comparison to other body parts. This affirms that the head, gaze, shoulder and posture forming the torso region of the human body, play an integral part in assessing road-crossing intentions. Additionally, the GradCAM visualisations for context show that the VE pays higher attention towards co-pedestrians in the vicinity that are influencing the pedestrian's crossing intent.

This work introduces a multimodal pedestrian intention prediction framework that adaptively fuses visual, motion, and interaction features using spatial, channel, and temporal attention mechanisms. A novel $MHA-AdjMa$t based GCN in Interaction Encoder leads to superior performance over state-of-the-art models on the JAAD and PIE datasets, predicting crossing intent up to 2.5 seconds in advance. However, limitations in capturing high-frequency temporal dependencies with GCNs persist. To address this, the subsequent section investigates transformer-based architectures for more robust modelling of pedestrian–environment interactions.

3.2 Predicting Pedestrian Intentions with Multimodal IntentFormer: A Co-Learning Approach

The prediction of pedestrian crossing intention is a crucial task in the context of autonomous driving to ensure traffic safety and reduce the risk of accidents without human intervention. Nevertheless, the complexity of pedestrian behaviour, which is influenced by numerous contextual factors in conjunction with visual appearance cues and past trajectory, poses a significant challenge. Several state-of-the-art approaches have recently emerged that incorporate multiple modalities. Nonetheless, the suboptimal modality integration techniques in these approaches fail to capture the intricate intermodal relationships and robustly represent pedestrian-environment interactions in challenging scenarios. To address these issues, a novel Multimodal IntentFormer architecture is presented. It works with three transformer encoders $\{TE_I, TE_{II}, TE_{III}\}$ which learn RGB, segmentation maps, and trajectory paths in a co-learning environment controlled by a Co-learning module. A novel Co-learning Adaptive Composite (CAC) loss function is also proposed, which penalizes different stages of the architecture, regularizes the model, and mitigates the risk of overfitting.

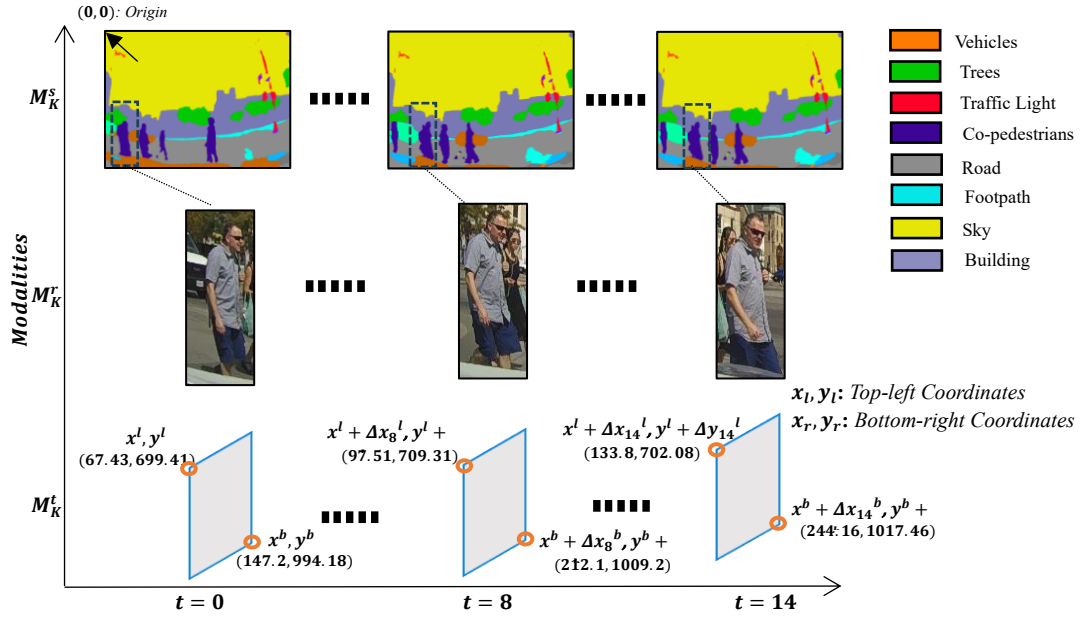


Fig. 3.10: Visualization of various input modalities for a sample input

Each encoder $\{TE_\eta\}$ applies the concept of the Multi-Head Shared Weight Attention (MHSWA) mechanism while learning three modalities in the proposed co-learning approach. The proposed architecture outperforms existing state-of-the-art approaches on benchmark datasets, PIE and JAAD, with 93% and 92% accuracy, respectively. Furthermore, extensive ablation studies demonstrate the efficiency and robustness of the architecture, even under varying Time-to-event (TTE) and observation lengths.

3.2.1 Proposed Methodology

Predicting pedestrian crossing intention is a challenging task with significant implications for pedestrian safety and developing advanced driver assistance systems. In this work, a brief window of 'K' timesteps is analysed from the ego vehicle's perspective, considering the pedestrian's RGB frames and trajectory coordinates. The objective is to ascertain the probability accurately $\rho \epsilon (0,1)$ of the pedestrian's intention to cross the road and, thus, classify the pedestrian as a crossing "1" or non-crossing "0" entity. To predict pedestrian crossing intention in traffic scenes, it is crucial to leverage a variety of modalities that can provide a comprehensive understanding of the pedestrian's surroundings. Therefore, the proposed approach combines three distinct modalities: RGB images, segmentation maps, and trajectory data.

RGB images capture the temporal variations of pedestrian appearance using a sequence of images cropped to the bounding box coordinates provided in the dataset. By analysing a sequence of images, changes in the pedestrian's pose, facial expression, and other visual cues can be tracked, which may indicate crossing intention. *Segmentation maps* provide a global context of the traffic scene surrounding the pedestrian. This facilitates the identification of areas that affect the pedestrian's crossing intention by segmenting the scene into distinct regions based on their visual characteristics. SegFormer [141] generates segmentation maps of the scene that encode different pixel regions in the road scene, including buildings, roads, vehicles,

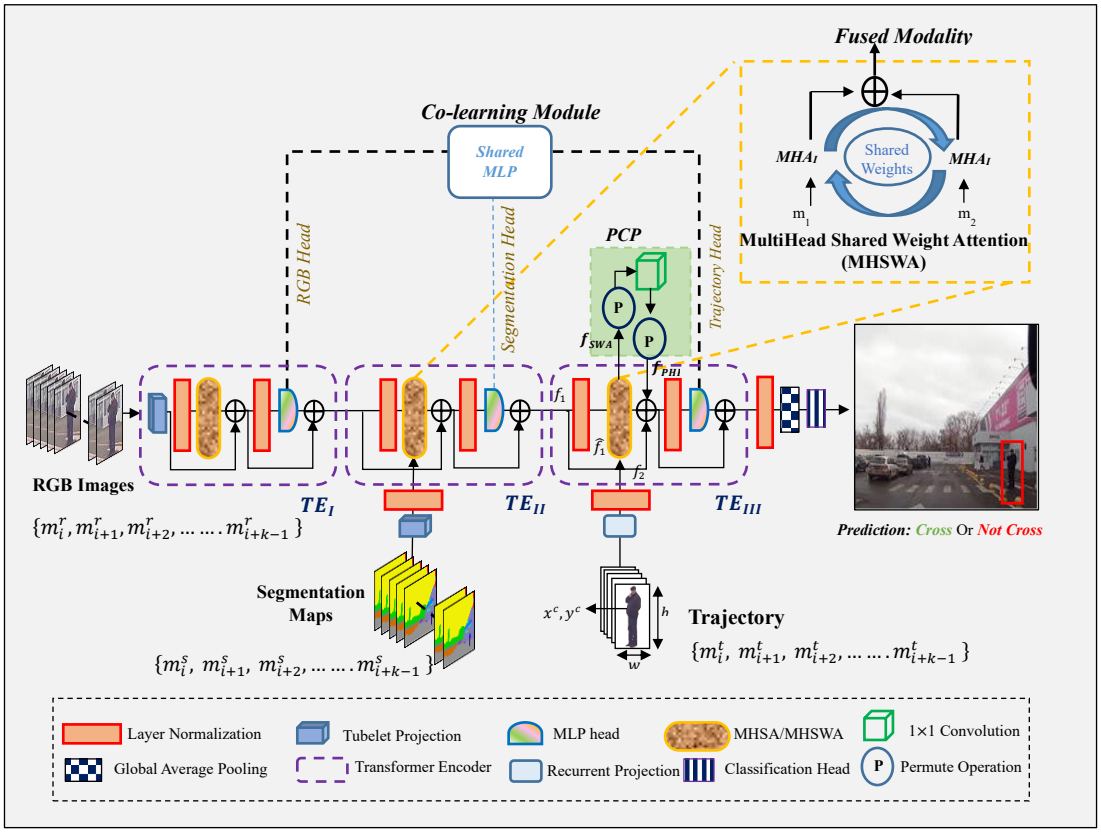


Fig. 3.11: Illustration of proposed IntentFormer architecture for pedestrian crossing intention prediction

and pedestrians, where each region is assigned a distinct label. SegFormer is pre-trained using the ADE20k dataset with 150 distinct classes, enabling effective segmentation of various road scene elements. Visualisations are also provided in Fig. 3.10 for a better understanding of segmentation maps. *Trajectory* provides the pedestrian's location in a 2D coordinate space, denoted by top-left (x^l, y^l) and bottom-right (x^b, y^b) pixel coordinates, enabling the tracking of their movement and predicting their future paths. Each coordinate is measured in the image frame with reference to the origin corner. Any amount of change in the top-left corner and bottom-right corner coordinates are measured as $(\Delta {x_k}^l, \Delta {y_k}^l)$, and $(\Delta {x_k}^b, \Delta {y_k}^b)$, at k^{th} timestep. The coordinates at the new time step k' are given by $(x^l + \Delta {x_{k'}}^l, y^l + \Delta {y_{k'}}^l)$and $(x^b + \Delta {x_{k'}}^b, y^b + \Delta {y_{k'}}^b)$. Fig. 3.10 illustrates the trajectory coordinates of a pedestrian in a sample trajectory.

Together, these modalities provide a comprehensive representation of the pedestrian and their surroundings, enabling the proposed architecture to accurately

predict their crossing intention and ultimately enhance pedestrian safety in traffic scenes. The mathematical representation of the modalities is as follows:

$$M_K^r = \{m_i^r, m_{i+1}^r, m_{i+2}^r, \dots \dots m_{i+k-1}^r\} \quad (3.9)$$

$$M_K^s = \{m_i^s, m_{i+1}^s, m_{i+2}^s, \dots \dots m_{i+k-1}^s\} \quad (3.10)$$

$$M_K^t = \{m_i^t, m_{i+1}^t, m_{i+2}^t, \dots \dots m_{i+k-1}^t\} \quad (3.11)$$

where M_K^r, M_K^s and M_K^t are RGB images, segmentation maps and trajectory data for a total of $'K'$ consecutive frames, respectively. Each modality is taken from i^{th} index to $i + k - 1^{th}$ frames where $'i'$ is the starting index number.

The architecture of the proposed Multimodal IntentFormer is illustrated in Fig. 3.11. The proposed architecture harnesses the power of three transformer encoder stages $\{TE_I, TE_{II}, TE_{III}\}$ to process a heterogeneous array of input modalities. The inputs are diligently sequentially fed to the encoder stages, conforming to the order in which they are presented. Notably, each encoder stage is endowed with Projection, Layer Normalization, Multi-head Attention (MHA), Multi-head Shared Weights Attention (MHSWA), and Multi-layer Perceptron layers that operate seamlessly in tandem to process the corresponding modality as represented in the Eqns. (3.12) - (3.25) as follows:

TE_I: $$\mathrm{PE}^{rgb} = Positional\_Encoder\left(Conv3d(M_K^r)\right) \quad (3.12)$$

$$\mathrm{Att}^{rgb} = MHA\left(LN(\mathrm{PE}^{rgb})\right) + \mathrm{PE}^{rgb} \quad (3.13)$$

$$Features^I = MLP_{shared}\left(LN(\mathrm{Att}^{rgb})\right) + \mathrm{Att}^{rgb} \quad (3.14)$$

TE_{II}: $$\mathrm{PE}^{seg} = Positional\_Encoder(Conv3d(M_K^s)) \quad (3.15)$$

$$LN^{seg} = Layer\_Normalization(\mathrm{PE}^{seg}) \quad (3.16)$$

$$LN^I = Layer\_Normalization(Features^I) \quad (3.17)$$

$$Att^{seg,\ I} = MHSWA(LN^{seg}, LN^I) + LN^{seg} \quad (3.18)$$

$$Features^{II} = MLP_{shared}\left(LN(Att^{seg,\ I})\right) + Att^{seg,\ I} \quad (3.19)$$

TE_{III}: $$\mathrm{PE}^{traj} = Positional\_Encoder(GRU(M_K^t)) \quad (3.20)$$

$$LN^{traj} = Layer\_Normalization\left(\mathrm{PE}^{traj}\right) \quad (3.21)$$

Algorithm 3.2: PCP Module

Input:

-Tensor X ($batch_{size}$**:** $\boldsymbol{b}$; $feature\_$ dim : $\boldsymbol{N}$; $embed\_dim$: $\boldsymbol{M}$**)**

Hyperparameters:

- N°, Convolution filter: 1×1 kernel

Output:

-Tensor $\boldsymbol{Y_{PCP}}$

Step 1: Permutation ($\boldsymbol{P_1}$**) Operation**

Perform permutation P_1 on the input tensor X, mathematically given as follows:

$X_{permuted} = P_1\,(X)$, where X $\in \mathbb{R}^{b\times N\times M}$, $X_{permuted} \in \mathbb{R}^{b\times M\times N}$

Step 2: Convolution filtering

Apply convolution operation with kernel size $[1 \times 1]$ using trainable filter weight as $W_{1\times1}$ on the permuted tensor $X_{permuted}$, obtained from step 1.

$Y_{conv} = Conv(X_{permuted}, W)$, where tensor $Y_{conv} \in \mathbb{R}^{b\times M\times N^{\circ}}$

Step 3: Permutation ($\boldsymbol{P_2}$**) Operation**

Perform permutation operation P_2 on the convolved tensor Y_{conv}, obtained from step 2.

$Y_{P_2} = P_2(Y_{conv})$, where final tensor $Y_{P_2} \in \mathbb{R}^{b\times N^{\circ}\times M}$

return $Y_{PCP} = Y_{P_2}$

$$LN^{II} = Layer\_Normalization(Features^{II}) \tag{3.22}$$

$$Att^{traj,\ II} = PCP(MHSWA(LN^{traj}, LN^{II})) + LN^{traj} \tag{3.23}$$

$$Features^{III} = MLP_{shared}\left(LN\left(Att^{traj,\ II}\right)\right) + Att^{traj,\ II} \tag{3.24}$$

Final output, $\hat{\Upsilon} = sigmoid(GAP(Layer\_Normalization(Features^{III})))$ (3.25)

Where $\mathrm{PE}^{rgb}, \mathrm{PE}^{seg}, \mathrm{PE}^{traj}$ represent positional encodings for RGB images, segmentation maps and trajectories, respectively. It is essential to underscore that two distinct types of projections are leveraged in this architecture: Tubelet projections $Conv3d(M_K^r)$ and recurrent projections $GRU(M_K^t)$. Firstly, the tubelet projections (TP) [142] given by $Conv3d(M_K^r)$ are deployed to assimilate both RGB pedestrian crops and segmentation maps, as utilised in Eqn. (3.12) and Eqn. (3.15). Secondly, recurrent projections (RP) given as $GRU(M_K^t)$, serve as a pivotal tool in processing complex trajectory data, as shown in Eqn. (20). $Features^{I}, Features^{II}$ and $Features^{III}$ represent the output feature vectors coming from the transformer encoder stages: TE_I, TE_{II} and TE_{III}. The objective of the Projection layer is to transform the input data into a latent representation space. The Layer Normalization layer is utilized to normalize the activations of the neurons in each layer, thereby facilitating the

optimization process. The mid-level fusion of different modalities commences with the second stage encoder inspired by [143]. The proposed architecture builds on it by employing a novel shared weight attention mechanism for cohesive learning of parameters. The following section explores the technical intricacies of the IntentFormer shedding light on the PCP (Permutation Convolution Permutation) module, Shared MLP (Multi-Layer Perceptron) layers, and Multi-Head Shared Weight Attention Module (MHSWA), and expounds on their functions and workings:

- *Co-learning Module*: It enables the integration of different modalities, i.e. RGB images, segmentation maps and trajectory in a unified framework, as illustrated in Fig. 3.11. This module is designed to share the MLP head across different layers, which helps to reduce the complexity of the framework while preserving the cross-modality relationships. In practice, this means that the module can simultaneously learn to map the input features to the correct pedestrian class using different modalities. It ensures that the learned representations are consistent across modalities, thus producing multi-modality enriched models for predicting pedestrian crossing intention.

- *Permutation-Convolution-Permutation (PCP) Module*: The PCP module, as shown in Fig. 3.11, facilitates the establishment of skip connections between two transformer layers despite the different dimensions of the output tensors: $f_{SWA(1,2)}$, and f_2. It performs a sequence of permutation operations, a 1×1 convolution operation followed by a permutation operation again. This sequence of operations ensures that the pattern of features stays unaltered without any parameter overhead, as observed when reshaping after dense operation. The steps of the algorithm are provided in Algorithm 3.2.

- *Multi-Head Shared Weight Attention Module (MHSWA)*: The proposed multi-head shared weight attention (MHSWA) module enables the simultaneous learning of attention matrices for heterogeneous modalities, fostering a more cohesive approach to modality fusion, as shown in Fig. 3.11. This module uses multiple instances of the same multi-head attention layer for different

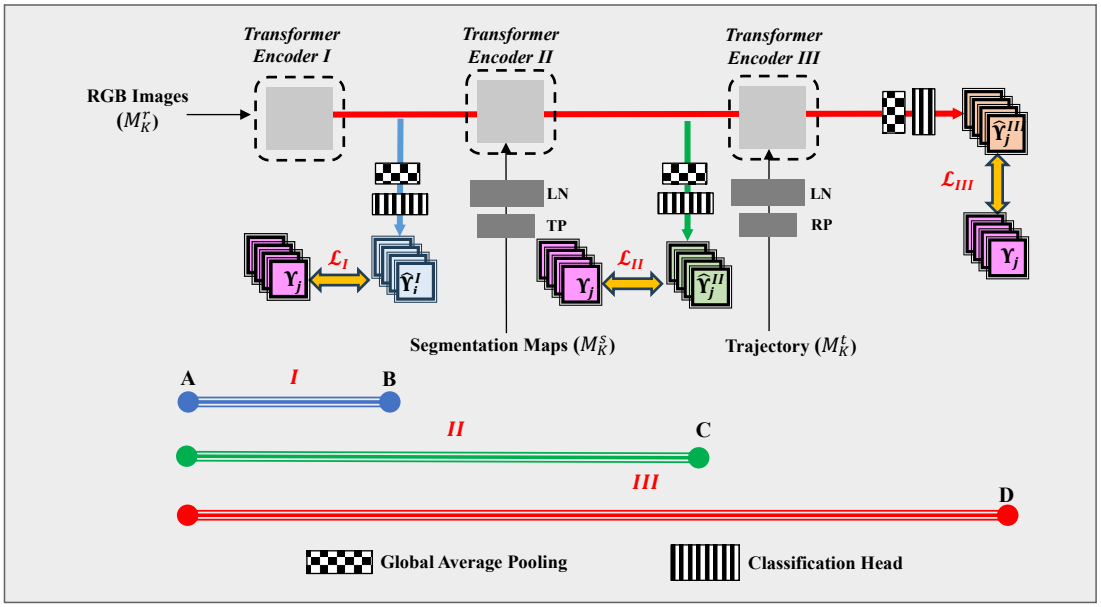


Fig. 3.12: Co-learning Composite (CAC) Loss Function

modalities, eliminating the need for separate attention layers and promoting efficient parameter usage. It employs key, query, and value matrices, which are computed by linearly projecting the inputs for each modality. These matrices are used to compute the attention weights for each modality. When multiple instances of the same multi-head attention layer are called for different modalities, the weights for each modality are adjusted simultaneously in the shared weight attention mechanism.

3.2.1.1 Co-learning Adaptive Composite (CAC) loss function

The most commonly used loss function for binary classification is the binary cross-entropy loss that measures the difference between predicted and true probability distributions. To achieve the goal of fine-tuning the training process and optimizing the model's performance in case of multiple modalities, this work presents a Co-learning Adaptive Composite (CAC) loss function to penalize different stages of the network's architecture, where $'\eta'$ denotes the stages of the architecture, namely RGB head, segmentation head and trajectory head, as described in Fig. 3.12. Υ_j and $\hat{\Upsilon}_j^{\eta}$ represents the ground truth values and predicted probabilities at stage $'\eta'$ respectively for $'j^{th}'$ pedestrian sample. The loss computations for the stages I, II and III follow a path $A \rightarrow B, A \rightarrow C$ and $A \rightarrow D$, respectively. The final loss function is an adaptive

summation of individual binary cross entropy loss terms calculated from various stages of the architecture for a total of $'m'$ samples in the dataset, as represented in Eqn. (18) and (19) as follows:

$$\mathcal{L}_{BCE}^{\eta} = -\sum_{j=1}^{m} \Upsilon_j \log\left(\widehat{\Upsilon}_j^{\eta}\right) + (1-\Upsilon_j)\log\left(1-\widehat{\Upsilon}_j^{\eta}\right) \tag{3.26}$$

$$\mathcal{L}_{final} = \lambda\mathcal{L}^{I} + \mu\mathcal{L}^{II} + \nu\mathcal{L}^{III} \tag{3.27}$$

3.2.2 Experimental Work and Results

This section presents the experimental evaluation of the proposed pedestrian intention prediction model. The implementation details, including architectural configurations, training procedures, and computational setup, are outlined, followed by a description of the datasets used for evaluation. A comparative analysis with state-of-the-art methods is then conducted to assess the effectiveness of the proposed models. Finally, an ablation study is performed to examine the contribution of individual components, providing insights into their impact on overall model performance.

3.2.2.1 Implementation Details

The proposed architecture is trained on a Google Colab Pro instance with access to a high-performance NVIDIA Tesla T4 GPU equipped with 16 GB of memory, running on the CUDA 12.0 platform. The model architecture is built using the TensorFlow 2.10.1 framework. The training regimen involves executing $28\,epochs$ and utilizing a batch size of 2 in conjunction with a tuning phase incorporating the $L2$ regularizer with a regularization factor of $1e^{-6}$. The ADAM optimizer is employed in these experiments, with learning rates $1e^{-4}$and $1e^{-5}$ for the PIE and JAAD datasets, respectively, that decay by 0.1 every $10\,epochs$. Early stopping callback is also employed to prevent overfitting by monitoring validation loss improvement and halting the training if no improvement is observed for the next $7\,epochs$. The benchmark protocol is followed to address the dataset imbalance, which involves adding flipped versions of underrepresented sequences and

subsampling from the overrepresented samples to balance the number of samples[17].

The computation of segmentation maps using Segformer [141] for the whole dataset has been executed before training. Each transformer block is configured to include 4 heads, a projection dimension of 64, and a shared MLP head consisting of 64×4 and 64 MLP heads. The patch size for inputting RGB and segmentation maps is set to $(2,8,8)$. The Tubelet Projection (TP), implemented as a 3D convolutional layer, efficiently extracts features by aligning the number of filters with the projection dimension, using a kernel size matching the specified patch size, and employing strides and padding configurations. The Recurrent Projection (RP), realized through a GRU layer with the number of hidden units equivalent to the projection dimension, is crucial in capturing temporal dependencies and patterns within the input data. The MLP layers are initialized using the HeNormal initializer, including a 50% dropout rate between layers to mitigate overfitting. The entire experiment is initialized with a random seed to ensure the reproducibility of results. Through empirical analysis, it has been determined that an observation length of 0.5 seconds and a time-to-event of 2.5 seconds represents an optimal configuration. Thus, the IntentFormer is trained with the number of observation frames fixed at 15, i.e. 0.5-second observation length at a frame rate of 30fps.

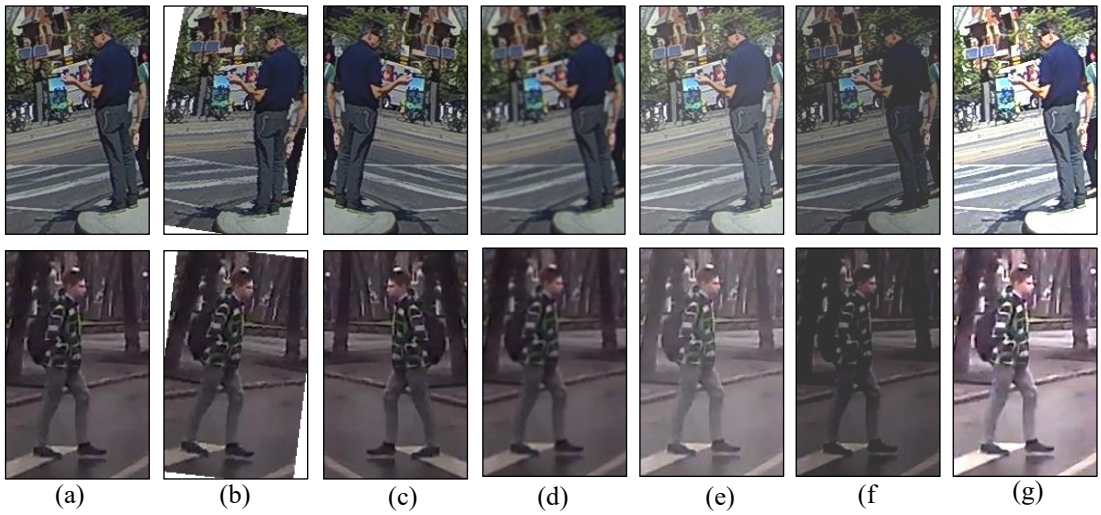


Fig. 3.13: Diverse data augmentations on pedestrian samples: (a)Original, (b)Rotation ±15°, (c) Horizontal flip, (d) Gaussian blur (0.9 kernel), (e) Intensity +50, (f) Intensity -50, (g) Intensity ×2.

Table 3.5: Evaluation of the Proposed Architecture in Comparison to Other Methods on the PIE Dataset

Methods	Year	PIE				
		Acc	AUC	F1	Prec	Rec
PIE_traj[6]	2019	0.79	-	0.87	-	-
SF-GRU[17]	2020	0.87	0.85	0.78	0.74	0.64
PCPA[96]	2021	0.87	0.86	0.77	-	-
TED[103]	2021	0.91	0.91	0.83	-	-
PG+[87]	2022	0.89	0.90	0.81	0.83	0.79
TAMFORMER[85]	2022	0.87	0.84	0.76	-	-
V-PedCross[86]	2022	0.89	0.88	0.67	0.74	0.84
MFFN[92]	2023	0.88	0.89	0.81	0.79	0.80
PedGNN[82]	2023	0.71	-	0.75	0.83	0.79
TrEP[105]	2023	0.93	0.94	0.87	0.89	0.88
PedFormer[93]	2023	0.93	0.90	0.87	0.89	0.88
VMI[91]	2023	0.92	0.91	0.87	0.86	0.88
IntentFormer(*Ours*)	-	**0.93**	**0.90**	**0.88**	**0.86**	**0.89**

Table 3.6: Evaluation of the Proposed Architecture in Comparison to Other Methods on the $JAAD_{beh}$ Dataset

Methods	Year	$JAAD_{beh}$				
		Acc	AUC	F1	Prec	Rec
PCPA[96]	2021	0.58	0.5	0.71	-	-
FFSTA[18]	2022	0.62	0.54	0.74	0.65	0.85
PG+[87]	2022	0.70	0.70	0.76	0.77	0.75
TAMFORMER[85]	2022	0.73	0.70	0.79	-	-
V-PedCross[86]	2022	0.64	0.66	0.76	0.70	0.89
STMA-GCN PedCross[101]	2023	0.69	0.58	0.80	0.68	0.97
IntentFormer(*Ours*)	-	**0.75**	**0.70**	**0.82**	**0.74**	**0.88**

Table 3.7: Evaluation of the Proposed Architecture in Comparison to Other Methods on the $JAAD_{all}$ Dataset

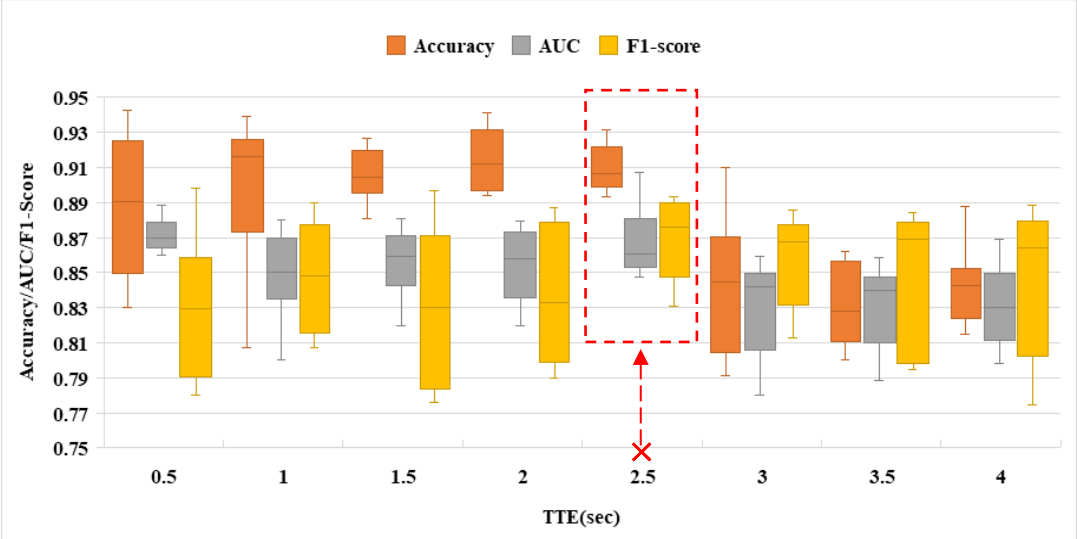
Methods	Year	$JAAD_{all}$				
		Acc	AUC	F1	Prec	Rec
SF-GRU[17]	2020	0.84	0.80	0.62	0.54	0.73
PCPA[96]	2021	0.85	0.86	0.68	-	-
FFSTA[18]	2022	0.83	0.82	0.63	0.51	0.81
PG+[87]	2022	0.86	0.88	0.65	0.58	0.75
TAMFORMER[85]	2022	0.89	0.82	0.7	-	-
V-PedCross[86]	2022	0.86	0.81	0.77	0.74	0.81
MFFN[92]	2023	0.91	0.90	0.81	0.80	0.81
PedGNN[82]	2023	0.86	-	0.77	0.96	0.86
TrEP[105]	2023	0.91	0.86	0.69	0.71	0.70
PedFormer[93]	2023	0.93	0.76	0.54	0.65	0.60
VMI[91]	2023	0.89	0.90	0.81	0.79	0.83
IntentFormer(*Ours*)	**-**	**0.92**	**0.90**	**0.83**	**0.81**	**0.85**

3.2.2.2 Datasets

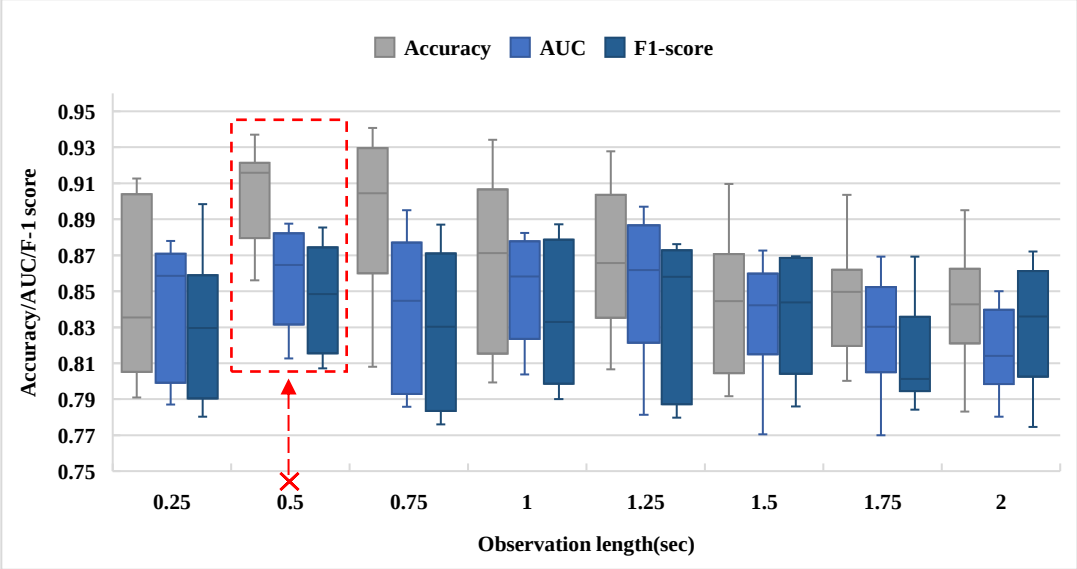
The proposed method is evaluated using two commonly used benchmark datasets, JAAD[144] and PIE[6]. The JAAD dataset consists of 346 high-resolution video clips depicting various driving scenarios in an urban setting, with pedestrians performing activities such as crossing the road, walking along the road, and waiting on the side. The dataset is split into two subsets, $JAAD_{all}$ and $JAAD_{beh}$, with the former containing 2100 visible pedestrians who are not crossing or near the end, and the latter comprising 495 crossings and 191 non-crossings. The PIE dataset offers a more extensive pedestrian data collection than JAAD, with 1,842 sections of the roadside annotated across different street structures and population densities. The dataset includes 1,842 behaviourally annotated pedestrians, with 519 crossings and 1323 non-crossings, as well as ego-vehicle speed annotations. Both datasets follow the same recommended training/validation/test split configuration for a thorough evaluation[6], [96]. Standard classification metrics such as Accuracy, AUC, F1 score, Precision, and Recall are employed to assess the proposed method's performance. Numerous pixel and geometric transformation techniques have been implemented to augment pedestrian crops to counteract overfitting. Fig. 3.13 showcases several data augmentation techniques applied to a subset of pedestrian crops from the dataset, including rotation by an angle of $\pm\theta$, horizontal flip, Gaussian blur, with a kernel σ , addition/subtraction by $\in$, and multiplication by a δ to pixel intensities.

3.2.2.3 Comparison with State-of-the-art Methods

The proposed architecture is evaluated against state-of-the-art methods as follows: PIE_traj[6], SF-GRU[17], PCPA[96], TED[103], PG+[87], TAMFORMER[85], V-PedCross[86], MFFN[92], PedGNN[82], TrEP[105], PedFormer[93], FFSTA[18], STMA-GCN PedCross[101] and VMI[91]. Table 3.5 and



(a)



(b)

Fig. 3.14: Performance evaluation of the proposed architecture across (a) Time-to-Event (TTE) and (b) Observation Length, sampled at 0.5s and 0.25s intervals, respectively.

3.7 illustrate that the proposed architecture, IntentFormer, achieves performance levels comparable to PedFormer [93] and TrEP [105]. This can be primarily attributed to integrating the Transformer encoder, a fundamental architectural component common to all these methods. Nonetheless, IntentFormer outperforms these methodologies [93], [105] on the $JAAD_{all}$ dataset, with a substantial improvement ranging from 14% to 54% in AUC, F1 score, precision, and recall. Moreover, while prior methodologies [87], [93], [105] typically confine time-to-event (TTE) predictions to 1-2 seconds,

IntentFormer attains superior results with the highest reported TTE of 2.5 seconds. On the $JAAD_{all}$ dataset, PedGNN [82] achieves the highest precision of 0.96; however, our proposed method compensates with superior accuracy and F1 score of 0.92 and 0.83, respectively, compared to 0.86 and 0.77 of PedGNN [82]. Furthermore, Table 3.6 demonstrates that IntentFormer exhibits the highest performance among methods evaluated on the $JAAD_{beh}$ dataset.

These findings indicate enhanced generalizability of the proposed IntentFormer across diverse datasets. This is attributed to an enriched understanding of pedestrian intentions facilitated by co-learning-induced shared training of the MLP layer. Incorporating Co-learning Adaptive Composite (CAC) loss has contributed to the model's generalizability by providing regularization. Moreover, deploying the Multi- Head Shared Weight Attention (MHSWA) module has effectively modelled intermodal relationships, further bolstering the model's superior performance.

3.2.2.3 Ablation Study

This section presents an ablation study to evaluate the impact of various design choices in the proposed framework. The effects of different Time to Event (TTE) and Observation Sequence Lengths (OSL) are examined, along with an analysis of modality fusion approaches, loss functions (CAC vs. BCE), and the comparison between co-learning and a vanilla architecture. Additionally, the contributions of individual modalities, their fusion order and combinations, and the effect of data augmentation are assessed. The relevance of different encoders, including the Motion Encoder, Interaction Encoder, and Visual Encoder, is also investigated. The analyses are as follows:

i. **Effect of Time-to-Event (TTE) and Observation Length**: The influence of time-to- event (TTE) and observation length on predictive performance is examined by considering various TTE points and observation lengths along the timeline of the crossing event. TTE points, ranging from 0 to 4 seconds, are sampled at intervals of 0.5 seconds, while observation lengths from 0 to 2 seconds are taken at intervals

of 0.25 seconds, as depicted in Fig. 3.14(a). TTE=0 represents the time of the crossing event. Performance improves as TTE approaches 0 seconds, indicating increased confidence in predicting crossing events. However, the variability in performance is also high, indicating that the performance at these timesteps does not consistently ensure high accuracy. For an efficient intention prediction model, the prediction confidence score should be high right before the crossing event, i.e., $TTE > 0$. At 2.5 seconds, the statistical measures of accuracy, AUC, and F1 score demonstrate high and relatively stable values with varying observation lengths, as depicted in Fig. 3.14(a). Beyond 2.5 seconds, there is a notable decline in overall performance, with accuracy decreasing by up to 6.5%.

Fig. 3.14(b) demonstrated that the optimal performance is observed within the 0.5-1.25 seconds observation length range, exhibiting minimal variation with changing TTE. The performance metrics peak at an observation length of 0.5 seconds and show minimal fluctuation. Hence, this observation length is ideal for achieving optimal performance, as the accuracy, AUC, and F1 scores remain consistently high within this range. Moreover, accuracy, the area under the curve (AUC), and the F1 score show a modest gain up to an observation length of 1.25 seconds since such a prolonged duration leads to higher information acquisition. However, beyond that, the performance drops as prolonged observation periods may contain irrelevant details about the scene dynamics that can undermine the prediction accuracy. Larger observation lengths signify a more significant number of frames required for analysing crossing intention, resulting in high computational

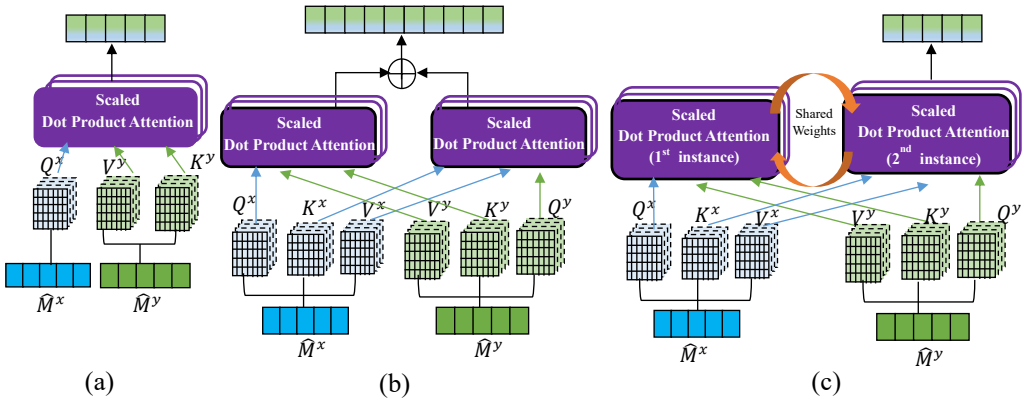


Fig. 3.15: Illustration of three Multi-Head Attention types: (a) Cross-Modal Attention (MHCMA), (b) Multimodal Attention (MHMMA), and (c) Shared-Weights Attention (MHSWA).

demands. Therefore, an efficient intention prediction model should make confident predictions with the least possible observation length. The proposed model demonstrates robustness by achieving optimal performance with an observation length of just 0.5 seconds, thereby minimizing computational demands and ensuring efficient prediction. These results highlight the efficiency of the proposed architecture in predicting crossing events even with fewer frames and high TTE (upto 3.5 secs), with performance metrics dropping by no more than 12-14%. This starkly contrasts the SF-GRU [17] method, which exhibited a substantial decline in performance metrics, reaching up to 33% when TTE is increased beyond 3 seconds. Furthermore, the PG+ [87] approach restricts TTE to 1-2 seconds, limiting its suitability for real-time scenarios. Notably, the proposed approach achieves superior accuracy compared to VMI [91] and comparable metrics, with the highest reported TTE to date while maintaining a significantly reduced computational footprint and inference time.

ii. **Analysis of modality fusion approaches:** In the field of multimodal deep learning, multi-head cross-modal attention (MHCMA) and multi-head multimodal attention (MHMMA) based fusion techniques have emerged as popular mid-level transformer-based approaches [143]. These attention mechanisms have unique characteristics and functionalities that may cater to specific application domains. The proposed model employs a multi-head shared weight attention (MHSWA) mechanism to facilitate the synergistic fusion of information across distinct modalities. The shared weight attribute capitalizes on the synergy of attention

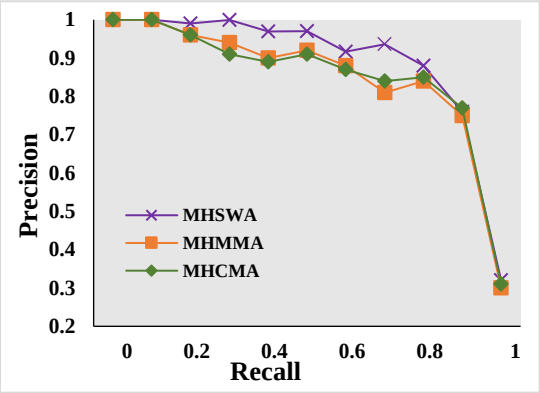


Fig. 3.16: Precision-Recall curves for different types of modality fusion attention mechanisms

weights from various heads to exploit cross-modal correlations efficiently. It comprises two scaled dot product attention instances tailored to specific modalities. The first instance, trained on the initial modality data, captures intricate interdependencies among elements. Subsequently, the second instance, initialized with learned weights from the first modality, refines its training on the subsequent modality, fostering a sequential, contextual understanding enriched by prior knowledge.

The distinct design characteristics of these three attention-based fusion strategies are elucidated in Fig. 3.15. An ablation study is conducted using the Precision-Recall curve, as depicted in Fig. 3.16, to assess the impact of the various fusion strategies on performance. The study's results revealed that the MHSWA method's precision-recall curve is notably closer to the ideal curve compared to the other two approaches. The varying behaviour of attention coefficients across the different stages of the proposed shared weight attention model is illustrated in Fig. 3.17. At stage *I*, Fig. 3.17(a), a high range of attention coefficients indicates that the model assigns varying levels of importance to different embeddings within the RGB data. This stage focuses on capturing fine-grained details and relationships specific to the RGB input, as it is the primary modality. A slight decrease in the attention coefficient range at this stage *II* is observed in Fig. 3.17(b), suggesting that the model focuses on commonalities and interactions between RGB and segmentation embeddings. The shared weight attention mechanism allows the model to emphasize cross-modal correlations and jointly process features from both modalities. In the last stage *III*, Fig. 3.17(c) highlights attention coefficients

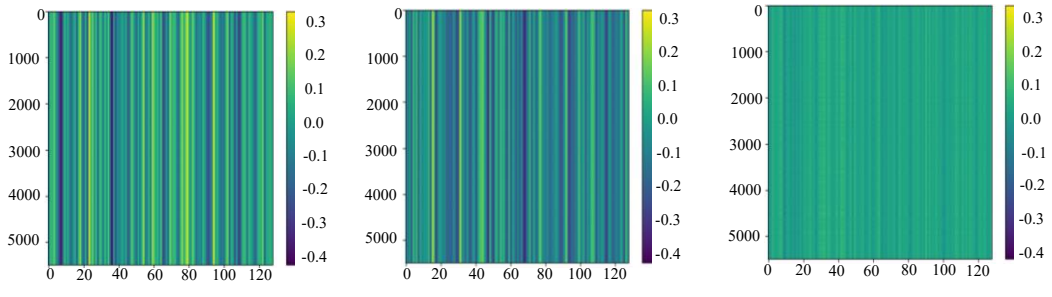


Fig. 3.17: Evolution of Attention Coefficients across Sequential Stages in the Proposed Shared Weight Attention Model

distinguished by a much shorter range, indicating that the model is assigning more consistent attention across embeddings from different modalities. It is inferred that the model is integrating information from previous stages (RGB and segmentation) and trajectory more uniformly. The change in attention behaviour from varying ranges to more uniform attention signifies that the model progressively shifts its emphasis from capturing modality-specific details to integrating multimodal information for decision-making. Hence, the observed behaviour aligns with the objective of multimodal learning: to learn robust representations that capture inter-modal relationships and produce consistent outputs despite the varied nature of the input sources.

In Fig. 3.18. Guided Integrated Gradient (IG) [145] Visualizations corresponding to individual attention map heads are presented for RGB sequences. It highlights the areas where Multi-head Shared Weights Attention (MHSWA) mechanisms positively influence the model's classification decision. This configuration comprises a total of four discrete attention map heads. The first attention map head primarily emphasizes the outline or shape of the target pedestrian. The second and third attention maps appear to capture details related to the target's immediate surroundings and the pedestrian's dynamic variations across the sequence of frames. The fourth attention map identifies contours and distinct patterns within the cropped image.

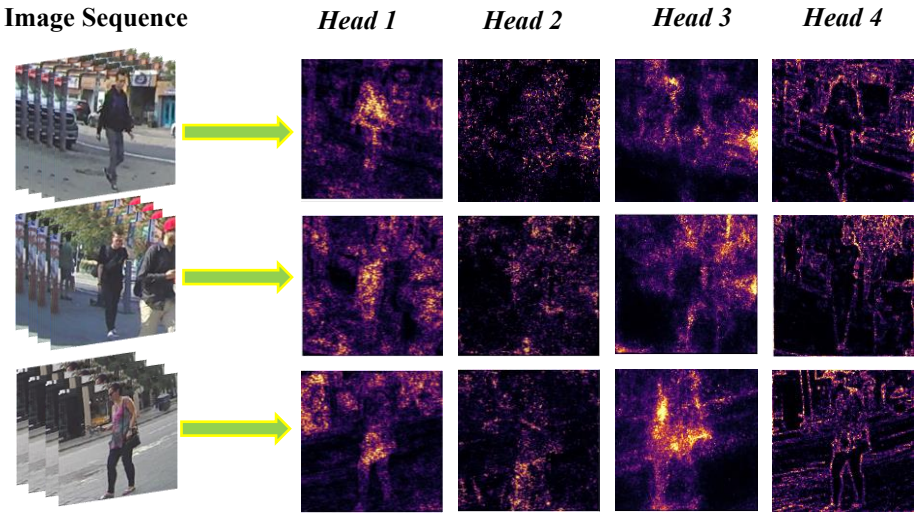
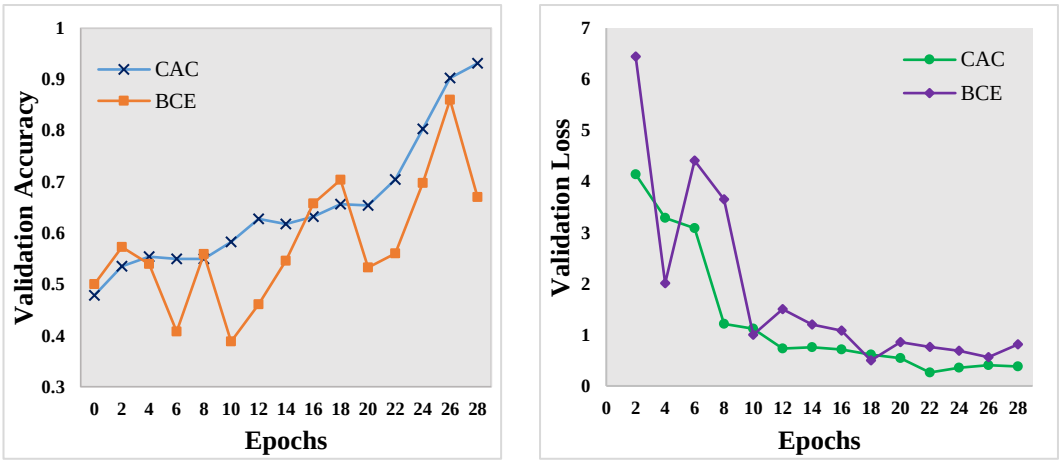


Fig. 3.18: Guided Integrated Gradient [145] Visualisation of IntentFormer

iii. **CAC vs BCE**: This section explores the impact of the proposed Co-learning Adaptive Composite (CAC) loss function on validation performance and the dynamic relationship between adaptive loss weights and training progress. Fig. 3.19(a) presents the validation accuracy curves for models trained using the standard Binary Cross-Entropy (BCE) and the proposed CAC loss function. The CAC loss function notably enhances the stability of validation accuracy throughout the training phase, reducing fluctuations compared to BCE and achieving superior validation accuracy. In Fig. 3.19(b), the validation loss curves show that BCE induces more frequent fluctuations than the CAC loss, leading to difficulties in convergence. In contrast, the CAC loss function achieves the lowest validation loss. These results indicate that the CAC loss function effectively mitigates overfitting



(a) (b)

Fig. 3.19: Effect of CAC and BCE loss functions on (a) validation accuracy and (b) validation loss curves.

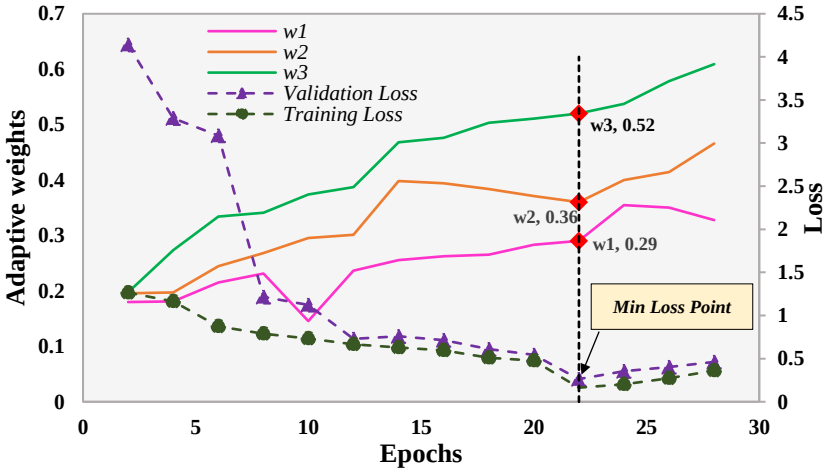


Fig. 3.20: Adaptive loss weights and training dynamics

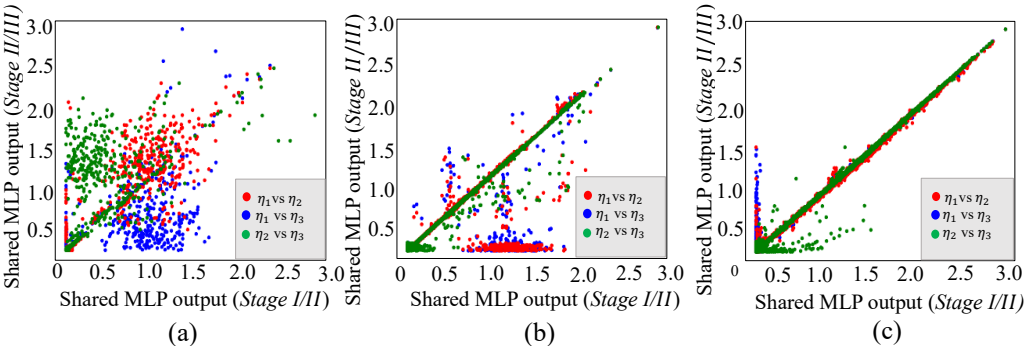


Fig. 3.21: Learned feature representations from the shared MLP layer in the co-learning architecture, across epochs (a) 3, (b) 15, and (c) 22

Fig. 3.22: Qualitative predictions on PIE/JAAD where IntentFormer correctly classifies intent, unlike the vanilla transformer. Red: non-crossing, Green: crossing.

during training, thereby enhancing the generalization capacity of the proposed architecture.

Fig. 3.20 illustrates the evolution of the training and validation loss alongside the changes in adaptive loss weights (w_1, w_2, w_3), throughout training epochs. The adaptive loss weights, initialised randomly, exhibit dynamic adjustments in response to the changing training landscape. Specifically, weights w_2 and w_3, exhibit a gradual and consistent increment throughout epochs culminating at respective maximal values of 0.46 and 0.60. Contrastingly, weight w_1 display more intricate behaviour, initially decreasing, followed by a gradual and consistent increase over epochs, reaching a maximum value of 0.35. This suggests that the w_1 loss term contributes significantly less to the overall loss as the model refines its representations. The vertical line denotes the epoch at which the minimum loss is attained, providing insight into the optimal point (w_1: 0.29, w_2: 0.36 and w_3: 0.52) in the training process. This allows us to fine-tune training

strategies and highlights the potential for adaptive loss weighting to enhance model training efficiency and performance.

iv. **Co-learning v/s Vanilla transformer architecture**: This section discusses the adaptive learning process of the proposed co-learning architecture that leverages shared MLP heads. The pairwise scatter plots in Fig. 21 illustrate a notable evolution in the alignment of learned representations across the three epochs (Epochs 3, 15, and 22) for the Co-learning multimodal architecture employing shared MLP heads. As training progresses, the shared MLP layers output at three stages increasingly converges along a linear trajectory. This highlights that the architecture effectively captures the shared semantics across modalities, allowing for improved feature extraction and cross-modal interaction. Furthermore, the dynamic alignment of representations over epochs suggests that the shared MLP layer effectively captures cross-modal relationships, allowing different modalities to learn and adapt coherently.

Furthermore, a comparative analysis of the proposed architecture with a vanilla transformer model without a shared MLP head is also carried out. The term "Vanilla transformer" here denotes a model variant in which the shared MLP in the co-learning architecture is substituted with three independent trainable MLPs, each assigned to a specific modality (RGB, segmentation, and trajectory). This modification facilitates a comparative analysis between the co-learning architecture

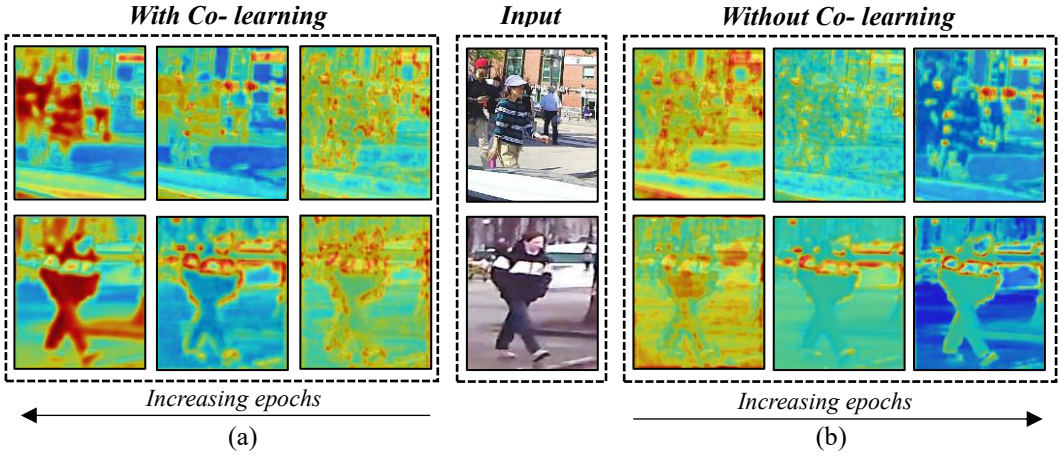


Fig. 3.23: Grad-CAM visualization of IntentFormer at 3, 15, and 22 epochs: (a) With co-learning (right to left), (b) Without co-learning (left to right).

utilizing shared MLPs and an alternative configuration employing non-shared, individual MLPs for each modality. The goal is to evaluate the influence of shared semantics across modalities on the learning dynamics. Qualitative results for the few samples from the JAAD/PIE dataset are presented in Fig. 3.22. Notably, Fig. 3.22 (d)-(e) depicts instances of no eye contact between the pedestrian and the camera, resulting in uncertainty regarding the direction in which the pedestrian would move. For instance, Fig. 3.22 (e) shows a pedestrian looking at a phone, making it difficult for the model to interpret intention from visual appearance cues such as gaze.

Conversely, Fig. 3.22 (f)-(g) illustrated examples of poor illuminations or reflections that tampered with the supposed appearance cues. Finally, Fig. 3.22 (h)-(i) showcases examples where the pedestrian sample is too small. The vanilla architecture does not perform well in these hard classification samples. However, the correct predictions by the proposed model can be attributed to the fact that it caters to the cross-modal relationships among visual appearance, segmentation maps and trajectory with consistent learned representations. Thus, even if one representation fails to capture the pedestrian's intention correctly, its relationship with the other two modalities strives to decipher it correctly, albeit with less confidence.

The Grad-CAM visualizations for the IntentFormer with and without the co-learning module (Vanilla transformer) are depicted in Fig. 3.23(a) and (b), respectively. Analysis of Fig. 3.23(a) reveals a progressive refinement in the Grad-CAM attention maps in the co-learning environment as the number of training epochs increases. Initially, at epoch 3, the Grad-CAM outputs are dispersed across the input image, lacking specific focus on any element. However, as training progresses, the importance weights become increasingly localized to image regions pertinent to classifying the pedestrian's intention. The attention maps become more precise, effectively highlighting the silhouette of the target pedestrian. Additionally, with the incorporation of segmentation maps and trajectory data in the second and third stages, respectively, it is observed that co-pedestrians and certain scene elements, such as road boundaries, also receive higher weightage as observed for models trained for epochs 15 and 22. This indicates an enhanced understanding of

the context and contributing factors to pedestrian intention prediction. Conversely, in Fig. 3.23(b), where IntentFormer is trained without the co-learning module, the pedestrian torso and some scene elements sparsely receive higher weights by the last training epoch. The input pixels are not highlighted precisely or comprehensively as in the co-learning training mode. This less effective localization of important features reduces the ability to identify the most relevant features for intention prediction.

v. **Impact of individual modalities, their combinations and fusion order**: This section investigates the impact of different modalities and fusion order permutations on the overall performance of pedestrian intention prediction. In our recent work[91], pedestrian appearance, scene context, pose, trajectory, and ego-vehicle speed were utilised for pedestrian intention prediction. The analysis demonstrated that context features achieved the highest performance metrics, followed by appearance features. In contrast, pose features contributed the least when utilized as graph node features to model the temporal relationships of pedestrian interactions. Based on these findings, the proposed work incorporates only RGB crops, trajectory, and segmentation maps for context as the primary modalities for the proposed intention prediction model. This approach minimizes

Table 3.8: Performance comparison of the IntentFormer model with different modalities, their combinations, and the order of fusion

Modalities	Accuracy		
	PIE	**$JAAD_{beh}$**	**$JAAD_{all}$**
T	0.56	0.41	0.55
R	0.59	0.45	0.60
S	0.43	0.39	0.40
T+R	0.63	0.52	0.64
T+S	0.58	0.48	0.61
R+S	0.66	0.54	0.69
T+S+R	0.78	0.68	0.82
T+R+S	0.76	0.67	0.80
S+T+R	0.88	0.69	0.86
S+R+T	0.89	0.70	0.88
R+T+S	0.90	0.69	0.85
R+S+T	**0.93**	**0.75**	**0.92**

**R: RGB Images, S=Segmentation Maps, T: Trajectory*

the additional memory footprint associated with pose features without significantly impacting the model's overall performance. It can be observed from Table 3.8 that individual modality (R: RGB pedestrian crops, S: Segmentation maps and T: Trajectory) achieve the lowest accuracy. When assessing single modality performance, input feed is given only through the first encoder stage; no other feed is given through subsequent encoder stages. In subsequent ablations involving combinations of two modalities, input feed is given through the first and second encoder stages. Combining these modalities leads to substantial performance improvements. For instance, combining T+R increases accuracy by 12.5% on PIE, T+S increases accuracy by 3.6%, and R+S increases accuracy by 16.4%, considering accuracy with only T as baseline. The highest accuracy is obtained with the combination R+S+T, resulting in a 66.1% increase in PIE, a 66.7% increase in $JAAD_{beh}$, and a 53.3% increase in $JAAD_{all}$ over baseline, demonstrating the effectiveness of integrating these modalities. In the case of a single modality in any of the encoder stages with no other modality feed, the modality in any of the encoder stages with no other modality feed, the MHSWA, designed for the fusion of two diverse modalities within the encoder, operates as standard MHA.

The experiments with different orders of fusion, as reported in Table 3.8, highlight that a noticeable dip in performance is observed when features such as

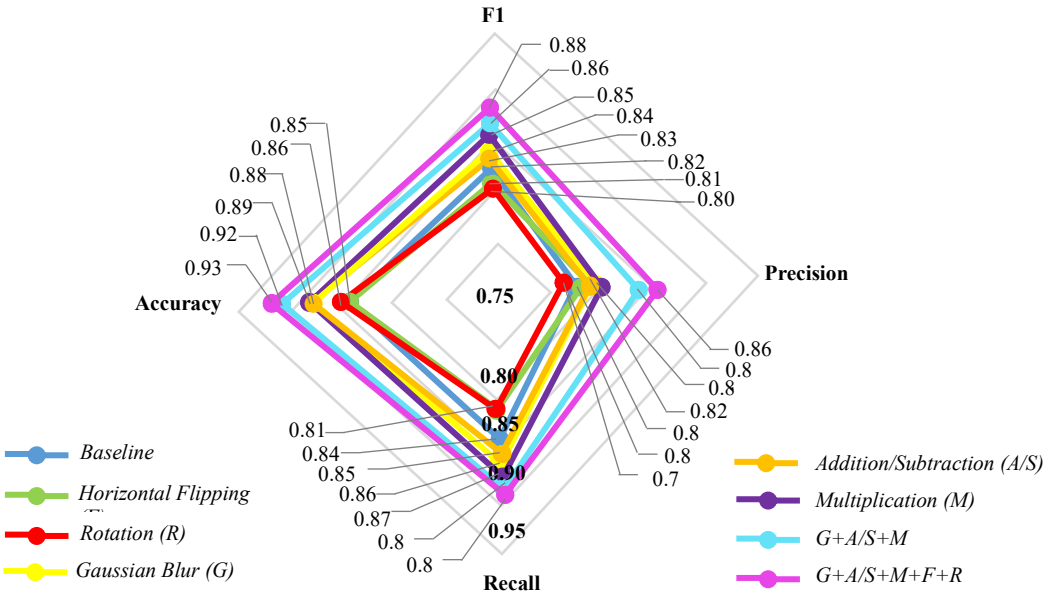


Fig. 3.24: Visual comparison of IntentFormer trained with different augmentations.

RGB images and segmentation maps are integrated at later stages of the network. By selecting the correct permutation by feeding trajectory at the last stage of the network, the accuracy performance improves by up to 9% on PIE, 5% on $JAAD_{beh,}$ and more than 8% on $JAAD_{all}$. This observation can be attributed to the proposed architecture's ability to leverage visual features in the earlier network stages effectively. The subsequent integration of dynamic features like trajectory coordinates at later stages optimally takes advantage of the enriched contextual understanding constructed by prior modalities. By aligning the integration order with the intrinsic complexity of features, the architecture maximizes the information captured by each modality. These findings highlight the pivotal role of the chosen sequence of feature integration in enhancing prediction accuracy.

Table 3.9: Quantitative Evaluation On The PIE/JAAD Dataset

Ablations	Model Variants							Accuracy		
	MLP Heads		Multi-Head Attention			Loss				
	MLP	MLP-shared	MHCMA	MHMMA	MHSWA	BCE	CAC	PIE	$JAAD_{beh}$	$JAAD_{all}$
1	✓	×	✓	×	×	✓	×	0.89	0.69	0.88
2	✓	×	×	✓	×	✓	×	0.89	0.70	0.87
3	✓	×	×	×	✓	✓	×	0.91	0.69	0.91
4	×	✓	✓	×	×	✓	×	0.90	0.70	0.90
5	×	✓	×	✓	×	✓	×	0.91	0.71	0.91
6	×	✓	×	×	✓	✓	×	0.90	0.70	0.90
7	✓	×	✓	×	×	×	✓	0.86	0.69	0.89
8	✓	×	×	✓	×	×	✓	0.87	0.65	0.88
9	✓	×	×	×	✓	×	✓	0.88	0.63	0.87
10	×	✓	✓	×	×	×	✓	0.91	0.70	0.88
11	×	✓	×	✓	×	×	✓	0.92	0.71	0.89
12	×	✓	×	×	✓	×	✓	**0.93**	**0.75**	**0.92**

Table 3.10: Comparison of IntentFormer with state-of-the-art models on the PIE, $JAAD_{beh}$, and $JAAD_{all}$ datasets, highlighting memory footprint, inference time, and highest achieved accuracy.

Model	Size (MB)	Inference time(ms)	Accuracy (PIE)	Accuracy ($JAAD_{beh}$)	Accuracy ($JAAD_{all}$)
PCPA[96]	118.8	38.6	86	50	70
FFSTA[18]	374.2	70.83	-	62	83
PG+[87]	0.28	5.47	89	70	86
TED[103]	12.8	2.76	91	-	-
V-PedCross[86]	4.8	-	89	64	86
PedGNN[82]	0.027	0.58	70.52	-	86.22
VMI[91]	19.07	11.03	92	-	89
IntentFormer	**2.13**	**3.8**	**93**	**75**	**92**

Table 3.11: Model Architecture and Hyperparameter Configuration

Modules/Layers/Encoders	Trainable Parameters			Hyperparameters
	Proposed IntentFormer		Vanilla Transformer	
	Non-Shared	Shared	Non-shared	
Tubelet/Recurrent Projection (TP/RP)	38K(*total*)	-	38K(*total*)	***TP**: Conv 3D: Filters-64, Kernel Size-(2,8,8)* ***RP:** GRU: Hidden units-64*
Positional Encoder_TP	351K	-	351K	*Embedding Layer Output Dimension- 64*
Positional Encoder_RP	896	-	896	*Embedding Layer Output Dimension- 64*
MHSA/MHSWA	16.6K		16.6K	*No. of heads (4), Size of each attention head (64), Dropout-50%*
PCP	82K	-	82K	*Conv 1D:* 1×1
Shared MLP	-	33K	-	*Two sequential MLPS with 64x4 and 64 neurons, Dropout-50%*
Layer Normalization (LN)	128			-
Classification Head	130	-	130	*Layer Normalization, GAP, Dropout-50%, MLP with 2 neurons*
$\boldsymbol{TE_I + TE_{II} + TE_{III}}$	132K	33K	231K	-
Total	522K	33K	621K	-

vi. **Effect of Data Augmentation**: Fig. 3.24 illustrates the impact of various augmentation techniques on the performance of our pedestrian intention prediction model. Among the techniques evaluated, horizontal flipping (F) and rotation (R) provided minimal enhancements compared to the baseline without augmentation. Additionally, Gaussian blur (G), addition/subtraction (A/S), and multiplication (M) demonstrated notable improvements, increasing overall performance metrics by 2.71%, 2.01%, and 3.63%, respectively, relative to the baseline. The combination of Gaussian blur, addition/subtraction, and multiplication (G + A/S + M) resulted in substantial enhancements, boosting accuracy by 8.24%, F1 score by 4.88%, precision by 5%, and recall by 4.76%. The inclusion of all five augmentations (G + A/S + M + F + R) yielded the highest overall improvements, with increases in accuracy by 9.41%, F1 score by 7.32%, precision by 7.50%, and recall by 5.95%.

These results demonstrate that complex augmentations such as Gaussian blur, addition/subtraction, and multiplication significantly enhance the model's ability to predict pedestrian intentions. Although primary augmentations like horizontal flipping and rotation are insufficient to capture the complexities of pedestrian movements and interactions, the synergistic effect observed from combining multiple augmentations highlights that diverse and comprehensive

augmentations can collectively enhance the model's robustness and accuracy in pedestrian intention prediction tasks.

vii. **Quantitative Analysis**: The analysis of model ablations in Table 3.9 reveals a notable 3-4% increase in accuracy for shared MLP configurations compared to their non-shared MLP counterparts. The multi-head attention configurations (MHCMA, MHMMA, MHSWA) demonstrate a systematic rise in accuracy across all datasets, with MHCMA exhibiting the lowest accuracy and the proposed MHSWA achieving the highest levels. This validates the impact of shared weight attention among diverse modalities (RGB images, Segmentation maps and trajectory) in a co-learning framework. The proposed Co-learning Adaptive Composite (CAC) loss also shows comparable performance to the widely used Binary Cross-Entropy (BCE) loss. It also introduces a significant improvement in regularization, leading to reduced fluctuations in validation accuracy. These collective findings underscore the effectiveness and efficiency of the proposed IntentFormer architecture in capturing intricate relationships among modalities for robust pedestrian intention prediction.

The IntentFormer model achieves superior accuracy of 93% on the PIE dataset, 75% on the $JAAD_{beh}$ dataset, and 92% on the $JAAD_{all}$ dataset while maintaining a competitive memory footprint of 2.13 MB and an inference time of 3.8 ms, as shown in Table 3.10. It consists of 555k parameters, showcasing a substantial decrease in parameters by approximately 11% compared to the vanilla transformer with 621K parameters, suggesting more parameter-efficient learning (Table 3.11). This parameter reduction also results in a 10% decrease in memory footprint. One of the key reasons behind the competitive memory footprint achieved for the proposed architecture is the co-learning module and the Multi-head Shared Weights Attention devised for model training that keeps the trainable parameters limited in numbers. Although the memory footprint is higher than that of PedGNN [88], IntentFormer offers a significant accuracy improvement, with a 27.62% increase on the PIE dataset and a 3.22% increase on the $JAAD_{all}$ dataset compared to PedGNN [82]. Thus, despite PedGNN's minimal memory footprint of 0.027 MB, it fails to adequately address the complex dynamics of real-time scenes compared

to the proposed IntentFormer. These results highlight the model's efficiency and effectiveness, making it well-suited for real-time applications in autonomous driving.

3.3 Conclusion and Future Scope

This chapter presented two successive approaches to pedestrian crossing intention prediction. The first work introduced a multimodal framework employing attention mechanisms across spatial, channel, and temporal dimensions, along with a novel multi-head-attention adjacency-matrix-based GCN ($MHA - AdjMat\ GCN$) to fuse visual, motion, and interaction features. This model demonstrated superior early intent prediction on the JAAD and PIE benchmarks (accurately anticipating crossing up to 2.5 s before the event).

Building on these insights, the second work proposed '*IntentFormer*', a multimodal transformer-based architecture. It integrates RGB images, semantic segmentation maps, and trajectory features through three co-trained transformer encoders. Each encoder uses a multi-head shared-weight self-attention mechanism, and the system is trained with a shared-MLP output head under a novel Co-learning Adaptive Composite (CAC) loss. This design excels with very short observation windows (0.5–1.25 s) and maintains strong prediction accuracy even up to 3.5 s before crossing, outperforming prior state-of-the-art methods.

Taken together, these sequential contributions illustrate that shifting from GCN-based feature fusion to transformer-based co-learning improves temporal modelling and cross-modal integration. To further enhance real-world robustness and applicability, future research should focus on modelling uncertainty and unpredictable behaviour and optimizing for real-time adaptation and generalization, enabling models to remain reliable across diverse urban scenarios and unseen conditions.

CHAPTER 4

LONG TERM INTENTION PREDICTION

Long-term pedestrian intention prediction is a critical task in fields such as autonomous driving, robotics, and smart city infrastructure. Accurately forecasting the future movements of pedestrians over extended periods involves significant challenges due to the complex, non-linear nature of human motion, sudden changes in behaviour, and the influence of environmental factors. These challenges are further compounded by the difficulty of capturing and modelling the wide range of contextual information that affects pedestrian decisions. Traditional trajectory prediction methods often struggle to account for these dynamic interactions and the inherent uncertainty in long-term predictions. This chapter addresses the challenges of long-term pedestrian intention prediction, focusing on the complexities of non-linear motion, sudden behavioral changes, and environmental interactions. A novel framework is introduced to enhance trajectory forecasting, incorporating mechanisms for adaptive learning, contextual integration, and uncertainty-aware prediction.

4.1 Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration

Despite the rapid advancement of highly automated vehicles poised to mitigate accidents caused by human errors, understanding the behaviours of road users, especially vulnerable pedestrians, remains a significant challenge. The evolution of pedestrian trajectory prediction, transitioning from early motion models to recent deep learning approaches, has highlighted persistent challenges in accurately predicting future trajectories, particularly in complex scenarios. To address this, this paper presents a Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration (PCTP-AGFL). The proposed method incorporates a dynamic progressive generator (DPG) comprising multiple LSTM layers that adapt progressively to pedestrian motion pattern complexities. The DPG is trained using a learned scheduled sampling strategy implemented through an Adaptive Gating Mechanism (AGM), allowing dynamic switching between teacher forcing and normal

mode. This is augmented with an Encoder-Decoder Contextual Attention (EDCA) module to enhance contextual awareness. A novel Adaptive Fuzzified Discriminator (AFD) is also introduced to enhance the model's capability to handle ambiguous trajectories. Experimental results on JAAD/PIE and ETH/UCY datasets demonstrate the method's superiority over baselines and state-of-the-art approaches. Furthermore, a comprehensive ablation study is carried out to tune the progression parameters, training strategy, and the type of classifier logic in the discriminator.

4.1.1 Proposed Approach

At time step n, the observed trajectory of a pedestrian in the last k timesteps is represented as $P_n = \{p_{n-k+1}, p_{n-k+2}, \dots, p_n\}$ where p_n includes its top-left (x_{tl}, y_{tl}) and bottom-right (x_{br}, y_{br}) bounding box coordinates. The primary objective is to predict its ν future coordinate positions $Q_n = \{y_{n+1}, y_{n+2}, \dots, y_{n+\nu}\}$. To address this challenge, a novel Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration is proposed as illustrated in Fig. 4.1. This architectural framework employs a learned scheduled sampling training strategy to provide essential guidance for pedestrian trajectory prediction. The Adaptive Fuzzified

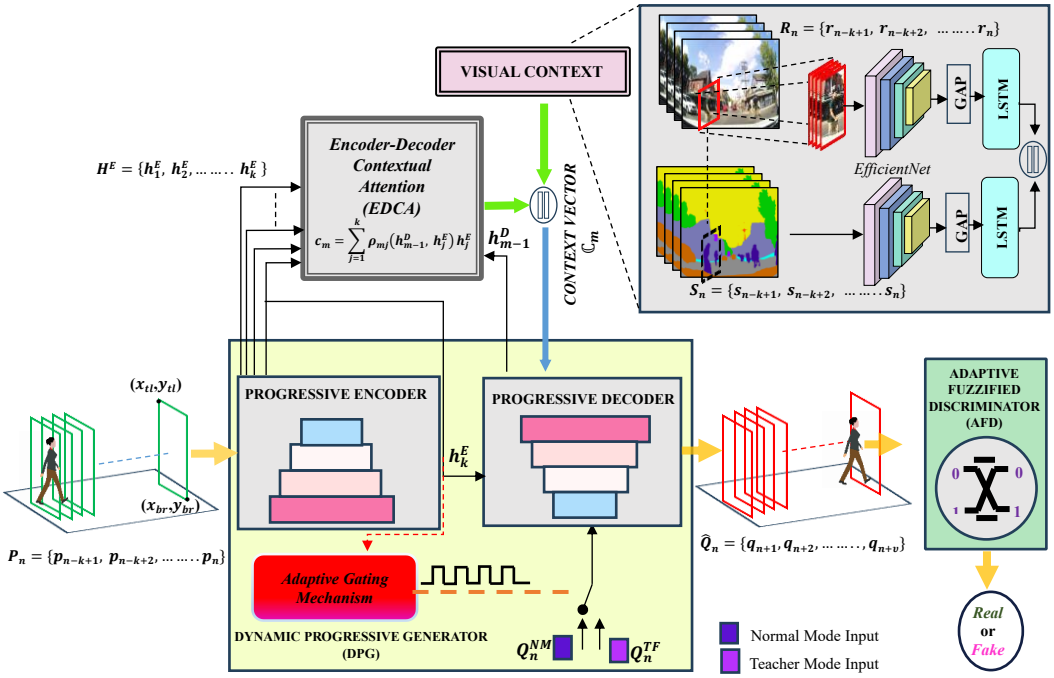


Fig. 4.1: Proposed trajectory prediction architecture.

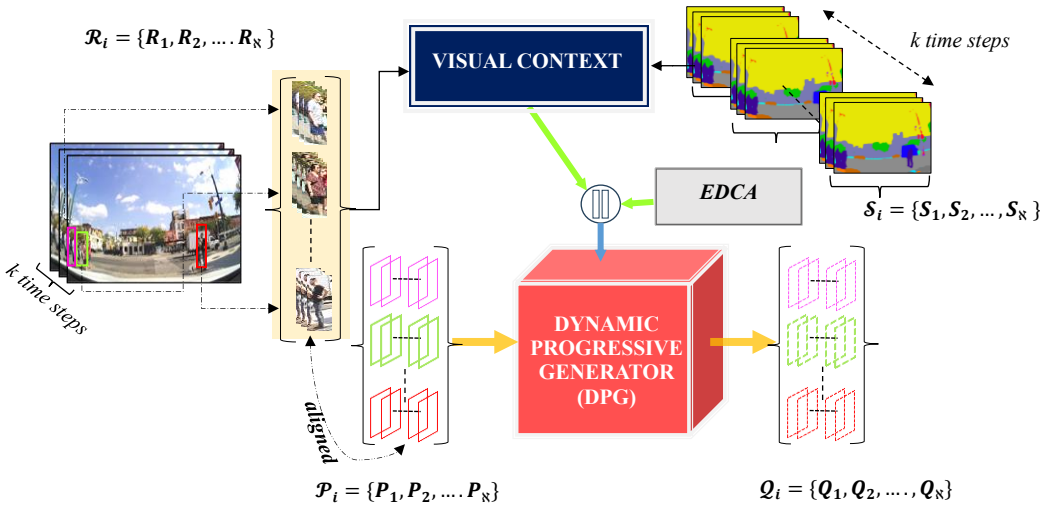


Fig. 4.2: Illustration of features alignment in multi-pedestrian frames for $\aleph$ pedestrians in a frame

Discriminator (AFD) enhances its ability to discriminate between real and fake trajectories by providing1 increased levels of nuanced confidence in its classifications. Moreover, the model leverages supplementary features derived from RGB images i.e. RGB crops $R_n = \{r_{n-k+1}, r_{n-k+2}, \ldots, r_n\}$ and segmentation maps $S_n = \{s_{n-k+1}, s_{n-k+2}, \ldots, s_n\}$ to capture contextual information, thereby enriching the understanding of the environment, which plays a pivotal role in improving the quality of trajectory predictions. Fig. 4.2 demonstrates how the model aligns trajectories with the visual context, leading to a richer representation of a pedestrian sample in a multi-pedestrian scenario. For all the pedestrians in a single frame, same segmentation maps

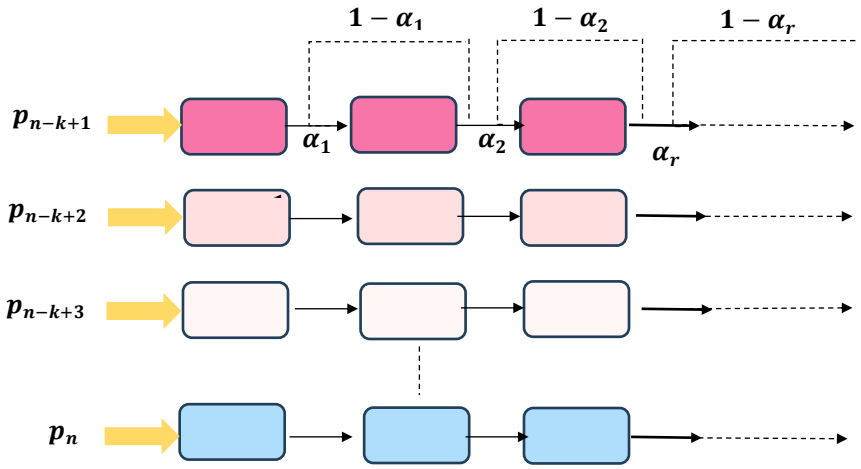


Fig. 4.3: Progressive encoder architecture

are used. Thus, segmentation map $\mathcal{S}_i$ for each i^{th} pedestrian in a frame is $S_1 = S_2 = S_3 \ldots\ldots = S_{\aleph}$.

4.1.1.1 Dynamic Progressive Generator (DPG)

The proposed dynamic progressive generator employs a hierarchical structure consisting of multiple layers of Long Short-Term Memory (LSTM) units, with the number of LSTM layers l, increasing progressively from 1 to r, to adapt to complex input patterns. A parameter governs the control over the progression α_l which is initialized and updated during training at each iteration within the range 0 to 1 as illustrated in Fig 4.3. This progressive growth strategy ensures that the model dynamically adjusts its depth to effectively capture the intricacies of the input data, effectively balancing model complexity with performance. A similar progressive layer is mirrored on the decoder side, preserving architectural design symmetry. The encoder captures the target pedestrian's motion pattern P_n as a latent vector using a recurrent cell i.e. LSTM. The new hidden state h^E_{k+1} at $(n-k+1)_{th}$ timestep is updated through an LSTM Cell given by Eqn (4.1).

$$h^E_{k+1} = LSTM_E(p_{n-k}, h^E_k) \tag{4.1}$$

In training an RNN encoder-decoder for sequence-to-sequence prediction tasks like machine translation, etc., different training strategies impact the learning process.

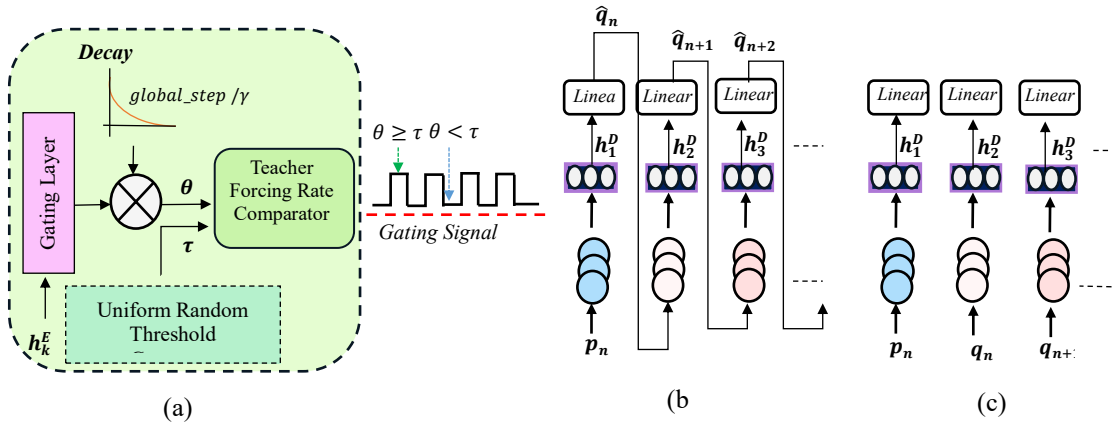


Fig. 4.4: (a) Adaptive Gating Mechanism (AGM) for learned scheduled sampling; (b) Normal mode ($\theta < \tau$),

Normal mode involves processing input sequences step by step, potentially leading to error propagation. The teacher-forcing strategy, on the other hand, utilizes ground truth target outputs from the training dataset as inputs during the training process to mitigate error accumulation over time. Chen et al.[146] successfully employed teacher forcing to address the speaker permutation problem, enhancing speaker embedding representation. Huang et al.[147] proposed teacher-forcing training strategy for image captioning. However, when using teacher forcing training strategy, the model might become overly reliant on the ground truth inputs and may not generalize well to unseen data during inference. To address this, another seminal work[148] employed a scheduled sampling strategy that gradually transfers the training phase from a teacher-forced manner to a normal training mode for video captioning. Leveraging these advancements of teacher-forcing framework in sequence-to-sequence modeling tasks, the proposed PCTP-AGFL presents a learned scheduled sampling strategy via Adaptive Gating Mechanism (AGM) as illustrated in Fig. 4.4. It allows dynamic switching between teacher forcing and normal mode training strategy, striking a balance between accuracy and generalization. This mitigates potential biases and errors associated with static training strategies. This mechanism utilizes the encoder's hidden state to decide whether to use teacher-forced input or the previous prediction as the input to the decoder at each timestep using a gating layer 'gL' defined as a fully connected layer with sigmoid activation. The gating factor ξ is computed as:

$$\xi = gL(h_k^E) \tag{4.2}$$

It defines a schedule that determines how the model switches between teacher forcing and using its predictions. The $intial\_teacher\_force\_rate\ \theta_O$ within a schedule typically starts with a high probability ($\sim$1) of using teacher forcing and gradually decreases this probability as training proceeds. The current teacher-forcing rate θ is defined as follows:

$$\theta = \theta_O\ * \xi * min(1.0, global\_step\ /\gamma) \tag{4.3}$$

Where $global\_step$ is the total count of the training steps executed, and γ is the teacher forcing decay rate. As training progresses, teacher forcing gradually decreases, and the probability of using the model's predictions increases. A randomly sampled number τ is generated between 0 and 1 as the threshold for deciding whether to use teacher-forcing or the model's predictions. It introduces stochasticity into the decision process and encourages the model to explore different behaviours during training. In the case of deterministic predictions, τ is set to 0.5. Finally, τ is compared with θ to assign the input to the decoder. During testing, the decoder is set to utilize its predictions due to the non-availability of ground truth trajectories.

Furthermore, an Encoder-Decoder Contextual Attention (EDCA) module is employed to regulate the attention allocation of the decoder at each time step towards the encoder's hidden states as shown in Fig. 4.5. It is mathematically denoted as

$$c_m = \sum_{j=1}^{k} \rho_{mj}\left(h_{m-1}^{D}, h_j^{E}\right) h_j^{E} \tag{4.4}$$

where m represents the current decoder timestep, h_j^E corresponds to the encoder's hidden state at the j^{th} timestep with j ranging from 1 to k; h_{m-1}^D denotes the input hidden state of the decoder at the $(m-1)^{th}$ timestep and ρ_{mj} is the attention coefficient. This coefficient assesses the influence and significance of prior encoder states on the current state of the decoder. The final context vector C_m is acquired through the concatenation of visual context features obtained from RGB images with Regions of Interest (ROIs) encompassing the pedestrian and full scene segmentation

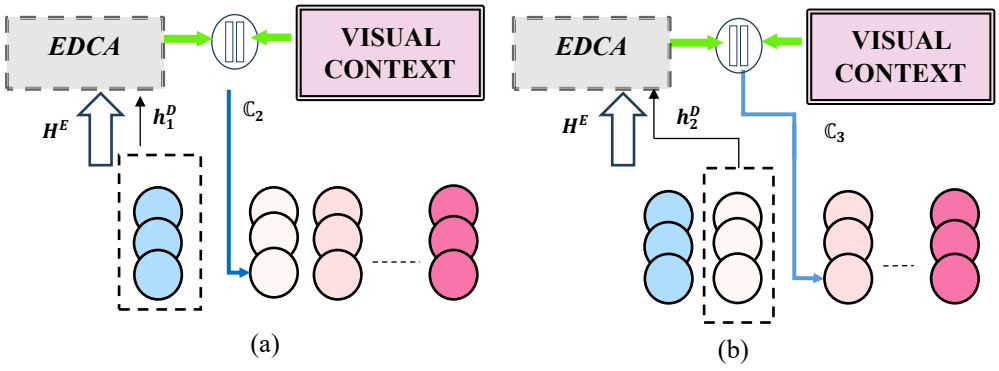


Fig. 4.5: Encoder-Decoder Contextual Attention at decoder timesteps, (a) $t = 2$ and (b) $t = 3$.

maps. This context vector is used to update the decoder's hidden state at m^{th} timestep. The spatio-temporal visual context features are derived through using EfficientNetB6, followed by Global Average Pooling and an LSTM layer. Therefore, the future hidden state h_{m+1}^{D} of the decoder at the $(m+1)^{th}$ timestep is given as:

$$h_m^D = \begin{cases} LSTM_D(q_{n+m-1}^{TF}, C_m), & if\ \theta \geq \tau \\ LSTM_D(q_{n+m-1}^{NM}, C_m), & if\ \theta < \tau \end{cases} \tag{4.5}$$

where q_{n+m-1}^{TF} and q_{n+m-1}^{NM} corresponds to ground truth coordinates and predicted coordinates from the previous decoder timestep, respectively.

4.1.1.2 Adaptive Fuzzified Discriminator (AFD)

In computer vision applications, binary logic classifiers in Convolutional Neural Networks (CNNs) excel at deterministic tasks like binary image categorization and object presence detection. Conversely, Fuzzy logic allows values to be represented as degrees of truth using membership functions to model uncertainty. It has demonstrated effectiveness in various computer vision classification applications, including image classification and reasoning problems [149], [150]. Capitalizing on the distinctive capabilities of Fuzzy logic, the proposed PCTP-AGFL strategically incorporates it into the discriminator. This integration augments prediction

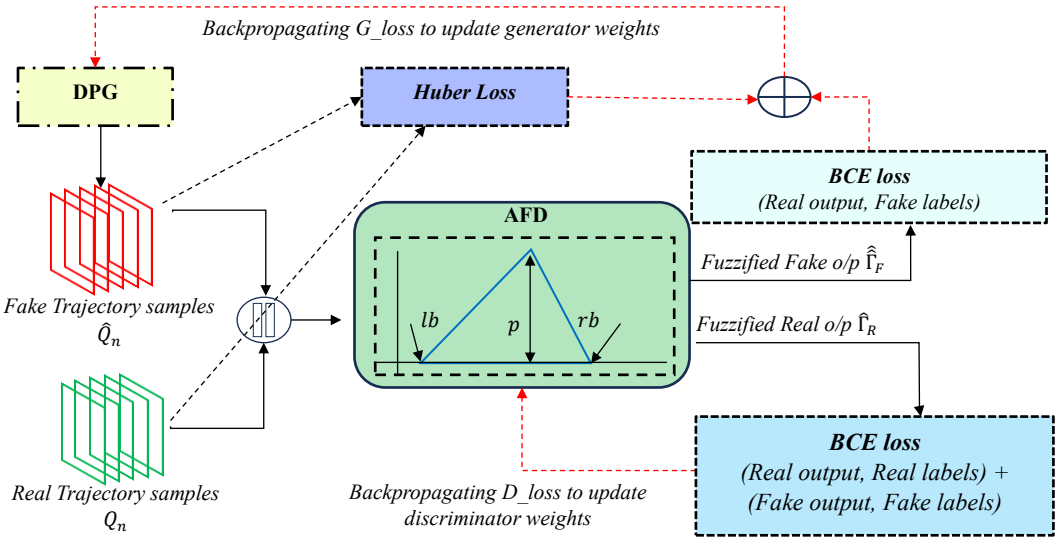


Fig. 4.6: Comprehensive Training Overview of the Proposed Architecture

Algorithm 4.1: Computation of Fuzzy Scores

Inputs:

- lb: Left boundary of the triangular function.
- p: Peak (center) of the triangular function.
- rb: Right boundary of the triangular function.
- $\hat{\mathfrak{Д}}$: Discriminator Output.

Output:

Fuzzy Scores $\hat{\Gamma}$

Initialization:

Initialize $lb, p,$ and rb during model training for real and fake samples, respectively.

Procedure:

for each batch of sample:

Step 1: Calculate real and fake membership scores using membership function $\mu_R(\hat{\mathfrak{Д}}) = \{\hat{\mathfrak{Д}}, \mu_R\}$ and $\mu_F(\hat{\mathfrak{Д}}) = \{\hat{\mathfrak{Д}}, \mu_F\}$ where μ_R and μ_F are degree of membership of the element x to the real class R and fake class F respectively.

for the left segment: if $lb \leq \hat{\mathfrak{Д}} < p$:

$$\mu_R(\hat{\mathfrak{Д}}) = (\hat{\mathfrak{Д}} - lb_R) \,/\, (p_R - lb_R)$$
$$\mu_F(\hat{\mathfrak{Д}}) = (\hat{\mathfrak{Д}} - lb_F) \,/\, (p_F - lb_F)$$

for the right segment: if $p \leq x \leq rb$:

$$\mu_R(\hat{\mathfrak{Д}}) = (rb_R - \hat{\mathfrak{Д}}) \,/\, (rb_R - p_R)$$
$$\mu_F(\hat{\mathfrak{Д}}) = (rb_F - \hat{\mathfrak{Д}}) \,/\, (rb_F - p_F)$$

Outside the function's range:

$$\mu_R(\hat{\mathfrak{Д}}), \mu_F(\hat{\mathfrak{Д}}) = 0$$

Step 2: Now the fuzzy rule is the union operation between real and fake sets:

$$\hat{\Gamma} = \mu_{RUF}(\hat{\mathfrak{Д}}) = \max(\mu_R(\hat{\mathfrak{Д}}), \mu_F(\hat{\mathfrak{Д}}))$$
$$\boldsymbol{If}\ max(\mu_R(\hat{\mathfrak{Д}}), \mu_F(\hat{\mathfrak{Д}}) == \mu_F(\hat{\mathfrak{Д}})$$
$$\hat{\Gamma} = \mu_F^C(\hat{\mathfrak{Д}}) = 1.0 - \mu_F(\hat{\mathfrak{Д}}),$$

where μ_F^C is a complement operation

return $\hat{\Gamma}$

capabilities, especially in challenging scenarios where discriminating between fake and real instances is ambiguous [150]. The complete training overview is illustrated in Fig. 4.6. It illustrates the interplay between real ground truth Q_n and generated fake trajectory samples $\hat{Q}_n$ fed to the discriminator. It is followed by adaptive fuzzification to address ambiguous cases where distinguishing between real and fake trajectories is challenging. Loss computation involves Huber loss for comparing real and fake trajectory samples, alongside binary cross-entropy (BCE) loss for adversarial training that are backpropagated to update generator and discriminator weights. Real and fake trajectory samples $S_n^{1:m} \sim p(Q_n^{1:m}, \hat{Q}_n^{1:m})$ are input to the discriminator, which generates real and fake output respectively as in Eqn (4.6):

$$\hat{\mathfrak{Ɔ}} = Linear(LSTM_{dis}(S_n^{1:m}, h_{dis}) \quad (4.6)$$

where $\hat{\mathfrak{Ɔ}}_R$ and $\hat{\mathfrak{Ɔ}}_F$ are the discriminator output labels corresponding to real (ground truth) and fake (generated) trajectory samples, respectively. $Linear$ serves as a fully connected layer with no activation. Subsequently, the output undergoes an adaptive fuzzification process, as detailed in Algorithm 4.1. For adaptive Fuzzification, two membership functions, $\mu_R(\hat{\mathfrak{Ɔ}})$ and $\mu_F(\hat{\mathfrak{Ɔ}})$, are defined for real and fake trajectory classes. These membership functions take the form of triangles with adaptable parameters, including left boundary (lb), right boundary (rb), and peak (p). These parameters are fine-tuned in each training epoch to ensure a clear separation between the distributions of real and fake trajectory samples. The algorithm computes the degree of membership, μ_R/μ_F, of the discriminator output $\hat{\mathfrak{Ɔ}}$ to the real/fake class, as defined in Step 1. The final fuzzy score denoted as $\hat{\Gamma}$, is determined through the union operation between the real and fake sets. This operation captures the maximum degree of membership of the input $\hat{\mathfrak{Ɔ}}$ to the real or fake class as described in Step 2. In cases where the maximum degree of membership is associated with the fake class, complementation is necessary. This allows the fuzzy scores to be employed for training and backpropagation with now $\hat{\Gamma}_R^{1:m}$ and $\hat{\Gamma}_F^{1:m}$ denoting the predicted fuzzy probability scores for real and fake samples with ground truth labels $\Gamma_R^{1:\mathrm{m}} = 1$ and $\Gamma_F^{1:\mathrm{m}} = 0$ respectively. This process ensures that the fuzzy scores effectively guide training by leveraging membership degrees and complementation, refining the discriminator's ability to distinguish real from fake samples.

4.1.2 Experimental Results and Works

In this section, the efficacy of the proposed method is assessed against several state-of-the-art approaches using two first-person view (FPV) datasets, JAAD [137] and PIE [6], and two bird's eye view (BEV) datasets, ETH [151] and UCY [152]. Furthermore, the evaluation entails a comprehensive comparative analysis and discussion on the impact of progression parameters, training strategies, and the type of discriminator logic, shedding light on the method's adaptability and performance

Algorithm 4.2: PCTP-AGFL Training Procedure

Inputs and Definitions:

(i) Mini-batch size: $'s'$

(ii) For $u \in (1, s)$,

- $Q^{(u)}$: Pedestrian ground truth trajectory space
- $\hat{Q}^{(u)}$: Pedestrian generated trajectory space
- $P^{(u)}$: Pedestrian historical trajectory space
- $R^{(u)}$: RGB image space
- $S^{(u)}$: Segmentation maps space
- $\varepsilon^{(u)}$: Random noise sampled from a normal distribution $\sim N(0,1)$
- $\mathbb{C}_{\boldsymbol{m}}^{(\boldsymbol{u})}$**:** Context Vector

(iii) Dynamic Progressive Generator (DPG) with model parameters ϑ_g:$\{\mathcal{W}_g, \mathcal{B}_g\}$ where $\mathcal{W}_g$ and $\mathcal{B}_g$ are set of weights and biases of layers constituting DPG

- Enc_{DPG} and Enc_{DPG} are encoders and decoders of DPG respectively
- $EDCA$: Encoder-Decoder Contextual Attention Module

(iv) Adaptive Fuzzified Discriminator (AFD) with model parameters ϑ_d:$\{\mathcal{W}_d, \mathcal{B}_d\}$ where $\mathcal{W}_d$ and $\mathcal{B}_d$ are set of weights and biases of layers constituting AFD

(v) $\alpha_g$and α_d: Learning rate for DPG and AFD respectively.

Outputs

Trained DPG and AFD model with updated parameters $\hat{\vartheta}_g$ and $\hat{\vartheta}_d$ respectively.

Procedure

for epochs 1, … , e ***do***

// **Train** Discriminator (AFD); **Freeze** Generator (DPG)

for discriminator steps 1, … , β ***do***

***Step* 1d:** Sample minibatch of size s from $\{Q^{(u)}, P^{(u)}, R^{(u)}, S^{(u)}\}$

***Step* 2d:** $P^{(u)} = P^{(u)} + \varepsilon^{(u)}$, where $\varepsilon^{(u)} \sim N(0,1)$

***Step* 3d:** Construct the input space: $Z^{(u)} = \{P^{(u)}, R^{(u)}, S^{(u)}\}$

***Step* 4d**: DPG generates the trajectory in three steps:

(i) $c_m^{(u)} = EDCA(Enc_{DPG}\{P^{(u)}\})$

(ii) $\mathbb{C}_{\boldsymbol{m}}^{(\boldsymbol{u})} = \left\{c_m^{(u)} \oplus R^{(u)} \oplus S^{(u)}\right\}$

(iii) ***if*** *teacher-forcing mode* **= = *True***

$\hat{Q}^{(u)} = Dec_{DPG}\left(\mathbb{C}_{\boldsymbol{m}}^{(\boldsymbol{u})}, Enc_{DPG}\{P^{(u)}\}, Q^{(u)}\right) = DPG(Z^{(u)})$

else: // *normal mode*

$\hat{Q}^{(u)} = Dec_{DPG}\left(\mathbb{C}_{\boldsymbol{m}}^{(\boldsymbol{u})}, Enc_{DPG}\{P^{(u)}\}\right) = DPG(Z^{(u)})$

***Step* 5d**: Update ϑ_d by ascending its stochastic gradient as:

(i) $\rho_d = \nabla_{\vartheta_d} \frac{1}{s} \sum_{\kappa=1}^{s} \left[\log AFD\left(Q^{(u)}\right) + \log\left(1 - AFD\left(\hat{Q}^{(u)}\right)\right)\right]$

(ii) $\hat{\vartheta}_d = \vartheta_d + \alpha_d . RMSProp(\vartheta_d, \rho_d)$

end for

//**Train** Generator (DPG); **Freeze** Discriminator (AFD)

for generator steps 1, … , β ***do***

Step* 1g−4g**: Repeat the ***Steps **(1d − 4d)** as in AFD training

***Step* 5g**: Update ϑ_g by descending its stochastic gradient as:

(i) $\rho_g = \nabla_{\vartheta_g} \frac{1}{s} \sum_{\kappa=1}^{s} \left[\log\left(1 - AFD\left(\hat{Q}^{(u)}\right)\right) + \mathcal{L}_{huber}\left(Q^{(u)}, \hat{Q}^{(u)}\right)\right]$

(ii) $\hat{\vartheta}_g = \vartheta_g - \alpha_g . Adam(\vartheta_g, \rho_g)$

end for

end for

return $\hat{\vartheta}_d$, $\hat{\vartheta}_a$

across varied experimental conditions.

4.1.2.1 Implementation Details

The proposed model's training is executed on a Google Colab Pro instance equipped with a high-performance NVIDIA Tesla T4 GPU, boasting 16 GB of memory, and operated within the CUDA 12.0 platform. The model architecture is constructed using the TensorFlow 2.10.1 framework. In terms of optimization, the Adam optimizer is applied to the Generator, with default parameters and an initial learning rate of 1×10^{-6}. In contrast, the Discriminator employs the RMSprop optimizer with an equivalent learning rate. The training procedure involves a batch size of 4, and training is concluded after 15 epochs. The segmentation maps are generated using state-of-the-art Segformer (MiT-B5)[141], a semantic segmentation model. RGB and segmentation features are precomputed using the EfficientNetB6 network. Consistently, the hidden size for all encoders and decoder LSTMs within the proposed method is set to 64 across all datasets. A dual Monte Carlo sampling strategy is employed in the generator implementation for stochastic predictions. It involves the introduction of random noise to the input data, particularly the past bounding box coordinates, simulating inherent uncertainties in observed data. Concurrently, random sampling is implemented at each iteration, incorporating a random threshold (τ) between 0 and 1 to switch between teacher-forcing and model predictions.

A comprehensive training procedure for the proposed PCTP-AGFL is presented in Algorithm 4.2. The algorithm entails alternating training phases over epochs, optimizing the DPG and AFD models iteratively. In phase I (Steps 1d-5d), AFD undergoes training while DPG parameters (ϑ_g) remain fixed. During this phase, batches of pedestrian trajectory data, along with corresponding images and segmentation maps are processed to update the discriminator's parameters, enhancing its ability to distinguish between ground truth trajectories and generated ones. Subsequently, in phase II (Steps 1g-5g), the AFD parameters (ϑ_d) are held constant while the DPG is trained. This phase improves the trajectory generation by minimizing the Huber loss between generated trajectories and ground truth ones, while the AFD provides adversarial feedback.

The following metrics are used for the evaluation of the proposed trajectory prediction algorithm for FPV datasets: MSE over bounding box coordinates, C_{MSE} and C_{FMSE} which are the MSEs of the centre of the bounding boxes averaged over the entire predicted sequence and only the last time step, respectively. The average displacement error (ADE), which measures accuracy along the whole trajectory, and the final displacement error (FDE), which measures accuracy only at the trajectory endpoint, are utilized for BEV datasets. All results metrics used for JAAD and the PIE dataset are in pixels, while for ETH and UCY, ADE and FDE are computed in Euclidean space.

Loss Function: The Huber loss, also known as the Huber penalty or smooth $L1$ loss is a function used here to measure how far the generated samples are from the ground truth. It is a compromise between $L1$ loss and the $L2$ loss and is less sensitive to outliers compared to the $L2$ loss. It is defined as follows:

$$\mathcal{L}_{huber}(Q_n, \hat{Q}_n) = \begin{cases} \frac{1}{2}(Q_n - \hat{Q}_n)^2, if\ |Q_n - \hat{Q}_n| \leq \delta \\ \delta|Q_n - \hat{Q}_n| - \frac{1}{2}\delta^2,\ otherwise \end{cases} \tag{4.7}$$

where δ is a threshold at which the loss function transitions from $L2$ loss to $L1$ loss. It is chosen to balance the trade-off between outliers' robustness and the loss function's smoothness. This experiment is empirically set at 1.

Another loss for adversarial training is defined in Eqn (4.8) as:

$$\mathcal{L}_{adversarial} = \min_{G}\max_{D} \mathrm{E}_{S_n \sim p(Q_n)}\left[\Gamma_{\mathrm{R}}^{1:\mathrm{m}} \log \hat{\Gamma}_{R}^{1:m}\right] + \mathrm{E}_{\hat{S}_n \sim p(\hat{Q}_n)}\left[(1\Gamma_{\mathrm{F}}^{1:m}) \log (1 - \hat{\Gamma}_{F}^{1:m})\right] \tag{4.8}$$

represents the min-max game where the generator minimizes the function while the discriminator maximizes it.

4.1.2.2 Datasets

The PIE dataset comprises 1,842 pedestrian trajectories annotated at 30 Hz, characterized by extended trajectory lengths and detailed annotations encompassing semantic intention, ego-motion, and neighbouring objects. The dataset includes 880, 243, and 719 pedestrian tracks in the train, validation, and test sets, respectively [6]. A sampling approach with a 0.5 overlap ratio ensures comprehensive coverage, excluding tracks below the minimum length of 2 seconds (observation + prediction) during trajectory prediction training.

JAAD features a comprehensive collection of 2,800 pedestrian trajectories captured from dash cameras, annotated at 30 Hz. The dataset is partitioned according to the specifications in [137], with divisions into 177, 117, and 29 clips for training, testing, and validation, respectively. Given its smaller sample size and shorter tracks, a sampling approach with an overlap ratio of 0.8 is implemented.

The ETH-UCY datasets comprise five sub-datasets, aggregating 1,536 annotated pedestrian trajectories across four unique scenes. Trajectories are observed for 3.2 seconds, with predictions extending for the subsequent 4.8 seconds, sampled at a rate of 2.5 Hz. Pedestrian centroids, featuring single x and y coordinates, are employed in line with the model's input requirements. Notably, visual context is omitted due to the absence of a first-person view. Following prior work [116], a leave-one-out strategy is applied to partition the train and test sets.

4.1.2.3 Comparison with SOTA methods

In this section, the proposed model undergoes a comprehensive comparison with state-of-the-art methods, including B-LSTM[153], PIE_traj[6], BiTrap[22], SGNet[23], DSCMP[108], PECNet[24], STAR[107], SIT[112], Trajectron++[110], LVTA[106], STI-GAN[94], S-DualCVAE[113], Y-Net[25], V2-Net[109], and NSP-SFM[114]. The evaluation is conducted under two distinct settings: deterministic,

Table 4.1: Deterministic Results on PIE/JAAD Dataset

Methods	PIE					JAAD				
	MSE			C_{MSE}	CF_{MSE}	MSE			C_{MSE}	CF_{MSE}
	0.5s	*1s*	*1.5s*	*1.5s*	*1.5s*	*0.5s*	*1s*	*1.5s*	*1.5s*	*1.5s*
B-LSTM[153]	101	296	855	811	3259	159	539	1535	1447	5615
PIE_{traj}[6]	58	200	636	596	2477	110	399	1248	1183	4780
BiTraP[22]	41	161	511	481	1949	93	378	1206	1105	4565
SGNet[23]	34	133	442	413	1761	82	328	1049	996	4076
Ours	**12**	**75**	**300**	**223**	**1299**	**35**	**205**	**825**	**784**	**3383**

Table 4.2: Stochastic Results on PIE/JAAD Dataset

Methods	PIE					JAAD				
	MSE			C_{MSE}	CF_{MSE}	MSE			C_{MSE}	CF_{MSE}
	0.5s	*1s*	*1.5s*	*1.5s*	*1.5s*	*0.5s*	*1s*	*1.5s*	*1.5s*	*1.5s*
BiTraP(GMM)[22]	38	90	209	171	368	53	250	585	501	998
BiTraP(NP)[22]	23	48	102	81	261	38	94	222	177	565
SGNet[23]	16	39	88	66	206	37	86	197	146	443
Ours	**6**	**21**	**59**	**45**	**138**	**19**	**55**	**147**	**105**	**301**

Table 4.3: Deterministic Results on ETH/UCY Dataset

Methods	ADE(4.8s)/FDE(4.8s)					
	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg
STAR[107]	0.56/1.11	0.26/0.50	0.52/1.15	0.41/0.90	0.31/0.71	0.41/0.87
SIT[112]	0.59/1.28	0.22/0.45	0.40/0.98	0.30/0.75	0.23/0.59	0.35/0.81
Trajectron++[110]	0.71/1.68	0.22/0.46	0.41/1.07	0.30/0.77	0.23/0.59	0.37/0.91
SGNet-ED[23]	0.63/1.38	0.27/0.63	0.40/0.96	0.26/0.64	0.21/0.53	0.35/0.83
LVTA[106]	0.57/1.10	0.42/0.69	0.55/1.19	0.42/0.92	0.35/0.75	0.46/0.92
Ours	**0.48/1.01**	**0.15/0.57**	**0.31/0.94**	**0.21/0.55**	**0.17/0.49**	**0.27/0.71**

Table 4.4: Stochastic Results on ETH/UCY Dataset

Methods	ADE(4.8s)/FDE(4.8s)					
	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg
STI-GAN[94]	0.77/1.53	0.70/0.73	0.53/1.20	0.33/0.66	0.33/0.66	0.53/0.96
S-DualCVAE[113]	0.66/1.18	0.34/0.61	0.39/0.74	0.27/0.48	0.24/0.42	0.38/0.69
DSCMP[108]	0.66/1.21	0.27/0.46	0.50/1.07	0.33/0.68	0.28/0.60	0.41/0.80
PECNet[24]	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48
STAR [107]	0.36/0.45	0.17/0.36	0.31/0.62	0.26/0.55	0.22/0.46	0.26/0.53
Trajectron++[110]	0.43/0.86	0.12/0.19	0.22/0.43	0.17/0.32	0.12/0.25	0.21/0.41
BiTrap-NP[22]	0.37/0.69	0.12/0.21	0.17/0.37	0.13/0.29	0.10/0.21	0.18/0.35
SGNet[23]	0.35/0.65	0.12/0.24	0.20/0.42	0.12/0.24	0.10/0.21	0.18/0.35
Y-Net[25]	0.28/0.33	0.10/0.14	0.24/0.41	0.17/0.27	0.13/0.22	0.18/0.27
V^{2}-Net[109]	0.23/0.37	0.11/0.16	0.21/0.35	0.19/0.30	0.14/0.24	0.18/0.28
NSP-SFM[114]	0.25/0.24	0.09/0.13	0.21/0.38	0.16/0.27	0.12/0.20	0.17/0.24
Ours	**0.26/0.54**	**0.05/0.17**	**0.11/0.33**	**0.07/0.15**	**0.07/0.17**	**0.11/0.27**

where the model yields a single trajectory, and stochastic, generating a set of $K = 20$ possible trajectories, with the best-performing sample subsequently reported.

Table 4.1 and 4.2 provide a comparative analysis of the proposed model's performance against prior baselines and state-of-the-art methods on first-person view (FPV) datasets. On JAAD, the proposed model demonstrates a substantial reduction in MSE by 57%, 38%, and 21% for prediction intervals of $0.5s$, $1.0s$, and $1.5s$, respectively, outperforming the previous state-of-the-art[23]. Similarly, on the PIE dataset, the proposed model exhibits MSE reductions of 65%, 44%, and 32% for prediction intervals of $0.5s$, $1.0s$, and $1.5s$, respectively, compared to the previous state-of-the-art[23]. Notably, as the prediction length extends, the proposed model showcases even more significant improvements when compared to prior work, particularly highlighting its efficacy in long-term prediction scenarios. To ensure a fair comparison with [23] on FPV datasets under stochastic settings, where $K = 20$ possible proposals are generated, and the best-performing sample is reported during evaluation. The proposed method consistently outperforms the state-of-the-art by an average of 35% on JAAD and 41% on PIE.

For the ETH/UCY dataset, deterministic and stochastic results are summarized in Table 4.3 and 4.4. These tables illustrate that, on average, the proposed model surpasses the state-of-the-art methods[23] by more than 23% and 12% in ADE and FDE, respectively. These outcomes highlight the model's ability to predict persistent and stable future trajectories. Compared with the FPV datasets, the improvements on the ETH/UCY dataset are relatively lower, attributed to the absence of a first-person view context and scene semantics. Nevertheless, the proposed method demonstrates remarkable efficacy, achieving a significant reduction of 36% and 19% in ADE and FDE, respectively, compared to the lowest ADE and FDE observed in the stochastic setting. It also achieves comparable performance in *ETH* and *Hotel* sets.

4.1.2.4 Ablation Study

This section presents an ablation study to assess the impact of key methodological choices in the proposed framework. The influence of progression parameters is examined, followed by comparing Normal, Teacher Forcing, and

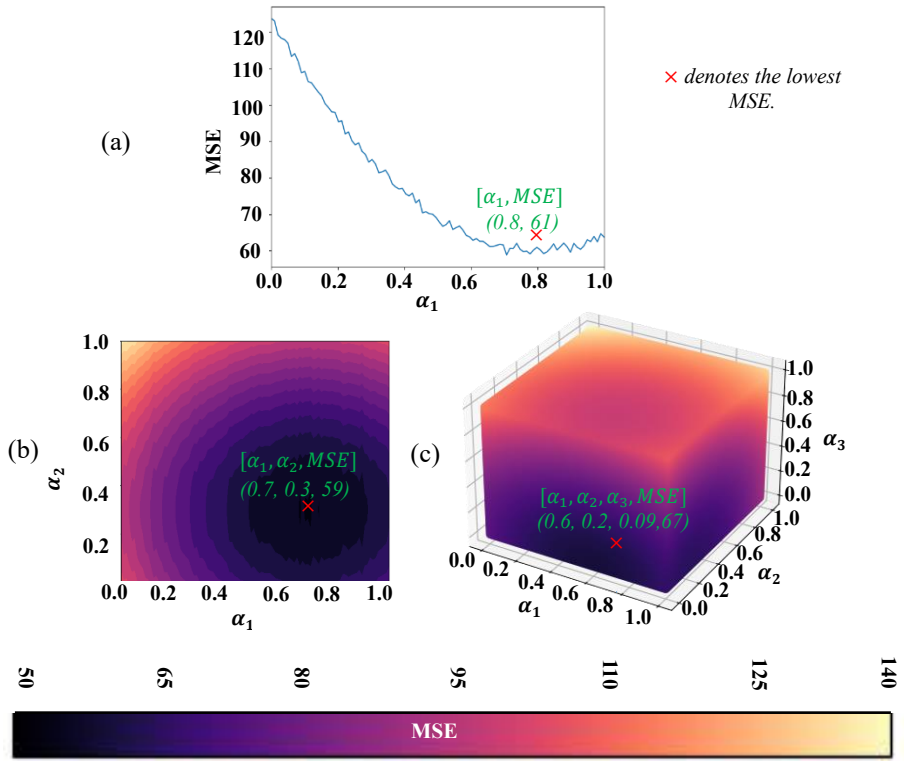


Fig. 4.7: Impact of Encoder-Decoder Progression where (a), (b), and (c) correspond to $r = 1, r = 2$ and $r = 3$ respectively.

Learned Scheduled Sampling strategies. Additionally, the effectiveness of an Adaptive Fuzzified Discriminator is evaluated against a Binary Discriminator, and the computational time cost is analysed. The analyses are as follows.

i. **Impact of Progression Parameters on Performance**: The investigation of progression parameters on the performance of the proposed approach reveals significant insights pertaining to predictive performance, as shown in Fig. 4.7. For $r = 1$, it is observed that the MSE experiences a notable reduction, with the best performance achieved when α_1 is approximately 0.8. Upon further increasing the model complexity to $r = 2$ with three LSTM layers, the MSE demonstrates a further decrease. In this scenario, the best performance is achieved with α_1 and α_2 values of approximately 0.7 and 0.3 for the respective LSTM layers. This trend suggests that enhancing model complexity, along with appropriately tuned α values, has a positive impact on predictive performance. However, the transition to $r = 3$ with four LSTM layers yields an unexpected deviation from the decreasing MSE

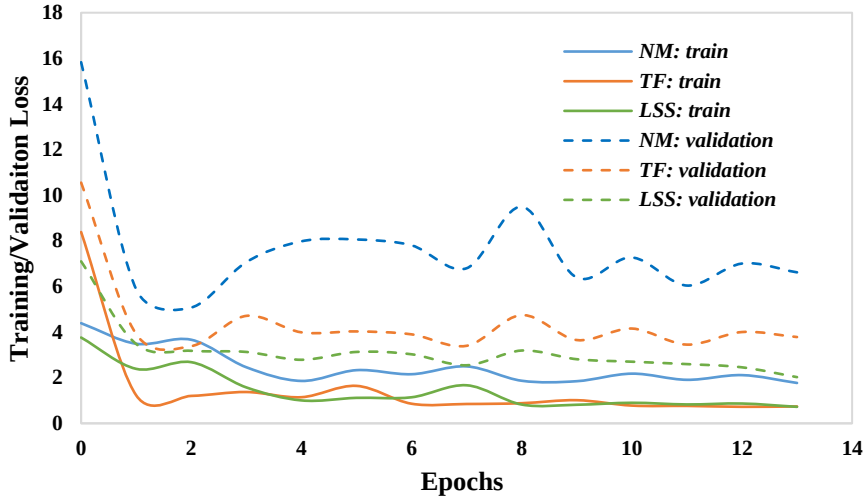
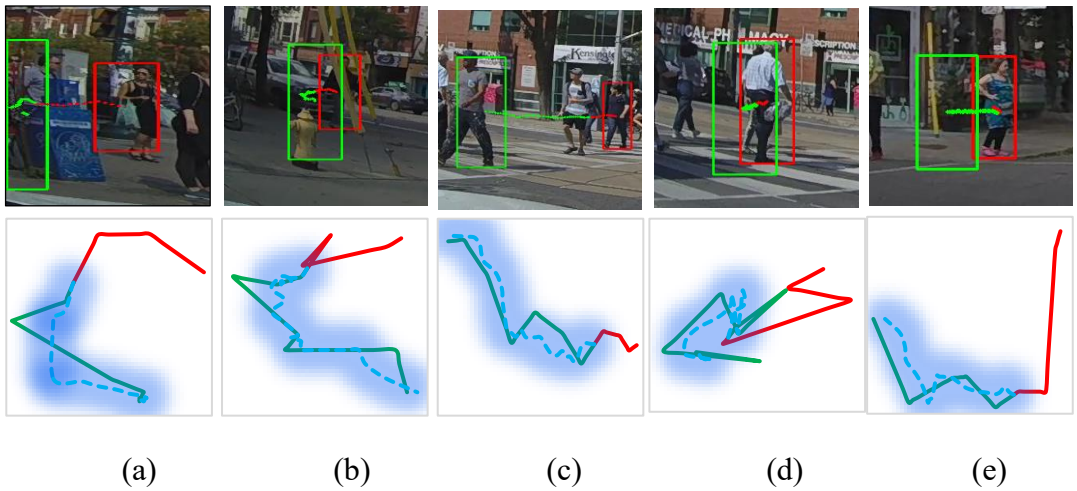


Fig. 4.8: Training and Validation Loss Curves for three training modes: learned scheduled sampling (LSS), teacher forcing (TF), and normal mode (NM).

trend. Contrary to expectations, the MSE increases despite the α_1, α_2 and α_3 values optimized to approximately 0.6, 0.2, and 0.09 adaptively for the respective LSTM layers. Furthermore, it is evident that α_3 does not contribute to the same extent as α_1 and α_2 , rendering it the least important in this configuration. It implies that while increasing the number of LSTM layers may enable the model to represent more intricate data features, it can potentially introduce overfitting or model complexity that fails to generalize effectively to unseen data. Through empirical



(a) (b) (c) (d) (e)

Fig. 4.9. Qualitative samples of complex trajectory patterns. *Row I* shows the pedestrian's trajectory with the start (***red*** box) and end (***green*** box) coordinates. The ***red*** line represents centre coordinates over 15 timesteps, while future ground-truth bounding boxes (next 45 timesteps) are in ***green***. *Row II* depicts the 2D spatial projection of bounding box centres in the x-y plane. The ***blue*** dashed line represents the average stochastic trajectory from 20 multimodal predictions, with the shaded region indicating the range of possible pathways.

analysis, it has been determined that the optimal configuration corresponds to $r = 2$ and $\{\alpha_1, \alpha_2\} = \{0.7, 0.3\}$ and $\{0.8, 0.1\}$ for FPV and BEV datasets, respectively.

ii. **Comparison of Normal vs Teacher Forcing vs Learned Scheduled Sampling Strategy**: The training and validation loss curves presented in Fig. 4.8 demonstrate the consistent outperformance of both the learned scheduled sampling mechanism and the teacher forcing mode in comparison to the conventional normal training mode within the context of trajectory prediction. It is inferred that both the normal training mode and the teacher forcing mode exhibit fluctuations in their validation loss, suggesting limitations in their capacity to generalize to previously unseen and complex data patterns effectively. In contrast, the learned scheduled sampling mechanism, which dynamically determines the probability of employing teacher-forced or normal training modes, maintains a remarkable level of stability in its validation loss curve.

Furthermore, the impact of the proposed methodology on addressing non-linear and intricate trajectory patterns is also evident in Fig 4.9 (a)-(e). These trajectories are characterized by multiple turns before ultimately reaching their respective destinations. For instance, in Fig. 4.9 (b), it is noticeable that the target pedestrian is somewhat occluded in preceding frames. In Fig. 4.9 (c) and (e), the visual is disconnected from the trajectory as the individual gazes in a different direction. Fig. 4.9 (d) introduces another dimension as the pedestrian initially has their back turned, as evident from the past trajectory, but subsequently executes an entirely distinct path. In all the cases, the average path predicted by our method closely matches the groundtruth motion pattern.

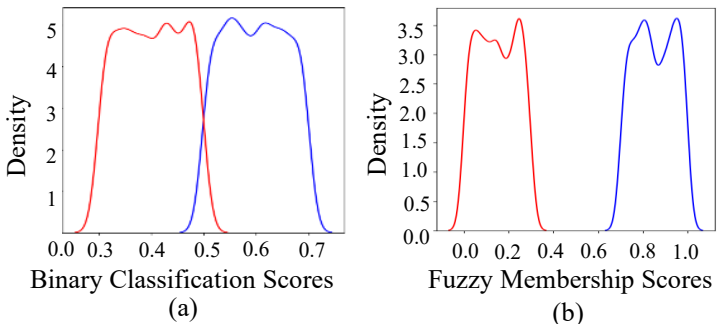


Fig. 4.10: KDE-based distribution of (a) binary classification and (b) fuzzy membership scores, with fake trajectories in red and real trajectories in blue.

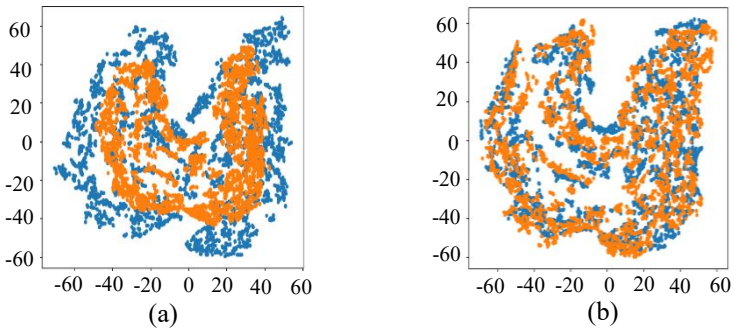


Fig. 4.11: t-SNE Visualization of Predicted vs. Ground Truth Trajectories: (a) Without AFD, predicted trajectories (***orange***) are confined, showing mode collapse. (b) With AFD, their distribution expands and aligns with ground truth (***blue***), mitigating mode collapse.

Table 4.5: Quantitative results of PCTP-AGFL Across FPV and BEV Datasets

Model Ablations	N	T	L	F	E	V_R	V_S	*CS*	IT(*ms*) 20/2000	MSE(1.5s) JAAD/PIE	ADE/FDE(4.8s) ETH/UCY(Avg)
Ablation_1	✓	×	×	×	×	×	×	1	77/80	186/86	0.21/0.36
Ablation_2	×	✓	×	×	×	×	×	1.2	77/80	191/90	0.26/0.39
Ablation_3	×	×	✓	×	×	×	×	1.09	77/80	181/79	0.18/0.31
Ablation_4	×	×	✓	✓	×	×	×	1.05	79/82	163/71	0.14/0.30
Ablation_5	×	×	✓	✓	✓	×	×	1.04	80/81	156/62	0.11/0.27
Ablation_6	×	×	✓	✓	✓	✓	×	0.95	82/85	151/60	-
Ablation_7	×	×	✓	✓	✓	✓	✓	0.90	84/87	147/59	-

**N: Normal Mode, T: Teacher-forcing, L: Learned Scheduled Sampling using AGM, F: AFD, E: EDCA, V_R: RGB context, V_S: Segmentation context, CS: Convergence Speed, IT: Inference Time*

iii. **Adaptive Fuzzified Discriminator vs Binary Discriminator**: KDE plots offer a robust means of depicting data distribution by estimating the probability density function. In the case of binary discriminator training with binary labels, the plot in Fig. 4.10(a) displays a certain degree of overlap between real and fake trajectories, implying a degree of ambiguity in classification. This convergence of curves underscores the limitations of binary classification, particularly when confronted with trajectories exhibiting intermediate characteristics. Conversely, the KDE plot for the fuzzy discriminator in Fig. 4.10(b) reveals two distinct, well-separated curves representing real and fake trajectories, highlighting the efficacy of the fuzzy approach in classifying trajectories.

The t-SNE plots in Fig. 4.11 visualize the high-dimensional trajectory embeddings, providing insights into the impact of AFD on the GAN's ability to capture and generate diverse trajectories. Without Adaptive Fuzzy Logic, the t-SNE plot in Fig. 4.11 (a) demonstrates a discernible divergence between the predicted and ground truth trajectory distributions. The generated trajectories exhibit a limited

coverage of the real distribution space, indicative of challenges in training the GAN model, resulting in a potential mode collapse. In contrast, the t-SNE plot in Fig. 4.11 (b) showcases a remarkable coverage of the entire trajectory space by the GAN model, even in its inherent complexity. The incorporation of adaptive Fuzzification mitigates mode collapse, enabling the generator to capture diverse modes within the data.

iv. **Computational time cost**: In the ablation study of PCTP-AGFL, various components are analysed for their impact on convergence speed (CS), inference time (IT), and accuracy, as shown in. Table 4.5. The convergence speed (CS) is computed relative to *Ablation 1*, where it is set as the baseline, while others are expressed as factors. The teacher forcing training strategy (*Ablation_2*) accelerates convergence by 1.2 times but exhibits overfitting, resulting in a 4% and 16% increase in error on FPV and BEV datasets compared to normal training (*Ablation_1*). Integration of AGM for learned scheduled sampling (*Ablation_3*) achieves a 5% reduction in errors across datasets, with a marginal 9% speed drop from *Ablation_2*. Adaptive Fuzzification causes a massive drop of 10% in MSE across FPV datasets. Furthermore, contextual awareness proves vital, as seen in *Ablation_6 and Ablation_7* $w.r.t.$ *Ablation_5*. It is observed that RGB Context (V_R) alone leads to a $3-4\%$ error reduction on FPV datasets, whereas incorporating segmentation context (V_S+V_R) results in a significant $5-6\%$ reduction in errors on average on FPV datasets, though with a slower relative convergence speed of 0.90. This emphasizes the importance of visual context in capturing complex spatio-temporal interactions, compensating for slower convergence with improved trajectory prediction accuracy. Even without visual context (*Ablation_5*), the proposed approach achieves a noteworthy reduction in ADE and FDE, as reflected in Table 4.5, while maintaining a reasonable relative convergence speed of 1.04.

Considering the computational efficiency, the proposed DPG (Dynamic Progressive Generator) and AFD (Adaptive Fuzzy Discriminator) utilize a minimal parameter count of 0.16M and 0.042M, respectively, leveraging precomputed EfficientNetB6 feature maps and segmentation maps. Furthermore, the number of

LSTM layers in DPG corresponds to the complexity of motion patterns, but with simple patterns, a single LSTM layer suffices. The AFD employs an adaptive membership function to mitigate decision ambiguity during classification, enhancing model interpretability and accelerating training convergence.

The computational times for 20 and 2000 samples are also compared in Table 4.5. Notably, the proposed PCTP-AGFL demonstrates minimal time differences between generating 20 and 2000 samples, with only a 2-3ms variation. In inference stage, the inference time remains independent of the training strategy adopted, i.e., teacher forcing training and normal mode training, as reflected in Table 4.5. The Adaptive Fuzzification (*Ablation_4*) incurs negligible impact on inference time. However, the inclusion of EDCA only (*Ablation_5*) and EDCA+V_R+V_s (*Ablation_7*) results in more accurate predictions, accompanied by a marginal increase of 5 ms in inference time. It is noteworthy that the inference time of *Ablation_5* is comparable with state-of-the-art [22], [116] as reported in [22].

4.2 Conclusion and Future Scope

This work introduces a novel Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration (PCTP-AGFL) and evaluates its effectiveness on both FPV and BEV datasets. The experimental results reveal our methodology's remarkable capability to closely emulate the complex trajectory patterns and their final destinations with significantly reduced mean squared error in comparison to other SOTA methods. Consequently, it simultaneously addresses the challenges associated with overfitting and generalization to complex data patterns that are often encountered in trajectory prediction methodologies. Furthermore, the Adaptive Fuzzified Discriminator (AFD) enhances discrimination in ambiguous cases. Future work includes exploring a combined short-term and long-term intention prediction approach for further advancements in trajectory prediction.

CHAPTER 5

UNIFIED SHORT-TERM AND LONG-TERM INTENTION PREDICTION

In the preceding chapters, we explored approaches to short-term and long-term pedestrian intention prediction as separate tasks, each with its own set of methodologies and challenges. Short-term prediction focuses on immediate actions such as crossing intentions, while long-term forecasting aims to anticipate extended trajectories based on historical motion patterns and environmental cues. Although treating these tasks independently has yielded meaningful insights, it often overlooks the interdependence between immediate pedestrian intent and longer-term behavioural outcomes.

Integrating short-term and long-term prediction within a unified framework leverage shared contextual and motion features, enabling more coherent and accurate predictions. Short-term cues, such as the decision to cross, provide strong signals that can influence and constrain long-term motion forecasts. To this end, this chapter introduces a dual-task approach that predicts short-term crossing intentions and long-term trajectories using pedestrian ROIs, scene attributes, and past trajectories. The framework addresses key limitations in feature fusion and adaptive prediction, contributing to more reliable pedestrian behaviour modelling across both short-term and long-term horizons.

5.1 Cross-Modal Pedestrian Behaviour Prediction: A Dual-Task Approach with Progressive Denoising Attention and CVAE

Pedestrian intention and trajectory prediction are crucial for advancing intelligent transportation systems and autonomous vehicles, significantly enhancing urban mobility's safety and efficiency. Traditional approaches have evolved from capturing pedestrian dynamics through image features and bounding box coordinates to leveraging multiple modalities and attention mechanisms. However, challenges in

robust cross-modal feature integration and adaptation to complex scenarios persist. This paper introduces a dual-task approach that simultaneously predicts short-term pedestrian crossing intentions and long-term trajectories by integrating features from pedestrian regions of interest (ROIs), scene attributes, and past trajectories. For crossing intention prediction, Progressive Denoising Attention (PDA) is developed, which iteratively refines cross-modal features to augment inter-class variations.

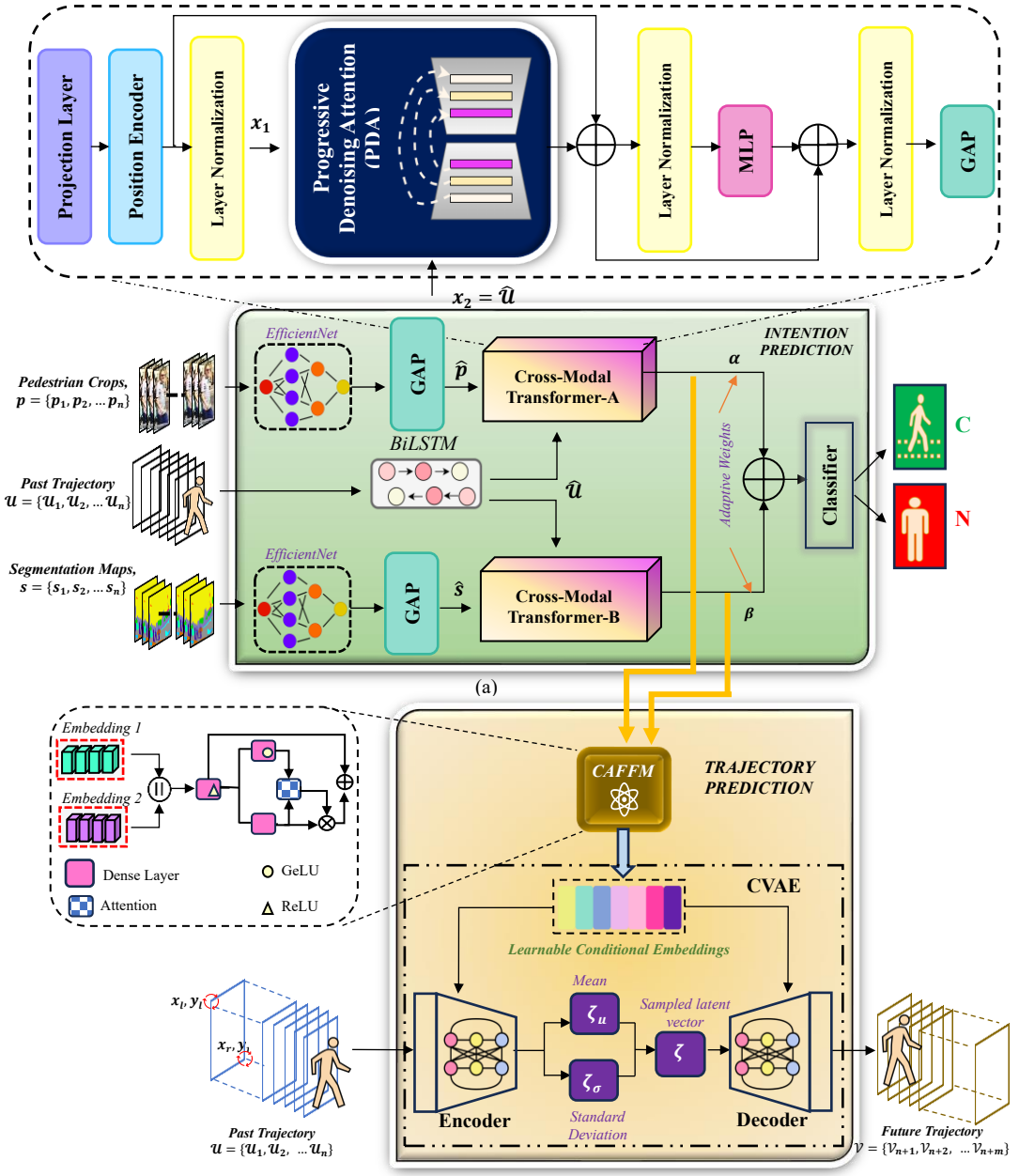


Fig. 5.1: Proposed Dual-task approach for pedestrian behaviour prediction. (a) Intention Estimation: (b) Trajectory Estimation

Additionally, a three-phase counterfactual training approach is employed that manipulates pedestrian ROIs and segmentation maps to further enhance model robustness in complex scenarios. For trajectory prediction, a Conditional Variational Autoencoder (CVAE) is implemented, guided by contextual embeddings from the novel Context-Aware Feature Fusion Module (CAFFM) to significantly reduce mean squared error by integrating rich spatiotemporal ROI and context information. Experimental results on benchmark datasets JAAD and PIE demonstrate the superior performance of the proposed approach in understanding and predicting pedestrian intent.

5.1.1 Proposed Approach

The proposed work introduces a dual-task approach designed to predict pedestrians' short-term and long-term intentions. The short-term intention anticipates the pedestrian crossing intention $\mathcal{I}$, where $(0 < \mathcal{I} < 1)$, while the long-term intention predicts the future trajectory $\mathcal{V}$ over future time steps m.

5.1.1.1 Intention Prediction

Pedestrian crossing intentions on the road are significantly influenced by past motion history. To address this, transformers are employed to process refined pedestrian and scene features conditioned on historical motion data, utilizing the proposed Progressive Denoising Attention (PDA). The input to the trajectory coordinates $\mathcal{U}$ comprising top-left (x_l, y_l) and bottom-right (x_r, y_r) coordinates are processed through a BiLSTM encoder to capture the trajectory information as illustrated in Fig. 5.1 (a).

Transformer A process the RGB pedestrian appearance features $\hat{p}$ conditioned on trajectory data. Similarly, Transformer B process segmentation map features $\hat{s}$ conditioned on trajectory data. The outputs of the two transformers and the encoded trajectory are dynamically weighted and concatenated to produce the final feature representation that is fed to the classifier for prediction. This transformer consists of layers such as Position Encoder and tokenization, Layer Normalization, MLP and the

Algorithm 5.1: Progressive Denoising Attention (PDA)

Inputs:

i) $\boldsymbol{x_1} \epsilon\ R^{N \times D_1}$: Input sequence 1
ii) $\boldsymbol{x_2} \epsilon\ R^{N \times D_2}$: Input sequence 2
iii) $\boldsymbol{\tau}$: Maximum number of iterations (*default*: 5)
iv) $\boldsymbol{\epsilon}$: Convergence tolerance (*default*: $1 \times\ 10^{-3}$)
v) $\boldsymbol{\sigma}$: Standard deviation of the noise (*default*: 0.1)

Output:

$\boldsymbol{Z} \epsilon\ R^{N \times D}$: Refined attention output

Steps:

1. **Initialization**:
   Initialize query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$:
   $$Q\ =\ Dense(D)(x_1), K = Dense(D)(x_2),$$
   $$V\ =\ Dense(D)(x_2)$$
2. **Self-Attention Calculation:**
   Compute initial attention scores and output:
   $$Z =\ softmax(\frac{QK^T}{\sqrt{D}})V$$
3. **Iterative Refinement:**
   - Set Z_{prev} = 0
   - $While\ i\ <\ \tau\ \&\&\ \left\|Z - Z_{prev}\right\| > \epsilon$:
     - Add Gaussian noise to Q, K and V:
       $$Q \leftarrow Q + \mathcal{N}(0, \sigma^2)$$
       $$K \leftarrow K + \mathcal{N}(0, \sigma^2)$$
       $$V \leftarrow V + \mathcal{N}(0, \sigma^2)$$
     - Update $Z_{prev} \leftarrow Z$
     - Apply self-attention: $Z =\ softmax(\frac{QK^T}{\sqrt{D}})V$
     - Denoise Z using the $Denoising\ UNet$:
       $Z \longleftarrow softmax(Denoising\ UNet(Z))$
     - Set $Q, K, V \leftarrow Z$
     - Increment i
4. **Return:** The final refined attention output $\boldsymbol{Z}$

novel PDA. The attention process incorporates the influence of pedestrian or scene attributes $\{x_1\}$ on the pedestrian's motion $\{x_2\}$ as shown in Algorithm 5.1. For instance, the model might focus more on sudden directional changes if the environment has obstacles or if the pedestrian is younger, indicating higher risk- taking behaviour. Conversely, in a clear environment with an elderly pedestrian, the model might reduce the influence of these signals, considering the lower likelihood of abrupt movements.

Attention: Diffusion-based denoising modules reduce noise while preserving essential structural features like edges and textures. Their adaptability to various data

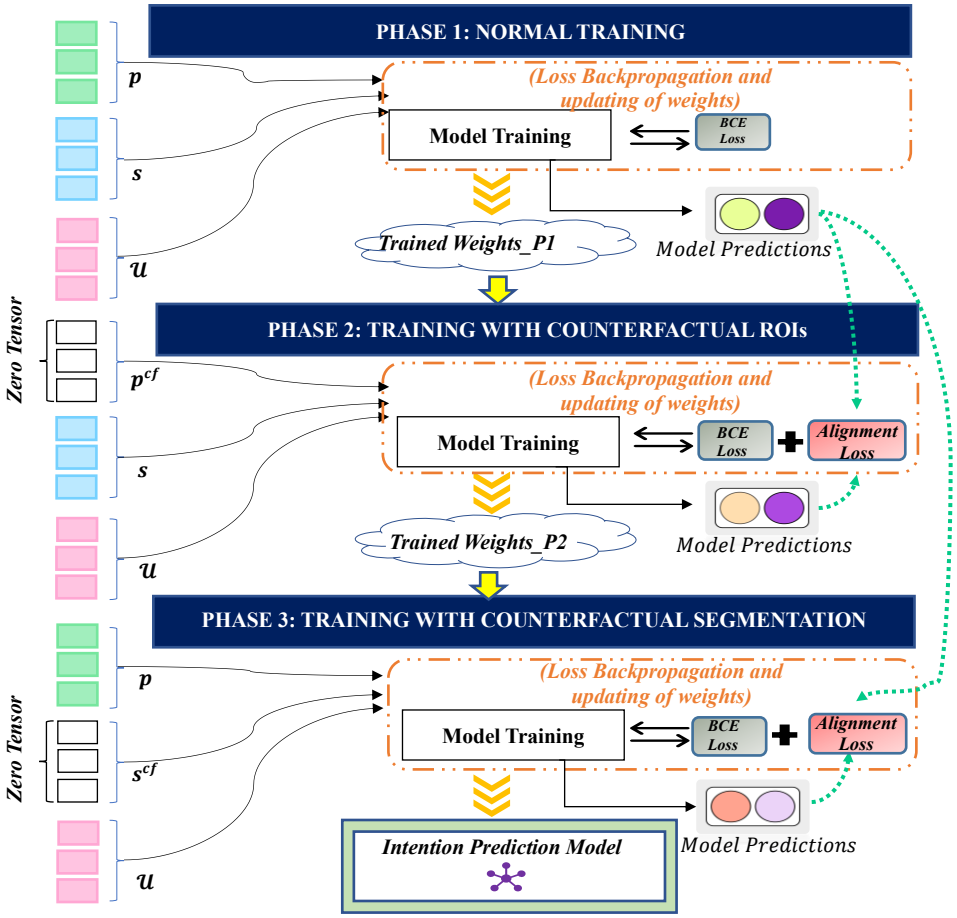


Fig. 5.2: Overview of the counterfactual training process.

and noise conditions makes them a robust choice for denoising tasks [154]. Inspired by their effectiveness, our proposed work introduces a diffusion-inspired attention mechanism to refine and denoise cross-modal features iteratively. Traditional attention mechanisms can struggle with initial misalignments between modalities. By incorporating a denoising process inspired by diffusion models, the proposed Progressive Denoising Attention (PDA) aims to enhance attention outputs iteratively, mimicking the human cognitive process of progressively improving understanding through successive refinements and reassessments. The key steps of PDA are outlined in Algorithm 5.1. Initially, query, key, and value matrices are initialized from the input sequences, as shown in step 1. Self-attention scores are then computed using these matrices, as demonstrated in step 2. The addition of Gaussian noise followed by denoising via U-Net architecture in Step 3 leads to more robust and accurate models, especially in complex environments where dynamic factors influence pedestrian

intentions. This iterative refinement allows the model to refine its focus over multiple steps, which is particularly beneficial for cross-modal tasks where initial alignments might be imprecise. Furthermore, convergence tolerance enables the mechanism to adapt dynamically to different sequences and contexts, ensuring that each scenario's attention mechanism is fine-tuned. The PDA updates attention outputs iteratively until reaching $\boldsymbol{\tau}$ iterations or meeting a convergence threshold as shown in Step 3. The noise addition step is skipped during testing.

Counterfactual Training: The proposed methodology advances the concept of counterfactual training [155] through a structured, multi-phase approach that manipulates pedestrian ROIs and segmentation maps. The proposed approach deepens the model's understanding of the causal relationships between contextual elements and pedestrian behaviour by systematically introducing counterfactual values in distinct phases and incorporating alignment losses to maintain consistency. The three phases of the counterfactual training are illustrated in Fig. 5.2. In Phase 1, the model is trained on pedestrian ROIs, segmentation maps, and past trajectories, establishing a baseline understanding by minimizing the binary cross-entropy loss between predictions and ground truth. In Phase 2, pedestrian ROIs are replaced with counterfactual values (p^{cf}) which are zero tensors, while segmentation maps and past trajectories remain intact. This phase forces the model to depend on structured scene information, enhancing abstract feature interpretation. The loss function includes both binary cross-entropy and an alignment loss to ensure consistency with Phase 1. Phase 3 builds on Phase 2 by replacing segmentation maps with counterfactual values (s^{cf}), also zero tensors, keeping pedestrian ROIs and trajectories unchanged, which refines the model's understanding of pedestrian appearance. Similar to Phase 2, alignment losses

Table 5.1: Performance using different counterfactual values on short term intention prediction

p^{cf}	s^{cf}	Accuracy		
		PIE	**JAAD$_{all}$**	**JAAD$_{beh}$**
Random	Random	0.75	0.79	0.40
Random	Zeros	0.83	0.85	0.52
Zeros	Random	0.88	0.86	0.67
Zeros	Zeros	**0.95**	**0.94**	**0.75**

Note: $\boldsymbol{p^{cf}}$ *and* $\boldsymbol{s^{cf}}$ *denote counterfactual values for pedestrian ROIs and segmentation maps, respectively.*

maintain consistency with Phase 1. Through exposure to these diverse counterfactual scenarios, the methodology significantly improves the model's generalization to new and unseen environments.

The goal of counterfactual training is to enable the model to learn to identify and reason about the impact of missing or altered information, without introducing noise or irrelevant features. Table 5.1 represents the ablation study conducted with different counterfactual values like random noise where values range from [-0.1, 0.1] and zeros as demonstrated. The counterfactual values are also switched between phases to assess their impact on performance.

Our findings revealed that maintaining zero tensors as counterfactual values for both Phase 2: Pedestrian ROIs (p^{cf}) and Phase 3: Segmentation Maps (s^{cf}) yielded the higher performance. Zero tensors represent the absence of the feature, allowing the model to focus solely on learning how the system behaves when that specific feature is absent. In contrast, introducing random values can introduce arbitrary and potentially distracting features, making it harder for the model to identify meaningful patterns, which leads to decreased performance. Based on these observations, zero tensors are chosen to preserve the integrity of the counterfactual training setup, ensuring that the model learns to deal with missing or altered information in a realistic and meaningful way.

5.1.1.2 Trajectory Prediction

The pedestrian's trajectory over the last n timesteps is represented as $\mathcal{U} = \{\mathcal{U}_1, \mathcal{U}_2, \ldots \mathcal{U}_n\}$ with each trajectory point comprising the top-left (x_l, y_l) and bottom-right (x_r, y_r) bounding box coordinates. The objective is to predict m future positions, denoted as $\mathcal{V} = \{\mathcal{V}_{n+1}, \mathcal{V}_{n+2}, \ldots \mathcal{V}_{n+m}\}$. To address this, the proposed model leverages a Variational Autoencoder (VAE) enhanced by a Context-Aware Feature Fusion Module (CAFFM) to accurately predict pedestrian trajectories by integrating both spatial and temporal contextual information as demonstrated in Fig 5.1 (b).

Let $\mathcal{S} = \{\mathcal{U}, \mathcal{V}\}$ be a training dataset of history trajectories $\mathcal{U}$ and future

trajectories $\mathcal{V}$; from a statistical perspective, the goal of multimodal human trajectory prediction is to draw a data distribution $\mathcal{p}(v|u)$ about target data $v \epsilon \mathcal{V}$, where $u \epsilon \mathcal{U}$ is known conditions. An encoder, parameterized by α, takes the input u and produces a distribution. $\mathcal{p}_\alpha(\zeta|u,c)$ where ζ is a latent variable, and c is the learned contextual embeddings. A decoder parameterized by β used u and samples from $\mathcal{p}_\alpha(\zeta|u,c)$ to produce $\mathcal{p}_\beta(y|u,\zeta,c)$. The latent variable is then marginalized out to obtain $\mathcal{p}(v|u)$,

$$\mathcal{p}(v|u,c) = \int \mathcal{p}_\beta(v|u,\zeta,c)\mathcal{p}_\alpha(\zeta|u,c)\, d\zeta \tag{5.1}$$

Variational inference is employed to approximate the intractable integral. A variational distribution $q_\delta(\zeta|u,c)$ to approximate the true posterior $\mathcal{p}_\alpha(\zeta|u)$. The evidence lower bound (ELBO) can be used to optimize the parameters α, β and δ through reconstruction and KL divergence loss:

$$\log \mathcal{p}(v|u,c) \geq \mathbb{E}_{q_\delta(\zeta|u,c)}\left[\log \mathcal{p}_\beta(v|u,\zeta,c)\right] - D_{KL}(q_\delta(\zeta|u,c)||\mathcal{p}_\alpha(\zeta|u,c)) \tag{5.2}$$

The optimization objective is to maximize the ELBO:

$$\mathcal{L}(\alpha,\beta,\delta) = \mathbb{E}_{q_\delta(\zeta|u,c)}\left[\log \mathcal{p}_\beta(v|u,\zeta,c)\right] - D_{KL}(q_\delta(\zeta|u,c)||\mathcal{p}_\alpha(\zeta|u,c)) \tag{5.3}$$

Content-Aware Feature Fusion Module (CAFFM): The modalities, pedestrian ROIs and segmentation maps are essential for capturing specific spatial information about the pedestrians and the contextual layout of the environment. Processing these inputs through transformers conditioned on past motion history allows the model to incorporate temporal aspects crucial for trajectory prediction. The proposed Context-Aware Feature Fusion Module (CAFFM) plays a pivotal role in guiding the Variational Autoencoder (VAE) for trajectory prediction, as shown in Fig. 5.1(b). By integrating raw features with attention-driven enhancements, the module ensures that the VAE is conditioned on a rich and informative context. This improves the accuracy of the predicted trajectories and ensures that the predictions are contextually relevant and dynamically responsive to changes in the environment.

Initially, the module concatenates the features from the MLP heads of transformers A and B, which process pedestrian ROIs and segmentation maps, respectively. These transformers are conditioned on past motion history to capture relevant spatiotemporal information. The concatenated inputs denoted as $a||b$, are passed through a dense layer followed by a $ReLU$ activation function to produce the output γ, expressed mathematically as:

$$\gamma = ReLU(Dense(a||b)) \tag{5.4}$$

The output γ is then divided into two branches. The first branch processes γ through a dense layer followed by a $GELU$ activation function, yielding γ_1:

$$\gamma_1 = GeLU(Dense(\gamma)) \tag{5.5}$$

The second branch processes γ_2 through a dense layer without an activation function:

$$\gamma_2 = Dense(\gamma) \tag{5.6}$$

An attention mechanism is subsequently applied, where the q is derived from γ_2, and the k and value v are derived from γ_1. This attention mechanism can be mathematically represented as:

$$A_{ped} = softmax\left(\frac{q.k^T}{\sqrt{d_k}}\right).v \tag{5.7}$$

The attention output is then multiplied element-wise with γ_2 to produce the attentive features. This can be denoted as:

$$F_{ped} = A_{ped} \odot \gamma_2 \tag{5.8}$$

Finally, a residual connection adds the original output γ back to the attentive features to yield the final refined feature:

$$\hat{F}_{ped} = \gamma + F_{ped} \tag{5.9}$$

The first branch, enhanced with a $GELU$ activation function, introduces non-linearity and captures more complex feature interactions, forming the key and value representations that define the contextual relevance. In contrast, the second branch maintains a more linear and less transformed representation of the features to keep the query closely aligned with the original feature space. This alignment allows the attention mechanism to modulate the linear representation of features effectively γ_2 based on the richer, non-linear context provided by γ_1. The residual connection ensures that the model considers both the raw information and the attention-driven adjustments, preserving long-term context (e.g., overall scene layout) along with dynamic changes captured by the attention mechanism. Consequently, this approach ensures that the refined feature output retains essential information from the original input while being enriched with contextually relevant modifications.

5.1.2 Experimental work and Results

This section outlines the implementation details for the dual-task approach focused on pedestrian intention and trajectory prediction including evaluation metrics and datasets used. It also presents comparisons with state-of-the-art methods and ablation studies on the impact of Progressive Denoising Attention (PDA), counterfactual training, the Context-Aware Feature Fusion Module (CAFFM), and the effects of different contextual embeddings along with memory footprint details.

5.1.2.1 Implementation details

Intention Prediction: The cross-modal transformer employs two projection mechanisms: $Conv1D$ for capturing local temporal relationships from EfficientNet features of pedestrian ROIs and segmentation maps and a Gated Recurrent Unit (GRU) for leveraging temporal dependencies across the past trajectory sequence. The output is then flattened to prepare it for further processing within the transformer. The $PositionalEncoder$ layer incorporates positional information into token embeddings initializing with an embedding dimension of 64. It assigns a unique positional

encoding to each token in the input sequence, ensuring the model can distinguish between tokens based on their position within the sequence.

The PDA utilizes a UNet architecture with a series of convolutional and deconvolutional blocks. The encoder starts with two convolutional blocks, first with 16 filters and then 32 filters, using $Conv1D$ layers with $ReLU$ activation to reduce spatial dimensions while increasing depth. At the core of the network lies a middle block, which further processes the encoded features using convolutions with 64 filters, maintaining the same structure but at a higher level of abstraction. Following this, the decoder part mirrors the encoder, utilizing deconvolutional ($Conv1DTranspose$) layers with $ReLU$ activation to upsample feature maps back to the original dimensions, with each block followed by a $Conv1D$ layer with the same number of filters.

$$L_{align(j)} = \left\| \mathcal{J}_{phase\ j} - \mathcal{J}^{*}_{phase\ 1} \right\|^2 \tag{5.10}$$

where $\mathcal{J}_{phase\ j}$ are the predictions from the j^{th} phase while $\mathcal{J}^{*}_{phase\ 1}$ denotes predictions derived from a model initialized with weights optimized during Phase 1. The asterisk (*) signifies that this model's inputs do not include counterfactuals. This loss ensures that the counterfactual manipulations introduced in later phases do not disrupt the model's learned representations. The total loss function utilized during training is expressed as:

$$L_{total} = \lambda_{BCE} L_{BCE} + \lambda_{align} L_{align} \tag{5.11}$$

Here, λ_{BCE} and λ_{align} are coefficients that balance the contribution of each loss, enabling the model to prioritize accurate predictions while maintaining consistency across phases. This combined loss approach reinforces causal relationships through counterfactual training, enhancing the model's ability to generalize to unseen scenarios. Optimal performance is achieved with $\lambda_{BCE} = 1$ and $\lambda_{align} = 0.36$.

Trajectory Prediction: The encoder architecture for the CVAE processes time-series data and an embedding vector. It begins with two inputs: a sequence of bounding box coordinates and a conditional embedding vector. The bounding box data is processed through two Bidirectional LSTM layers, with 32 units in the first layer and 16 units in the second, capturing both past and future dependencies while reducing dimensionality. The encoded sequence is concatenated with the embedding vector, forming a combined feature set. This set is further refined through two Dense layers with 256 and 128 units, respectively, activated by ReLU. Dropout and batch normalization are applied to improve generalization. The network outputs two Dense layers representing the mean and log variance of the latent space distribution, each with 64 units. These parameters are passed to a Lambda layer for reparameterization, producing the latent variable h.

The decoder architecture is designed to reconstruct sequences from a latent representation and conditional embedding. The inputs are concatenated first and processed through two Dense layers with 128 and 256 units, respectively, both activated by $ReLU$. Dropout and batch normalization are applied after each Dense layer to enhance generalization. The processed features are then replicated across 45 timesteps using a $RepeatVector$ layer, preparing the data for sequence generation. The sequence is generated through two $LSTM$ layers, with 16 units in the first and 32 units in the second, each set to return sequences. Finally, a $TimeDistributed\ Dense$ layer with 4 units and a linear activation function is applied to reconstruct the output sequence.

The intention prediction model is trained independently; however, the trajectory prediction model utilizes the pretrained intention model to generate the embedding vector. The intention and trajectory models are trained using the RMSProp optimizer with learning rates of 10^{-5} and 10^{-2}, respectively. The intention model is trained for 100 epochs with a batch size of 128 and L2 regularization of 0.001. The trajectory model is trained for 60 epochs with a batch size of 64 and L2 regularization of 0.0001. All the experiments are conducted on a Google Colab Pro instance with

Table 5.2: Deterministic Results on PIE/JAAD Dataset

(**CMSE** and **CFMSE** are the mean square error between the predicted and ground truth centres of bounding boxes, over all future time steps and the final predicted time step, respectively)

Methods	PIE					JAAD				
	MSE			C_{MSE}	CF_{MSE}	MSE			C_{MSE}	CF_{MSE}
	0.5s	*1s*	*1.5s*	*1.5s*	*1.5s*	*0.5s*	*1s*	*1.5s*	*1.5s*	*1.5s*
PIE_traj[6]	58	200	636	596	2477	110	399	1248	1183	4780
BiTraP[22]	41	161	511	481	1949	93	378	1206	1105	4565
SGNet[23]	34	133	442	413	1761	82	328	1049	996	4076
MlgtNet[156]			459	418	1629			1002	938	3489
PCTP-AGFL[157]	12	75	300	223	1299	35	205	825	784	3383
Ours	**8**	**71**	**225**	**201**	**1132**	**34**	**199**	**789**	**756**	**3121**

Table 5.3: Stochastic Results on PIE/JAAD Dataset

Methods	PIE					JAAD				
	MSE			C_{MSE}	CF_{MSE}	MSE			C_{MSE}	CF_{MSE}
	0.5s	*1s*	*1.5s*	*1.5s*	*1.5s*	*0.5s*	*1s*	*1.5s*	*1.5s*	*1.5s*
BiTraP(GMM)[22]	38	90	209	171	368	53	250	585	501	998
BiTraP(NP)[22]	23	48	102	81	261	38	94	222	177	565
SGNet[23]	16	39	88	66	206	37	86	197	146	443
PCTP-AGFL[157]	6	21	59	45	138	19	55	147	105	301
Ours	**5**	**16**	**51**	**42**	**128**	**11**	**40**	**126**	**99**	**289**

access to an NVIDIA Tesla T4 GPU (16 GB memory), running on the CUDA 12.0 platform. The implementation is done using TensorFlow 2.10.1.

5.1.2.2 Datasets

Intention: The proposed method is evaluated using the JAAD [137] and PIE [6] benchmark datasets. JAAD includes 346 high-resolution video clips of urban driving scenarios, with two subsets: $JAAD_{all}$ (2,100 visible pedestrians not near crossings) and $JAAD_{beh}$ (495 crossings and 191 non-crossings). PIE offers a broader dataset with 1,842 roadside sections at 30 Hz, including 519 crossings, 1,323 non-crossings, and ego-vehicle speed annotations. Both datasets follow the recommended training/validation/test split for comprehensive evaluation [6], [137]. It is evaluated using standard classification metrics: Accuracy, AUC, F1 score, Precision, and Recall.

Table 5.4: Performance of the proposed method on short term intention prediction on PIE dataset

Methods	PIE				
	Acc	**AUC**	**F1**	**Prec**	**Rec**
PedGNN[82]	0.71	-	0.75	0.83	0.79
PIE_traj[6]	0.79	-	0.87	-	-
TAMFORMER[85]	0.87	0.84	0.76	-	-
IPIPF[158]	0.88	0.85	0.80	0.82	0.78
V-PedCross[86]	0.89	0.88	0.67	0.74	0.84
PG+[87]	0.89	0.90	0.81	0.83	0.79
TED[103]	0.91	0.91	0.83	-	-
VMI[91]	0.92	0.91	0.87	0.86	0.88
Biped84]	0.92	0.91	0.86	0.83	-
MTMGN[159]	0.90	0.87	0.92	0.95	0.90
TrEP[105]	0.93	0.94	0.87	0.89	0.88
PedFormer[93]	0.93	0.90	0.87	0.89	0.88
IntentFormer[160]	0.93	0.90	0.88	0.86	0.89
Ours	**0.95**	**0.94**	**0.92**	0.94	**0.93**

Table 5.5: Performance of the proposed method on short term intention prediction on JAADall/JAADbeh dataset

Methods	$JAAD_{all}/JAAD_{beh}$				
	Acc	**AUC**	**F1**	**Prec**	**Rec**
FFSTA[18]	0.83/0.62	0.82/0.54	0.63/0.74	0.51/0.650	0.81/0.85
Biped[84]	0.84/-	0.79/-	0.61/-	0.54/-	-
V-PedCross[86]	-/0.64	-/0.66	-/0.76	-/0.70	-/0.89
PG+[87]	0.86/0.70	0.88/0.70	0.65/0.76	0.58/0.77	0.75/0.75
IPIPF[158]	0.86/-	0.84/-	0.69/-	0.74/-	0.66/-
TAMFORMER[85]	0.89/0.73	0.82/0.70	0.70/0.79	-	-
VMI[91]	0.89/-	0.90/-	0.81/-	0.79/-	0.83/-
MTMGN[159]	0.89/0.70	0.89/0.70	0.73/0.83	0.66/0.79	0.89/0.87
TrEP[105]	0.91/-	0.86/-	0.69/-	0.71/-	0.70/-
PedFormer[93]	0.93/-	0.76/-	0.54/-	0.65/-	0.60/-
IntentFormer[160]	0.92/0.75	0.90/0.70	0.83/0.82	0.81/0.74	0.85/0.88
Ours	**0.94/0.75**	**0.91/0.71**	**0.81/0.85**	**0.80/0.81**	0.82/**0.89**

Trajectory: Trajectory prediction is assessed using MSE over bounding box coordinates and C_{MSE} and CF_{MSE}, which measure the MSE of the bounding box centre over the entire sequence and the final time step, respectively. All metrics for the JAAD and PIE datasets are reported in pixels. It is assessed using MSE over bounding box coordinates and C_{MSE} and CF_{MSE}, which measure the MSE of the bounding box centre over the entire sequence and the final time step, respectively. All metrics for the JAAD and PIE datasets are reported in pixels. JAAD [137] features 2,800 pedestrian trajectories captured at 30 Hz, divided into 177 training, 117 testing, and 29 validation clips, using a 0.8 overlap ratio for sampling. PIE [6] contains 880, 243, and 719 pedestrian

tracks in the training, validation, and test sets, respectively, with a 0.5 overlap ratio, excluding tracks shorter than 2 seconds during trajectory prediction training.

5.2.2.3 Comparison with SOTA methods

The proposed method exhibits superior performance in short-term intention prediction across multiple datasets, including PIE, $JAAD_{all}$, and $JAAD_{beh}$. On the PIE dataset (Table 5.2), our work achieves the highest accuracy (0.95) and an AUC of 0.94, comparable to leading model IntentFormer[160]. With an F1 score of 0.92, the method surpasses all other approaches in precision (0.94) and recall (0.93), demonstrating a robust and reliable solution for pedestrian intention prediction in dynamic environments. Similarly, on the $JAAD_{all}$ dataset (Table 5.3), the proposed method attains the highest accuracy (0.94) and AUC (0.91), outperforming models like PedFormer[93]. On the $JAAD_{beh}$ dataset, our work matches the highest accuracy (0.75) and achieves the top F1 score (0.85) with strong precision (0.81) and recall (0.89). Compared to other methods, the proposed method consistently demonstrates superior performance, particularly in challenging conditions, underscoring its robustness and reliability across diverse scenarios.

The trajectory prediction assessment is performed under two distinct settings: deterministic, where a single trajectory is predicted, and stochastic, where a set of $K = 20$ potential trajectories is generated, with the best-performing sample reported. The proposed approach significantly improves deterministic trajectory prediction, as detailed in Table 5.4. On the PIE dataset, the method achieves a 25% reduction in MSE at the 1.5-second interval compared to PCTP-AGFL[157], and demonstrates a 10% improvement in C_{MSE} and a 13% improvement in CF_{MSE}, indicating superior accuracy over extended prediction periods. Similarly, on the JAAD dataset, the method reports an 18.4% reduction in MSE, a 3.6% improvement in C_{MSE}, and a 7.7% reduction in CF_{MSE} compared to PCTP-AGFL [157].

Table 5.5, which presents results for stochastic trajectory prediction, further highlights that the proposed approach achieves an average reduction of 16.3% in MSE

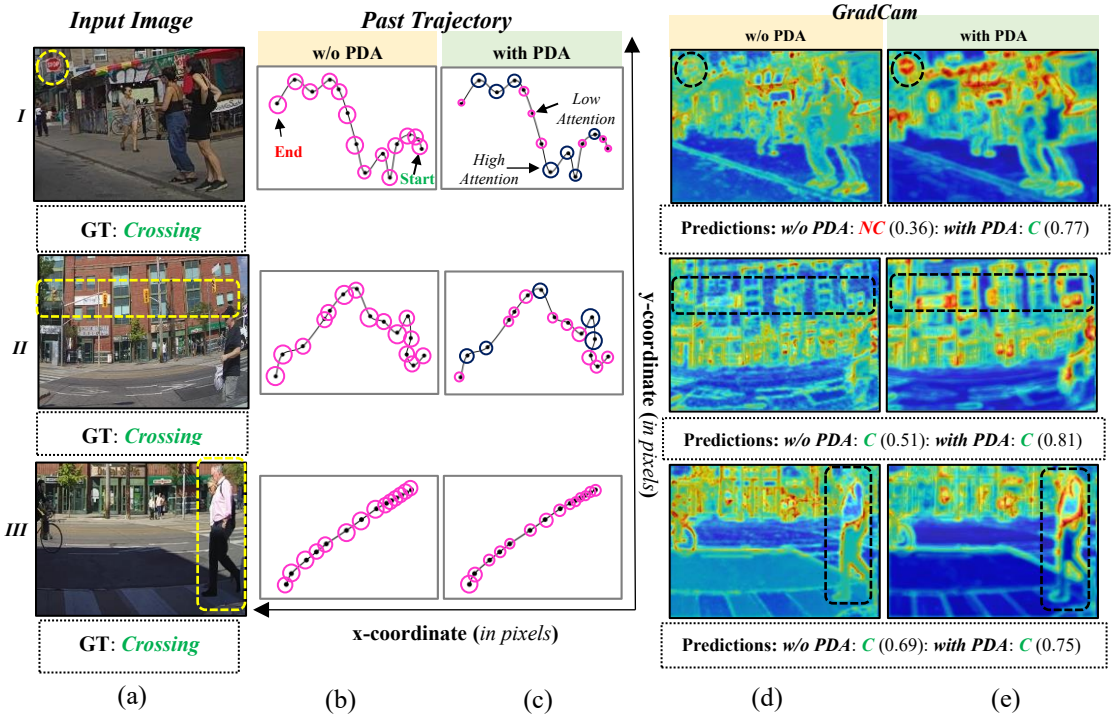


Fig. 5.3: Case studies illustrating how attention mechanisms influence prediction outcomes across three scenarios (Rows I-III).

on the PIE dataset and 20.2% on the JAAD dataset relative to PCTP-AGFL. Additionally, the approach shows an average improvement of 7% in CMSE and CFMSE combined on the PIE dataset and 4.8% on the JAAD dataset. These findings collectively underscore the effectiveness of the proposed work in reducing prediction errors and enhancing overall performance across a range of datasets.

5.2.2.4 Ablation Study

This section presents an ablation study to evaluate the impact of key components in the proposed framework. The effectiveness of the progressive denoising mechanism and counterfactual training is examined, along with an analysis of memory footprint and computational complexity. Additionally, the role of alignment loss is investigated to assess its contribution to model performance. These analyses offer deeper insights into the trade-offs and benefits of the proposed design choices. The analyses are as follows.

i. **PDA**: From a cognitive perspective, pedestrians adjust their behaviour based on environmental cues and their attributes. The proposed Progressive Denoising

Attention analyses pedestrian crossing intentions on the road, leveraging iterative refinement of attention scores based on historical motion data, pedestrian visual appearance, and semantic scene features. Fig. 5.3(a) illustrates input scenes with salient contextual cues highlighted using yellow dashed lines; (b–c) Temporal attention weights along the pedestrian's past trajectory where solid dots denote discrete time steps and circles radii indicate corresponding attention weights, shown without PDA (b) and with PDA (c). Attention weights corresponding to sharp turns and directional changes are highlighted in black. (d–e) Grad-CAM visualisation without (d) and with PDA (e). Ground truth and predicted pedestrian intention labels: Crossing (***C***) or Not Crossing (***NC***) are shown along with associated confidence scores in each row. Rows I and II depict scenarios involving pedestrian interaction with traffic infrastructure such as stop signs and traffic lights. These environmental cues result in changes in pedestrian motion—such as halts, starts, or turning behaviour which is clearly visible in the past trajectory segment over the last 15 timesteps (Fig. 5.3(b–c)). Without PDA, the temporal attention weights assigned by the cross-modal transformer remain relatively uniform, showing little sensitivity to such behavioural transitions. In contrast, with PDA, higher attention weights are allocated specifically to the turning or decision-critical points along the path, reflecting the model's increased responsiveness to contextual cues (Fig. 5.3(c)). The Grad-CAM maps further demonstrate that, without PDA (Fig. 5.3(d)), the attention tends to diffuse across less relevant areas, reducing the alignment between visual cues and behavioural outcomes. However, in the presence of PDA (Fig. 5.3(e)), the model concentrates more accurately on semantically meaningful regions—specifically traffic signals and pedestrian appearance. This shift in focus leads to improved predicted label accuracy and higher confidence scores.

Row III in Fig. 3 presents a scenario involving an elderly pedestrian following a smooth and linear trajectory. Here, the temporal attention weights remain uniform in both model variants (Fig. 5.3(b-c)), reflecting the low variability and predictability of motion typically associated with elderly individuals. The Grad-CAM visualizations (Fig. 5.3(d–e)) show that, in both cases, the spatial focus remains consistently centred on the pedestrian, suggesting minimal dependence on

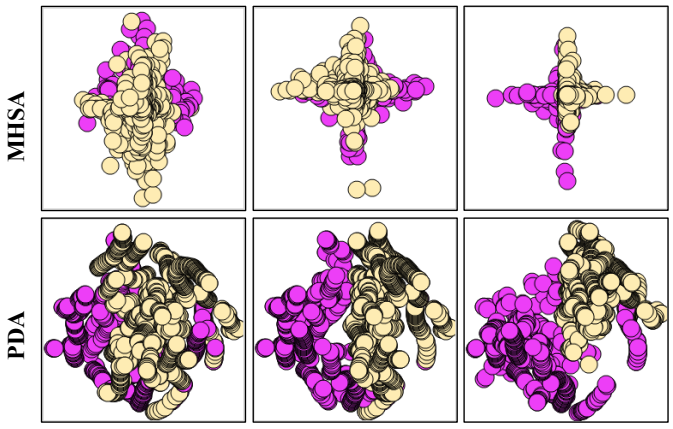


Fig. 5.4: t-SNE embeddings of the attention outputs from MHSA (Row I) and PDA (Row II)

additional contextual cues. Accordingly, both models yield similar prediction outcomes and confidence levels.

The t-SNE embeddings of attention outputs for MHSA (Row I) and PDA (Row II) are presented in Fig. 5.4. In the case of MHSA, the embeddings initially appear crowded (Row-I (a)), with some separation emerging in later epochs, as shown in Row-I (b) and (c). However, this separation remains poorly defined, potentially leading to less confident predictions. In contrast, interclass variation increases for PDA as training progresses, and by the final epoch, the distinction between the two classes becomes significantly more apparent. This illustrates PDA's iterative refinement process, where the model dynamically adjusts its

Table 5.6: Impact of iterative denoising and number of iterations on convergence and performance in PDA-based cross-modal feature refinement

Ablations	Attention	τ	$\mathcal{N}$	$\mathcal{E}$	T(mins)	Accuracy
A_1	SHA(*Baseline*)	--	×	200	4	86
A_2	PDA	1	×	180	5	88.5
A_3	PDA	3	×	150	6.5	89.2
A_4	PDA	5	×	130	8.5	90
A_5	PDA	10	×	110	12	90.5
A_6	PDA	Dynamic	×	105	10	91
A_7	PDA	1	✓	160	5.5	90
A_8	PDA	3	✓	130	7	91.5
A_9	PDA	5	✓	110	9	92
A_10	PDA	10	✓	100	13	92.8
A_11	PDA	Dynamic	✓	100	8	95

τ=Total iterations per step; T= Average training time/epochs; $\mathcal{E}$ = Total number of epochs to convergence: $\mathcal{N}$: Noise Injection

attention outputs over multiple steps, enhancing prediction accuracy and ensuring efficient convergence.

Table 5.6 shows the impact of PDA on training time and model performance by varying the number of iterations (τ) and the presence of noise injection. The introduction of PDA significantly enhances accuracy while reducing the total number of epochs required for convergence ($\mathcal{E}$). The baseline model (A_1) requires 200 epochs to achieve 86.0% accuracy, whereas PDA in its optimal configuration (A_11) reduces $\mathcal{E}$ by 50% (100 epochs) while improving accuracy by 10%, demonstrating its effectiveness in cross-modal feature representation.

The impact of τ on training efficiency is evident in Table 5.6 where increasing τ generally improves accuracy but also raises the average training time per epoch (T). For instance, A_5 ($\tau = 10$, no noise) achieves 90.5% accuracy but requires 12 min/epoch, whereas A_3 (τ = 3, no noise) reaches 89.2% accuracy at a reduced computational cost of 6.5 min/epoch. However, dynamic iteration control, as implemented in A_6 and A_11, consistently outperforms fixed τ settings by achieving better accuracy with lower training overhead. Specifically, A_6 (dynamic τ, no noise) converges in 105 epochs, reaching 91.0% accuracy with $T = 10$ min/epoch, demonstrating improved efficiency.

Noise injection (N) further enhances accuracy while maintaining efficiency. Comparing A_3 (τ = 3, no noise) and A_8 (τ = 3, noise) in Table 5.6, the latter achieves 2.6% higher accuracy with only a 0.5 min increase in T, highlighting its role in improving cross-modal alignment. This suggests that noise injection helps refine feature representations while adding minimal computational cost.

The optimal PDA configuration, A_11 (dynamic τ, noise ✓), achieves 95.0% accuracy, reduces $\mathcal{E}$ to 100 epochs, and maintains T at 8.0 min/epoch, making it the most effective balance between computational cost and performance. While PDA introduces additional computational complexity per epoch, its ability to accelerate convergence offsets this overhead, demonstrating its efficiency in cross-modal feature refinement.

ii. **Counterfactual Training**: In this study, an ablation analysis is conducted to evaluate the effectiveness of a three-phase counterfactual training methodology in

Fig. 5.5: Qualitative Samples: Crossing intention confidence scores for NM single-phase training vs. three-phase training with counterfactual samples. ***Green***: crossing, ***Red***: non-crossing.

Table 5.7: Performance metrics across different phases of counterfactual training with and without PDA

Training Modes	PDA	PIE			$JAAD_{all}$/JAADbeh		
		Accuracy	AUC	F1	Accuracy	AUC	F1
Phase-1 (NM)	×	0.85	0.86	0.80	0.87/0.67	0.83/0.66	0.75/0.76
Phase-2 (CF^R)	×	0.88	0.89	0.85	0.87/0.68	0.86/0.67	0.75/0.80
Phase-3 (CF^S)	×	0.86	0.90	0.88	0.89/0.71	0.87/0.68	0.76/0.78
Phase-1 (NM)	✓	0.94	0.93	0.83	0.93/0.74	0.91/0.70	0.74/0.79
Phase-2 (CF^R)	✓	0.94	0.94	0.89	0.94/0.73	0.90/0.71	0.77/0.80
Phase-3 (CF^S)	✓	0.95	0.94	0.92	0.94/0.75	0.91/0.71	0.81/0.85

enhancing pedestrian intention prediction. The methodology is designed to improve the model's robustness to challenging visual conditions, such as blurred and noisy images, by encouraging a deeper understanding of contextual cues. As illustrated in Fig. 5.5, the model trained with the three-phase counterfactual approach achieves significantly higher confidence scores than a conventional single-phase training model. This suggests that the counterfactual training enhances the model's focus on causal relationships among various contextual elements while reducing reliance on compromised visual information, thus lowering the risk of overfitting to specific noise patterns.

Table 5.7 further supports these findings by comparing model performance across different training phases, specifically on the PIE and $JAAD_{all}$/$JAAD_{beh}$ datasets. The results show a clear and consistent improvement in Accuracy, AUC, and F1 scores from Phase-I (normal training) to Phase-3 (counterfactual training with segmentation maps). In Phase-I, the model establishes a baseline performance but struggles with more complex scenarios, as indicated by the relatively lower F1 scores. However, in Phase-2, where counterfactual training with pedestrian ROIs is introduced, there is a notable enhancement in all metrics, reflecting a refined

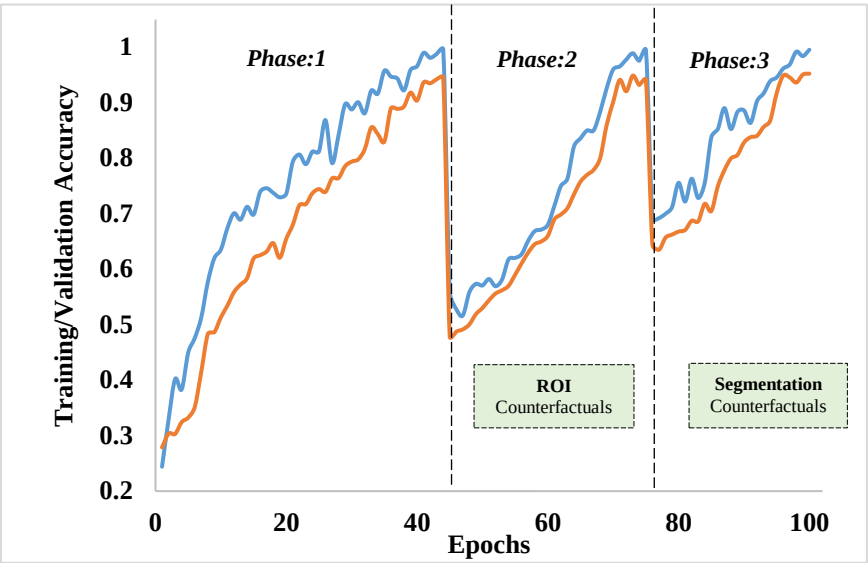


Fig. 5.6: Training (***blue***) and validation (***orange***) accuracy over epochs during counterfactual training.

understanding of contextual elements and improved robustness to variations in pedestrian appearances. Finally, Phase-3, with the implementation of counterfactual training using segmentation maps, leads to significant improvements in the F1 score (PIE: 0.92; $JAAD_{all}$/$JAAD_{beh}$=0.81/0.85) and incremental gains in other performance metrics. These results show that incrementally adding counterfactual scenarios during training significantly improves the model's resilience to real-world data challenges.

Fig. 5.6 illustrates the training progression of the proposed intention prediction model across different phases. Training progresses in each phase until the validation accuracy plateaus. Early stopping is triggered when the validation accuracy improvement is ≤2% over five consecutive epochs. Phase transitions are denoted by vertical dashed lines. Introducing counterfactuals in Phases 2 and 3 leads to temporary accuracy dips, but the final validation accuracy stabilizes at 95% in Phase 3, indicating successful convergence.

The transition from Phase 1 (Baseline) to Phase 2 (ROI Counterfactuals) occurs at epoch 44, when validation accuracy stabilizes at 94% with no further improvement. The transition from Phase 2 to Phase 3 (Segmentation Counterfactuals) occurs at epoch 75, following validation accuracy stabilization at 94% after introducing ROI counterfactuals. Training concludes at epoch 100, when validation accuracy plateaus at 95%, reflecting the model's adaptation to

Table 5.8: Evaluation of trajectory prediction performance using different contextual embeddings and fusion strategies

Training Mode	Contextual Embeddings	Fusion Strategy	MSE	
			PIE	JAAD
Normal	No context	-	390	887
	Only RGB	-	350	862
	Only Segmentation	-	369	859
	RGB+ Segmentation	Concatenation	350	824
	RGB+ Segmentation	CAFFM	333	803
Counterfactual	No context	-	389	885
	Only RGB	-	290	795
	Only Segmentation	-	320	850
	RGB+ Segmentation	Concatenation	250	800
	RGB+ Segmentation	CAFFM	225	789

segmentation counterfactuals. Notably, the training convergence in Phases 2 and 3 occurs more quickly than in Phase 1. Furthermore, a 5–6% improvement is observed in performance metrics, due to PDA's iterative refinement, resulting in more precise and confident predictions in complex cross-modal scenarios.

iii. **CAFFM**: The impact of different contextual embeddings and fusion methods on trajectory prediction is analysed on the PIE and JAAD datasets, as summarized in Table 5.8. The results indicate that the model's performance improves by including contextual information and using the Context-Aware Feature Fusion Module (CAFFM). Notably, the CAFFM achieves the lowest MSE across both datasets, with 333 on PIE and 803 on JAAD in the baseline model embeddings and further reduction to 225 on PIE and 789 on JAAD under counterfactual model embeddings. This suggests that the CAFFM effectively leverages spatial and temporal contexts, enhancing the accuracy of trajectory prediction.

The analysis reveals that the counterfactual training significantly improves the model's performance. When no context is used, the baseline models show higher MSE values (MSE(PIE): 389, MSE(JAAD): 885), indicating lower prediction accuracy. Incorporating RGB and segmentation embeddings separately reduces the MSE, showing that each modality contributes valuable contextual information. The concatenation of these embeddings further improves the performance, suggesting a more comprehensive representation of the scene. The most substantial performance gains are observed when using the CAFFM embeddings derived from the counterfactual model. The MSE values decrease significantly (MSE(PIE): 225,

Table 5.9: Comparison of computational efficiency of DPITRA-short term intention model with SOTA methods

Model	Size (MB)	Inference time(ms)	Accuracy		
			PIE	$JAAD_{beh}$	$JAAD_{all}$
PCPA[102]	118.8	38.6	86	50	70
FFSTA[18]	374.2	70.83	-	62	83
PG+[93]	0.28	5.47	89	70	86
TED[109]	12.8	2.76	91	-	-
V-PedCross[92]	4.8	-	89	64	86
PedGNN[88]	**0.027**	**0.58**	70.52	-	86.22
VMI[97]	19.07	11.03	92	-	89
IntentFormer[162]	2.13	3.8	93	**75**	92
DPITRA	4.46	2.53	**95**	**75**	**94**

Table 5.10: Comparison of computational efficiency of DPITRA-long term with SOTA methods

Model	IT(*ms*) 20/2000	MSE(1.5s) JAAD/PIE
PCTP-AGFL[163]	84/87	147/59
DPITRA	80/82	126/51

Table 5.11: Inference Time Per Batch Breakdown for Trajectory Prediction

Batch Size	Intention Prediction Module	CAFFM	CVAE Encoder	CVAE Decoder	Total Time
20	5.05 ms	4.99 ms	20.22 ms	50.19 ms	80.15 ms
500	5.06 ms	5.01 ms	20.73 ms	50.28 ms	81.08 ms
2000	5.07 ms	5.03 ms	21.01 ms	50.99 ms	82.10 ms

MSE(JAAD): 789. This suggests that the embeddings generated from the counterfactual training capture richer, more nuanced information, which enhances the VAE's ability to produce accurate trajectory predictions.

iv. **Memory footprint and computational complexity**: The proposed dual-task approach for intention and trajectory prediction effectively balances memory usage and inference speed. The model has a total memory footprint of 14.41 MB, including a 4.46 MB short-term intention module. This short-term model delivers a peak accuracy of 95% on the PIE dataset, outperforming smaller footprint models like PedGNN[88] (accuracy: 70.52%) and PG+[93] (accuracy: 89%) while maintaining a minimal inference time of 2.53 ms as shown in Table 5.9.

Furthermore, the trajectory prediction model exhibits minimal time variation between processing 20 and 2000 samples, with an inference time of 80 and 82ms, respectively, demonstrating superior efficiency compared to models like PCTP-

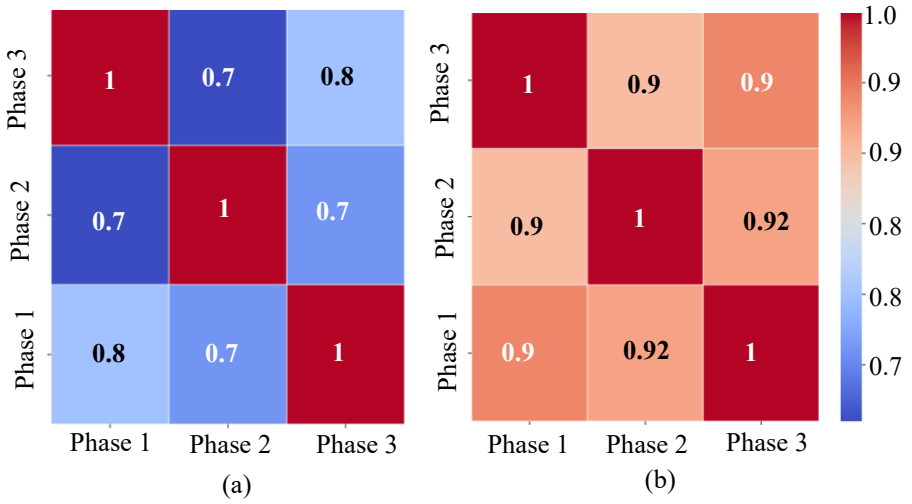


Fig. 5.7: Prediction Correlation Matrix demonstrating the role of alignment loss in counterfactual training. (a) Without alignment loss. (b) With alignment loss

AGFL[157] as reported in Table 5.10. Additionally, Table 5.11 provides a component-wise breakdown of inference time per batch for trajectory prediction. The observed consistency in per-batch inference time, regardless of sample size, can be attributed to the use of custom TensorFlow data pipeline that minimizes data loading and preprocessing overhead while maximizing GPU utilization. This setup enables efficient parallel processing of all samples within a batch, preventing computational overhead from scaling with batch size.

v. **Role of Alignment loss:** Alignment loss enforces consistency by penalizing deviations between predictions from later phases (Phases 2 and 3) and the baseline Phase 1. This regularization ensures that the counterfactual manipulations introduced in later phases do not disrupt the model's learned representations. The model maintains stable and coherent predictions across phases by minimising the alignment loss, enabling it to generalize better to unseen scenarios. Fig. 5.7 illustrates the effect of alignment loss using prediction correlation matrices. Panel (a) shows the correlation between predictions across phases when alignment loss is excluded. The lower correlation values indicate inconsistencies in predictions across phases. In contrast, panel (b) displays the correlation matrix when alignment loss is included. The significantly higher correlation values demonstrate that alignment loss maintains consistent predictions, even with counterfactual modifications in Phases 2 and 3. These results validate the role of alignment loss in

maintaining consistency in predictions across phases, enhancing the model's ability to generalize to unseen data, and improving both robustness and reliability.

5.2 Conclusion and Future Scope

The work presents a dual-task approach excels in short-term pedestrian intention and long-term trajectory forecasting, as demonstrated by its superior benchmark performance. The iterative refinement through Progressive Denoising Attention (PDA) enhanced the inter-class separation between crossing and non-crossing samples, improving prediction accuracy. Moreover, the three-phase counterfactual training improved significantly on noisy and blurred samples. The Context-Aware Feature Fusion Module (CAFFM) embeddings further reduced MSE in trajectory predictions by leveraging spatial and temporal information from pedestrian ROI and scene context. The proposed model also achieved an optimal balance between performance and computational complexity, surpassing existing solutions while maintaining a minimal inference time. Future work should focus on further minimizing the model's memory footprint and computational complexity without compromising performance.

CHAPTER 6

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

In this thesis, we addressed two key tasks: short-term pedestrian intention prediction, specifically crossing behaviour, and long-term trajectory forecasting in complex urban environments. The comprehensive methodologies introduced are robust and demonstrate strong predictive performance across both temporal horizons. A detailed evaluation confirms the potential of the proposed approach for deployment in safety-critical applications such as autonomous driving and intelligent transportation systems. Section 6.1 presents a summary of the contributions made in this thesis, followed by a discussion on future research directions in Section 6.2 and the broader societal impact of this work in Section 6.3.

6.1 Summary of the Work Done in the Thesis

This thesis presented four major approaches to pedestrian intention prediction, each addressing distinct challenges across short-term intention recognition and long-term trajectory forecasting. Together, these approaches contribute to a comprehensive understanding of pedestrian behaviour in complex, real-world environments.

The first approach introduced a multimodal pedestrian intention prediction framework that adaptively fuses rich visual, motion, and interaction features. By applying attention mechanisms across spatial, channel, and temporal dimensions, and incorporating a novel Multi-Head Attention with Adjacency Matrix-based Graph Convolutional Network (MHA-AdjMat GCN) in the interaction encoder, the model significantly enriched pedestrian feature representations. This framework demonstrated superior performance in predicting pedestrian crossing intentions up to 2.5 seconds in advance on the JAAD and PIE datasets, outperforming several state-of-the-art (SOTA) baselines. Despite its effectiveness, the model exhibited limitations in capturing high-frequency temporal dependencies, a common challenge when using GCNs in dynamic sequence modelling.

The second approach proposed a novel transformer-based architecture, ‘IntentFormer’, which predicts pedestrian crossing intentions in a co-learning environment. This architecture integrates RGB features, segmentation maps, and pedestrian trajectories, enabling robust multimodal learning. Three key innovations characterize this model: a shared-MLP head for collaborative co-learning, Multi-Head Shared Weights Attention (MHSWA) for efficient inter-modal representation learning, and a Co-learning Adaptive Composite (CAC) loss function designed to reduce overfitting by penalizing intermediate prediction errors. ‘IntentFormer’ performs optimally within a 0.5 to 1.25-second observation window, requiring fewer frames while maintaining high time-to-event (TTE) accuracy. Nonetheless, the model encounters challenges in scenarios involving abrupt or erratic pedestrian behaviors—such as sudden direction changes or variable speeds—limiting its robustness in highly dynamic environments.

In the third approach, a Progressive Contextual Trajectory Prediction framework with Adaptive Gating and Fuzzy Logic Integration (PCTP-AGFL) was developed to address the complexity of long-term trajectory prediction. Evaluated on both first-person view (FPV) and bird’s eye view (BEV) datasets, the proposed model demonstrated its ability to accurately emulate complex trajectory patterns and predict final destinations, achieving a significantly lower mean squared error compared to existing methods. This framework effectively tackles overfitting and generalization issues, which are common in trajectory forecasting. Additionally, the integration of the Adaptive Fuzzified Discriminator (AFD) improves performance in ambiguous scenarios by enhancing the model's ability to distinguish subtle variations in motion intent.

Finally, the thesis introduced a unified dual-task framework capable of jointly performing short-term pedestrian intention prediction and long-term trajectory forecasting. The proposed model achieved strong benchmark performance through iterative refinement enabled by the Progressive Denoising Attention (PDA)

mechanism, which enhanced inter-class separation between crossing and non-crossing intentions. The incorporation of a three-phase counterfactual training strategy further improved the model's robustness, especially when dealing with noisy or blurred visual inputs. Furthermore, the Context-Aware Feature Fusion Module (CAFFM) leveraged spatial and temporal cues from pedestrian regions of interest (ROIs) and the surrounding scene, substantially reducing prediction error while maintaining computational efficiency. This approach achieves a balance between accuracy, memory footprint, and inference speed, making it highly suitable for real-time deployment in intelligent transportation systems.

In summary, the methodologies proposed in this thesis advance the field of pedestrian intention prediction by effectively capturing the complex interactions between pedestrians and their surrounding environment, enabling accurate short-term crossing intention recognition and long-term trajectory forecasting in dynamic traffic scenes.

6.2 Future Research Scope

Building upon the advancements made in this study, several key research directions can be pursued to further refine pedestrian intention prediction and its applications in autonomous navigation. The integration of Reinforcement Learning (RL) presents a promising avenue for enhancing the adaptability of multimodal pedestrian intention models. While the proposed Co-Learning Transformer and Interaction Encoder effectively capture pedestrian-environment interactions, incorporating RL-based mechanisms can enable adaptive decision-making in dynamic and unseen scenarios, improving the ability of autonomous vehicles (AVs) to respond to unpredictable pedestrian movements.

Another crucial direction is the expansion of datasets to include diverse urban and rural settings, varied weather conditions, and cultural contexts. The current study has demonstrated strong performance across benchmark datasets; however, models often struggle with generalization due to limited dataset diversity. Extending training data to encompass a broader spectrum of pedestrian behaviours, environmental

influences, and scene complexities can significantly enhance robustness. The proposed counterfactual training approach and context-aware feature fusion techniques offer a foundation for handling missing and noisy data, which can be further extended to adapt models to diverse real-world conditions.

For real-time applications, efficient feature extraction remains a key challenge. While the proposed Multi-Head Shared Weight Attention Mechanism and Progressive Denoising Attention (PDA) have optimized inference time and computational efficiency, further improvements can be made to ensure real-time deployment in AV systems. Future research can explore lightweight, hardware-efficient feature extraction techniques that reduce computational load while preserving accuracy, making pedestrian intention models more practical for real-world AV implementation.

Additionally, real-time scene semantic map generation can significantly improve contextual awareness in pedestrian prediction. The Encoder-Decoder Contextual Attention (EDCA) mechanism and Interaction Encoder (IE) with Graph Convolutional Networks have shown effectiveness in modelling pedestrian interactions, but incorporating real-time scene understanding through dynamic semantic mapping can further enhance decision-making capabilities. By integrating spatial-temporal pedestrian behaviours with road semantics, traffic signals, and environmental cues, models can achieve higher predictive accuracy and adaptability in complex urban environments.

Lastly, lightweight architectures optimized for AV hardware are essential to ensure the seamless integration of pedestrian intention models into autonomous navigation systems. While the proposed Multimodal IntentFormer and Dynamic Progressive Generator (DPG) with Adaptive Fuzzified Discriminator (AFD) have successfully minimized model complexity without sacrificing performance, further advancements in model compression, quantization, and efficient transformer-based architectures can improve inference speed and energy efficiency. Optimizing models to operate under real-world AV constraints will ensure their practical applicability, enabling safer and more intelligent pedestrian-aware navigation. By addressing these

future directions, pedestrian intention prediction can be further refined, leading to more robust, adaptable, and computationally efficient models that enhance the safety and decision-making capabilities of AVs in real-world environments.

6.3 Social Impact

Beyond safety, the proposed multimodal intention prediction frameworks also have profound implications for urban mobility and traffic efficiency. With the rise of smart cities and intelligent transportation networks, integrating pedestrian behaviour prediction into traffic management systems, crosswalk automation, and vehicle-to-infrastructure (V2I) communication can lead to smoother traffic flow, reduced congestion, and optimized pedestrian crossings. The ability to accurately predict pedestrian intent ensures that AVs and human-driven vehicles can coexist more harmoniously, minimizing abrupt stops, reducing fuel consumption, and lowering carbon emissions associated with traffic inefficiencies.

Moreover, the focus on lightweight architectures and real-time deployment ensures that these solutions are accessible and scalable. Many regions, particularly in developing countries, struggle with the adoption of high-end autonomous technologies due to hardware and computational constraints. By optimizing model efficiency without compromising accuracy, this research ensures that pedestrian safety solutions can be deployed in a wide range of settings, including low-cost AVs, public transportation systems, and surveillance networks, making roads safer for all pedestrians, regardless of technological infrastructure.

This work also has broader applications in assistive technologies. The ability to predict human movement and intent can be leveraged for mobility assistance in elderly care facilities, smart navigation for visually impaired individuals, and robotic assistance in crowded public spaces. The counterfactual training approach and context-aware feature fusion developed in this research enhance the robustness of human motion understanding, which can be extended to improve human-robot interaction, healthcare monitoring, and public safety surveillance in various social contexts.

REFERENCES

[1] D. M. Research, “Autonomous Vehicles Market is expected to reach a revenue of USD 337.2 Bn by 2033, at 20.2% CAGR: Dimension Market Research,” GlobeNewswire News Room. Accessed: Mar. 28, 2025. [Online]. Available: https://www.globenewswire.com/news-release/2024/11/18/2983034/0/en/Autonomous-Vehicles-Market-is-expected-to-reach-a-revenue-of-USD-337-2-Bn-by-2033-at-20-2-CAGR-Dimension-Market-Research.html

[2] A. Rasouli, "The Role of Context in Understanding and Predicting Pedestrian Behavior in Urban Traffic Scenes," Ph.D. dissertation, Department of Electrical Engineering and Computer Science, York University, Toronto, Ontario, Canada, 2020. [Online]. Available: https://yorkspace.library.yorku.ca/xmlui/handle/10315/37753

[3] “Global Status Report On Road Safety 2018 Summary,” 2018, Accessed: Nov. 10, 2021. [Online]. Available: http://apps.who.int/bookorders.

[4] "Global Status Report on Road Safety" 2023, 1st ed. Geneva: World Health Organization, 2023.

[5] R. Q. Minguez, I. P. Alonso, D. Fernandez-Llorca, and M. A. Sotelo, “Pedestrian Path, Pose, and Intention Prediction Through Gaussian Process Dynamical Models and Pedestrian Activity Recognition,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1803–1814, 2019.

[6] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, “PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction,” In *International Conference on Computer Vision (ICCV)*, pp. 6261–6270, 2019.

[7] C. Zhang, C. Berger, and M. Dozza, “Social-IWSTCNN: A Social Interaction-Weighted Spatio-Temporal Convolutional Neural Network for Pedestrian Trajectory Prediction in Urban Traffic Scenarios,” In *IEEE Intelligent Vehicles Symposium (IV)*, pp. 1515–1522, 2021.

[8] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social LSTM: Human Trajectory Prediction in Crowded Spaces,” In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–971, 2016

[9] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, “SR-LSTM: State Refinement for LSTM towards Pedestrian Trajectory Prediction,” In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 12077–12086, 2019.

[10] Y. Zhu, D. Qian, D. Ren, and H. Xia, “StarNet: Pedestrian Trajectory Prediction using Deep Neural Network in Star Topology,” In *IEEE International Conference on Intelligent Robots and Systems*, pp. 8075–8080, 2019.

[11] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “Understanding Pedestrian Behavior in Complex Traffic Scenes,” *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 1, pp. 61–70, 2018.

[12] N. Sharma, C. Dhiman, and S. Indu, “Pedestrian Intention Prediction for Autonomous Vehicles: A Comprehensive Survey,” *Neurocomputing*, vol. 508, pp. 120–152, Oct. 2022.

[13] D. Ridel *et al.*, “A Literature Review on the Prediction of Pedestrian Behavior in Urban Scenarios,” In *Proc. IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 3105–3112, 2018.

[14] J. Xue, J.-W. Fang, and P. Zhang, "A Survey of Scene Understanding by Event Reasoning in Autonomous Driving," *International Journal of Automation and Computing*, vol. 15, no. 3, pp. 249–266, 2018.

[15] P. Pandey and J. V. Aghav, "Pedestrian–Autonomous Vehicles Interaction Challenges: A Survey and A Solution to Pedestrian Intent Identification," In *Proc. International Conference on Data Intelligence and Security (ICDIS),* pp. 283-292, 2019.

[16] A. Bighashdel and G. Dubbelman, "A Survey on Path Prediction Techniques for Vulnerable Road Users: From Traditional to Deep-Learning Approaches," In *IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 1039–1046, 2020.

[17] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian Action Anticipation using Contextual Feature Fusion in Stacked RNNs," In *British Machine Vision Conference (BMVC),* pp. 1–13, 2020.

[18] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and Ü. Özguner, "Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 221–230, 2022.

[19] F. Piccoli *et al.*, "FuSSI-Net: Fusion of Spatio-temporal Skeletons for Intention Prediction Network," In *Proc. Asilomar Conference on Signals, Systems and Computers*, pp. 68–72, 2020.

[20] H. Razali, T. Mordan, and A. Alahi, "Pedestrian Intention Prediction: A Convolutional Bottom-Up Multi-Task Approach," *Transportation Research Part C: Emerging Technologies*, vol. 130, p. 103259, 2021.

[21] F. Bartoli, G. Lisanti, L. Ballan, and A. Del Bimbo, "Context-Aware Trajectory Prediction," In *Proc. International Conference on Pattern Recognition (ICPR)*, pp. 1941–1946, 2018.

[22] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, "BiTraP: Bi-directional Pedestrian Trajectory Prediction with Multi-modal Goal Estimation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1463–1470, 2020.

[23] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, "Stepwise Goal-Driven Networks for Trajectory Prediction," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2716–2723, 2022.

[24] K. Mangalam *et al.*, "It Is Not the Journey but the Destination: Endpoint Conditioned Trajectory Prediction," In *European Conference on Computer Vision (ECCV),* pp. 759-776, 2020.

[25] K. Mangalam, Y. An, H. Girase, and J. Malik, "From Goals, Waypoints & Paths to Long Term Human Trajectory Forecasting," In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15213–15222, 2021.

[26] Y. Xu, Z. Piao, and S. Gao, "Encoding Crowd Interaction with Deep Neural Network for Pedestrian Trajectory Prediction," In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5275–5284, 2018.

[27] H. Xue, D. Q. Huynh, and M. Reynolds, "SS-LSTM: A Hierarchical LSTM Model for Pedestrian Trajectory Prediction," In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1186–1194, 2018.

[28] J. Gesnouin, S. Pechberti, B. Stanciulcscu, and F. Moutarde, "Trouspi-Net: Spatio-Temporal Attention On Parallel Atrous Convolutions And U-Grus For Skeletal Pedestrian Crossing

Prediction," In *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 01–07, 2021.

[29] T. Chen, R. Tian, and Z. Ding, "Visual Reasoning using Graph Convolutional Networks for Predicting Pedestrian Crossing Intention," In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3096–3102, 2021.

[30] A. Bertugli, S. Calderara, P. Coscia, L. Ballan, and R. Cucchiara, "AC-VRNN: Attentive Conditional-VRNN for Multi-Future Trajectory Prediction," *Computer Vision and Image Understanding*, vol. 210, pp. 1–16, 2021.

[31] R. Chandra *et al.*, "Forecasting Trajectory and Behavior of Road-Agents Using Spectral Clustering in Graph-LSTMs," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4882–4890, 2020.

[32] A. T. Schulz and R. Stiefelhagen, "Pedestrian Intention Recognition using Latent-dynamic Conditional Random Fields," In *Proc. IEEE Intelligent Vehicles Symposium,* pp. 622–627, 2015.

[33] Y. Yoo, K. Yun, S. Yun, J. Hong, H. Jeong, and J. Y. Choi, "Visual Path Prediction in Complex Scenes with Crowded Moving Objects," In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2668–2677, 2016.

[34] Lamberto Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese, "Knowledge Transfer for Scene-specific Motion," In *European Conference on Computer Vision (ECCV)*, pp. 1–16, 2016.

[35] S. Neogi, M. Hoy, W. Chaoqun, and J. Dauwels, "Context Based Pedestrian Intention Prediction Using Factored Latent Dynamic Conditional Random Fields," In *Proc. IEEE Symposium Series on Computational Intelligence (SSCI),* pp. 1–8, 2018.

[36] K. Saleh, M. Hossny, and S. Nahavandi, "Contextual Recurrent Predictive Model for Long-Term Intent Prediction of Vulnerable Road Users," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3398–3408, 2020.

[37] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto, "Intent-aware Long-Term Prediction of Pedestrian Motion," In *Proc. IEEE International Conference on Robotics and Automation,* pp. 2543–2549, 2016.

[38] A. Vemula, K. Muelling, and J. Oh, "Modeling Cooperative Navigation in Dense Human Crowds," In *Proc. IEEE International Conference on Robotics and Automation*, pp. 1685–1692, 2017.

[39] R. Hug, S. Becker, W. Hübner, and M. Arens, "Particle-based Pedestrian Path Prediction using LSTM-MDL Models," In *Proc. IEEE Conference on Intelligent Transportation Systems (ITSC),* pp. 2684–2691, 2018.

[40] A. Monti, A. Bertugli, S. Calderara, and R. Cucchiara, "DAG-Net: Double Attentive Graph Neural Network for Trajectory Forecasting," In *Proc. International Conference on Pattern Recognition (ICPR)*, pp. 2551–2558, 2020.

[41] E. Rehder, F. Wirth, M. Lauer, and C. Stiller, "Pedestrian Prediction by Planning Using Deep Neural Networks," In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5903–5908, 2018.

[42] Z. Pei, X. Qi, Y. Zhang, M. Ma, and Y.-H. Yang, “Human Trajectory Prediction in Crowded Scene using Social-Affinity Long Short-Term Memory,” *Pattern Recognition*, vol. 93, pp. 273–282, 2019.

[43] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, “Soft + Hardwired Attention: An LSTM Framework for Human Trajectory Prediction and Abnormal Event Detection,” *Neural Networks*, vol. 108, pp. 466–478, 2017.

[44] Manh, H., & Alaghband, G. (2018). Scene-LSTM: A Model for Human Trajectory Prediction. *arXiv preprint arXiv:1808.04018*.

[45] K. Saleh, M. Hossny, and S. Nahavandi, “Intent Prediction of Pedestrians via Motion Trajectories Using Stacked Recurrent Neural Networks,” *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 4, pp. 414–424, 2018.

[46] B. Völz, K. Behrendt, H. Mielenz, I. Gilitschenski, R. Siegwart, and J. Nieto, “A Data-Driven Approach for Pedestrian Intention Estimation,” In *IEEE Conference on Intelligent Transportation Systems, Proceedings (ITSC)*, pp. 2607–2612, 2016..

[47] O. Ghori *et al.*, “Learning to Forecast Pedestrian Intention from Pose Dynamics,” In *Proc. IEEE Intelligent Vehicles Symposium*, pp. 1277–1284, 2018.

[48] D. O. Pop, "Multi-Task Cross-Modality Deep Learning for Pedestrian Risk Estimation," Ph.D. dissertation, Department of Computer Science, Normandie Université, Institut National des Sciences Appliquées Rouen, Rouen, France, 2019. [Online] Available: https://inria.hal.science/tel-02997196/.

[49] M. Hoy, Z. Tu, K. Dang, and J. Dauwels, “Learning to Predict Pedestrian Intention via Variational Tracking Networks,” In *IEEE Conference on Intelligent Transportation Systems, Proceedings (ITSC)*, pp. 3132–3137, 2018.

[50] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, F. Galasso, and M. Cristani, “MX-LSTM: Mixing Tracklets and Vislets to Jointly Forecast Trajectories and Head Poses,” In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6067–6076, 2018.

[51] A. Sadeghian, F. Legros, M. Voisin, R. Vesel, A. Alahi, and S. Savarese, “CAR-Net: Clairvoyant Attentive Recurrent Network,” In *Proc. European conference on computer vision (ECCV),* pp. 151-167, 2018.

[52] Y. Feng, T. Zhang, A. P. Sah, L. Han, and Z. Zhang, “Using Appearance to Predict Pedestrian Trajectories through Disparity-Guided Attention and Convolutional LSTM,” *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 7480–7494, 2021.

[53] Q. Ma, Q. Zou, Y. Huang, and N. Wang, “Dynamic Pedestrian Trajectory Forecasting with LSTM-Based Delaunay Triangulation,” *Applied Intelligence*, vol. 52, no. 3, pp. 3018–3028, 2022.

[54] Z. Fang and A. M. López, “Is the Pedestrian going to Cross? Answering by 2D Pose Estimation,” In *Proc. IEEE Intelligent Vehicles Symposium,* pp. 1271–1276, 2018.

[55] J. Gesnouin, S. Pechberti, G. Bresson, B. Stanciulescu, and F. Moutarde, “Predicting Intentions of Pedestrians from 2d Skeletal Pose Sequences with a Representation-Focused Multi-Branch Deep Learning Network,” *Algorithms*, vol. 13, no. 12, pp. 1–23, 2020.

[56] N. Shafiee, T. Padir, and E. Elhamifar, "Introvert: Human Trajectory Prediction via Conditional 3D Attention," In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16810–16820, 2021.

[57] L. Rossi, M. Paolanti, R. Pierdicca, and E. Frontoni, "Human Trajectory Prediction and Generation using LSTM Models And GANs," *Pattern Recognition*, vol. 120, p. 108136, 2021.

[58] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction," In *European Conference on Computer Vision (ECCV)*, pp. 507–523, 2020.

[59] Y. Ma, X. Zhu, X. Cheng, R. Yang, J. Liu, and D. Manocha, "AutoTrajectory: Label-Free Trajectory Extraction and Prediction from Videos Using Dynamic Points," In *European Conference on Computer Vision (ECCV),* pp. 646-662, 2020.

[60] R. Yu and Z. Zhou, "Towards Robust Human Trajectory Prediction in Raw Videos," In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8059–8066, 2021.

[61] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer Networks for Trajectory Forecasting," In *Proc. International Conference on Pattern Recognition (ICPR)*, pp. 10335–10342, 2020.

[62] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: Modeling Spatial-temporal Interactions for Human Trajectory Prediction," In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 6271–6280, 2019.

[63] D. Varytimidis, F. Alonso-Fernandez, B. Duran, and C. Englund, "Action and Intention Recognition of Pedestrians in Urban Traffic," In *International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 676–682, 2018.

[64] A. Kalatian and B. Farooq, "A Context-Aware Pedestrian Trajectory Prediction Framework for Automated Vehicles," *Transportation Research Part C: Emerging Technologies*, vol. 134, p. 103453, 2021.

[65] A. Vemula, K. Muelling, and J. Oh, "Social Attention: Modeling Attention in Human Crowds," In *Proc. IEEE International Conference on Robotics and Automation*, pp. 4601–4607, 2018.

[66] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks," In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2255–2264, 2018.

[67] W. C. Ma, D. A. Huang, N. Lee, and K. M. Kitani, "Forecasting Interactive Dynamics Of Pedestrians With Fictitious Play," In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4636–4644, 2017.

[68] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning Social Etiquette: Human Trajectory Understanding in Crowded Scenes," In *European Conference on Computer Vision (ECCV),* pp. 549-565, 2016.

[69] J. Li, H. Ma, Z. Zhang, and M. Tomizuka, "Social-WaGDAT: Interaction-aware Trajectory Prediction via Wasserstein Graph Double-Attention Network," 2020, [Online]. Available: http://arxiv.org/abs/2002.06241

[70] L. Shi *et al.*, “SGCN:Sparse Graph Convolution Network for Pedestrian Trajectory Prediction,” In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8990–8999, 2021.

[71] R. Zhou, H. Zhou, H. Gao, M. Tomizuka, J. Li and Z. Xu, "Grouptron: Dynamic Multi-Scale Graph Convolutional Networks for Group-Aware Dense Crowd Trajectory Forecasting", In *International Conference on Robotics and Automation (ICRA)*, pp. 805-811, 2022.

[72] J. Sun, Q. Jiang, and C. Lu, “Recursive Social Behavior Graph for Trajectory Prediction,” In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 657–666, 2020.

[73] A. Sadeghian, A. Alahi, and S. Savarese, “Tracking the Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies,” In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 300–311, 2017.

[74] F. Schneemann and P. Heinemann, “Context-Based Detection of Pedestrian Crossing Intention for Autonomous Driving in Urban Environments,” In *IEEE International Conference on Intelligent Robots and Systems*, pp. 2243–2248, 2016.

[75] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, “DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents,” In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2165–2174, 2017.

[76] J. Li, F. Yang, M. Tomizuka, and C. Choi, “EvolveGraph: Multi-Agent Trajectory Prediction With Dynamic Relational Reasoning,” In *Advances in Neural Information Processing Systems (NeuroIPS)*, vol. 33, pp.19783-19794, 2020.

[77] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, S. H. Rezatofighi, and S. Savarese, “SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints,” In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1349–1358, 2019.

[78] L. Shi *et al.*, “SGCN: Sparse Graph Convolution Network for Pedestrian Trajectory Prediction,” In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8990–8999, 2021.

[79] Y. Fang, Z. Jin, Z. Cui, Q. Yang, T. Xie, and B. Hu, “Modeling Human–Human Interaction with Attention-Based High-Order GCN for Trajectory Prediction,” *The Visual Computer 2021*, pp. 1–13, 2021.

[80] Z. Yin, R. Liu, Z. Xiong, and Z. Yuan, “Multimodal Transformer Network for Pedestrian Trajectory Prediction,” In *IJCAI International Joint Conference on Artificial Intelligence*, pp. 1259–1265, 2021

[81] B. Liu *et al.*, “Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.

[82] M. N. Riaz, M. Wielgosz, A. G. Romera, and A. M. López, “Synthetic Data Generation Framework, Dataset, and Efficient Deep Model for Pedestrian Intention Prediction,” In *International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2742–2749, 2023

[83] S. S. Monfort and B. C. and Mueller, “Pedestrian Injuries from Cars and Suvs: Updated Crash Outcomes from The Vulnerable Road User Injury Prevention Alliance (VIPA),” *Traffic Injury Prevention*, vol. 21, no. sup1. S165–S167, 2020.

[84] A. Rasouli, M. Rohani, and J. Luo, “Bifold and Semantic Reasoning for Pedestrian Behavior Prediction,” In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15580–15590, 2021

[85] N. Osman, G. Camporese, and L. Ballan, “TAMformer: Multi-Modal Transformer with Learned Attention Mask for Early Intent Prediction,” In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[86] J. Bai, X. Fang, J. Fang, J. Xue, and C. Yuan, “Deep Virtual-to-Real Distillation for Pedestrian Crossing Prediction,” In *International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1586–1592, 2022.

[87] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, “Pedestrian Graph +: A Fast Pedestrian Crossing Prediction Model Based on Graph Convolutional Networks,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 21050–21061, 2022.

[88] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, “Coupling Intent and Action for Pedestrian Crossing Behavior Prediction,” In *Proc. Joint Conference on Artificial Intelligence*, pp. 1238–1244, 2021.

[89] O. Hamed and H. J. Steinhauer, “Pedestrian Intention Recognition and Action Prediction Using a Feature Fusion Deep Learning Approach,” In *Proc. International Conference on Modeling Decisions for Artificial Intelligence (MDAI)*, pp. 89–100, 2021.

[90] A. Singh and U. Suddamalla, “Multi-Input Fusion for Practical Pedestrian Intention Prediction,” In *International Conference on Computer Vision Workshops (ICCVW)*, pp. 2304–2311, 2021.

[91] N. Sharma, C. Dhiman, and S. Indu, “Visual-Motion-Interaction-Guided Pedestrian Intention Prediction Framework,” *IEEE Sensors Journal*, vol. 23, no. 22, pp. 27540–27548, 2023.

[92] R. Ni, B. Yang, Z. Wei, H. Hu, and C. Yang, “Pedestrians Crossing Intention Anticipation Based on Dual-Channel Action Recognition and Hierarchical Environmental Context,” *IET Intelligent Transport Systems*, vol. 17, pp. 1–15, 2022.

[93] A. Rasouli and I. Kotseruba, “PedFormer: Pedestrian Behavior Prediction via Cross-Modal Attention Modulation and Gated Multitask Learning,” In *International Conference on Robotics and Automation (ICRA),* pp. 9844–9851, 2023

[94] L. Huang, J. Zhuang, X. Cheng, R. Xu, and H. Ma, “STI-GAN: Multimodal Pedestrian Trajectory Prediction Using Spatiotemporal Interactions and a Generative Adversarial Network,” *IEEE Access*, vol. 9, pp. 50846–50856, 2021.

[95] S. K. Jayaraman, L. P. Robert, X. Jessie Yang, and D. M. Tilbury, “Multimodal Hybrid Pedestrian: A Hybrid Automaton Model of Urban Pedestrian Behavior for Automated Driving Applications,” *IEEE Access*, vol. 9, pp. 27708–27722, 2021.

[96] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, “Benchmark for Evaluating Pedestrian Action Prediction,” in *Winter Conference on Applications of Computer Vision (WACV)*, pp. 1257–1267, 2021.

[97] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[98] W. Zhu, Y. Liu, P. Wang, M. Zhang, T. Wang, and Y. Yi, “Tri-HGNN: Learning Triple Policies Fused Hierarchical Graph Neural Networks For Pedestrian Trajectory Prediction,” *Pattern Recognition*, vol. 143, p. 109772, 2023.

[99] A. Y. Naik, A. Bighashdel, P. Jancura, and G. Dubbelman, “Scene Spatio-Temporal Graph Convolutional Network for Pedestrian Intention Estimation,” In *Intelligent Vehicles Symposium (IV)*, pp. 874–881, 2022.

[100] X. Zhang, P. Angeloudis, and Y. Demiris, “Dual-Branch Spatio-temporal Graph Neural Networks for Pedestrian Trajectory Prediction,” *Pattern Recognition*, vol. 142, p. 109633, 2023.

[101] Y. Ling, Q. Zhang, X. Weng, and Z. Ma, “STMA-GCN_PedCross: Skeleton Based Spatial-Temporal Graph Convolution Networks with Multiple Attentions for Fast Pedestrian Crossing ntention Prediction,” In *International Conference on Intelligent Transportation Systems (ITSC)*, pp. 500–506, 2023.

[102] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, “A Comprehensive Survey on Graph Neural Networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2019.

[103] L. Achaji, J. Moreau, T. Fouqueray, F. Aioun, and F. Charpillet, “Is Attention to Bounding Boxes All You Need for Pedestrian Action Prediction?” In *Intelligent Vehicles Symposium (IV)*, pp. 895–902, 2022.

[104] Y. Zhou, G. Tan, R. Zhong, Y. Li, and C. Gou, “PIT: Progressive Interaction Transformer for Pedestrian Crossing Intention Prediction,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, 2023.

[105] Z. Zhang, R. Tian, and Z. Ding, “TrEP: Transformer-Based Evidential Prediction for Pedestrian Intention with Uncertainty,” In *Proc. of the AAAI Conference on Artificial Intelligence*, pp. 3534–3542, 2023.

[106] H. Xue, D. Q. Huynh, and M. Reynolds, “A Location-Velocity-Temporal Attention LSTM Model for Pedestrian Trajectory Prediction,” *IEEE Access*, vol. 8, pp. 44576–44589, 2020.

[107] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, “Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction,” In *European Conference on Computer Vision (ECCV),* pp. 507-523, 2020.

[108] C. Tao, Q. Jiang, L. Duan, and P. Luo, “Dynamic and Static Context-Aware LSTM for Multi-agent Motion Prediction,” In *European Conference on Computer Vision,* pp. 547-563, 2020.

[109] C. Wong, B. Xia, Z. Hong, Q. Peng, W. Yuan, Q. Cao, Y. Yang and X. You, “View Vertically: A Hierarchical Network for Trajectory Prediction via Fourier Spectrums,” In *European Conference on Computer Vision (ECCV)*, pp. 682-700, 2022.

[110] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, “Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data,” In *European Conference on Computer Vision (ECCV),* pp. 683-700, 2020.

[111] B. Ivanovic and M. Pavone, “The Trajectron: Probabilistic Multi-Agent Trajectory Modeling with Dynamic Spatiotemporal Graphs,” In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2375–2384, 2019.

[112] T. Su, Y. Meng, and Y. Xu, "Pedestrian Trajectory Prediction via Spatial Interaction Transformer Network," In *IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, pp. 154–159, 2021.

[113] J. Gao, X. Shi, and J. J. Q. Yu, "Social-DualCVAE: Multimodal Trajectory Forecasting Based on Social Interactions Pattern Aware and Dual Conditional Variational Auto-Encoder," *CoRR*, vol. abs/2202.03954, 2022.

[114] J. Yue, D. Manocha, and H. Wang, "Human Trajectory Prediction via Neural Social Physics," In *European Conference on Computer Vision (ECCV),* pp. 376-394, 2022.

[115] J. T. Chauhan, "Comparative Study of GAN and VAE," *International Journal of Computer Applications*, vol. 182, no. 22, pp. 1–5, 2018.

[116] A. Rasouli and J. K. Tsotsos, "Autonomous Vehicles that Interact with Pedestrians A Survey of Theory and Practice," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, 2020.

[117] F. Fang, P. Zhang, B. Zhou, K. Qian, and Y. Gan, "Atten-GAN: Pedestrian Trajectory Prediction with GAN Based on Attention Mechanism," *Cognitive Computation*, vol. 14, no. 6, pp. 2296–2305, 2022.

[118] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and U. Ozguner, "Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-temporal Attention," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 221–230, 2022.

[119] A. Rasouli, T. Yau, M. Rohani, and J. Luo, "Multi-Modal Hybrid Architecture for Pedestrian Action Prediction," In *IEEE Intelligent Vehicles Symposium (IV)*, pp. 91–97, 2022

[120] S. Lu, Z. Zhu, J. M. Gorriz, S. H. Wang, and Y. D. Zhang, "NAGNN: Classification Of COVID-19 Based on Neighboring Aware Representation from Deep Graph Neural Network," *International Journal of Intelligent Systems*, vol. 37, no. 2, pp. 1572–1598, 2022.

[121] Y. Zhang *et al.*, "Deep Learning in Food Category Recognition," *Information Fusion*, vol. 98, p. 101859, 2023.

[122] S. Lu, S. H. Wang, and Y. D. Zhang, "Detection of Abnormal Brain in MRI via Improved Alexnet and ELM Optimized by Chaotic Bat Algorithm," *Neural Computing and Applications*, vol. 33, no. 17, pp. 10799–10811, 2021.

[123] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," In *International Conference on Machine Learning (ICML),* pp. 10691–10700, 2019.

[124] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

[125] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, 2022.

[126] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," In *Proc. European Conference on Computer Vision (ECCV),* pp. 3-19, 2018

[127] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep High-Resolution Representation Learning for Human Pose Estimation,” In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5686–5696, 2019.

[128] T. Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” In *European Conference on Computer Vision (ECCV),* pp. 740-755, 2014.

[129] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[130] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, “1D Convolutional Neural Networks and Applications: A Survey,” *Mechanical Systems and Signal Processing*, vol. 151, pp. 1–20, 2021.

[131] B. Xu and H. Yin, “Graph Convolutional Networks in Feature Space for Image Deblurring and Super-resolution,” In *Proc. International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, 2021.

[132] Z. Tao, C. Ouyang, Y. Liu, T. Chung, and Y. Cao, “Multi-Head Attention Graph Convolutional Network Model: End-To-End Entity and Relation Joint Extraction Based on Multi-Head Attention Graph Convolutional Network,” *CAAI Transactions on Intelligence Technology*, vol. 8, no. 2, pp. 468-477, 2023.

[133] T. N. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” In *Proc. International Conference on Learning Representations (ICLR),* 2017

[134] D. Hendrycks and K. Gimpel, “Gaussian Error Linear Units (GELUs),” Jun. 2016, Accessed: Jun. 09, 2022. [Online]. Available: http://arxiv.org/abs/1606.08415

[135] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance On ImageNet Classification,” In *Proc. IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

[136] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior,” In *IEEE International Conference on Computer Vision Workshops, (ICCVW)* pp. 206–213, 2017.

[137] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “It’s Not All About Size: On the Role of Data Properties in Pedestrian Detection,” In *Proc. European Conference on Computer Vision (ECCV) Workshops,* 2018.

[138] J. Lian, W. Ren, L. Li, Y. Zhou, and B. Zhou, “PTP-STGCN: Pedestrian Trajectory Prediction Based on a Spatio-temporal Graph Convolutional Neural Network,” *Applied Intelligence*, vol. 53, no. 3, pp. 2862–2878, May 2023.

[139] Y. Zhou *et al.*, “Social Graph Convolutional LSTM For Pedestrian Trajectory Prediction,” *IET Intelligent Transport Systems*, vol. 15, no. 3, pp. 396–405, 2021.

[140] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2016.

[141] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021.

[142] R. Girdhar, J. J. Carreira, C. Doersch, and A. Zisserman, “Video Action Transformer Network,” In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 244–253, 2019.

[143] Z. Zhong, D. Schneider, M. Voit, R. Stiefelhagen, and J. Beyerer, “Anticipative Feature Fusion Transformer for Multi-Modal Action Anticipation,” In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[144] Kotseruba, I., Rasouli, A., & Tsotsos, J. K. (2016). Joint Attention in Autonomous Driving (JAAD). *arXiv preprint arXiv:1609.04741*.

[145] A. Kapishnikov, S. Venugopalan, B. Avci, B. Wedin, M. Terry, and T. Bolukbasi, “Guided Integrated Gradients: An Adaptive Path Method for Removing Noise,” In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5048–5056, 2021.

[146] Z. Chen, B. Han, S. Wang, and Y. Qian, “Attention-based Encoder-Decoder Network for End-to-End Neural Speaker Diarization with Target Speaker Attractor,” In *Proc. INTERSPEECH*, pp. 3552–3556, 2023.

[147] Y. Huang, J. Chen, H. Ma, H. Ma, W. Ouyang, and C. Yu, “Attribute Assisted Teacher-Critical Training Strategies for Image Captioning,” *Neurocomputing*, vol. 506, pp. 265–276, 2022.

[148] H. Chen, K. Lin, A. Maye, J. Li, and X. Hu, “A Semantics-Assisted Video Captioning Model Trained With Scheduled Sampling,” *Frontiers in Robotics and AI*, vol. 7, 2020.

[149] M. Sinambela, T. Rahayu, E. Darnila, and T. Limbong, “A Review of Digital Image Classification Based on Fuzzy Logic,” *MEANS (Media Informasi Analisa dan Sistem)*, vol. 5, no. 1, pp. 37–40, 2020.

[150] T.-L. Nguyen, S. Kavuri, and M. Lee, “A Multimodal Convolutional Neuro-Fuzzy Network for Emotion Understanding of Movie Clips,” *Neural Networks*, vol. 118, pp. 208–219, 2019.

[151] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, “You’ll Never Walk Alone: Modeling Social Behavior For Multi-Target Tracking,” In *IEEE International Conference on Computer Vision (ICCV),* pp. 261–268, 2009.

[152] A. Lerner, Y. Chrysanthou, and D. Lischinski, “Crowds by Example,” *Computer Graphics Forum*, vol. 26, no. 3, pp. 655–664, Sep. 2007.

[153] A. Bhattacharyya, M. Fritz, and B. Schiele, “Long-Term On-Board Prediction of People in Traffic Scenes under Uncertainty,” *Archivos De Economía*, vol. 457, pp. 4194–4202, 2017.

[154] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” In *Advances in Neural Information Processing Systems (NeuroIPS)*, pp. 6840-6851, 2020.

[155] R. Huang, J. Ding, M. Pagnucco, and Y. Song, “Fully Decoupling Trajectory and Scene Encoding for Lightweight Heatmap-oriented Trajectory Prediction,” *IEEE Robotics and Automation Letters*, pp. 1–8, 2024.

[156] A. Feng, C. Han, J. Gong, Y. Yi, R. Qiu, and Y. Cheng, “Multi-Scale Learnable Gabor Transform for Pedestrian Trajectory Prediction From Different Perspectives,” *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2024.

[157] N. Sharma, C. Dhiman, and S. Indu, “Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration,” IEEE Transactions on Intelligent Vehicles, vol. 9, no. 11, pp. 6960-6970 pp. 1–11, 2024.

[158] N. Sharma, C. Dhiman, and S. Indu, “Intelligent Pedestrian Intention Prediction Framework,” In *IEEE International Conference on Service Operations and Logistics, and Informatics, SOLI,* pp. 1-5, 2022.

[159] B. Yang, Z. Fan, H. Hu, C. Hu, and R. Ni, “Explainable Pedestrian Crossing Intention Prediction Based On Multi-Task Mutual Guidance Network,” *IEEE Transactions on Intelligent Vehicles*, pp. 1–11, 2024.

[160] N. Sharma, C. Dhiman, and S. Indu, “Predicting Pedestrian Intentions with Multimodal IntentFormer: A Co-learning approach,” *Pattern Recognition*, vol. 161, p. 111205, 2025.

LIST OF PUBLICATIONS AND THEIR PROOFS

- **N. Sharma**, C. Dhiman, and S. Indu, "Pedestrian Intention Prediction for Autonomous Vehicles: A Comprehensive Survey," ***Neurocomputing***, vol. 508, pp. 120–152, 2022, doi: https://doi.org/10.1016/j.neucom.2022.07.085. **Impact Factor: 6.5.**

Neurocomputing 508 (2022) 120–152

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Survey paper

Pedestrian Intention Prediction for Autonomous Vehicles: A Comprehensive Survey

Neha Sharma, Chhavi Dhiman *, S. Indu

Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

ARTICLE INFO

Article history:
Received 10 March 2022
Revised 15 July 2022
Accepted 24 July 2022
Available online 5 August 2022
Communicated by Zidong Wang

Keywords:
Pedestrian intention
Autonomous vehicles
Vulnerable road user
Long-term intention
Short-term intention
Dynamical-based models
Goal-driven models
Path-planning models

ABSTRACT

Lately, Autonomous vehicles (AV) have been gaining traction globally owing to their huge social, economic and environmental benefits. However, the rising safety apprehensions for vulnerable road users (VRU) alongside have become a stumbling block to the large-scale implementation of AVs. Deciphering intention in advance considering all social norms and context surrounding the VRU is a highly challenging task due to the huge amount of variability in their motion, actions and end goals. Based on this pensiveness, this paper extensively surveys the variety of techniques applied to anticipate pedestrian intention and classifies them from multiple perspectives. Some newly introduced datasets with added complexities of human behaviour on road have also been outlined. It also provides a comparative analysis of the performance of pedestrian intention prediction approaches on several benchmark datasets as per the various assessment parameters available. In addition to this, several potential challenges and their possible solutions paving way for future research initiatives have also been thoroughly analysed in this endeavour.

1. Introduction

The autonomous vehicle market is expected to have enormous growth potential by 2023 at a compound annual growth rate (CAGR) of around 17 per cent even after a sudden unexpected halt owing to worldwide lockdown in wake of containment measures for the coronavirus pandemic [1]. Autonomous vehicular technology comes with enormous economic benefits for society ranging from reducing costs of driving to increasing fuel efficiency and many more. The absence of humans from driver seats is sought to make the driving experience error-free, stress-free both for the driver and passenger, and thus reducing human errors which consequently leads to alleviating accident rates. This ensures a safe traffic environment for both car and non-car users. Moreover, the level of comfort that it brings with it allows one to engage in other productive work or recreation while on their way to the destination without sparing attention to the road traffic [2].

Despite the highly promising future of AVs, and its booming economic ventures, creating a fully autonomously working car remains an unfulfilled desire of many tech giants even after garnering huge success now and then in Advanced Driving Assistance System (ADAS) by the research community. According to The Global Status Report on Road Safety published by the World Health Organization (WHO) [3], the number of deaths on roads globally has reached an unprecedented high of 1.35 million deaths annually. Out of which, victims of nearly half of the road accidents are vulnerable road users (VRU). Huge challenges persist when it comes to developing appropriate infrastructure and proper safety traffic regulations to facilitate the harmonious coexistence of AVs and VRUs in urban traffic scenarios. Therefore, a high level of precision and accuracy is required as several lives are involved which can't be risked in the name of technological advancements [4].

One of the most challenging issues faced by autonomous vehicles is mimicking the perception that humans have in an understanding multitude of social cues in everyday traffic scenarios to avoid fatal vehicle-to-VRU collisions [5]. This is to prevent any severe injury to the latter as they don't have any special protective equipment. Additionally, it creates a secure and more congenial atmosphere for every road user agent. Hence, early anticipation of VRU's intention is desired so that AVs get adequate time to design their manoeuvres accordingly [6]. A variety of approaches are employed for this challenging task which includes interpreting the forthcoming actions of vul-

* Corresponding author.
E-mail addresses: nehashrm013@gmail.com (N. Sharma), chhavi.dhiman@dtu.ac.in (C. Dhiman), s.indu@dtu.ac.in (S. Indu).

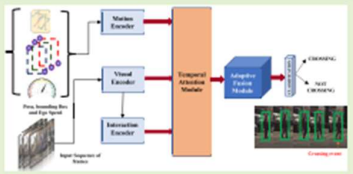
https://doi.org/10.1016/j.neucom.2022.07.085
0925-2312/

- **N. Sharma**, C. Dhiman, and S. Indu, "Visual–Motion–Interaction-Guided Pedestrian Intention Prediction Framework," ***IEEE Sensors Journal***, vol. 23, no. 22, pp. 27540–27548, 2023**,** doi: 10.1109/JSEN.2023.3317426. **Impact Factor: 4.5**

Visual–Motion–Interaction-Guided Pedestrian Intention Prediction Framework

Neha Sharma, *Member, IEEE*, Chhavi Dhiman, *Member, IEEE*, and S. Indu, *Senior Member, IEEE*

***Abstract*—The capability to comprehend the intention of pedestrians on the road is one of the most crucial skills that the current autonomous vehicles (AVs) are striving for, to become fully autonomous. In recent years, multi-modal methods have gained traction employing trajectory, appearance, and context for predicting pedestrian crossing intention. However, most existing research works still lag rich feature representational ability in a multimodal scenario, restricting their performance. Moreover, less emphasis is put on pedestrian interactions with the surroundings for predicting short-term pedestrian intention in a challenging ego-centric vision. To address these challenges, an efficient visual–motion–interaction-guided (VMI) intention prediction framework has been proposed. This framework comprises visual encoder (VE), motion encoder (ME), and interaction encoder (IE) to capture rich multimodal features of the pedestrian and its interactions with the surroundings, followed by temporal attention and adaptive fusion (AF) module (AFM) to integrate these multimodal features efficiently. The proposed framework outperforms several SOTA on benchmark datasets: Pedestrian Intention Estimation (PIE)/Joint Attention in Autonomous Driving (JAAD) with accuracy, AUC, *F*1-score, precision, and recall as 0.92/0.89, 0.91/0.90, 0.87/0.81, 0.86/0.79, and 0.88/0.83, respectively. Furthermore, extensive experiments are carried out to investigate different fusion architectures and design parameters of all encoders. The proposed VMI framework predicts pedestrian crossing intention 2.5 s ahead of the crossing event. Code is available at: https://github.com/neha013/VMI.git.**



***Index Terms*— Autonomous vehicles (AVs), intention prediction, pedestrians.**

I. Introduction

ACCORDING to the Global Status Report on Road Safety 2018, vehicle crashes cause numerous annual deaths, particularly impacting vulnerable road users [1]. Pedestrians, lacking protective gear, face high vulnerability, and substantial injury risk in collisions. Consequently, the growing advancement of autonomous vehicle (AV) technology is being explored to enhance road safety and convenience for all users. AV technology holds the potential to reduce accidents attributed to human errors like fatigue, misperception, and inattention. Leading automotive manufacturers and tech giants like Bayerische Motoren Werke (BMW), Tesla, and Google are actively advancing AV technology in this pursuit.

Predicting pedestrians' road-crossing decisions is pivotal for achieving a reliable driverless experience through AVs. Initial studies emphasized pedestrian dynamics to anticipate crossing intent [2]. Yet, analyzing merely the trajectory proves inadequate for understanding underlying intentions [3]. Beyond trajectory, various factors impact pedestrian road-crossing decisions. These factors fall into three primary modalities: pedestrian-specific (encompassing pose and appearance), context-specific (involving scene infrastructure and social interaction with co-pedestrians), and hybrid modality encompassing comprehensive human cognitive aspects while observing a pedestrian on the road [3].

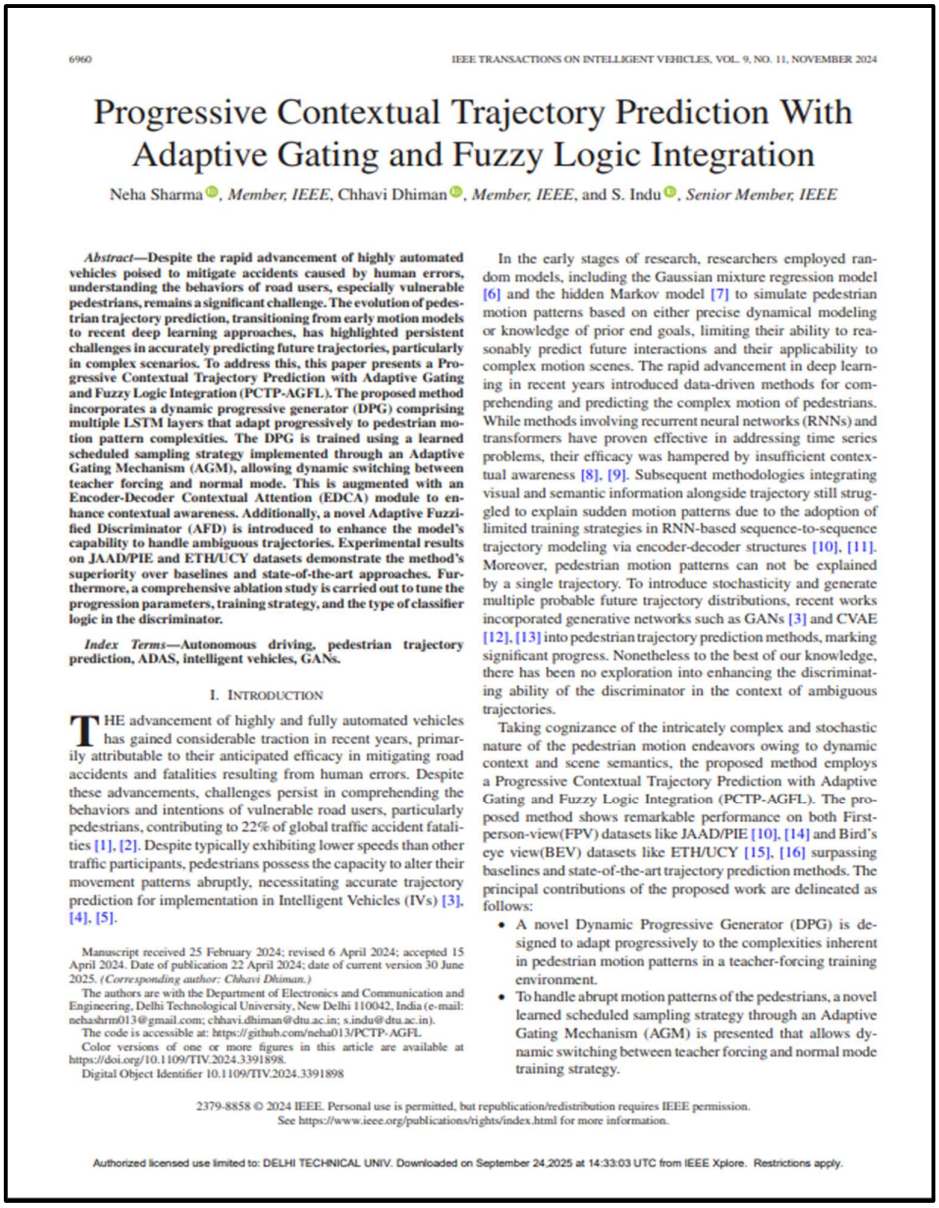
Inspired by how human drivers interact with pedestrians and make decisions intuitively, recent endeavors [4], [5] aim to decipher pedestrians' crossing intentions by analyzing a varied combination of pedestrian-specific and context-specific features. Nonetheless, dealing with such diverse modalities necessitates an efficient multimodal fusion framework that can capture adequate discriminatory features for classification. Moreover, interpreting pedestrian interactions with the surrounding environment is highly challenging in a dynamic ego-centric setting. Quite a few approaches [6], [7] could exploit the social interaction features for short-term intention prediction, so far. Therefore, to address these issues, the proposed multimodality fusion framework for pedestrian intention prediction focuses on a holistic understanding of the

Manuscript received 22 August 2023; accepted 16 September 2023. Date of publication 26 September 2023; date of current version 14 November 2023. The associate editor coordinating the review of this article and approving it for publication was Prof. Yu-Dong Zhang. *(Corresponding author: Chhavi Dhiman.)*

The authors are with the Department of Electronics and Communication and Engineering, Delhi Technological University (DTU), Delhi 110042, India (e-mail: nehashrm013@gmail.com; chhavi.dhiman@dtu.ac.in; s.indu@dtu.ac.in).

Digital Object Identifier 10.1109/JSEN.2023.3317426

- **N. Sharma**, C. Dhiman, and S. Indu, "Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration," ***IEEE Transactions on Intelligent Vehicles***, vol. 9, no. 11, pp. 6960-6970, 2024, doi: 10.1109/TIV.2024.3391898. **Impact Factor: 14.3**



Progressive Contextual Trajectory Prediction With Adaptive Gating and Fuzzy Logic Integration

Neha Sharma, *Member, IEEE*, Chhavi Dhiman, *Member, IEEE*, and S. Indu, *Senior Member, IEEE*

***Abstract*—Despite the rapid advancement of highly automated vehicles poised to mitigate accidents caused by human errors, understanding the behaviors of road users, especially vulnerable pedestrians, remains a significant challenge. The evolution of pedestrian trajectory prediction, transitioning from early motion models to recent deep learning approaches, has highlighted persistent challenges in accurately predicting future trajectories, particularly in complex scenarios. To address this, this paper presents a Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration (PCTP-AGFL). The proposed method incorporates a dynamic progressive generator (DPG) comprising multiple LSTM layers that adapt progressively to pedestrian motion pattern complexities. The DPG is trained using a learned scheduled sampling strategy implemented through an Adaptive Gating Mechanism (AGM), allowing dynamic switching between teacher forcing and normal mode. This is augmented with an Encoder-Decoder Contextual Attention (EDCA) module to enhance contextual awareness. Additionally, a novel Adaptive Fuzzified Discriminator (AFD) is introduced to enhance the model's capability to handle ambiguous trajectories. Experimental results on JAAD/PIE and ETH/UCY datasets demonstrate the method's superiority over baselines and state-of-the-art approaches. Furthermore, a comprehensive ablation study is carried out to tune the progression parameters, training strategy, and the type of classifier logic in the discriminator.**

***Index Terms*—Autonomous driving, pedestrian trajectory prediction, ADAS, intelligent vehicles, GANs.**

I. Introduction

THE advancement of highly and fully automated vehicles has gained considerable traction in recent years, primarily attributable to their anticipated efficacy in mitigating road accidents and fatalities resulting from human errors. Despite these advancements, challenges persist in comprehending the behaviors and intentions of vulnerable road users, particularly pedestrians, contributing to 22% of global traffic accident fatalities [1], [2]. Despite typically exhibiting lower speeds than other traffic participants, pedestrians possess the capacity to alter their movement patterns abruptly, necessitating accurate trajectory prediction for implementation in Intelligent Vehicles (IVs) [3], [4], [5].

Manuscript received 25 February 2024; revised 6 April 2024; accepted 15 April 2024. Date of publication 22 April 2024; date of current version 30 June 2025. *(Corresponding author: Chhavi Dhiman.)*

The authors are with the Department of Electronics and Communication and Engineering, Delhi Technological University, New Delhi 110042, India (e-mail: nehashrm013@gmail.com; chhavi.dhiman@dtu.ac.in; s.indu@dtu.ac.in).

The code is accessible at: https://github.com/neha013/PCTP-AGFL.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2024.3391898.

Digital Object Identifier 10.1109/TIV.2024.3391898

In the early stages of research, researchers employed random models, including the Gaussian mixture regression model [6] and the hidden Markov model [7] to simulate pedestrian motion patterns based on either precise dynamical modeling or knowledge of prior end goals, limiting their ability to reasonably predict future interactions and their applicability to complex motion scenes. The rapid advancement in deep learning in recent years introduced data-driven methods for comprehending and predicting the complex motion of pedestrians. While methods involving recurrent neural networks (RNNs) and transformers have proven effective in addressing time series problems, their efficacy was hampered by insufficient contextual awareness [8], [9]. Subsequent methodologies integrating visual and semantic information alongside trajectory still struggled to explain sudden motion patterns due to the adoption of limited training strategies in RNN-based sequence-to-sequence trajectory modeling via encoder-decoder structures [10], [11]. Moreover, pedestrian motion patterns can not be explained by a single trajectory. To introduce stochasticity and generate multiple probable future trajectory distributions, recent works incorporated generative networks such as GANs [3] and CVAE [12], [13] into pedestrian trajectory prediction methods, marking significant progress. Nonetheless to the best of our knowledge, there has been no exploration into enhancing the discriminating ability of the discriminator in the context of ambiguous trajectories.

Taking cognizance of the intricately complex and stochastic nature of the pedestrian motion endeavors owing to dynamic context and scene semantics, the proposed method employs a Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration (PCTP-AGFL). The proposed method shows remarkable performance on both First-person-view(FPV) datasets like JAAD/PIE [10], [14] and Bird's eye view(BEV) datasets like ETH/UCY [15], [16] surpassing baselines and state-of-the-art trajectory prediction methods. The principal contributions of the proposed work are delineated as follows:

- A novel Dynamic Progressive Generator (DPG) is designed to adapt progressively to the complexities inherent in pedestrian motion patterns in a teacher-forcing training environment.
- To handle abrupt motion patterns of the pedestrians, a novel learned scheduled sampling strategy through an Adaptive Gating Mechanism (AGM) is presented that allows dynamic switching between teacher forcing and normal mode training strategy.

- **N. Sharma**, C. Dhiman, and S. Indu, "Predicting Pedestrian Intentions with Multimodal IntentFormer: A Co-learning Approach," ***Pattern Recognition***, vol. 161, p. 111205, 2025**,** doi: https://doi.org/10.1016/j.patcog.2024.111205. **Impact Factor: 7.6**

Pattern Recognition 161 (2025) 111205

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Predicting pedestrian intentions with multimodal IntentFormer: A Co-learning approach

Neha Sharma, Chhavi Dhiman *, Sreedevi Indu

Department of Electronics and Communication Engineering, Delhi Technological University, India

ARTICLE INFO

Keywords:
Pedestrians intention prediction
Co-learning
Weight sharing intent transformer
Autonomous driving

ABSTRACT

The prediction of pedestrian crossing intention is a crucial task in the context of autonomous driving to ensure traffic safety and reduce the risk of accidents without human intervention. Nevertheless, the complexity of pedestrian behaviour, which is influenced by numerous contextual factors in conjunction with visual appearance cues and past trajectory, poses a significant challenge. Several state-of-the-art approaches have recently emerged that incorporate multiple modalities. Nonetheless, the suboptimal modality integration techniques in these approaches fail to capture the intricate intermodal relationships and robustly represent pedestrian-environment interactions in challenging scenarios. To address these issues, a novel Multimodal IntentFormer architecture is presented. It works with three transformer encoders $\{TE_I, TE_M, TE_{SI}\}$ which learn RGB, segmentation maps, and trajectory paths in a co-learning environment controlled by a Co-learning module. A novel Co-learning Adaptive Composite (CAC) loss function is also proposed, which penalizes different stages of the architecture, regularizes the model, and mitigates the risk of overfitting. Each encoder $\{TE_x\}$ applies the concept of the Multi-Head Shared Weight Attention (MHSWA) mechanism while learning three modalities in the proposed co-learning approach. The proposed architecture outperforms existing state-of-the-art approaches on benchmark datasets, PIE and JAAD, with 93 % and 92 % accuracy, respectively. Furthermore, extensive ablation studies demonstrate the efficiency and robustness of the architecture, even under varying Time-to-event (TTE) and observation lengths. The code is available at https://github.com/neha013/IntentFormer

1. Introduction

As of 2021, the global market size for autonomous vehicles(AVs) is estimated to reach $125.67 billion by 2030, according to a recent report published by Market Research Future [1]. The market is expected to witness substantial growth in the coming years, with key players such as Alphabet, Uber, Tesla, and General Motors investing heavily in the development of autonomous vehicle technology. This growth is attributed to the increasing demand for advanced technologies in the automotive industry to increase safety, improve efficiency, and provide comfort and accessibility to all, irrespective of age, disability, or other factors. Furthermore, the unfortunate COVID-19 outbreak that curbed mobility for the risk of infection transmission highlighted the relevance of AVs as they can effectively reduce any human interaction. These could transport anti-epidemic food items, medicines, and other essentials without posing a risk of death due to virus infection.

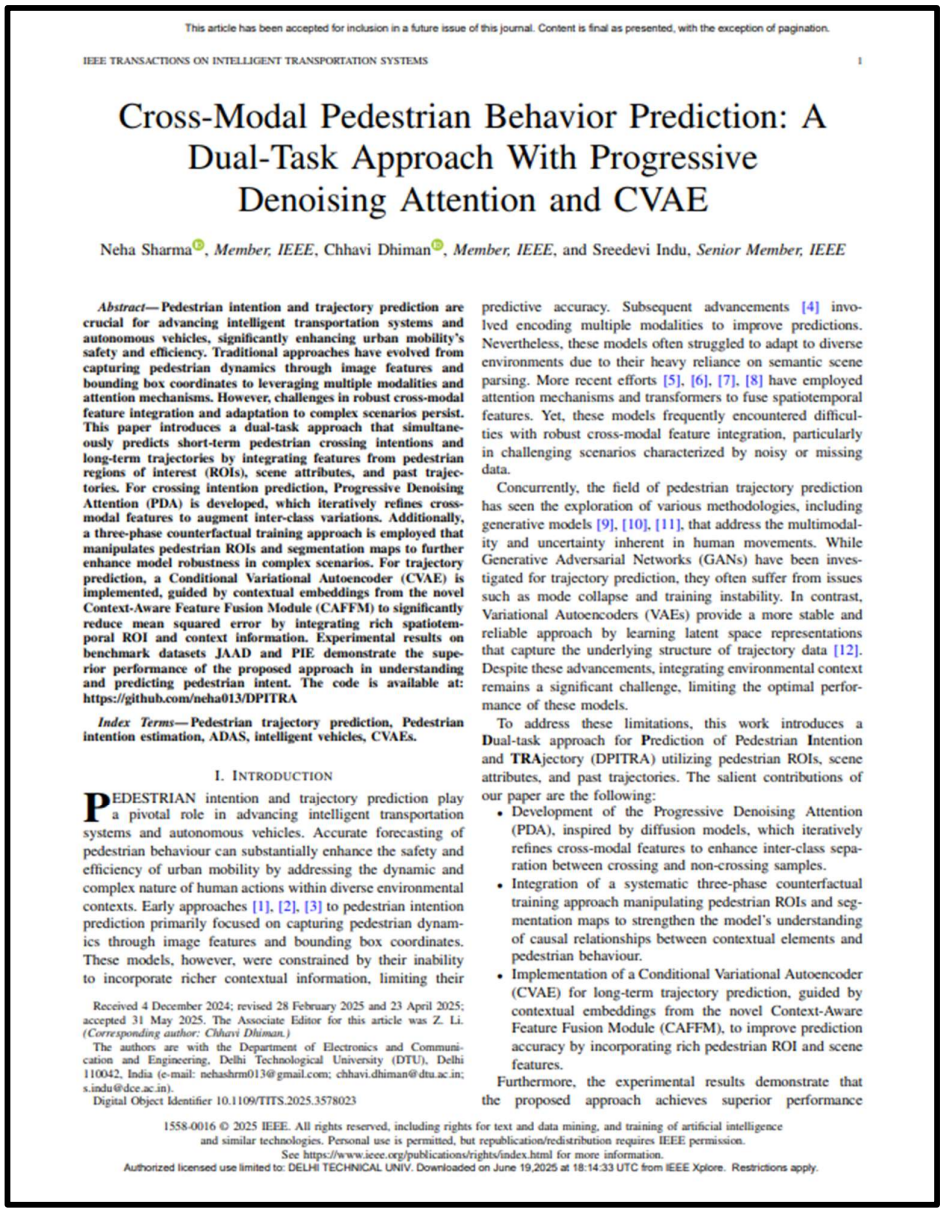
Nonetheless, the full realization of these benefits is still several years away, as the challenges posed by the complexity of urban traffic environments can impede the reliable operation of AVs. The pedestrian-vehicle conflicts in a crowded urban road environment are one of the most crucial problems that have recently elicited enormous attention from the AV research community. Pedestrians crossing the street are susceptible to vehicle conflicts, leading to safety concerns. Therefore, a comprehensive understanding of pedestrian crossing behaviour cues can assist in interpreting pedestrian intentions and expected crossing actions, leading to improved road safety and fewer conflicts between pedestrians and vehicles.

A combination of visual, dynamic, and motion cues is exhibited by pedestrians when they intend to cross the road, offering valuable clues to their crossing behaviour. For instance, a pedestrian may cross the road if he/she is approaching the crosswalk and looking at the incoming vehicle to ask for a passage. On the other hand, a person standing still at the curb, showing no signs of motion or visual gait towards the crossing action, is less likely to cross the street in a short while. Hence, the pedestrian's positive crossing intent refers to observable behaviour and cues exhibited by a pedestrian, indicating a deliberate intention to cross

* Corresponding author.
E-mail addresses: nehashrm013@gmail.com (N. Sharma), chhavi.dhiman@dtu.ac.in (C. Dhiman), s.indu@dce.ac.in (S. Indu).

https://doi.org/10.1016/j.patcog.2024.111205
Received 7 November 2023; Received in revised form 6 September 2024; Accepted 15 November 2024
Available online 26 November 2024

- **N. Sharma**, C. Dhiman, and S. Indu, "Cross-Modal Pedestrian Behavior Prediction: A Dual Task Approach with Progressive Denoising Attention and CVAE". ***IEEE Transactions on Intelligent Transportation Systems,*** doi: 10.1109/TITS.2025.3578023. **Impact Factor: 8.4**



Cross-Modal Pedestrian Behavior Prediction: A Dual-Task Approach With Progressive Denoising Attention and CVAE

Neha Sharma, *Member, IEEE*, Chhavi Dhiman, *Member, IEEE*, and Sreedevi Indu, *Senior Member, IEEE*

***Abstract*— Pedestrian intention and trajectory prediction are crucial for advancing intelligent transportation systems and autonomous vehicles, significantly enhancing urban mobility's safety and efficiency. Traditional approaches have evolved from capturing pedestrian dynamics through image features and bounding box coordinates to leveraging multiple modalities and attention mechanisms. However, challenges in robust cross-modal feature integration and adaptation to complex scenarios persist. This paper introduces a dual-task approach that simultaneously predicts short-term pedestrian crossing intentions and long-term trajectories by integrating features from pedestrian regions of interest (ROIs), scene attributes, and past trajectories. For crossing intention prediction, Progressive Denoising Attention (PDA) is developed, which iteratively refines cross-modal features to augment inter-class variations. Additionally, a three-phase counterfactual training approach is employed that manipulates pedestrian ROIs and segmentation maps to further enhance model robustness in complex scenarios. For trajectory prediction, a Conditional Variational Autoencoder (CVAE) is implemented, guided by contextual embeddings from the novel Context-Aware Feature Fusion Module (CAFFM) to significantly reduce mean squared error by integrating rich spatiotemporal ROI and context information. Experimental results on benchmark datasets JAAD and PIE demonstrate the superior performance of the proposed approach in understanding and predicting pedestrian intent. The code is available at: https://github.com/neha013/DPITRA**

***Index Terms*— Pedestrian trajectory prediction, Pedestrian intention estimation, ADAS, intelligent vehicles, CVAEs.**

I. Introduction

PEDESTRIAN intention and trajectory prediction play a pivotal role in advancing intelligent transportation systems and autonomous vehicles. Accurate forecasting of pedestrian behaviour can substantially enhance the safety and efficiency of urban mobility by addressing the dynamic and complex nature of human actions within diverse environmental contexts. Early approaches [1], [2], [3] to pedestrian intention prediction primarily focused on capturing pedestrian dynamics through image features and bounding box coordinates. These models, however, were constrained by their inability to incorporate richer contextual information, limiting their predictive accuracy. Subsequent advancements [4] involved encoding multiple modalities to improve predictions. Nevertheless, these models often struggled to adapt to diverse environments due to their heavy reliance on semantic scene parsing. More recent efforts [5], [6], [7], [8] have employed attention mechanisms and transformers to fuse spatiotemporal features. Yet, these models frequently encountered difficulties with robust cross-modal feature integration, particularly in challenging scenarios characterized by noisy or missing data.

Concurrently, the field of pedestrian trajectory prediction has seen the exploration of various methodologies, including generative models [9], [10], [11], that address the multimodality and uncertainty inherent in human movements. While Generative Adversarial Networks (GANs) have been investigated for trajectory prediction, they often suffer from issues such as mode collapse and training instability. In contrast, Variational Autoencoders (VAEs) provide a more stable and reliable approach by learning latent space representations that capture the underlying structure of trajectory data [12]. Despite these advancements, integrating environmental context remains a significant challenge, limiting the optimal performance of these models.

To address these limitations, this work introduces a **D**ual-task approach for **P**rediction of Pedestrian **I**ntention and **TRA**jectory (DPITRA) utilizing pedestrian ROIs, scene attributes, and past trajectories. The salient contributions of our paper are the following:

- Development of the Progressive Denoising Attention (PDA), inspired by diffusion models, which iteratively refines cross-modal features to enhance inter-class separation between crossing and non-crossing samples.
- Integration of a systematic three-phase counterfactual training approach manipulating pedestrian ROIs and segmentation maps to strengthen the model's understanding of causal relationships between contextual elements and pedestrian behaviour.
- Implementation of a Conditional Variational Autoencoder (CVAE) for long-term trajectory prediction, guided by contextual embeddings from the novel Context-Aware Feature Fusion Module (CAFFM), to improve prediction accuracy by incorporating rich pedestrian ROI and scene features.

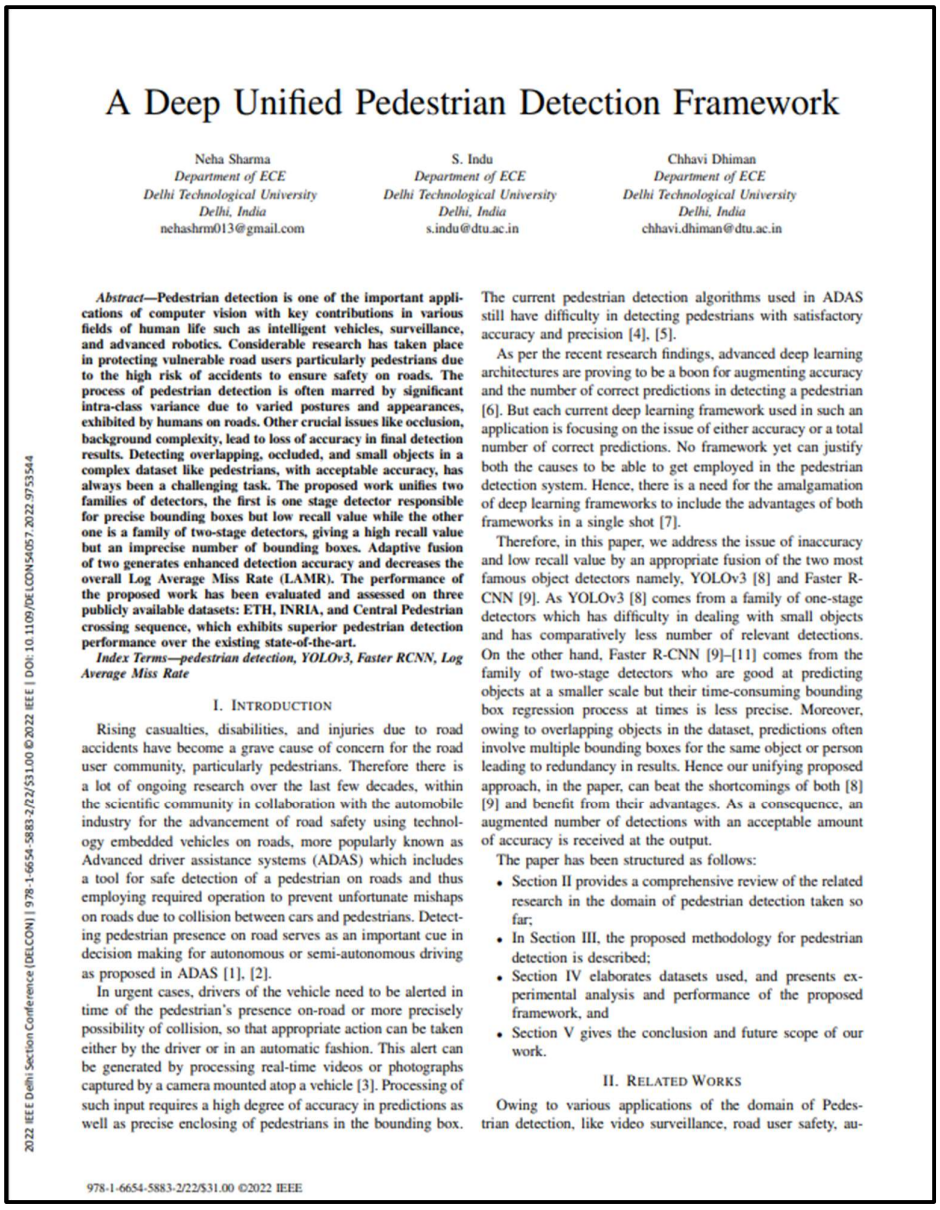
Furthermore, the experimental results demonstrate that the proposed approach achieves superior performance

Received 4 December 2024; revised 28 February 2025 and 23 April 2025; accepted 31 May 2025. The Associate Editor for this article was Z. Li. *(Corresponding author: Chhavi Dhiman.)*

The authors are with the Department of Electronics and Communication and Engineering, Delhi Technological University (DTU), Delhi 110042, India (e-mail: nehashrm013@gmail.com; chhavi.dhiman@dtu.ac.in; s.indu@dce.ac.in).

Digital Object Identifier 10.1109/TITS.2025.3578023

- **N. Sharma**, C. Dhiman, and S. Indu, "A Deep Unified Pedestrian Detection Framework", In *IEEE Delhi Section Conference (DELCON),* pp. 1-6, 2022, doi: 10.1109/DELCON54057.2022.9753544



A Deep Unified Pedestrian Detection Framework

Neha Sharma
Department of ECE
Delhi Technological University
Delhi, India
nehashrm013@gmail.com

S. Indu
Department of ECE
Delhi Technological University
Delhi, India
s.indu@dtu.ac.in

Chhavi Dhiman
Department of ECE
Delhi Technological University
Delhi, India
chhavi.dhiman@dtu.ac.in

2022 IEEE Delhi Section Conference (DELCON) | 978-1-6654-5883-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/DELCON54057.2022.9753544

***Abstract*—Pedestrian detection is one of the important applications of computer vision with key contributions in various fields of human life such as intelligent vehicles, surveillance, and advanced robotics. Considerable research has taken place in protecting vulnerable road users particularly pedestrians due to the high risk of accidents to ensure safety on roads. The process of pedestrian detection is often marred by significant intra-class variance due to varied postures and appearances, exhibited by humans on roads. Other crucial issues like occlusion, background complexity, lead to loss of accuracy in final detection results. Detecting overlapping, occluded, and small objects in a complex dataset like pedestrians, with acceptable accuracy, has always been a challenging task. The proposed work unifies two families of detectors, the first is one stage detector responsible for precise bounding boxes but low recall value while the other one is a family of two-stage detectors, giving a high recall value but an imprecise number of bounding boxes. Adaptive fusion of two generates enhanced detection accuracy and decreases the overall Log Average Miss Rate (LAMR). The performance of the proposed work has been evaluated and assessed on three publicly available datasets: ETH, INRIA, and Central Pedestrian crossing sequence, which exhibits superior pedestrian detection performance over the existing state-of-the-art.**

***Index Terms*—pedestrian detection, YOLOv3, Faster RCNN, Log Average Miss Rate**

I. Introduction

Rising casualties, disabilities, and injuries due to road accidents have become a grave cause of concern for the road user community, particularly pedestrians. Therefore there is a lot of ongoing research over the last few decades, within the scientific community in collaboration with the automobile industry for the advancement of road safety using technology embedded vehicles on roads, more popularly known as Advanced driver assistance systems (ADAS) which includes a tool for safe detection of a pedestrian on roads and thus employing required operation to prevent unfortunate mishaps on roads due to collision between cars and pedestrians. Detecting pedestrian presence on road serves as an important cue in decision making for autonomous or semi-autonomous driving as proposed in ADAS [1], [2].

In urgent cases, drivers of the vehicle need to be alerted in time of the pedestrian's presence on-road or more precisely possibility of collision, so that appropriate action can be taken either by the driver or in an automatic fashion. This alert can be generated by processing real-time videos or photographs captured by a camera mounted atop a vehicle [3]. Processing of such input requires a high degree of accuracy in predictions as well as precise enclosing of pedestrians in the bounding box. The current pedestrian detection algorithms used in ADAS still have difficulty in detecting pedestrians with satisfactory accuracy and precision [4], [5].

As per the recent research findings, advanced deep learning architectures are proving to be a boon for augmenting accuracy and the number of correct predictions in detecting a pedestrian [6]. But each current deep learning framework used in such an application is focusing on the issue of either accuracy or a total number of correct predictions. No framework yet can justify both the causes to be able to get employed in the pedestrian detection system. Hence, there is a need for the amalgamation of deep learning frameworks to include the advantages of both frameworks in a single shot [7].

Therefore, in this paper, we address the issue of inaccuracy and low recall value by an appropriate fusion of the two most famous object detectors namely, YOLOv3 [8] and Faster R-CNN [9]. As YOLOv3 [8] comes from a family of one-stage detectors which has difficulty in dealing with small objects and has comparatively less number of relevant detections. On the other hand, Faster R-CNN [9]–[11] comes from the family of two-stage detectors who are good at predicting objects at a smaller scale but their time-consuming bounding box regression process at times is less precise. Moreover, owing to overlapping objects in the dataset, predictions often involve multiple bounding boxes for the same object or person leading to redundancy in results. Hence our unifying proposed approach, in the paper, can beat the shortcomings of both [8] [9] and benefit from their advantages. As a consequence, an augmented number of detections with an acceptable amount of accuracy is received at the output.

The paper has been structured as follows:

- Section II provides a comprehensive review of the related research in the domain of pedestrian detection taken so far;
- In Section III, the proposed methodology for pedestrian detection is described;
- Section IV elaborates datasets used, and presents experimental analysis and performance of the proposed framework, and
- Section V gives the conclusion and future scope of our work.

II. Related Works

Owing to various applications of the domain of Pedestrian detection, like video surveillance, road user safety, au-

978-1-6654-5883-2/22/$31.00 ©2022 IEEE

- **N. Sharma**, C. Dhiman, and S. Indu, "Intelligent Pedestrian Intention Prediction Framework" In *IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, pp. 1-5, 2022 doi: 10.1109/SOLI57430.2022.10295014.

16th IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI 2022)

Intelligent Pedestrian Intention Prediction Framework

Neha Sharma, Chhavi Dhiman, S. Indu
Department of ECE
Delhi Technological University
Delhi, India
nehashrm013@gmail.com, chhavi.dhiman@dtu.ac.in, s.indu@dtu.ac.in

2022 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI) | 979-8-3503-3224-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/SOLI57430.2022.10295014

***Abstract*—One of the most critical tasks in autonomous driving is anticipating pedestrian crossing intention on roads to ensure safe and reliable driving. This will instil trust in the road user community in driving assistance endeavours from Advanced Driving Assistance Systems (ADAS) to Autonomous Vehicles (AVs) encouraging their co-existence. In this paper, a cascade of three modules is employed, the convolution module that acts as a feature extractor, the recurrent module that is used for sequential tasks followed by classification module. It is shown that with the help of information regarding the past trajectory, appearance of pedestrians and the ego-vehicle speed, the proposed data-driven approach is able to predict pedestrian crossing intention reliably. The proposed algorithm is able to anticipate crossing intention in two publicly available benchmark datasets, JAAD and PIE with an accuracy of 88% and 86% respectively.**

***Index Terms*—pedestrian intention, autonomous vehicles, advanced driver assistance systems, data-driven, deep learning**

I. Introduction

During the span of the last few years, there seems to be substantial evolution in ADAS and AV technology when it comes to object classification and localisation [1]. This can be attributed to the success of advanced deep learning architectures that are improving every day with the amount of increasing data offered to them. However, comprehension of the environmental stimulus and human response to it is still a daunting task for AV systems. Human intuition helps in manoeuvring in ambiguous and challenging real-life driving scenarios where there is a continuous passage of pedestrians, vehicles and other road users. Autonomous driving systems, however, programmed well, fail to understand this inherent complexity of human behaviour. Therefore, an autonomous vehicle must imbibe this behavioural and contextual understanding of the surrounding situation to achieve a smooth and human-like driving experience.

The comprehension of pedestrian intention requires deeper analysis and understanding of prior action states. Several pivotal works [2] employ motion history to anticipate the pedestrian's forthcoming crossing intention or even the whole trajectory. However, such works fail to capture pedestrians' unpredictable motion dynamics such as an abrupt change in motion, direction or velocity and even irregular walking patterns that confuse the learning models leading to erroneous predictions. Therefore, the need arises to employ other vital sources of information like visual features and also the ego-vehicle speed that shall help make the model learning more robust and unaffected by variations in pedestrian motion dynamics with respect to the onboard camera [3].

In this paper, a data-driven deep learning-based pedestrian intention prediction model is proposed employing three modules in succession that are convolution, recurrent fusion and classification. The convolution module process visual features that can be extracted from the image while the recurrent module processes non-visual features like trajectory. The ego-vehicle speed is also taken into consideration since the relative distance and speed of the pedestrian with respect to the ego-vehicle is not constant but dynamic in nature. The rest of the paper is structured in the following fashion: Section II introduces several state-of-the-art methods and their noteworthy contributions; Section III elaborates on the proposed methodology for pedestrian crossing intention prediction, Section IV covers the discussion of datasets, the experimental setup, results, and their analysis. Finally, in Section V, the paper's conclusions and potential future directions are presented."

II. Related Works

This section outlines several state-of-the-art methods proposed for pedestrian crossing intention predictions in the last few years. The preliminary works [4], [5] in the field of designing pedestrian intention frameworks required dynamic motion modelling. These works fail to be robust in case of abrupt motion variations of a pedestrian. The next category of research works [6], [7] demanded a set of prior end goals that becomes a bottleneck provided the difficulty in assessing endpoints from the onboard camera. The last and most recent category is data-driven approaches [8], [9] that do not require either motion modelling or prior end goal estimates. Motivated by the success of deep learning approaches in several other fields and their increasing utilisation in predicting human-computer interaction, this section limits the discussion to data-driven approaches provided the demerits of former techniques as discussed before.

Quite a few pioneer research works in the field of pedestrian intention prediction utilised multiple information related to pedestrians and their surroundings like pose [10], [11], motion history [2], [12], head orientation [13], ego-vehicle speed

- N. Sharma, C. Dhiman, and S. Indu, “LLM-Guided Visual Reasoning for Scene-Aware Pedestrian Intention Prediction” In *International Conference on Pattern Recognition and Machine Intelligence (PReMI)*, 2025 (**Accepted**).

Gmail — Neha Sharma <nehashrm013@gmail.com>

PReMI 2025 – Final Acceptance Notification

Sent on behalf of Dr. Sriparna Saha <meteor.support@springernature.com> Mon, Sep 22, 2025 at 9:44 PM
Reply-To: "Dr. Sriparna Saha" <sriparna@iitp.ac.in>
To: nehashrm013@gmail.com

Dear Author,

We are delighted to inform you that your submission entitled "LLM-Guided Visual Reasoning for Scene-Aware Pedestrian Intention Prediction" has been accepted for publication at the 11th International Conference on Pattern Recognition and Machine Intelligence (PReMI 2025) going to be held at IIT Delhi from December 11-14, 2025. Congratulations!

STATISTICS: As for every edition of the conference, the reviewing process was thorough and highly selective. We received a total of 426 submissions and accepted 167 of them, corresponding to an acceptance rate of 39% The list of accepted papers is available at the conference website.

REVIEWS: Every submission was reviewed by two or more experts in the field. In case of conflicts, meta reviews were considered. The reviews were then discussed by Program Committee Chairs and final decisions were taken. The competition was intense, and regrettably, we had to reject several good papers simply because of limited space.

CAMERA-READY: You are now required to revise your paper and prepare the camera-ready version for inclusion in the conference proceedings. While doing so, we encourage you to carefully consider the reviews received. Even if you may not fully agree with certain feedback or feel that a reviewer has misunderstood a point, such comments can still help you refine and improve the clarity of your presentation. A detailed email with instructions for submitting the camera-ready paper will be sent to you shortly.

You will find your final reviews at the METEOR website. We hope that you will find the feedback received helpful.

We are looking forward to seeing you at IIT Delhi this December!

Congratulations once again!

Program Chairs

DELHI TECHNOLOGICAL UNIVERSITY

Formerly Delhi College of Engineering

Shahbad Daulatpur, Main Bawana Road, Delhi –42

PLAGIARISM VERIFICATION

Title of the Thesis: **Pedestrian Intention Prediction for Autonomous Vehicles**

Total Pages: **154**

Name of the Scholar: **Neha Sharma**

Supervisor: **Prof. S. Indu**

Co-Supervisor: **Dr. Chhavi Dhiman**

Department: **Electronics and Communication Engineering**

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software used: **Turnitin**

Submission ID: **trn:oid:::27535:96977641**

Similarity Index: **30%**

Self-Publication(s) Similarity Index: **24%**

Final Total Similarity Index: **6%**

Total Word Count: **45642**

Date: **May 27, 2025**

Candidate's Signature **Signature of Supervisor(s)**

CURRICULUM VITAE OF MS. NEHA SHARMA

Education

❖ **Ph.D., ECE (Computer Vision and Deep Learning)** 01/2021-09/2025
Thesis title: Pedestrian Intention Prediction for Autonomous Vehicles
Supervisor: Prof. S. Indu
Co-Supervisor: Dr. Chhavi Dhiman
Dept. of ECE, Delhi Technological University (DTU), Bawana Road, Shahbad Daulatpur Village, Delhi 110042, India

❖ **M.Tech., ECE (Signal Processing and Digital Design)** 08/2017-06/2019

Thesis title: Statistical watermarking approach for 3D mesh using local curvature estimation
Supervisor: Prof. J. Panda
Dept. of ECE, Delhi Technological University (DTU), Bawana Road, Shahbad Daulatpur Village, Delhi 110042, India

❖ **B.Tech., Electronics and Communication Engineering** 08/2017-06/2019

Thesis title: Face Feature Extraction using Gabor Wavelet Transform and Local Binary Pattern
Supervisor: Dr. Jasdeep Kaur Dhanoa
Dept. of ECE, Indira Gandhi Delhi Technical University for Women (IGDTUW), Kashmere Gate, New Delhi 110006, India

Skills

✓ Speaking languages: English and Hindi
✓ Coding: Python, MATLAB, LaTeX, C++, VHDL and Verilog
✓ Software/Tools: MATLAB, Arduino, vlabs COA Simulator, MPLAB IDE, PROTEUS, PIC simulator IDE and Mendeley/Zotero
✓ Research Interests: Pedestrian Intention Prediction, Computer Vision, Machine learning, Pattern Recognition

Professional Experiences:

- ML System Architect, CraftifAI (*full-time*) 06/2025-*present*
- Full-time DTU Fellow (teaching assistant) 01/2021-05/2025
  (UAS Lab, Dept. of ECE, DTU)
- Assistant Professor 06/2019-03/2021
  (Dept. of Electrical and Electronics Engineering, KIET Group of Institutions, Ghaziabad, Uttar Pradesh)

Awards/ Honors

❖ Research Excellence Award 2021, 2023, 2024
Published three SCI-indexed journal paper in Dept. of ECE, DTU Delhi.