DEVELOPMENT OF GRAPH NEURAL NETWORK APPROACH FOR PEDESTRIAN TRAJECTORY PREDICTION

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

MASTER OF TECHNOLOGY

in

Artificial Intelligence

by

Vedant Dhanyamraju
(Roll No. 2K24/AFI/11)

Under the Supervision of

Prof. Anil Singh Parihar
Professor, Department of Computer Science and Engineering
Delhi Technological University

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

May, 2026

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

ACKNOWLEDGEMENTS

I wish to express my sincerest gratitude to **Prof. Anil Singh Parihar** for their continuous guidance and mentorship that they provided me during the project. They showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this project as well as its industrial relevance. They were always ready to help me and clear my doubts regarding any hurdles in this project. Without their constant support and motivation, this project would not have been successful.

Place: Delhi Vedant Dhanyamraju

Date: May, 2026

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

CANDIDATE'S DECLARATION

I, **Vedant Dhanyamraju**, Roll No – **2K24/AFI/11**, student of M.Tech (**Artificial Intelligence**), hereby declare that the project dissertation titled "**DEVELOPMENT OF GRAPH NEURAL NETWORK APPROACH FOR PEDESTRIAN TRAJECTORY PREDICTION**," which is submitted by me to the **Department of Computer Science and Engineering**, Delhi Technological University, Delhi, in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Associate ship, Fellowship, or other similar title or recognition.

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.

Signature of Supervisor **Signature of External Examiner**

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

CERTIFICATE BY THE SUPERVISOR(s)

Certified that **Vedant Dhanyamraju** (Roll No. **2K24/AFI/11**) has carried out the research work presented in this thesis entitled **"DEVELOPMENT OF GRAPH NEURAL NETWORK APPROACH FOR PEDESTRIAN TRAJEC- TORY PREDICTION"** for the award of Master of Technology from the Department of **Computer Science and Engineering**, Delhi Technological University, Delhi, under my supervision.

The thesis embodies results of original work, and studies are carried out by the student themselves, 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.

Signature

Prof. Anil Singh Parihar
Professor
Department of Computer Science and Engineering
Delhi Technological University

Date: May, 2026

DEVELOPMENT OF GRAPH NEURAL NETWORK APPROACH FOR PEDESTRIAN TRAJECTORY PREDICTION

Vedant Dhanyamraju

ABSTRACT

Pedestrian trajectory prediction is a fundamental problem in autonomous systems, robotics, and self-driving vehicles, where accurately forecasting future human motion is critical for safe navigation and interaction. The task is particularly challenging due to the highly dynamic, socially interactive, and inherently multimodal nature of pedestrian behavior.

This thesis first presents a comprehensive study of the pedestrian trajectory prediction domain, ranging from physics-based models and classical machine learning techniques to modern deep learning and multimodal generative frameworks. Existing methods are analyzed using commonly adopted evaluation metrics such as Average Displacement Error (ADE) and Final Displacement Error (FDE). The review highlights several persistent limitations in current systems, including poor generalization across scenes, limited robustness to irregular crowd behavior, computational complexity, and challenges in modeling diverse future intentions.

Motivated by these observations, this thesis proposes two novel deep learning architectures for multimodal pedestrian trajectory forecasting. The first proposed model, SWIFT-Net (Slot-guided Wavelet Interaction Fusion for Trajectories), introduces a hierarchical framework for capturing both local and global interaction patterns. Separate spatial and temporal interaction graphs are constructed and decomposed using wavelet-based multi-scale feature extraction, enabling the model to learn short-range and long-range dependencies simultaneously. An adaptive fusion mechanism together with slot attention disentangles distinct behavioral patterns within crowds, allowing the model to generate multiple socially plausible future trajectories. The second proposed model, TRACE-Net (Trajectory Reasoning via Adaptive Competitive Encodings), focuses on adaptively modeling irregular crowd dynamics through deformable spatio-temporal reasoning. The architecture employs asymmetric and deformable 3D convolutions to dynamically adjust receptive fields based on interaction patterns. In addition, a Competitive Attention mechanism is introduced to separate latent multimodal intentions into distinct behavioral encodings. These representations are further processed using an Attention-Gated Temporal Decoder to generate diverse and socially consistent future trajectories.

Extensive experiments conducted on the ETH and UCY benchmark datasets demonstrate the effectiveness of the proposed approaches in improving trajectory prediction accuracy while maintaining diversity and social plausibility. Overall, this thesis contributes both a detailed understanding of the current state of pedestrian trajectory prediction research and novel architectures designed to address key limitations in existing methods.

LIST OF PUBLICATIONS

Conference Proceedings

1. **Vedant Dhanyamraju**, Kavinder Singh, and Anil Singh Parihar, 2026. SWIFT-Net: Slot-Guided Wavelet Interaction Fusion for Trajectories In: *Proceedings of the 2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)*, Manila, Philippines, June 11-14, 2026. (Registered for Oral Presentation, Paper ID: 1571282710).

2. **Vedant Dhanyamraju**, Kavinder Singh, and Anil Singh Parihar, 2026. Beyond Fixed Kernels: Dynamic Spatiotemporal Interaction Graphs for Trajectory Forecasting In: *Proceedings of the 2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)*, Manila, Philippines, June 11-14, 2026. (Registered for Oral Presentation, Paper ID: 1571283568).

TABLE OF CONTENTS

Acknowledgements **ii**

Candidate's Declaration **iii**

Certificate by the Supervisor(s) **iv**

Abstract **v**

List of Publications **vi**

List of Tables **ix**

List of Figures **x**

List of Symbols, Abbreviations and Nomenclature **xii**

1 INTRODUCTION **1**

2 LITERATURE REVIEW **7**
2.1 Physics-Based and Classical Machine Learning Approaches 7
2.2 Emergence of RNN-Based Sequence Modelling 8
2.3 Transformer-Based Techniques of Global Interaction 8
2.4 Graph Neural Networks for Social and Spatio-Temporal Reasoning . . 9
2.5 Multi-Scale Interaction Learning and Deformable Modelling 9
2.6 Multi-Source Feature Fusion and Wavelet-Based Modelling 10
2.7 Multimodal Prediction 10
2.8 Interaction-Aware Recurrent Models and Social Context 11
2.9 Goal-, Endpoint-, and Scene-Conditioned Prediction 12
2.10 Dynamic Graphs and Hierarchical Relational Reasoning 12
2.11 Multi-Scale and Deformable Modelling 13
2.12 Multi-Stream Feature Learning and Hybrid Representations 14
2.13 Multimodal Prediction and Uncertainty Modelling 14
2.14 Evaluation Protocols and Practical deployment considerations 15
2.15 Open Challenges and Research Motivation 16

3 METHODOLOGY **17**
3.1 PROBLEM FORMULATION AND EVALUATION METRICS . . . 17
3.2 SWIFT-Net 18

3.2.1 Constructing Spatial and Temporal Graphs 18
3.2.2 Hierarchical Wavelet Interaction Module 19
3.2.3 Graph Convolution for Feature Extraction 20
3.2.4 Adaptive Spatio-Temporal Fusion Module 20
3.2.5 Slot Attention Module 21
3.2.6 Temporal Convolutional Network for Multi-trajectory Prediction 22
3.3 TRACE-Net 22
3.3.1 Constructing the Spatial Graph 23
3.3.2 Adaptive Spatio-Temporal Interaction Module 23
3.3.3 Graph Convolution and Feature Representation 24
3.3.4 Competitive Attention Module 24
3.3.5 Attention-Gated Temporal Decoder 25
3.3.6 Latin Hypercube Sampling 26
3.3.7 Training Using Negative Log-Likelihood 26

4 RESULTS AND DISCUSSION 27
4.1 Results of SWIFT-Net 27
4.2 Results of Trace-Net 28

5 CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT 31
5.1 Conclusion 31
5.2 Future Scope 31
5.3 Social Impact 32

References 37

A PROOF OF PUBLICATION 38

List of Tables

4.1 Results of SWIFT-Net : ADE/FDE Metrics (in meters) of Our Implementations on ETH and UCY Datasets. 28
4.2 Results of TRACE-Net : ADE/FDE Metrics (in meters) of Our Implementations on ETH and UCY Datasets. 29

List of Figures

3.1 Architecture of SWIFT-Net 18
3.2 Architecture of TRACE-Net 23

4.1 Prediction error at different prediction times of the different methods on the ETH/UCY datasets. 28
4.2 The average ADE and FDE curves on the ETH and UCY datasets using the different sampling methods at different sampling numbers 29

LIST OF SYMBOLS

i	Index of a pedestrian
t	Time index
N	Number of pedestrians in a scene
n	Number of pedestrians in the loss formulation
K	Number of slots / candidate trajectories
S	Number of trajectory samples
T_{obs}	Number of observed time steps
T_{pred}	Number of predicted time steps
(x_i^t, y_i^t)	2D position of pedestrian i at time t
p_t^i	Position vector of pedestrian i at time t
o_t^i	Relative displacement of pedestrian i between consecutive frames
$\mathbf{X}_i$	Observed trajectory of pedestrian i
$\mathbf{Y}_i$	Future trajectory of pedestrian i
$\mathbf{y}_i^t$	Ground-truth position of pedestrian i at time t
$\hat{\mathbf{y}}_i^t$	Predicted position of pedestrian i at time t
ADE	Average Displacement Error
FDE	Final Displacement Error
G_S	Spatial graph
G_T	Temporal graph
V_t	Node set at time t
U_t	Edge set / adjacency matrix at time t
$R_{S/T}$	Initial spatial / temporal interaction score matrix
R_{QK}	Initial spatial interaction score matrix
$R^{(l)}_{S/T,c_0c_1}$	Wavelet coefficient at level l
$C^{(l)}_{S/T,c_0c_1}$	Processed wavelet coefficient at level l
$A_{S/T}$	Reconstructed adjacency matrix
LL, LH, HL, HH	Wavelet sub-bands (low-low, low-high, high-low, high-high)
$\downarrow 2$	Downsampling operator
$\uparrow 2$	Upsampling operator
E_S	Graph embedding of the spatial graph
Q, K	Query and key matrices
$\phi(\cdot)$	Feature embedding function
δ	Non-linear activation function

ψ_1, ψ_2, ψ_3	Learnable convolution operators / kernels
W_{GS}, W_{GT}, W_{ST}	Graph convolution weight matrices
W_S, W_C, W_T	Spatial, channel, and feature attention weights
$\mathbf{H}_S$	Spatial interaction features
$\mathbf{H}_T$	Temporal interaction features
$\mathbf{H}_{ST}$	Spatio-temporal feature representation
$\mathbf{H}_{STF}$	Fused spatio-temporal feature representation
H'	Slot-enhanced feature tensor
X	Per-pedestrian feature embedding after temporal compression
A	Slot attention matrix
U	Weighted aggregation of input features
S_0	Initial slot vectors
S_t	Slot representation at iteration t
S_{agg}	Aggregated slot representation
GRU	Gated Recurrent Unit
MLP	Multi-Layer Perceptron
ΔP	Deformable offset tensor
P^*	Shifted feature tensor after offset addition
γ_Q	Query-dependent transformed feature representation
R_{ST}^l	Refined spatio-temporal representation at layer l
A_{ST}	Refined spatio-temporal adjacency matrix
$\mu_{x,ti}, \mu_{y,ti}$	Mean components of the predicted 2D Gaussian
$\sigma_{x,ti}, \sigma_{y,ti}$	Standard deviation components of the predicted 2D Gaussian
ρ_{ti}	Correlation coefficient of the predicted 2D Gaussian
LHS	Latin Hypercube Sampling
Z	Standard normal transformed sample
L	Cholesky factor
L_i^W	Negative log-likelihood loss for sample i
L_w	Winner-takes-all loss
L_o	Overall prediction loss
L	Total training loss
$\mathbb{R}$	Set of real numbers
$\mathcal{N}(\mu, \sigma)$	Gaussian distribution with mean μ and standard deviation σ
$\Phi^{-1}(\cdot)$	Inverse Gaussian cumulative distribution function

CHAPTER 1

INTRODUCTION

Pedestrian trajectory prediction is critical for a wide range of intelligent systems, including self-driving cars, social robots and smart surveillance. Stand at a busy street for some minutes and two things becomes clear , people are constantly making split-second decisions about where to go next and that collisions are very rare. We read social cues and anticipate group flows while dodging strangers , and also coordinate with dozens of people all without thinking. Teaching a machine to do this is an interesting problem in compute vision and intelligent systems. Accurate pedestrian trajectory prediction is important for many applications: autonomous vehicles need it to navigate safely and surveillance systems rely on it to flag unusual behaviour while social robots need it to move through human spaces without being awkward or obstructing flow.

The challenge extends far beyond estimating a simple walking direction. Human movement is continuously shaped by immediate neighbors, broader crowd patterns, and the physical constraints of the scene—dynamics that are notoriously difficult to model. Human locomotion is continually informed by an interplay between individual intent, social pressure and environmental constraints. As such, trajectory prediction is a problem of predicting behavior: a single past can lead to multiple futures based on la-tent social context.

Early solutions relied on hand-crafted rules, often treating pedestrians as particles driven by social forces [1]. While interpretable, these techniques were far too rigid. Human behavior is highly diverse and context-dependent; it cannot be locked into a static set of mathematical rules. These methods were basic, yet insufficient to model the complex and varied nature of crowd movement in the real world [2, 3]. The advent of deep learning marked a breakthrough. Recurrent architectures like Social-LSTM [2] and Social-GAN [3] proved that data-driven interaction modeling yields significantly more realistic trajectory predictions. Recurrent networks [4] and, more recently, Graph Neural Networks (GNNs) [5, 6] demonstrate that data-driven interaction models lead to much more realistic predictions by learning spatial and temporal interactions. More recently, Transformers have been adopted to capture global attention across all agents in a scene, while Graph Neural Networks (GNNs) provide a native way to model the continuously evolving spatial relationships between pedestrians [4, 5, 6, 7, 8].

Despite this progress, important bottlenecks remain in current architectures. Most existing models evaluate pedestrian interactions at a single scale. They either lock onto localized, pairwise interactions or zoom out to capture global crowd dynamics,

but they rarely handle both at the same time. This is a severe limitation, as real-world human behavior is inherently multi-scale. Someone navigating a dense crowd is simultaneously reacting to the person directly in front of them and adjusting to the general flow of the group twenty meters ahead. Capturing both levels of detail and understanding how they interact is essential for generating truly plausible predictions.

At the same time, existing models often operate on spatial and temporal features separately or rely on a fixed convolution kernel for spatial aggregation. Yet, human interactions are dynamic and intrinsically geometric; people don't move on fixed grids or interact at fixed temporal resolutions. Furthermore, although the need for multimodal prediction is well acknowledged - often resolved through Generative Adversarial Networks (GANs) [3] or a set of trajectory anchors [9, 10] - these approaches often fail to explicitly separate multimodal intent. They are prone to yield unrealistic, mixed-mode trajectories due to mode entanglement, or are inherently limited by the low spatial coverage of their anchor sets.

To address these limitations, this thesis proposes two complementary frameworks for multimodal pedestrian trajectory prediction. The first framework, SWIFT-Net, is built around the insight that pedestrian interactions must be modeled hierarchically across multiple spatial and temporal scales. SWIFT-Net processes pedestrian interactions through a hierarchical wavelet decomposition module that explicitly separates and analyzes features at multiple spatial and temporal scales. Rather than discarding fine-grained details in favor of global patterns or vice versa, it retains both and learns how to fuse them adaptively. A spatio-temporal fusion block modulates how much weight to place on spatial versus temporal cues depending on the local context giving more weight to social proximity in dense crowds, for instance, and leaning on temporal continuity when a pedestrian is moving freely. A slot attention module then breaks it into latent behavioral modes present in the fused features, allowing the model to explicitly represent the multi-modal nature of future motion. Finally, a temporal convolutional network directly predicts multiple candidate trajectories, trained to be both accurate and diverse. We evaluate SWIFT-Net on the standard ETH and UCY benchmark datasets and show consistent, significant improvements over prior state-of-the-art methods.

The second framework, TRACE-Net (Trajectory Reasoning via Adaptive Competitive Encodings), focuses on explicitly disentangling multimodal human intention while modeling fluid social interactions. Rethinking the conventional spatial pooling, TRACE-Net introduces an Adaptive Spatio-Temporal Interaction Module. This module employs deformable 3D convolution with asymmetrically-shaped kernels to adaptively reshape its receptive field. This enables the network to learn the exact spatial and temporal offsets for the interactions, naturally adapting to the geometry of crowd dynamics.

In addition, to account for multimodal intent without explicitly sampling from anchor points, TRACE-Net introduces a Competitive Attention module. This module explicitly disentangles the joint spatio-temporal features into multiple latent "slots" which compete to represent the different modes of crowd behavior. By disentangling the movements into isolated actions, such as yielding, turning and accelerating, the model is able to plan independent motion patterns. Finally, disentangled encodings are

fed to an Attention-Gated Temporal Decoder which, in combination with Latin Hypercube Sampling, yields a statistically confident and highly diverse set of long-term predictions.

Together, the proposed frameworks aim to improve the accuracy, diversity and realism of pedestrian trajectory prediction by addressing key limitations in existing spatio-temporal and multimodal interaction modeling approaches.

In addition to social interactions, scene understanding is also a crucial component of the pedestrian trajectory prediction. The geometry and semantics of the environment have a strong constraint on human movement. It is natural for people to prefer walking on sidewalks over walking on roads, avoid static obstacles, follow entrances and exits, and change their paths as per the available free space. Therefore, the problem of trajectory prediction cannot be considered as a problem of sequence forecasting alone, but also needs to reflect the structure of an environment where the motion occurs. In most early trajectory prediction systems, only position coordinates were used, and they were often unrealistic, such as going through walls or ignoring scene semantics. In recent works, semantic maps, occupancy grids, bird's-eye-view representation and visual scene encoding are thus combined to enhance the environmental awareness [4, 7]. Modern architectures are capable of generating physically plausible trajectories directly from the spatial context, alongside being mathematically accurate.

Uncertainty is one of the other big difficulties for pedestrian trajectory prediction. Human behaviour is inherently multimodal: when a pedestrian approaches an intersection, she may go straight, turn left, stop or change her direction based on subtle factors that are not directly observable in the context. What this implies is that there are more than one socially acceptable futures for a single observed history. Traditional deterministic predictors are not very successful in such cases, looking for a "mean" trajectory, in the process sometimes generating unrealistic trajectories that do not represent any valid human behaviour. Given this, multi-modal forecasting has been the subject of recent attention, including methods like Generative Adversarial Networks (GANs), latent variable models, and trajectory sampling strategies (see [3]). The models attempt to produce a range of potential and not a single deterministic outcome. Multimodal prediction has therefore dropped the emphasis on the minimization of the positional error and moved towards modelling uncertainty, diversity and social plausibility of trajectories.

In the field of architecture, the prediction of trajectories has undergone very marked developments during the last decade. Early models used Long Short-Term Memory (LSTM) networks as they were able to model temporal dependencies in sequential data, as mentioned in [2]. These methods encoded trajectories as a function of time and then autoregressively decoded future coordinates. These worked well, but there were problems with modelling long-range dependence and the sequential computation overhead in recurrent models. This led to the use of convolutional and graph-based techniques which would allow for more efficient processing of interactions in parallel. A natural solution for encoding pedestrians as nodes and their interactions as dynamically changing edges was proposed by Graph Neural Networks (GNNs) [5, 6]. Relational reasoning is explicitly modeled in such approaches and they offer better scalability in dense scenes with complex interactions.

Later, transformer-based methods extended this capability by introducing self-attention mechanisms that can capture global dependencies across all agents and timesteps at once across all timesteps and agents [8, 7]. Self-attention enables each pedestrian to provide dynamic attention to the most relevant surrounding agents (independent of spatial distance or temporal ordering). This can be very helpful when there are a lot of people around and meaningful interactions are not necessarily happening with adjacent neighbors. Moreover, the parallelization support of transformer model training and inference are high, which is suitable for large-scale modelling. This success in NLP and computer vision has motivated the development of a large number of research projects investigating attention-based forecasting methods for human motion understanding.

Despite all these advances, predicting the trajectories of people in dense crowds is still a difficult task, since it is determined by a number of interacting factors at various spatial and temporal scales. Local interactions can be distinguished from global interactions by collision avoidance and by short range coordination with surrounding pedestrians, while collective crowd motion and long term navigation goals give rise to global interactions. Older methods tend to be one-sided and pay attention to one part of the process and ignore the other. The locally oriented approaches might not account for global flows of crowds and the short-range behavioural adjustments might not be detected in the globally oriented models. This imbalance motivates the need for hierarchical and multi-scale representations capable of simultaneously modelling fine-grained interpersonal interactions and large-scale collective dynamics.

Temporal complexity adds to the challenge. Human navigational decisions take place on multiple time scales; some responses are moment-to-moment decisions, while others are over seconds. The current forecast methods typically rely on a fixed temporal window and a fixed temporal representation that might not be suitable to describe such varying temporal dependencies. Multi-scale temporal reasoning is thus gaining increasing relevance in the latest research. The goal of Hierarchical temporal encoders and multi-resolution interaction modules is to concisely encode short-term reactive behaviour and long-term intentional movement. In dense urban areas where human behaviors and actions are constantly adapting to their environment and surroundings, capturing the temporal dynamics is especially crucial.

Outside of the modelling challenges, the computational demands of practical deployment are also challenging. Predictions need to be computed in real time under limited computational budgets in the case of autonomous vehicles and robotic systems. Many high-tech models use expensive attention mechanisms, large graphs, or iterative sampling approaches that are hard to scale for large scenes, however. Therefore, a trade-off between prediction accuracy and inference efficiency is of interest. Some recent efforts (e.g., past and present) try to solve this problem by using lightweight interaction modules, simplified graphs, or feed-forward architectures trying to strike a balance between reducing the computational complexity and retaining competitive performance. Achieving a balance between efficiency, scalability and prediction quality is a significant hurdle in the real world deployment.

Benchmark datasets have also been at the core of pedestrian trajectory prediction approaches. ETS and UCY were selected as foundational datasets since they offer labeled pedestrian trajectories in real-world settings, where there are rich social inter-

actions. These datasets allowed these to be evaluated consistently according to parameters like Average Displacement Error (ADE) and Final Displacement Error (FDE) used widely in the literature. More recent datasets have added more complex urban scenes, heterogeneous agents, semantic maps, and larger crowds to provide more diversity and realism of the evaluation scenarios. The increasing amount of data has thus enabled the emergence of increasingly complex learning-based models that are able to capture nuances of human behavior.

There is already a lively debate about how to evaluate. ADE and FDE reveal only the positional accuracy, but not the quality of multimodal predictions or the social realism of generated trajectories. Low displacement error does not guarantee that the prediction will respect the scene constraints or generate socially implausible motion. In addition to the standard error measures, recent work also focuses on the measures of diversity, uncertainty calibration, collision rates and socially acceptable behavior. This is a general trend in the field from geometric to behaviourally meaningful forecasting.

One of the new trends is explainability and interpretability. Deep learning methods can forecast very accurately, but the reasoning used is not easily understood as these methods are usually hard to interpret because they are black-box systems. Certain applications are safety-critical, like autonomous driving, which are especially important for interpretability as prediction failures can incur major consequences. Researchers are thus starting to look for ways of delivering interpretable interaction reasoning, attention visualization, or physically grounded forecasting mechanisms. Particularly, hybrid solutions that leverage data-driven learning and include explicit physical constraints or social priors have been in the spotlight as a potential approach to achieve both robustness and explainability.

The significant advances in the field of trajectory prediction has also grown the applications beyond autonomous driving. Attention is a key factor in human-aware navigation, which has gained attention in the field of service robots, crowd management, augmented reality, intelligent surveillance, sports analytics and smart city infrastructure. One field in which trajectory forecasting plays a crucial role in robotic navigation is in developing socially compliant motion planning that allows for smoother interaction with humans. Surveillance systems: predicting abnormal trajectories can assist in anomaly detection and public safety monitoring. Likewise, crowd simulation systems make use of realistic motion prediction to model evacuation situations and optimize the design of infrastructure. The variety of these applications demonstrates the interdisciplinary concepts and importance of pedestrian trajectory prediction in today's smart systems.

Despite significant progress in trajectory prediction methods, there are still a number of open research questions that have yet to be addressed. Current models have yet to achieve good generalization on unseen environments, robustness to distribution shifts, few observations, and dense crowds. Numerous methods are based on the training using benchmarks, and do not easily transfer to real-world deployments that have different environmental conditions or pedestrian behaviours. In addition, most data sets are relatively small and do not reflect the variety of urban environments. The existing models lack generalised interaction reasoning, are computationally expensive, and do not provide behaviour interpretation.This is a problem that needs to be addressed by

models that can learn generalised interaction reasoning, yet are computationally efficient and behaviourally interpretable.

These limitations are the motivation for the present work, which aims to further advance the understanding of multi-scale pedestrian interaction modelling. This thesis aims in particular to advance trajectory forecasting in architectures that are able to capture both local interpersonal behaviours and the overall crowd dynamics. The work proposes a new approach of hierarchical interaction representations and dynamic relational modelling to create predictions that are accurate, realistic, scalable and adaptable to complex environments. The general goal is thus towards the development of strong pedestrian trajectory prediction systems that can be effectively used in real-world intelligent transportation systems and autonomous navigation applications.

CHAPTER 2

LITERATURE REVIEW

Pedestrian trajectory prediction methods can be broadly categorised based on how they model motion and uncertainty. In this work, we group them into four categories: physics-based, classical machine learning-based, deep learning-based, and multimodal approaches. This taxonomy captures the progression from rule-based to data-driven and multimodal methods, which we analyze in detail in the following sections.

2.1 Physics-Based and Classical Machine Learning Approaches

Physics-based methods rely on hand-crafted rules and physical intuition, such as the Social Force Model [1] and geometric or kinematic approaches like ORCA [11] and constant velocity models [12]. Earlier work in this domain was mostly physics modelling where motion was described using rules and pedestrian interaction. A well known example being the Social Force Model [1] which modelled pedestrian behaviour using attraction and repulsion forces between the pedestrians and the surroundings. Other works built on this to add more details to reflect crowd behaviour [13]. Other works took a different approach like ORCA [11] handled collisions between pedestrians by modelling velocities to predict which movement will lead to a crash. Simpler baselines like the constant velocity model (CVM) [12] assume linear motion and use only the previous movement to predict the future. These models are quite simple, yet they perform reasonably well in some cases. Some other methods also use acceleration and avoidance between pedestrians [14], while others combine motion planning with physics to predict long-term motion [15].

These methods are easy to interpret and efficient to run, but they have strong assumptions about pedestrian movement, hence they struggle in scenes which are heavily crowded or scenes with complex pedestrian interactions. Classical machine learning methods introduce probabilistic formulations, including switching dynamical systems [16], Gaussian processes [17], and Bayesian models [18], but remain limited in modelling complex interactions. Later work explored using classical Machine Learning approach-es with probabilistic formulations to deal with some of these issues. Methods based on switched linear dynamical systems [16], Gaussian processes [17],

and Bayesian models [18] try to model uncertainty and temporal variation in a more data-driven way. Even so, capturing complex interactions between multiple agents is still difficult for these mod-els, which is why more recent work has moved toward deep learning-based approach-es.

2.2 Emergence of RNN-Based Sequence Modelling

The shift to deep learning fundamentally changed pedestrian trajectory prediction. Deep learning-based methods learn motion patterns directly from data and form the dominant paradigm. Deep learning methods are now used quite a lot for pedestrian trajectory prediction, mainly because they can pick up motion patterns and interactions directly from data. Unlike earlier approaches, they don't rely on hand-crafted rules, but learn temporal behaviour, interactions, and scene context in an end-to-end manner.

Recurrent neural networks brought about a big change in prediction as they could learn how movement happens over time (Temporally). Some of the earlier work in this area used recurrent neural networks (RNNs), treating trajectories as sequences. Social-LSTM [2] is one of the first examples, where a pooling mechanism is used to bring in information from nearby pedestrians, so the model can account for local interactions. Further works built on this idea to improve pedestrian interactions are handled by using more flexible architectures and attention mechanisms [19], allowing the model to attend to relevant pedestrians instead of treating all the neighbours equally.

Generative approaches came in later to handle the variability of human motion. Social-GAN helped by creating many different but realistic paths [3]. Social GAN [3] used adversarial training by using GANs to produce socially possible trajectories and So-Phie [20] used both social and temporal attention. MG-GAN [21] uses multiple generators to increase the diversity of predictions and reduce mode collapse. These methods move from predicting only one trajectory to instead model a distribution over trajectories. Other methods focused on where the person wanted to go to make better long-term guesses [22]. Nevertheless, the issues with RNNs remain. The sequential nature restricted the speed of training, long-range dependencies might degrade over time, and sometimes the uncertainty representation was limited. These constraints inspired a shift to more parallelizable and socially conscious architectures.

2.3 Transformer-Based Techniques of Global Interaction

Transformers, which have a global receptive field, and attention mechanisms, were of interest because of their ability to model long-term temporal association. Recently, transformer-based models have explored to better capture long-term dependency and complex interactions. Models that used transformers performed highly in multi-agent prediction tasks because they allowed every pedestrian to focus on the rest of individuals in a scene to attend to the others, which are known as transformer-based models [8].

To achieve the fusion of social and temporal cues, architectures such as AgentFormer proposed agent-aware relational attention to predict better fusion results by incorporating social and temporal information [7]. AgentFormer [7] used attention to model both temporal and social relationships, while Social-Implicit [23] used a lighter architecture to get strong performance, showing that efficiency and performance can go hand in hand. Subsequent studies added semantics or scene layouts, or even physics-based priors, to the attention module to enhance awareness of the contexts of their surrounding environment [24].

In spite of such benefits, transformer models were associated with high computational costs since attention is quadratically related to the size of the group of agents and the time steps. Besides, most of these models reasoned independently through time and space, something that restricted them to interactions that develop together. Such issues led to the inclusion of structured relational reasoning through graph neural network modeling.

2.4 Graph Neural Networks for Social and Spatio-Temporal Reasoning

Researchers also explored graph modelling, where the pedestrians are represented as nodes of a graph and the interactions between these pedestrians are the edges. The graph representation gives a natural structure to model multi-agent motion: the pedestrians are nodes and the moving dynamic interactions are represented by the edges. Graph-based models exhibited high flexibility to different scene densities and agent configurations of multi-agents [6, 25].

Methods like Social-BiGAT [6] and Trajectron++ [25] use graph neural networks to handle interactions between pedestrians, making it easier to model interactions, even in crowded scenes. Sparse graph convolution methods significantly simplified the com-putations, but retained significant information on relational aspects of the data set [5]. Social-STGCNN proposed convolutional spatiotemporal propagation as its efficient and high-quality accurate model, and Dynamic-Evolving GCNs modified edge relations with time to track the changes of interactions in the network [26, 27]. Hierarchical and group-conscious networks also enhanced predictions by learning social substructures in large crowds of people [28, 29].

Despite its great success, most of the models could not cope with local and global interactions at the same time. There were also a number of works that used immobile receptive fields that were not easily adaptable to changes in scene topology [7, 24].

2.5 Multi-Scale Interaction Learning and Deformable Modelling

As deformable convolution was found to be successful and effective in video tasks, the researchers proposed spatial temporal deformability in trajectory prediction. Adaptive

deformable kernels are able to move sampling positions according to motion structure, and this allows more informative features to be aggregated better [30]. Specifically, the first application of deformable spatial temporal convolution in forecasting by DSTCNN shows its potentiality in the first place [31]. DSTIGCN built on these concepts and added the interaction-aware GCN layers with deformable sampling on the spatial and temporal domains [18]. This enables the model to view the change of social patterns at various levels, and remains adaptable in the way it learns. Tri-HGNN and hierarchical transformers apply layers to blend tiny details, where larger tendencies, such as preventing a crash to the movement of the whole group are involved in it.

At the same time, existing models often operate on spatial and temporal features separately or rely on fixed convolution kernels for spatial aggregation. Human interactions are dynamic and intrinsically geometric; people do not move on fixed grids or interact at fixed temporal resolutions. These limitations motivated the development of adaptive and deformable interaction modelling approaches capable of learning dynamic spatial-temporal relationships more effectively.

2.6 Multi-Source Feature Fusion and Wavelet-Based Modelling

Prediction does not only concern movement but also surroundings such as the arrangement of the scene. Models that incorporate all of these factors are used in new studies to make better predictions [25]. Other works focused on using surrounding scenes, for example, Car-Net [4] and Y-Net [32] use the scene information to predict motion. This shows that motion alone is not enough and the scene context is also important. The use of wavelet-based modeling gained popularity due to its ability to isolate the large and the small [33] and WTGCN was developed [34]. They apply this approach to mute down noise and put emphasis on the interactions that do count. These methods demonstrate that multi-scale representations can better capture both fine-grained local interactions and broader crowd dynamics simultaneously.

2.7 Multimodal Prediction

A common limitation in many deep learning-based methods is that they still predict only a single future trajectory, even though pedestrian behaviour is not really deterministic. For the same observed trajectory, there are often multiple possible ways a person might move, especially in open or crowded environments. Because of this, recent work has started looking at multimodal prediction, where the idea is to generate more than one possible future instead of committing to a single path. More recent work shifts toward multimodal prediction, where instead of a single output, multiple possible futures are generated. This includes anchor-based methods like CoverNet and TNT [9, 35], as well as generative approaches such as MG-GAN and MotionDiffuser [21, 36].

One way to do this is through generative models that try to produce diverse trajecto-

ries. For instance, MG-GAN [21] builds on earlier GAN-based approaches like Social GAN [3] and generates multiple plausible samples. This helps in improving diversity, but mode collapse is still an issue and the outputs end up looking similar or miss some possible behaviours. Researchers explored using predefined trajectory anchors, where methods such as CoverNet, MultiPath, TNT, and DenseTNT [9, 10, 35, 37] sample multiple trajectories and then try to pick or rank them to get the most likely paths. This is easier to train but it depends heavily on the anchors that are being chosen, which is a limit on how diverse the predictions are.

Diffusion-based approaches like MotionDiffuser [36] model the trajectories by starting from noise and then refining these over multiple steps. These methods are trained without relying on anchors, increasing the diversity, producing realistic and varied outputs. But these methods are harder and more computationally expensive to train.

The most common approach to evaluate is the Best-of-K strategy, where these models generate multiple trajectories, generally 20 and the closest to the ground truth trajectory is used for evaluation [3, 9]. While this captures diversity, it also gives a very optimistic picture, as the worse predictions are ignored. This sort of evaluation carries a trade-off between diversity and accuracy, the models that generate more options can look reliable in BoK evaluation even if they arent reliable. Overall, multimodal prediction is a step towards more realistic predictions of pedestrians as it accounts for uncertainty. Evaluation and computational cost are open issues and there is still room for improvement here.

2.8 Interaction-Aware Recurrent Models and Social Context

A significant step beyond early sequence-to-sequence forecasting was the inclusion of explicit social context in recurrent models. Instead of treating each pedestrian as an isolated trajectory, these methods consider the scene as a small interactive system in which nearby agents influence one another. Social-LSTM is one of the landmark models in this direction, where a pooling mechanism aggregates hidden states from surrounding pedestrians so that each agent is predicted with awareness of nearby motion patterns [2]. Later work refined this idea by changing the way social influence is represented and filtered. Social Attention introduced a learnable attention mechanism that assigns different importance to different neighbours, rather than combining all surrounding agents uniformly [38]. Likewise, STGAT models spatial-temporal interactions using graph attention to improve the representation of local social dynamics [39].

Beyond pooling and attention, several studies tried to improve the hidden-state evolution inside recurrent architectures. SR-LSTM introduced state refinement so that the latent representation is repeatedly corrected before decoding the future trajectory [40]. IA-LSTM also follows this line by explicitly embedding interaction awareness into the recurrent update, which helps the model better distinguish between informative and irrelevant neighbours [19]. These models are important because they show that the main

challenge is not only temporal modelling, but also deciding which social cues matter at each time step. At the same time, the dependence on fixed neighbourhood definitions can make these approaches sensitive to density changes, scene scale, and the number of pedestrians present in the observation window.

2.9 Goal-, Endpoint-, and Scene-Conditioned Prediction

Another important direction in pedestrian forecasting is to separate high-level intent from low-level path generation. In many scenes, the future motion of a pedestrian becomes easier to reason about once a likely destination, waypoint, or goal has been identified. Goal-conditioned models therefore try to infer an endpoint first and then decode a full trajectory that is consistent with that endpoint. PECNet is a representative example of this idea, as it predicts endpoint information and then uses it to produce plausible future paths [22]. Similar endpoint-aware or destination-oriented strategies also appear in MultiPath, TNT, and DenseTNT, where a set of possible goals or anchors is generated and then refined into full trajectories [10, 35, 37].

This decomposition is attractive because it reduces the difficulty of long-horizon prediction and gives the model a clearer notion of intent. However, the approach also introduces a second prediction problem: the endpoint itself must be inferred correctly. When the goal estimate is wrong, the downstream trajectory can quickly drift away from the true motion. For this reason, scene-aware reasoning is often combined with goal prediction. CAR-Net and Y-Net incorporate environmental context so that predicted paths remain aligned with walkable space and scene semantics [4, 32]. These models illustrate an important principle: pedestrian motion cannot be explained by kinematics alone, because the environment shapes where people can and cannot move. In practice, combining destination cues with scene context often produces more realistic trajectories than using either source of information in isolation.

2.10 Dynamic Graphs and Hierarchical Relational Reasoning

Graph-based modelling has become a dominant framework for representing multi-agent interaction because it provides a natural way to encode who influences whom. In a graph formulation, pedestrians are nodes and pairwise relations are edges, which allows the model to exchange information across the crowd in a structured manner. Trajectron++ is a foundational example of this strategy, as it uses a dynamic graph representation to model heterogeneous interactions and forecast future motion under social constraints [25]. Social-BiGAT extends this idea by combining graph attention with a multimodal generator, which helps the model capture both interactions and uncertainty [6].

Later work explored different ways of making graph reasoning more efficient and more adaptive. Social-STGCNN replaces recurrent processing with spatio-temporal

graph convolution, which makes temporal evolution and social influence easier to model jointly [26]. SGCN uses sparse graph convolution to reduce unnecessary message passing while retaining useful social structure [5]. More recent methods attempt to update graph relations over time rather than assuming a fixed interaction topology. DERGCN models evolving graphs so that interaction edges can change as agents move through the scene [27], while STAGP focuses on adaptive spatio-temporal graph pooling to filter the most informative relations at different scales [41].

A related theme is hierarchical interaction modelling. In crowded scenes, pedestrian motion is not only determined by pairwise proximity, but also by group structure and collective behaviour. GCHGAT and Tri-HGNN introduce hierarchical attention or layered graph reasoning to distinguish between individual-level and group-level movement patterns [29, 28]. Similarly, SKGACN injects social knowledge into graph attention, showing that prior knowledge about social behaviour can be beneficial when the raw interaction signal is ambiguous [42]. RDGCN further explores the balance between graph density and predictive quality by reasoning over reasonably dense neighbourhood structures [43]. Taken together, these works suggest that graph topology is one of the most important design choices in pedestrian forecasting: too sparse a graph may miss relevant influences, while too dense a graph can introduce noise and unnecessary computation.

2.11 Multi-Scale and Deformable Modelling

One common problem of the fixed-kernel models is that the interactions of the pedestrians are not necessarily organized by a fixed spatial grid or a fixed time scale. Individuals can go fast or slow, form groups and split up, or respond to an obstacle in an unpredictable manner. To overcome this, researchers have started to incorporate multi-scale and deformable mechanisms into the trajectory prediction. Another advantage to deformable convolutions is that they have the ability to sample points that move depending on the motion structure, not just a fixed grid. To address this issue, trajectory prediction models based on deformable spatio-temporal convolution and interaction-aware graph extensions have been proposed, including DSTCNN [31] and DSTIGCN [44]. These models try to match the receptive field with the geometry of a pedestrian's motion, which is useful when there are very few interactions in the scene.

An example of the value of multi-scale reasoning is if a model has to include both local avoidance and crowd flow. While fine-grained local cues are crucial for collision avoidance, longer-range dependencies are required for collective behaviour such as lane-like motion, queue formation and turning. Deformable and multi-scale approaches can better adapt to changes in crowd density by using both local and global aggregation, which is more natural than rigid architectures. One of the reasons is that recent work has begun to shift from single scale temporal encoding, to representations that are scalable with motion patterns at different resolutions. Deformability is not just a parameter-efficiency gimmick—in fact it is a means for aligning the inductive bias of the model with the irregularity of human motion.

2.12 Multi-Stream Feature Learning and Hybrid Representations

In this section, we examine Multi-Stream Feature Learning and Hybrid Representations.In this section, we will review Multi-Stream Feature Learning and Hybrid Representations. One field of literature that has emerged in recent times is the application of several feature streams that represent complementary aspects of motion. Rather than using only coordinates, these models can include separate branches to process separate pieces of information: trajectory history, social context, scene semantics, and motion cues, before fusing them together. The advantage of multi-stream learning is that pedestrian forecasting is inherently heterogeneous; one stream could be used for short-term kinematics, another for scene layout, and a third for social proximity. It has been shown that using multiple complementary signals can lead to better predictive robustness, especially when some are noisy or incomplete [45].

Other models that use graph reasoning, attention, or knowledge-guided priors, also include hybrid representations. The usage of, for instance, the scene images, goal hypotheses, and socially structured graphs, for instance, may be understood as different perspectives on the same motion process. The combinations have also been shown to perform better than purely homogeneous architectures, by keeping the burden of explaining all the aspects of the future in the single architecture. Concurrently, hybrid designs have their own practical issues: What proportion of the information should be allocated to each branch, how should the two branches be integrated, and which modality should be predominant when cues conflict? Questions focused on the real world are particularly important if a driver is only partially visible, or partially obscured in some way, and if the tracking is imperfectly noisy.

2.13 Multimodal Prediction and Uncertainty Modelling

Predicting and modelling uncertainty in a multimodal manner. One future path is often too limited for pedestrian forecasting since human motion is inherently indeterminate and can have multiple possible future paths. As a result, the prediction of multiple modes has emerged as a key challenge in the recent research on trajectory forecasting. The previous methods mentioned in the early days of generative models, like Social GAN, have shown that using adversarial learning to generate a variety of trajectories is feasible, rather than just one deterministic trajectory.Early generative methods have proved that it is possible to create multiple trajectories with adversarial learning, not a single deterministic trajectory, such as Social GAN. In contrast, MG-GAN takes this step by employing multiple generators to minimize mode collapse and thereby promote diversity in the predicted futures [21]. These methods are useful in that they explicitly recognize uncertainty but diversity is not sufficient—the generated paths need to be socially acceptable and physically feasible.

The anchor-based methods solve this issue by forecasting over a limited number of possible futures. The three categories of prediction space structures that are used in CoverNet, MultiPath, TNT, and DenseTNT are: candidate set, goal proposal, and

anchor hypothesis [9, 10, 35, 37]. This is easier to train and can often enhance the stability of the task, but also constrains the model to the quality of the anchors or proposal set. With an incomplete candidate set, the candidate set that the candidate chooses may not be representative of the true future. The other approach is to begin with noise and refine samples into plausible futures; the latter is diffusion-based trajectory prediction. In this family, MotionDiffuser exemplifies the use of diffusion for making controllable multi-agent motion predictions, with great diversity [36]. The downside is computational burden; typically, diffusion models need multiple denoising steps.

As far as evaluation is concerned, multimodal prediction reveals the shortcomings of Best-of-K reporting. The best possible trajectory selected from many samples can help one to have a better ADE and FDE, but one could also be misled by the selection that most of the samples are poor or unrealistic. Social-Implicit explicitly states its limitations in evaluating this prediction protocol and makes the case for careful treatment of the quality of the predictions as well as the estimation of the likelihood of the model. This is why, the literature is slowly shifting to evaluation approaches not only for accuracy but also for diversity, calibration, collision avoidance and the probability assigned to the ground truth.

2.14 Evaluation Protocols and Practical deployment considerations

The result of pedestrian trajectory prediction methods can vary significantly depending on the metric used for evaluation. The most popular measures are still ADE and FDE, which are readily calculable and readily comparable between models. These metrics, however, are mostly related to the geometric proximity of the GT, and do not necessarily reflect the realism, social acceptability, or usefulness of a trajectory for downstream planning. In multimodal settings, minimum ADE and minimum FDE under a Best-of-K protocol are often reported, but these minimums may be too high since they do not consider any minimums except the best sample. In particular, if the model is used for autonomous driving, only one trajectory or one probabilistic estimate may be useful for the planner.

In real life deployment, there are further constraints that aren't represented in the benchmarks. When a model needs to run in real time, it's important it runs quickly, doesn't use too much memory, and can be scaled to cope with the number of pedestrians. While slightly less accurate, a method that is significantly slower might not be very useful in systems with strict latency budgets, such as constant velocity prediction [12]. Another issue is resistance to partial history, occlusions and tracking noise. Clean, densely labelled datasets may not be applicable in the real world when sensors are not as accurate as they are during training, or when pedestrian behaviour varies from the training domain. This is the reason for the necessity of architectures that are stable with respect to missing observations and domain shift.

The interpretability is also essential in safety-critical applications. Physics based and knowledge based methods are easier to understand since the parameters of these methods often have semantic meaning [1, 11]. Unlike deep learning models, proba-

bilistic methods can better explain the rationale behind a particular trajectory and are more accurate at longer time horizons. One of the many reasons that makes hybrid approaches attractive is that they can integrate the flexibility of a learned representation with the transparency of a rule-based representation prior. In this respect, the literature is coming together towards a common objective: to develop models that are accurate in benchmark datasets, reliable, interpretable, and efficient enough for real transportation systems.

2.15 Open Challenges and Research Motivation

There are several open challenges and motivations for further research. In spite of the rapid development in trajectory forecasting, there are still a number of open problems that are not solved. First, models are still unable to generalize well over scenes with varying layouts or sizes of pedestrians, and conventions of behavior. Second, long-horizon forecasting is still challenging, due to uncertainty building up over time and the number of potential futures increasing as time passes. Third, current models fall short when it comes to successfully modelling complex social interaction, including group splitting, merging, yielding, and non-verbal communication. Fourth, the gap between the performance of the benchmark and the deployment readiness is still present, particularly if the real-time constraints and sensor imperfections are considered.

The challenges indicate that future studies should not only concentrate on the single metric improvement, but should be more complex. Instead, prediction systems should be made to be designed wider, with uncertainty estimation, goal reasoning, scene awareness, interaction modelling, computational efficiency and interpretability built-in. In this regard, multi-scale graph reasoning, goal-conditioned decoding and hybrid physics-learning strategies hold promise. Literature thus indicates that a new generation of adaptive, explainable and applicable pedestrian trajectory prediction models are needed for safety-critical autonomous driving applications.

CHAPTER 3

METHODOLOGY

3.1 PROBLEM FORMULATION AND EVALUATION METRICS

Given a set of observed pedestrian trajectories, the task is to predict how those pedestrians will move over some future time horizon. Let the observed trajectory of pedestrian i be

$$\mathbf{X}_i = \{(x_i^t, y_i^t)\}_{t=1}^{T_{obs}}, \tag{3.1}$$

and the future trajectory to be predicted as

$$\mathbf{Y}_i = \{(x_i^t, y_i^t)\}_{t=T_{obs}+1}^{T_{obs}+T_{pred}}. \tag{3.2}$$

In addition to the target pedestrian, most models also take into account nearby pedestrians to model interactions, and sometimes include scene information like obstacles or walkable areas. Earlier approaches handled these explicitly using hand-crafted rules[1, 11, 12, 14], whereas more recent methods try to learn these effects directly from data[2, 25, 7].

Depending on how the problem is set up, a model can either predict a single trajectory or multiple possible futures. Single-trajectory prediction is still quite common, but multimodal approaches try to deal with the uncertainty in human motion by generating several candidate trajectories or even a distribution over outcomes[3, 9, 36].

Displacement-based metrics are usually used to evaluate performance. The Average Displacement Error (ADE) measures the mean Euclidean distance between the predicted trajectory and the ground truth over the prediction horizon:

$$\text{ADE} = \frac{1}{T_{pred}} \sum_{t=1}^{T_{pred}} \left\| \hat{\mathbf{y}}_i^t - \mathbf{y}_i^t \right\|_2, \tag{3.3}$$

The Final Displacement Error (FDE), on the other hand, looks only at the error at the last predicted time step:

$$\text{FDE} = \left\| \hat{\mathbf{y}}_i^{T_{pred}} - \mathbf{y}_i^{T_{pred}} \right\|_2. \tag{3.4}$$

Evaluation is done using a Best-of-K setup for multimodal models, where multiple

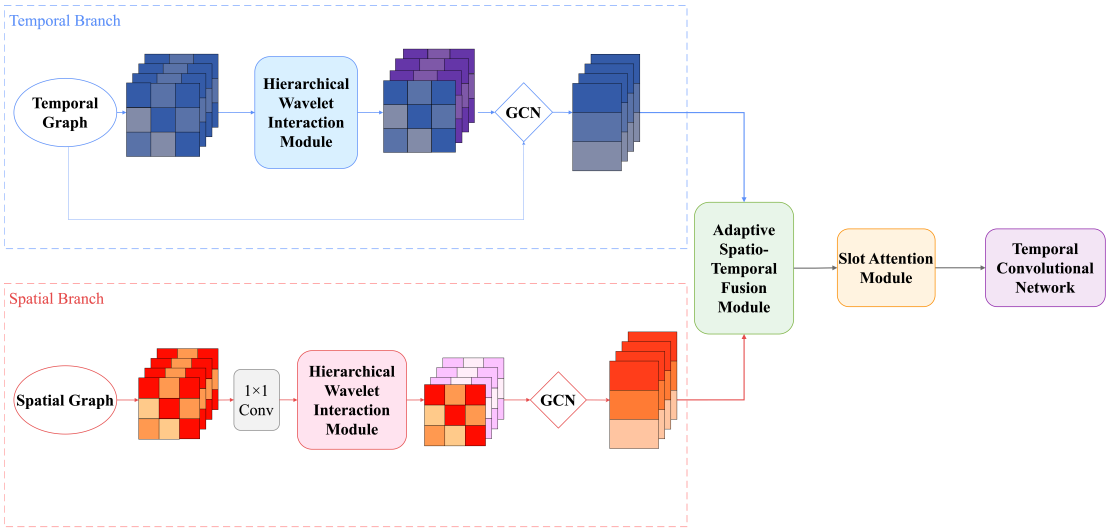

Figure 3.1: Architecture of SWIFT-Net

predictions are generated, and the one closest to the ground truth is selected[3, 9]. This tends to reward diversity, but it can also make results look better than they actually are, since poorer predictions are ignored and the overall distribution quality is not fully reflected.

Overall, this formulation brings out two main difficulties: capturing interactions and scene context in a meaningful way, and dealing with the uncertainty that comes from having multiple plausible futures.

3.2 SWIFT-Net

SWIFT-Net (Slot-guided Wavelet Interaction Fusion for Trajectories) addresses a different limitation in trajectory prediction. Many models fail to capture the full range of interactions because they operate at a single scale. Human behaviour, however, emerges from both local interactions and broader crowd patterns. SWIFT-Net is designed to capture multi-scale interaction features and fuse them intelligently.

3.2.1 Constructing Spatial and Temporal Graphs

SWIFT-Net constructs two separate graphs:

- A spatial graph representing social interactions at each frame
- A temporal graph representing motion along the time axis

Both graphs are passed through attention layers to obtain initial interaction score matrices for spatial and temporal dimensions.

$$R_{S/T} = \text{Softmax}\left(\frac{Q_{S/T}K_{S/T}^T}{\sqrt{d_{S/T}}}\right) \tag{3.5}$$

$$R_S = \text{Conv}_{1\times1}(R_S) \tag{3.6}$$

3.2.2 Hierarchical Wavelet Interaction Module

This module of SWIFT-Net: multi-scale decomposition. Wavelet transforms break a signal into:

- **Low-frequency components**, representing global patterns
- **High-frequency components**, capturing local changes and details

SWIFT-Net applies three levels of wavelet decomposition to both spatial and temporal interaction matrices. At each level, the decomposition produces four components:

- LL (low–low)
- LH (low–high)
- HL (high–low)
- HH (high–high)

These represent different directions and scales of interaction.
After decomposition, each component is processed using asymmetric convolution (1×5 and 5×1 kernels). This allows the network to pick up directional motion patterns and fine-grained relational cues.
An inverse wavelet transform then reconstructs the final adjacency matrices.

For space and time:

$$R^{(l+1)}_{S/T,c_0c_1} = (\downarrow 2)\left(f^{(l)}_{c_0c_1} * R^{(l)}_{S/T,LL}\right), \quad c_0, c_1 \in \{L,H\} \tag{3.7}$$

$$C^{(l)}_{S/T,c_0c_1} = \text{Conv}_{5\times1}(R^{(l)}_{S/T,c_0c_1}) + \text{Conv}_{1\times5}(R^{(l)}_{S/T,c_0c_1}) + R^{(l)}_{S/T,c_0c_1} \tag{3.8}$$

$$C^{(l)}_{S/T,LL} \leftarrow C^{(l)}_{S/T,LL} + \sum_{c_0,c_1\in\{L,H\}} f^{(l+1)}_{c_0c_1} * (\uparrow 2)\, C^{(l+1)}_{S/T,c_0c_1} \tag{3.9}$$

$$A_{S/T} = \text{Softmax}\left(C^{(0)}_{S/T,LL}\right) + I \tag{3.10}$$

3.2.3 Graph Convolution for Feature Extraction

The spatial and temporal adjacency matrices are each passed through graph convolution layers to obtain:

- social interaction features (Hs)
- temporal interaction features (Ht)

Both contain complementary information, but their importance varies depending on the scenario.

$$H_S = \delta(A_S G_S W_{GS}) \tag{3.11}$$
$$H_T = \delta(A_T G_T W_{GT}) \tag{3.12}$$

3.2.4 Adaptive Spatio-Temporal Fusion Module

Instead of simply adding Hs and Ht, SWIFT-Net introduces a guided attention-based fusion block. The module computes:

- spatial attention
- channel attention
- feature attention

These help determine how much weight to assign to spatial or temporal information for each pedestrian. For example, in a dense crowd, spatial cues may dominate, whereas temporal cues may be more important when a pedestrian is walking alone. The output is a fused feature representation HstF that is more balanced and context-aware.

Combine features:

$$H_{ST} = H_S + H_T \tag{3.13}$$

Spatial attention:

$$W_S = \text{Conv}_{7\times7}\left([\text{AvgPool}_S(H_{ST}), \text{MaxPool}_S(H_{ST})]\right) \tag{3.14}$$

Channel attention:

$$W_C = \text{CNN}\left([\text{AvgPool}_C(H_{ST}), \text{MaxPool}_C(H_{ST})]\right) \tag{3.15}$$

Feature attention:

$$W_S = \sigma(\text{Conv}_{7\times7}([H_{ST}; W_C + W_S])) \tag{3.16}$$

Final fused features:

$$W_T = 1 - W_S \tag{3.17}$$

$$H_{STF} = (1 + W_S) H_S + (1 + W_T) H_T \tag{3.18}$$

3.2.5 Slot Attention Module

While the spatial–temporal fusion module produces rich interaction-aware features, it does not explicitly disentangle different underlying motion patterns or interaction modes. In real-world scenarios, pedestrian behavior is inherently multi-modal, where multiple plausible future trajectories may exist depending on social context.

To address this, we incorporate a Slot Attention module to decompose fused features into a set of latent representations, referred to as *slots*. Each slot is designed to capture a distinct interaction pattern or behavioral mode.

Given the spatial–temporal feature representation:

$$H \in \mathbb{R}^{N \times H \times T \times C}$$

we first compress the temporal dimension to obtain per-pedestrian features:

$$X \in \mathbb{R}^{N \times H \times C}$$

These features are normalized and passed into the Slot Attention module, which initializes a fixed number of learnable slot vectors:

$$S_0 \sim \mathcal{N}(\mu, \sigma)$$

At each iteration, attention is computed between input features and slots:

$$A = \text{Softmax}\left(\frac{K(X)Q(S)^T}{\sqrt{d}}\right) \tag{3.19}$$

The attention weights are normalized across inputs to ensure competition between slots:

$$\tilde{A}_{ij} = \frac{A_{ij}}{\sum_i A_{ij}} \tag{3.20}$$

Each slot is then updated using weighted aggregation of input features:

$$U = \tilde{A}^T V(X) \tag{3.21}$$

The slots are iteratively refined using a GRU-based update mechanism:

$$S^{t+1} = \text{GRU}(U, S^t) \tag{3.22}$$

followed by a residual MLP refinement:

$$S^{t+1} = S^t + \text{MLP}(\text{LayerNorm}(S^t)) \tag{3.23}$$

After multiple iterations of this, the slots capture distinct aspects of interaction dynamics. Instead of treating them independently, we aggregate slot information to obtain a compact representation:

$$S_{\text{agg}} = \frac{1}{K} \sum_{k=1}^{K} S_k \tag{3.24}$$

This aggregated slot representation is then broadcast and fused back into the original spatial–temporal features:

$$H' = H + S_{\text{agg}} \tag{3.25}$$

This process enhances the feature representation by incorporating global interaction patterns and latent behavioral modes, improving the model's ability to generate diverse and socially consistent trajectory predictions.

3.2.6 Temporal Convolutional Network for Multi-trajectory Prediction

SWIFT-Net does not sample trajectories from a Gaussian distribution. Instead, it directly predicts multiple future trajectories using a temporal convolution network. Each predicted trajectory represents a possible future mode.
To train this, SWIFT-Net uses a mix of losses to balance accuracy and diversity:

- a **winner-takes-all loss**, which selects the closest predicted trajectory
- an **overall prediction loss**, encouraging all predictions to be meaningful

$$\begin{aligned} L_w = &\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=t_{\text{obs}}+1}^{t_{\text{pred}}} \min_k \left\| \hat{p}_{i,t}^{(k)} - p_{i,t} \right\|_2 \\ &+ \frac{1}{n} \sum_{i=1}^{n} \min_k \left\| \hat{p}_{i,t_{\text{pred}}}^{(k)} - p_{i,t_{\text{pred}}} \right\|_2 \end{aligned} \tag{3.26}$$

$$\begin{aligned} L_o = &\frac{1}{nTS} \sum_{i=1}^{n} \sum_{t=t_{\text{obs}}+1}^{t_{\text{pred}}} \sum_{k=1}^{S} \left\| \hat{p}_{i,t}^{(k)} - p_{i,t} \right\|_2 \\ &+ \frac{1}{nS} \sum_{i=1}^{n} \sum_{k=1}^{S} \left\| \hat{p}_{i,t_{\text{pred}}}^{(k)} - p_{i,t_{\text{pred}}} \right\|_2 \end{aligned} \tag{3.27}$$

$$L = L_w + \lambda L_o \tag{3.28}$$

3.3 TRACE-Net

TRACE-Net (Trajectory Reasoning via Adaptive Competitive Encodings) is built around the idea that pedestrian interactions unfold jointly in space and time. Most earlier models treated spatial and temporal features separately, then fused them afterward. TRACE-Net attempts to directly learn these dependencies in a unified manner.

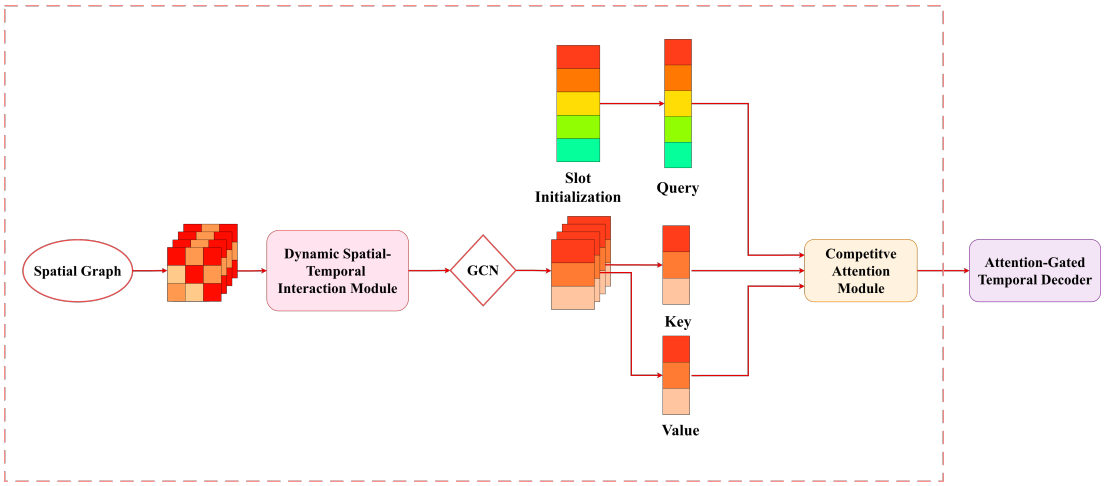

Figure 3.2: Architecture of TRACE-Net

3.3.1 Constructing the Spatial Graph

Pedestrians in each frame are represented as nodes, and edges initially connect all pairs of pedestrians. Instead of using only raw positions $p_t^i = (x_t^i, y_t^i)$, the model uses relative motion (displacement) between consecutive frames, defined as $o_t^i = p_t^i - p_{t-1}^i$. This helps normalise the scale of the scene and allows the graph to represent changes in behaviour more effectively.

The node set is constructed from these displacements, $V_t = \{o_t^i \mid i = 1, \ldots, N\}$, and the initial edges are fully connected, $U_t = \mathbf{1}$. Thus, the spatial graph at time t is formulated as $G_S = (V_t, U_t)$.

An attention mechanism is used to compute an initial spatial interaction score matrix, which provides a rough estimate of how strongly each pair of pedestrians influences each other. First, the graph features are embedded via $E_S = \phi(G_S, W_E)$. These embeddings are then linearly projected into query matrices $Q = \phi(E_S, W_Q)$ and key matrices $K = \phi(E_S, W_K)$. Finally, the interaction score matrix R_{QK} is calculated using scaled dot-product attention:

$$R_{QK} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{QK}}}\right) \tag{3.29}$$

3.3.2 Adaptive Spatio-Temporal Interaction Module

This is the main contribution of the TRACE-Net paper. The core concept of Adaptive convolution is that, rather than sampling features from a fixed grid, the kernel can shift to accommodate the pattern of interactions. Human interactions rarely follow strict patterns, so an adaptive kernel is useful in modelling realistic interactions. TRACE-Net builds on regular 3D deformable convolution by incorporating three asymmetric kernels: $3 \times 1 \times 1$, $1 \times 3 \times 1$, $1 \times 1 \times 3$. We model temporal transitions, horizontal spatial shifts, and vertical spatial shifts. As such, the model can learn interactions across dimensions without overfitting or mixing them up. The deformable offsets are predicted

using tiny convolution layers and then each kernel is applied to features sampled from shifted locations. The model can learn more complex spatial-temporal interactions by stacking three layers of this module.

$$\Delta P = \text{Conv3D}(R_{QK}) \tag{3.30}$$

$$P^* = P_K + \Delta P \tag{3.31}$$

$$\gamma(Q) = \sum_{P_K \in R} W(P_K) \cdot R_{QK}(Q + P^*) \tag{3.32}$$

$$R_{ST}^{(l+1)} = \delta\left(\psi_1(R_{ST}^{(l)}) + \psi_2(R_{ST}^{(l)}) + \psi_3(R_{ST}^{(l)})\right) + R_{ST}^{(l)} \tag{3.33}$$

$$A_{ST} = \text{Softmax}\left((\sigma(R_{ST}^{(3)}) + I) \odot R_{ST}^{(3)}\right) \tag{3.34}$$

3.3.3 Graph Convolution and Feature Representation

With a refined spatial-temporal adjacency matrix, TRACE-Net uses a graph convolution layer to produce the final representation of a trajectory. The graph convolution updates the features of each pedestrian by aggregating information from its neighbours, as defined by the learned adjacency matrix.

$$H_{ST} = \delta(A_{ST}\, G_S\, W_{ST}) \tag{3.35}$$

3.3.4 Competitive Attention Module

While the spatial-temporal fusion backbone certainly makes it possible to extract rich representations with information about interactions, it necessarily mixes different patterns of motion. Pedestrian motion is inherently multimodal; given a social context, there are multiple possible motion outcomes.

To disentangle this mixture in a holistic manner, we propose a competitive attention module. This decomposition breaks the feature fusion into distinct latent features, called *slots*, that each have a specialisation in modelling a particular behavioral mode or interaction context.

Specifically, given the spatial–temporal feature tensor H_{ST} produced by the graph convolution (we will refer to it as H in the following for simplicity):

$$H \in \mathbb{R}^{N \times H \times T \times C}$$

we initially compress the temporal dimension to extract localized, per-pedestrian feature embeddings:

$$X \in \mathbb{R}^{N \times H \times C}$$

Following normalization, these embeddings are processed by the Competitive Attention module, which initializes a fixed number of learnable slot vectors sampled from a standard distribution:

$$S_0 \sim \mathcal{N}(\mu, \sigma)$$

During each refinement iteration, cross-attention is calculated between the input features and the latent slots:

$$A = \text{Softmax}\left(\frac{K(X)Q(S)^T}{\sqrt{d}}\right) \tag{3.36}$$

Crucially, these attention coefficients are normalized over the input dimension, enforcing a strict competition among the slots to claim specific feature representations:

$$\tilde{A}_{ij} = \frac{A_{ij}}{\sum_i A_{ij}} \tag{3.37}$$

The slots are subsequently updated via a weighted aggregation of the input features:

$$U = \tilde{A}^T V(X) \tag{3.38}$$

To capture evolving dynamics, the slots undergo iterative refinement driven by a Gated Recurrent Unit (GRU):

$$S^{t+1} = \text{GRU}(U, S^t) \tag{3.39}$$

which is followed by a residual Multi-Layer Perceptron (MLP) for non-linear projection:

$$S^{t+1} = S^t + \text{MLP}(\text{LayerNorm}(S^t)) \tag{3.40}$$

Upon completing the iterative updates, the slots successfully encapsulate distinct, disentangled facets of the interaction dynamics. To leverage this information holistically, the slot states are averaged to form a consolidated, global representation:

$$S_{\text{agg}} = \frac{1}{K}\sum_{k=1}^{K} S_k \tag{3.41}$$

Finally, this aggregated context is broadcast and residually fused back into the original spatial–temporal feature tensor:

$$H' = H + S_{\text{agg}} \tag{3.42}$$

This iterative update explicitly fuses disentangled behavioral modes and social context in the baseline feature space, thus enabling the downstream decoder to generate a wide range of diverse, socially acceptable multimodal trajectories.

3.3.5 Attention-Gated Temporal Decoder

To predict future trajectories, TRACE-Net uses an Attention-Gated Temporal Convolution Network. This module takes the slot-enhanced feature representation H' and processes it across time, while a gating mechanism adjusts how much information should pass through at each time step. It outputs the parameters of a 2D Gaussian

distribution for each future frame:

$$\text{mean:} \quad \mu_t^i = (\mu_{x,t}^i, \mu_{y,t}^i) \tag{3.43}$$

$$\text{standard deviation:} \quad \sigma_t^i = (\sigma_{x,t}^i, \sigma_{y,t}^i) \tag{3.44}$$

$$\text{correlation:} \quad \rho_t^i \tag{3.45}$$

3.3.6 Latin Hypercube Sampling

Most models rely on Monte Carlo sampling, which tends to produce clustered predictions and ignores rare behaviours. TRACE-Net instead uses Latin Hypercube Sampling (LHS), which forces samples to be spread across the entire distribution. The sampled points are generated as:

$$\text{lhs}_j^{(d)} = \frac{j + r_j}{m} \tag{3.46}$$

These samples are then mapped to the standard normal distribution using the inverse Gaussian cumulative distribution function, Φ^{-1}. The covariance structure is applied using the Cholesky decomposition $E_S = LL^T$ (and similarly $P = LL^T$), yielding the transformed variable $Z = \Phi^{-1}(\text{LHS})$. The final diverse predicted trajectories are computed as:

$$\hat{p}_t^i = \hat{\mu}_t^i + LZ \tag{3.47}$$

3.3.7 Training Using Negative Log-Likelihood

Since the model predicts Gaussian parameters, the loss function is the **Negative Log-Likelihood (NLL)** of the ground truth trajectory under the predicted distribution. This encourages the model to predict both accurate means and realistic uncertainty.

$$L_i(W) = -\sum_{t=t_{\text{obs}}+1}^{t_{\text{pred}}} \log P\big(p_t^i \mid \hat{\mu}_t^i, \hat{\sigma}_t^i, \hat{\rho}_t^i\big) \tag{3.48}$$

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Results of SWIFT-Net

We tested our model on the ETH and UCY pedestrian trajectory datasets, which capture real world crowd behavior across five different scenes. These datasets include annotated trajectories and reflect a variety of interaction patterns, from simple movements to more crowded and complex situations. Following common practice, we use 8 time steps as input and predict the next 12, evaluating the model in a leave-one-out setting.

We implement out model in PyTorch and trained using the Adam optimizer with a learning rate of 0.01 and weight decay of 1e-4. We train the model for 150 epochs with gradient clipping applied to stabilize training. We use a multi-step learning rate scheduler to reduce the learning rate at predefined milestones.

The architecture of our model consists of a spatial-temporal graph convolution backbone with learnable interaction graphs, followed by a temporal convolutional network (TCN) for sequence prediction. We use an embedding dimension of 64 and stack multiple temporal convolution layers to model future trajectories.

To capture multiple behaviour patterns, we include a slot attention module that disentangles multiple modes of pedestrian motion. The model outputs multiple trajectories (20 samples per agent), and training is done using a best-of-K style objective that encourages at least one prediction to match the ground truth closely.

During evaluation, the predicted relative trajectories are converted to absolute coordinates and compared against ground truth using standard metrics the Average Displacement Error (ADE) and the Final Displacement Error (FDE)
Evaluating pedestrian trajectory prediction requires observing how models handle extended time horizons. As the prediction window expands from 4 to 24 frames, the inherent uncertainty of human movement naturally compounds, leading to an increase in displacement errors across all models. Figure 4.1 demonstrates the ADE and FDE metrics under these extended horizons. We compare SWIFT-Net against leading baseline architectures, including Social-STGCNN [26], SGCN [5], and WTGCN [34]. While traditional models degrade rapidly as the horizon stretches, SWIFT-Net exhibits a significantly slower rate of error accumulation. Even at $t = 24$ frames, the model maintains an ADE of 0.43 meters and an FDE of 0.72 meters, substantially outperforming

Table 4.1: Results of SWIFT-Net : ADE/FDE Metrics (in meters) of Our Implementations on ETH and UCY Datasets.

Method	**ETH**	**HOTEL**	**UNIV**	**ZARA1**	**ZARA2**
STGAT[39]	0.65/1.12	0.35/0.66	0.52/1.10	0.34/0.69	0.29/0.60
SR-LSTM[40]	0.63/1.25	0.37/0.74	0.51/1.10	0.41/0.90	0.32/0.70
Social-STGCNN[26]	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48
SGCN[5]	0.63/1.03	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48
GCHGAT[29]	0.63/1.10	0.38/0.73	0.55/1.16	0.33/0.66	0.30/0.64
Social-Implicit[23]	0.66/1.44	**0.20/0.36**	0.31/0.60	0.25/0.50	0.22/0.43
Tri-HGNN[28]	0.62/0.86	0.38/0.65	0.49/0.88	0.27/0.44	0.25/0.40
RDGCN[43]	0.58/0.94	0.30/0.45	0.35/0.65	0.28/0.48	0.25/0.44
STAGP[41]	0.65/1.21	0.41/0.73	0.38/0.68	0.28/0.46	0.25/0.44
SKGACN[42]	0.55/0.83	0.30/0.50	0.39/0.75	0.30/0.51	0.26/0.45
WTGCN[34]	0.60/0.95	0.25/0.37	0.36/0.65	0.27/0.46	0.23/0.39
SWIFT-Net(Ours)	**0.36/0.57**	0.24/0.45	**0.25/0.43**	**0.18/0.31**	**0.16/0.29**

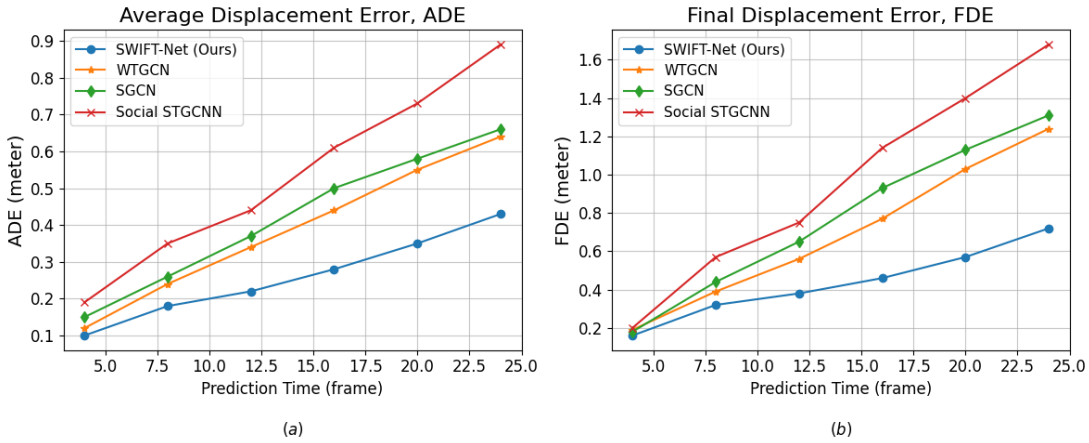

Figure 4.1: Prediction error at different prediction times of the different methods on the ETH/UCY datasets.

WTGCN. This stability highlights the advantage of our hierarchical wavelet decomposition and adaptive fusion mechanisms, which successfully capture and sustain long-range temporal dependencies without discarding critical interaction context.

4.2 Results of Trace-Net

To ensure robustness in the predictive ability of the proposed architecture, we test the model architecture on the popular ETH and UCY datasets of pedestrian trajectories. These benchmarks consist of five different scenes, which exhibit a wide range of social interactions from simple unidirectional linear motions to highly crowded and complex interactions. Following standard evaluation procedures, 8 time steps are used to predict the next 12 time steps, with a rigorous leave-one-out cross-validation protocol applied across the five different scenes. The model is developed using the PyTorch library. The

Table 4.2: Results of TRACE-Net : ADE/FDE Metrics (in meters) of Our Implementations on ETH and UCY Datasets.

Method	**ETH**	**HOTEL**	**UNIV**	**ZARA1**	**ZARA2**
STGAT[39]	0.65/1.12	0.35/0.66	0.52/1.10	0.34/0.69	0.29/0.60
SR-LSTM[40]	0.63/1.25	0.37/0.74	0.51/1.10	0.41/0.90	0.32/0.70
Social-STGCNN[26]	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48
SGCN[5]	0.63/1.03	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48
GCHGAT[29]	0.63/1.10	0.38/0.73	0.55/1.16	0.33/0.66	0.30/0.64
Social-Implicit[23]	0.66/1.44	**0.20/0.36**	0.31/0.60	0.25/0.50	0.22/0.43
Tri-HGNN[28]	0.62/0.86	0.38/0.65	0.49/0.88	0.27/0.44	0.25/0.40
RDGCN[43]	0.58/0.94	0.30/0.45	0.35/0.65	0.28/0.48	0.25/0.44
STAGP[41]	0.65/1.21	0.41/0.73	0.38/0.68	0.28/0.46	0.25/0.44
SKGACN[42]	0.55/0.83	0.30/0.50	0.39/0.75	0.30/0.51	0.26/0.45
WTGCN[34]	0.60/0.95	0.25/0.37	0.36/0.65	0.27/0.46	0.23/0.39
TRACE-Net(Ours)	**0.46/0.74**	0.21/0.37	**0.28/0.48**	**0.21/0.38**	**0.16/0.30**

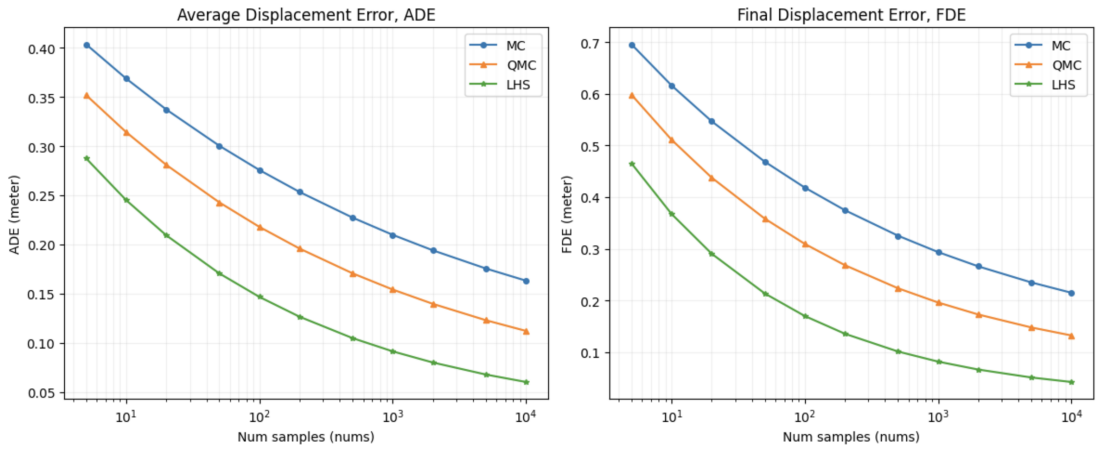

Figure 4.2: The average ADE and FDE curves on the ETH and UCY datasets using the different sampling methods at different sampling numbers

base model is built upon a spatial-temporal graph convolutional network with learnable interaction graphs that is then fed to a Attention-Gated Temporal Decoder for sequence prediction. The embedding dimension of the features is maintained at 64, and several temporal convolution layers are applied to predict future.The network is trained for 100 epochs using Adam optimizer with an initial learning rate of 0.01 and weight decay of 1×10^{-4}. To enhance the stability of the training process, gradient clipping is applied and a multi-step learning rate scheduler is used to decrease the learning rate at specific epochs. To explicitly tackle the multimodal nature of human movement, the model includes a Competitive Attention module, designed to separate different modes of human movement. As a result, the model outputs 20 samples for each agent. A best-of-Kstyle loss is used to train the model, ensuring that at least one of the predicted trajectories is close to the ground truth.

At test time, all relative trajectories are transformed back to absolute coordinates. The

model's performance is evaluated against ground truth data using two commonly used metrics for trajectory prediction: Average Displacement Error (ADE) and Final Displacement Error (FDE).

Effect of Latin Hypercube Sampling: To confirm the generative efficiency of the proposed architecture, we perform an ablation study comparing our proposed Latin Hypercube Sampling (LHS) strategy with conventional Monte Carlo (MC) and Quasi-Monte Carlo (QMC) techniques. As shown in Fig. 4.2, while the error bounds ADE and FDE of winner-takes-all objective decrease as the number of generated samples increases, LHS consistently exhibits superior performance. In particular, LHS displays a much faster decay in error for a reduced number of samples, and reaches a lower bound in error. This experimental result motivates the inclusion of LHS in the TRACE-Net pipeline to ensure multiple diverse predictions, without the computational burden of large sample sizes.

CHAPTER 5

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

5.1 Conclusion

In this work, we introduce SWIFT-Net, a multi-scale framework for pedestrian trajectory prediction that models both spatial and temporal interactions. By combining graph-based interaction modeling with wavelet decomposition, the model is able to capture patterns at different levels of granularity. The adaptive fusion mechanism helps maintain a balance of spatial and temporal cues depending on the scene, while the inclusion of slot attention allows the model to represent multiple possible modes of human behavior. Unlike approaches that rely on sampling from a distribution, our method directly predicts a set of candidate trajectories, making it both efficient and interpretable. This paper also presents TRACE-Net, a multi-scale architecture to overcome the rigidity of spatio-temporal modeling structures and achieve multi-modal predictions in pedestrian trajectory forecasting. Through the use of an Adaptive Spatio-Temporal Interaction Module, the framework adeptly models non-linear social interactions via adaptive kernels. A Competitive Attention mechanism explicitly disentangles latent motion modes, to generate diverse and socially acceptable trajectories. Detailed experiments on ETH and UCY benchmarks show that TRACE-Net offers a sound mathematical framework for multi-modal reasoning on crowded scenes. Together, the proposed frameworks demonstrate the importance of multi-scale interaction modeling, adaptive spatio-temporal reasoning, and explicit multimodal representation for realistic pedestrian trajectory prediction. By addressing limitations in existing approaches related to rigid interaction modeling, limited contextual reasoning, and uncertainty in future motion, this thesis contributes robust and interpretable approaches for socially-aware trajectory forecasting in complex environments.

5.2 Future Scope

There are still several directions worth exploring. For example, some obstacles, fixed landmarks or semantic maps could be added to the scene to make the trajectory constraints more accurate and enhance performance. Furthermore, the models would be more useful in the real world, like autonomous navigation, intelligent surveillance, and human-aware robotics, if they were adaptable to larger crowds, high-density scenes, or

real-time deployment situations. Future work will also examine how the models can be adapted online during inference, adding online learning components to the models. There are several considerations for ongoing research, such as scalable multimodal reasoning, modeling dynamic interactions, and more robust evaluation methods for various trajectory prediction tasks.

5.3 Social Impact

Pedestrian trajectory prediction has many direct applications to human environment intelligent systems. Understanding human movement is essential for enhancing the safety and reliability of autonomous vehicles, particularly in complex or high-traffic environments.The ability to accurately predict human movement can help to improve the safety and reliability of autonomous vehicles, especially in complex or high-traffic environments. Likewise, for service robots or assistive robotic systems in hospitals, airports, shopping centres and other public areas, where it is important to understand human motion to avoid collision and to behave naturally in human crowds, socially-aware navigation is crucial. The objectives of the proposed frameworks in this thesis are to help make intelligent systems safer and more adaptive to human systems. The proposed approaches explicitly model social interaction dynamics, multi-scale crowd dynamics and multimodal future behaviours and make an effort to capture the uncertainty and variability in real-world human movement. This can be useful to enhance decision making for applications in which the wrong trajectory prediction can result in unsafe or inefficient behaviour.

In addition to autonomous navigation, pedestrian trajectory prediction also finds application in intelligent surveillance, crowd management, urban planning and disaster response. By analyzing crowd movement patterns, identifying unusual motion patterns, and optimizing the layout of public infrastructure to enhance safety and accessibility, predictive systems can help to improve the overall efficiency and effectiveness of the system.

Meanwhile, the use of trajectory prediction systems needs to be responsible. Models built from smaller data sets might not be able to replicate performance in all settings, in all cultures or in all crowd behaviours. Moreover, there are significant privacy, surveillance, and ethical issues with large-scale monitoring systems. Future research should not only prioritize further enhancing prediction accuracy and robustness but also ensure fairness, transparency, privacy protection, and responsible deployment of intelligent trajectory prediction systems in real-world scenarios.

BIBLIOGRAPHY

[1] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] A. Sadeghian, F. Legros, M. Voisin, R. Vesel, A. Alahi, and S. Savarese, "Carnet: Clairvoyant attentive recurrent network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[5] L. Shi, L. Wang, C. Long, S. Zhou, M. Zhou, Z. Niu, and G. Hua, "Sgcn: Sparse graph convolution network for pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 8994–9003.

[6] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," *Advances in neural information processing systems*, vol. 32, 2019.

[7] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9813–9823.

[8] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 10 335–10 342.

[9] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff, "Covernet: Multimodal behavior prediction using trajectory sets," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[10] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," *arXiv preprint arXiv:1910.05449*, 2019.

[11] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics research: the 14th international symposium ISRR*. Springer, 2011, pp. 3–19.

[12] C. Schöller, V. Aravantinos, F. Lay, and A. Knoll, "What the constant velocity model can teach us about pedestrian motion prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1696–1703, 2020.

[13] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *CVPR 2011*, 2011, pp. 1345–1352.

[14] I. Karamouzas, P. Heil, P. Van Beek, and M. H. Overmars, "A predictive collision avoidance model for pedestrian simulation," in *International workshop on motion in games*. Springer, 2009, pp. 41–52.

[15] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras, "People tracking with human motion predictions from social forces," in *2010 IEEE International Conference on Robotics and Automation*, 2010, pp. 464–469.

[16] V. Pavlovic, J. M. Rehg, and J. MacCormick, "Learning switching linear models of human motion," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. MIT Press, 2000.

[17] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 283–298, 2007.

[18] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "Nonparametric bayesian learning of switching linear dynamical systems," in *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., vol. 21. Curran Associates, Inc., 2008.

[19] J. Yang, Y. Chen, S. Du, B. Chen, and J. C. Principe, "Ia-lstm: Interaction-aware lstm for pedestrian trajectory prediction," *IEEE Transactions on Cybernetics*, vol. 54, no. 7, pp. 3904–3917, 2024.

[20] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[21] P. Dendorfer, S. Elflein, and L. Leal-Taixé, "Mg-gan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 13 158–13 167.

[22] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 759–776.

[23] A. Mohamed, D. Zhu, W. Vu, M. Elhoseiny, and C. Claudel, "Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 463–479.

[24] W. Zhang, H. Cheng, F. T. Johora, and M. Sester, "Forceformer: Exploring social force and transformer for pedestrian trajectory prediction," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, 2023, pp. 1–7.

[25] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *European conference on computer vision*. Springer, 2020, pp. 683–700.

[26] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 412–14 420.

[27] J. Mi, X. Zhang, H. Zeng, and L. Wang, "Dergcn: Dynamic-evolving graph convolutional networks for human trajectory prediction," *Neurocomputing*, vol. 569, p. 127117, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223012407

[28] 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.

[29] L. Zhou, Y. Zhao, D. Yang, and J. Liu, "Gchgat: Pedestrian trajectory prediction using group constrained hierarchical graph attention networks," *Applied Intelligence*, vol. 52, no. 10, pp. 11 434–11 447, 2022.

[30] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3d convolution for video super-resolution," *IEEE Signal Processing Letters*, vol. 27, pp. 1500–1504, 2020.

[31] W. Chen, H. Sang, J. Wang, and Z. Zhao, "Dstcnn: Deformable spatial-temporal convolutional neural network for pedestrian trajectory prediction," *Information Sciences*, vol. 666, p. 120455, 2024.

[32] K. Mangalam, Y. An, H. Girase, and J. Malik, "From goals, waypoints & paths to long term human trajectory forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 15 233–15 242.

[33] T. Yao, Y. Pan, Y. Li, C.-W. Ngo, and T. Mei, "Wave-vit: Unifying wavelet and transformers for visual representation learning," in *European conference on computer vision*. Springer, 2022, pp. 328–345.

[34] W. Chen, H. Sang, J. Wang, and Z. Zhao, "Wtgcn: wavelet transform graph convolution network for pedestrian trajectory prediction," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 12, pp. 5531–5548, 2024.

[35] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid, C. Li, and D. Anguelov, "Tnt: Target-driven trajectory prediction," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 895–904. [Online]. Available: https://proceedings.mlr.press/v155/zhao21b.html

[36] C. ". Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, and D. Anguelov, "Motiondiffuser: Controllable multi-agent motion prediction using diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 9644–9653.

[37] J. Gu, C. Sun, and H. Zhao, "Densetnt: End-to-end trajectory prediction from dense goal sets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 15 303–15 312.

[38] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4601–4607.

[39] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "Stgat: Modeling spatial-temporal interactions for human trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[40] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 085–12 094.

[41] Z. Liu, L. He, L. Yuan, K. Lv, R. Zhong, and Y. Chen, "Stagp: Spatio-temporal adaptive graph pooling network for pedestrian trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2001–2007, 2023.

[42] K. Lv and L. Yuan, "Skgacn: Social knowledge-guided graph attention convolutional network for human trajectory prediction," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[43] H. Sang, W. Chen, J. Wang, and Z. Zhao, "Rdgcn: Reasonably dense graph convolution network for pedestrian trajectory prediction," *Measurement*, vol. 213, p. 112675, 2023.

[44] W. Chen, H. Sang, J. Wang, and Z. Zhao, "Dstigcn: Deformable spatial-temporal interaction graph convolution network for pedestrian trajectory prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 5, pp. 6923–6935, 2025.

[45] Y. Wu, L. Wang, S. Zhou, J. Duan, G. Hua, and W. Tang, "Multi-stream representation learning for pedestrian trajectory prediction," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 3, 2023, pp. 2875–2882.

Appendix-A

PROOF OF PUBLICATION

The following page(s) contain the proof of acceptance of the research work carried out as part of this thesis. The papers were accepted for presentation in the conference:

"Beyond Fixed Kernels: Dynamic Spatiotemporal Interaction Graphs for Trajectory Forecasting"
Author(s): Vedant Dhanyamraju, Kavinder Singh and Anil Singh Parihar
Accepted in: *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)*
Location: Philippine Women's University, Manila, Philippines
Acceptance Date: *17 May 2026*
PAPER ID : 1571283568

"SWIFT-Net: Slot-Guided Wavelet Interaction Fusion for Trajectories"
Author(s): Vedant Dhanyamraju, Kavinder Singh and Anil Singh Parihar
Accepted in: *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)*
Location: Philippine Women's University, Manila, Philippines
Acceptance Date: *17 May 2026*
PAPER ID : 1571282710

The acceptance letters / emails of the accepted articles are attached below.

24/AFI/11 VEDANT DHANYAMRAJU <vedantdhanyamraju_24afi11@dtu.ac.in>

CHOICE of Presentation [CICN 2026]

gstomar=ieee.org@edas.info <gstomar=ieee.org@edas.info> Sun, May 17, 2026 at 11:13 AM
Reply-To: Geetam Singh Tomar <gstomar@ieee.org>
To: Vedant Dhanyamraju <vedantdhanyamraju_24afi11@dtu.ac.in>, Kavinder Singh <kavinder@dtu.ac.in>, Anil Singh Parihar <parihar.anil@gmail.com>

Dear Mr. Vedant Dhanyamraju:

Congratulations -
On behalf of organising committee of 2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN), I am happy to inform you that, your paper #1571283568, Title: ('Beyond Fixed Kernels: Dynamic Spatiotemporal Interaction Graphs for Trajectory Forecasting') has been registered for Oral Presentation. The presentation schedule is online on https://cicn.in/Important_Dates.html you are requested to details of your paper and if any change or change in your paper you may contact on email.
[Quoted text hidden]

24/AFI/11 VEDANT DHANYAMRAJU <vedantdhanyamraju_24afi11@dtu.ac.in>

CHOICE of Presentation [CICN 2026]

gstomar=ieee.org@edas.info <gstomar=ieee.org@edas.info> Sun, May 17, 2026 at 11:13 AM
Reply-To: Geetam Singh Tomar <gstomar@ieee.org>
To: Vedant Dhanyamraju <vedantdhanyamraju_24afi11@dtu.ac.in>, Kavinder Singh <kavinder@dtu.ac.in>, Anil Singh Parihar <parihar.anil@gmail.com>

Dear Mr. Vedant Dhanyamraju:

Congratulations -
On behalf of organising committee of 2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN), I am happy to inform you that, your paper #1571282710, Title: ('SWIFT-Net: Slot-Guided Wavelet Interaction Fusion for Trajectories') has been registered for Oral Presentation. The presentation schedule is online on https://cicn.in/Important_Dates.html you are requested to details of your paper and if any change or change in your paper you may contact on email.
If you have not signed copyright then sign the copyright within 2 days. procedure is:
login to your edas account, click on (c) just below status of paper and when author information is seen edit it remove Mr. Dr etc and may correct author names and then save and then next, ... and submit

The conference will be held on June 11-14, 2026 at Philippine Women's University, Manila, Philippines. Should you have any questions, please feel free to inform us through email. kindly visit website http://cicn.in for details.

With kind regards,

Chair, Technical Programming Committee
18th IEEE CICN- 2026
Conference email: cicnmir@gmail.com

ICICI Bank

E-Receipt

Transaction Reference Number	FDY3872435
Transaction Type	Within Bank
Date of Transaction	2026-04-30-T-20:13:24
From ICICI Bank Account	003101034198-S U M PRASAD DHANYAMRAJU
Payment/Payee	019701005475-MACHINEINTEL
Transaction Amount(Rs.)	8067
Remarks	ID1571283568

ICICI Bank

E-Receipt

Transaction Reference Number	FDY3871601
Transaction Type	Within Bank
Date of Transaction	2026-04-30-T-20:16:04
From ICICI Bank Account	003101034198-S U M PRASAD DHANYAMRAJU
Payment/Payee	019701005475-MACHINEINTEL
Transaction Amount(Rs.)	8067
Remarks	ID1571282710

DELHI TECHNOLOGICAL UNIVERSITY

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

PLAGIARISM VERIFICATION

Title of the Thesis **DEVELOPMENT OF GRAPH NEURAL NETWORK APPROACH FOR PEDESTRIAN TRAJECTORY PREDICTION**

Total Pages **45** Name of the Scholar **Vedant Dhanyamraju**

Supervisor (s)

(1) **"Prof. Anil Singh Parihar"**

Department **of Computer Science and Engineering**

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

Software used: **"Turnitin"** Similarity Index: **8%**, Total Word Count: **14281**

Date: **May, 2026**

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

vedant dhanyamraju

Modified Thesis

Thesis 2

Document Details

Submission ID

trn:oid:::27535:140545513

Submission Date

May 26, 2026, 8:49 PM GMT+5:30

Download Date

May 26, 2026, 8:57 PM GMT+5:30

File Name

Modified Thesis.pdf

File Size

5.2 MB

45 Pages

14,281 Words

81,534 Characters

8% Overall Similarity

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

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 8 words)

Exclusions

- 3 Excluded Sources

Match Groups

90 Not Cited or Quoted 7%
Matches with neither in-text citation nor quotation marks

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

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

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

Top Sources

4% Internet sources
5% Publications
4% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

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

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

Match Groups

- **90** Not Cited or Quoted 7%
  Matches with neither in-text citation nor quotation marks
- **12** Missing Quotations 1%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

4% Internet sources
5% Publications
4% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	arxiv.org	1%
2	Publication	Xiaobo Chen, Fengbo Luo, Feng Zhao, Qiaolin Ye. "Goal-Guided and Interaction-A...	<1%
3	Internet	edwinhernandez.com	<1%
4	Internet	www.mdpi.com	<1%
5	Internet	export.arxiv.org	<1%
6	Publication	Biao Yang, Shicheng Lu, Shaobo Shen, Chuan Hu, Rongrong Ni. "MMFlow: Multim...	<1%
7	Publication	Jiejun Hao, Ting Wang, Yuanzi Xu, Rongjun Cheng. "TSI-GCN: Temporal-Spectral In...	<1%
8	Internet	assets-eu.researchsquare.com	<1%
9	Internet	bura.brunel.ac.uk	<1%
10	Student papers	Indian Institute of Technology, Bombay on 2026-02-18	<1%

vedant dhanyamraju

Modified Thesis

Thesis 2

Document Details

Submission ID

trn:oid:::27535:140545513

Submission Date

May 26, 2026, 8:49 PM GMT+5:30

Download Date

May 26, 2026, 8:58 PM GMT+5:30

File Name

Modified Thesis.pdf

File Size

5.2 MB

45 Pages

14,281 Words

81,534 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.