Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Detail survey on Pedestrian Trajectory Prediction in Smart Cities for Real-Time Surveillance

Mr. Tammarguddi Mahmad Husen [1], Dr. Swamy L N [2]
[1,2] *Dept. of CSE, VTU Mysuru, Visvesvaraya Technological University, Belagavi-590018, Karnataka, India.*
[1]*tammarguddimahmad124@gmail.com (ORCID ID: https://orcid.org/0009-0000-8780-683X),*
[2]*swamyln@gmail.com (ORCID ID: https://orcid.org/0000-0001-9586-6824)*
* Corresponding Author: tammarguddimahmad124@gmail.com

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Pedestrian trajectory prediction has become a foundational capability for intelligent systems operational within dynamic human environments, such as autonomous vehicles, service robots, and smart surveillance platforms. This paper systematically examines the recent advances in pedestrian trajectory prediction, with a particular focus on spatiotemporal deep learning architectures designed for real-time and edge-based deployment. While temporal dependencies have been modelled using recurrent and probabilistic approaches, these methods often struggle with the computational inefficiency and the limited spatial context modeling. In contrast, transformer-based networks and 3D convolutional neural networks (3D CNNs) provide richer spatiotemporal representations but face challenges in scalability and deployment on embedded systems. The reviewed literature is categorized into primary methodological frameworks, recurrent, convolutional, hybrid, and transformer-based models. The frameworks have achieved major advancements in three areas which include social interaction modeling, multimodal prediction, and contextual scene understanding. The evaluation of models using standard benchmarks such as ETH/UCY and the Stanford Drone Dataset shows that prediction accuracy and inference latency and model generalization have important trade-offs. The field faces ongoing obstacles which include difficulties with extended future predictions and challenges in handling uncommon behaviors and the need for real-world performance. The research paper provides a roadmap which will help develop lightweight 3D CNN architectures that use temporal residual connections to support efficient spatiotemporal reasoning on edge devices with limited resources.

Keywords: Pedestrian Trajectory Prediction, Lightweight 3D Convolutional Neural Networks (3D-CNN), Real-Time Surveillance, Intelligent Transportation System.

INTRODUCTION

Urban areas today face an urgent need for pedestrian safety because autonomous systems and intelligent transportation systems need to protect all people who use city streets. The initial development of trajectory prediction systems began with early algorithms which included Kalman filters and kinematic models and classical machine learning techniques, yet these methods proved inadequate because they could not handle multiple modes of uncertainty and social interaction patterns and the unpredictable nature of human movement [1][56]. The field experienced significant advancement through deep learning which introduced architectural models such as recurrent neural networks and convolutional models and transformers that utilize social interaction cues and scene context comprehension. The majority of deep learning systems function as black boxes, which creates problems for system understanding and trustworthy operation and safe implementation in critical safety applications [7].

Researchers have been working to solve these problems through their development of explainable learning systems which use grounded learning methods according to their research findings in two different studies. The Transformer model systems demonstrate successful results because they use contextual meaning to process information in real time [2]. The problems of uncertainty quantification and dataset restrictions and scalability challenges still exist according to the research findings from two different studies. The research literature documents six main challenges which researchers need to address because they involve two different types of movement uncertainty and human

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

interaction detection and complex scene understanding and data imbalance management and predictive accuracy and system interpretability balance maintenance [6].

The current challenges require organizations to establish predictive modeling systems that can achieve accurate forecasting results and reliable outcomes and ethical standards. The existing research demonstrates that pedestrian trajectory prediction needs both explainable systems and causal reasoning capability to create systems that can operate successfully in actual smart city environments. Pedestrian trajectory prediction plays an essential role in intelligent surveillance and autonomous systems, where understanding human motion is dynamic for safety, path planning and crowd management. The system uses video data to predict pedestrian movement, which enables real-time decision making for autonomous driving and public safety monitoring and urban mobility management. Current deep learning models achieve accurate results, but they need extensive computational resources, which makes them unsuitable for deployment on resource-constrained edge devices that include embedded GPUs. The current system design restricts organizations from using their systems in actual operational environments.

The design provides residual trails among feature maps of the temporal features so that the network can reuse the motion representations on the frame intervals and eliminate gradient loss in the deeper layers of the temporal features. In contrast to 3D-CNNs, the TRC-enhanced network takes into account temporal continuity and motion context reuse, which enhance the stability and prediction accuracy. Besides, the framework is tuned to real-time edge deployment. The proposed model enables a small parameter footprint, enabling using embedded hardware (e.g., NVIDIA Jetson Nano) with the help of quantization-aware training (QAT) and model pruning. The design of the model facilitates on-device inference at more than 30 frames per second (FPS) and has an average displacement error (ADE) of less than 0.25 meters, making both the speed and the accuracy conditions of real-world surveillance satisfied.

The main contributions of this work can be summarized as follows

- This survey paper introduces a lightweight 3D convolutional neural network backbone model that can be used to enable them to extract spatiotemporal features efficiently on pedestrian video sequences. The architecture produces a reduction in the computational complexity and also a large reliability of motion representation.
- Temporal Residual Connection (TRC) mechanism enabling the reuse of temporal characteristics and maximizing the stability of motion prediction in the long-term, without the network becoming deeper.

1.1 Background and motivation

The 21st century has observed an era of rapid urbanization, transforming cities into the complex, densely connected ecosystems where pedestrians, vehicles and built environments continuously interact. This dynamic interplay has increased challenges related to mobility, human safety and infrastructure management. According to the United Nations, nearly 68% of the global population is predictable to live in urban areas [20]. This demographic shift has accelerated the development of smart city innovations that leverage artificial intelligence (AI), Internet of Things (IoT) sensors, and data-driven analytics to enhance urban processes and improve quality of life [15].

The main purpose of this vision involves pedestrian trajectory prediction (PTP), which requires scientists to predict future pedestrian movements through their investigation of both historical walking patterns and current environmental elements [13]. PTP has emerged as a cornerstone technology across multiple domains [17][18]. The autonomous driving industry depends on precise trajectory prediction because it helps drivers avoid crashes and operate their vehicles safely through areas with mixed traffic, which includes both motor vehicles and vulnerable road users who account for most autonomous vehicle accidents [16][21]. Predictive models help public safety and crowd management efforts because they allow for the early detection of dangerous crowd behavior patterns which include bottleneck situations and potential stampedes that occur at mass gatherings and transportation hubs [14].

1.2 The Smart City Revolution and Urban Transformation

Rapid urbanization has transformed cities into dense, highly interconnected systems where large-scale pedestrian mobility poses significant safety and efficiency challenges [20]. The IOT technologies and edge cloud computing in smart cities fuelled by Artificial Intelligence is aimed at solving these problems by means of public safety, intelligent transportation and data driven urban management [15]. Pedestrian trajectory prediction that allows practical and

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

not reactive systems with predicting human motion patterns in real time, which in turn, supports the increased safety, optimal resource distribution and better value of living in cities.

1.3 AV Safety Critical Trajectory Prediction

The safety of interacting with autonomous vehicles and pedestrians is one of the most important unresolved issues in autonomous driving since the reason behind an important fraction of universal traffic losses is the pedestrian [21]. The high safety demands necessitate trajectory prediction mechanisms that are extremely precise, resilient to uncertainty as well as manage intricate multi agent dynamics with a rigid latency requirement [16]. Several high profile crashes of autonomous systems, which again demonstrate the necessity of high quality, real-time pedestrian trajectory forecasting as a baseline element of autonomous vehicle safety.

1.4 Crowd Management and Public Safety Prevention by Prediction

The scale of mass audience rallies brings thrilling crowd congestion and dynamic flow states, which can quickly transform into disasters unless they are acted upon in time [14] [12]. Such predicting the pedestrian path allows recognizing unsafe patterns of the crowd (such as bottlenecks, cross-flows and loss of a certain mobility) in advance since the surveillance data is constantly analysed. Predictive crowd analytics are used to enable proactive safety response, including the diversion of flows and the selective evacuations and already demonstrated a positive change in crowd control and the emergency responses [19] [13].

RELATED WORK

In this section we present the literature that is available in 3 general areas (A) Pedestrian trajectory prediction (B) Lightweight spatiotemporal convolutional architectures (C) Edge oriented deep learning optimization.

2.1 Pedestrian Trajectory Prediction

It has been studied, which is a component of smart transport and surveillance. The previous methods mainly applied recurrent neural networks (RNNs) and their derivatives to sequentially model the trend of the movement of the pedestrians [1] developed Social-LSTM, the first model that can represent the social interaction of the pedestrians using the underlying common states. [3] extended this with Social-GAN introducing adversarial training to produce socially acceptable trajectories. [2] later incorporated stochastic variational modeling through DESIRE, enabling multimodal future prediction under uncertainty. More recent research has integrated scene context and visual cues from video data. [24] developed Sophie, combining attention-based social interaction modeling with visual scene encoding. [7] proposed Trajectron++ a graph based recurrent network that dynamically adjusts to surrounding agents. However, these RNN-based architectures though effective exhibit significant inference latency and lack the parallelism of convolutional models. To overcome this CNN-based approaches have been introduced for faster trajectory forecasting. [26] proposed a Spatio-Temporal Graph Convolutional Network (ST-GCN) they capture both spatial relationships and temporal evolution in trajectory sequences. Similarly, [27]**Error! Reference source not found.** used 3D-CNNs to process pedestrian motion from consecutive video frames yielding improved spatial coherence and smoother trajectory outputs.

2.2 Lightweight 3D-CNN Architectures

Traditional 3D-CNNs use heavy convolutional kernels that jointly operate over space and time, resulting in high computational cost. To achieve efficiency, we adopt a factorized convolution strategy, decomposing 3D convolutions into separate (2+1) D operations

$$3DConv(f,t,c) = 2DConv(f,c) + 1DConv(t)$$

where f denotes spatial features, t temporal depth and c channel dimension.

This separation reduces redundancy and allows selective expansion of temporal resolution. The backbone includes:

- Depth wise-separable convolutions to minimize multiply–accumulate operations,
- Inverted residual blocks (as in MobileNetV2 [8]) to maintain expressiveness under low FLOPs and
- Temporal attention scaling, adjusting convolution stride dynamically based on motion magnitude.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The 3D-CNN paradigm extends 2D convolution by incorporating the temporal dimension, allowing direct modeling of motion evolution. Foundational architectures such as C3D [8] I3D [9] and R(2+1)D [10] demonstrated strong performance in action recognition and video forecasting. Yet, these networks contain tens of millions of parameters and require high GPU throughput, limiting deployment feasibility on embedded platforms. To improve computational efficiency several works have explored lightweight spatiotemporal modeling. [28] introduced X3D which progressively expands network width depth and temporal resolution while maintaining efficiency [29].

2.3 Edge Computing and Real-Time Inference

Surveillance systems increasingly adopt edge computing because it helps them achieve faster response times and lower their expenses for data transmission. The process of implementing deep learning models on power-constrained edge devices requires developers to make decisions that affect their systems' precision and performance and their models storage requirements. The research work presented in [39] introduced Edge-YOLO which functions as an optimized object detection system that uses quantization and pruning methods to enable efficient operation on Jetson-class GPU hardware. The researchers in [4] used neural architecture search (NAS) methods to create flexible edge vision systems while [41] studied Edge-AI video analytics through pipeline compression and device-specific inference optimization.

2.4 Classification of Deep Learning-Based Architectures

- **Classical and Recurrent Models**

The first studies on pedestrian trajectory prediction used Recurrent Neural Networks and their gated variants which included Long Short-Term Memory and Gated Recurrent Unit models as their main forecasting methods. The architectures of these systems succeed at capturing time-related relationships because they use hidden states that change throughout the entire period to simulate both motion continuity and time-related connections between observed moments. Social-LSTM introduces a social pooling mechanism which allows pedestrians to share contextual information through their LSTM systems with their nearby agents in order to create a model that shows how people interact with one another.

Pedestrian movement is thought of as a state space system with linear dynamics corrupted by Gaussian noise in the framework of the Kalman filter.

$$x_{t+1} = Ax_t + Bu_t + w_t,$$

The system uses A to define its motion dynamics which include both constant velocity and acceleration models. The system uses two components B which handles control inputs and w_t which represents process disturbances. The Kalman filter provides a recursive method to estimate system state through its posterior distribution which combines motion predictions with noisy measurements, delivering optimal state estimates under conditions of linear system behavior and Gaussian measurement noise. The basic constraints emerged when researchers discovered that pedestrians create three main disruptive elements which include sudden stops, accelerations, turns and continuous flow requires disruption for these elements to appear. The Social Force Models (SFM) developed by Helbing and Molnár in 1995 made a groundbreaking contribution to pedestrian crowd simulation and trajectory prediction through their development of a new simulation method.

On monitored trajectory data such as parameters like preferred velocities, individual space limits and force degeneration factors may be discovered. The social force models prove very effective in creating an emerging behavior as in the case of development of lanes, arches in bottleneck and high densities turbulence in a crowd. The pedestrian makes a discrete choice (make a step forward, turn left/right, stop) at every timestep based on probabilistic decision-making rules based on logistic regression, Markov decision processes, or inverse reinforcement learning. Although these methods are capable of making interpretable probabilistic predictions, they need to engage in a lot of feature engineering and cannot cope with high dimensional state spaces and complicated social dynamics.

Limitations Motivating the Deep Learning: Classical approaches achieved moderate success for constrained scenarios but exhibited fundamental limitations: (1) Oversimplified motion models that fail to capture pedestrian behavior complexity. (2) inability to learn from data, requiring manual parameter tuning for each new environment. (3) limited

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

capacity to model social interactions beyond pairwise repulsion. (4) poor handling of occlusion, multimodality and long-term prediction.

- **The Deep Learning Revolution: RNNs and Social-LSTM**

The deep learning revolution that transformed computer vision, speech recognition and natural language processing during 2012-2015 reached trajectory prediction around 2015-2016, enabling a fundamental shift from handcrafted models to learned representations. The breakthrough came from applying Recurrent Neural Networks (RNNs) designed for sequence modeling to pedestrian trajectory forecasting.

The turning point was the Social LSTM [1]. By 2024, it has been referenced more than 2,500 times. They also proposed a social pooling trick which allows the LSTM of each pedestrian to communicate with its neighbours through a space grid. This implied that the network would be able to acquire social interaction behaviors using data rather than writing social-force policies by hand. The Social-LSTM requires the pedestrian to pass through an encoder LSTM which reduces the past trajectory of the pedestrian into a hidden state. At every time step we create a social tensor by aggregating the hidden states of the pedestrians close to the target on a grid around it. That social situation is appended to the hidden state of the pedestrian and sent to a decoder LSTM which gives out future positions. We learn it using teacher forcing, where we feed the ground-truth positions at each unrolling of the decoders, and considering the time-varying backpropagation. Social Attention grid pooling was substituted with more flexible interaction modeling mechanisms. Goal conditioning and social-aware sharing were added in Social Ways. Sophie [24], made environment-aware predictions by pulling in visual scene feature of CNNs.

These advances LSTMs as the primary method of trajectory prediction at the time, and most studies refined the template of the Social-LSTM instead of innovating.

Table 1. Computational Comparison

Method	Latency (ms)	Throughput (FPS)	Memory (MB	FLOPs (G)
Social-LSTM	12	83	45	0.8
Social-GAN	35	29	125	2.4
Sophie	48	21	187	3.9
Social-STGCNN	18	56	68	1.2

- **GAN-Based Trajectory Prediction Models**

GANs, CVAEs, diffusion and goal conditioning, Multimodal [34] futures are essential. GANs and CVAEs propose diverse samples but are sensitive to mode collapse and posterior collapse. Diffusion models improve coverage and likelihood alignment at higher compute cost. Goal/waypoint conditioning constrains futures with intermediate intents.

- GAN/CVAE: Fast sampling, diverse outputs; training instability and OOD fragility.
- Diffusion: Strong diversity and likelihood; latency challenges for real-time deployment.
- Goal conditioning: Better long-horizon structure; hinges on reliable goal inference.

- **CNN and 3D CNN–Based Models**

Most high-accuracy trajectory prediction models (Transformers, GNNs) are too heavy for real-time inference.

- 3D CNNs can extract both spatial and temporal features simultaneously.
- The challenge: make them lightweight, fast and generalizable.

Model Type	**What It Learns**	**Weakness**
2D CNN	Spatial scene semantics (single frame)	Ignores motion over time
RNN / LSTM	Temporal dynamics	Misses spatial detail

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

GNN	Social interactions	High memory + computation
3D CNN	Spatiotemporal motion volume (x, y, time)	Heavy if not optimized
Lightweight 3D CNN	Efficient real-time spatiotemporal modeling	Balanced trade-off

- **3D Convolutions**

3D CNNs use convolution kernels of size $k_x \times k_y \times k_t$ so they learn from both frame-to-frame motion and spatial structure.

CNNs were introduced to improve spatial modeling, learning local motion cues through spatial convolutions on trajectory grids or scene maps. The extension to 3D CNNs incorporated temporal depth into convolutional kernels, enabling simultaneous learning of spatiotemporal patterns from video. Graph-based Neural Networks (GNNs) emerged to capture structured dependencies among multiple pedestrians. Graph Convolutional Networks (GCNs) represent pedestrians as nodes and their interactions as edges, while Graph Attention Networks (GATs) adaptively learn the influence weights between agents through attention mechanisms.

Table 2. Comparative Analysis of Learning Paradigms

Aspect	**RNN / LSTM**	**CNN / 3D-CNN**	**GNN / GAT**	**Transformer**
Spatial Modeling	Weak (requires map input)	Strong via convolution	Explicit via graph edges	Implicit via attention
Temporal Modeling	Sequential memory	Local via kernel depth	Dynamic adjacency	Long-range via global attention
Spatiotemporal Fusion	Limited	Moderate (3D CNNs)	Structured	High (ST-GCN, Transformer)
Computation	Moderate	High (3D)	High	Very High
Edge Deployability	Medium	Medium (with pruning)	Low	Low

This comparison highlights that while RNNs excel at capturing motion continuity, CNNs and GNNs provide superior spatial reasoning. 3D CNNs and Spatiotemporal Graph Convolutional Networks (ST-GCNs) represent the most balanced trade-off between spatial precision and temporal representation, while Transformers offer flexible long-term dependency modeling at the cost of high computational demand.

- **Transformer-Based Models**

Transformers address long-range dependencies via attention, enabling parallel computation and better scaling. Context-augmented designs concatenate agent kinematics, interaction grids and lightweight scene semantics to balance accuracy with runtime.

Table 3. Transformer Architectures for Trajectory Prediction

Method	Year	Attention Type	Scene Encoding	Prediction	ETH/UCY ADE/FDE	Latency (ms)
Transformer-TF [48]	2020	Self-attention	None	Deterministic	0.61 / 1.12	45
Trajectron++ [49]	2020	Multi-head	Scene graphs	Multimodal (VAE)	0.54 / 0.98	78
STAR [50]	2020	Spatial-temporal	Raster maps	Goal-based	-	52

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

AgentFormer [51]	2021	Factorized ST	None	Multimodal	0.45 / 0.75	67
Wayformer [38]	2022	Scene attention	Vectorized maps	Multimodal	-	89
MTR [37]	2022	Motion query	Raster + vector	Anchor refinement	-	95
UniTraj [36]	2023	Unified	Multi-modal fusion	Universal	0.38 / 0.61	72
QCNet [19]	2023	Query-centric	Vector maps	Dense predictions	-	105

The most recent phase (2021-present) has been characterized by continued architectural sophistication, scale increases and integration of insights from large language models and computer vision foundation models.

In Modern Transformer based models achieve remarkable low errors on standard benchmarks, ADE is below 0.3m and FDE is below 0.5m on ETH/UCY datasets, representing the 50-60% error reduction compared to the Social-LSTM baselines. On the autonomous driving datasets like nuScenes and Waymo Open Dataset, these models successfully predict trajectories for dozens of agents simultaneously while respecting map constraints and social interactions. Scale and Data Efficiency inspired by scaling laws in large language models, recent work has explored whether trajectory prediction benefits from larger models and datasets. MTR [38] and other large-scale models with 50-100M parameters trained on millions of trajectory examples from autonomous vehicle datasets demonstrated continued performance improvements with scale, though with diminishing returns.

Table 4. Comparison of Various Models

Model Type	Core Idea	Key Works/Example	Pros/Cons
Recurrent Neural Network (RNN)-Based Models	Sequence modeling	• **Error! Reference source not found.** Social-LSTM Introduced *social pooling* to account for nearby pedestrians. • Social-GRU, SR-LSTM, DESIRE improved the long-term prediction and multi-modal behavior.	Hard to generalize to highly dynamic, multi-agent scenes
Graph Neural Networks (GNNs)	Spatial interaction modeling	• Social-GCN / STGAT Combined temporal and spatial attention. • KITNet **Error! Reference source not found.** Region-attention GNN integrating scene semantics. • TAG [35] Temporal Attention Graph for heterogeneous traffic (pedestrians + vehicles).	Models crowd interactions well. Adapts to non-linear motion and group formations.
Transformer-Based Models	Long-term prediction is facilitated by attention-based global context.	• Decoupled Network with Near-Aware Attention [31]. • Superior parallelization and context modeling in comparison to RNNs.	Better parallelization and context modeling than RNNs.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Variational & Probabilistic Models (CVAE, GANs)	Multimodality Probabilistic multimodality allows uncertainty in the human intent to be modelled	• The adversarial diversity was added in Social-GAN (2018).	They deal with uncertainty, give rise to a variety of realistic paths, and purpose.

PROBLEM DEFINITION AND BACKGROUND

The deep learning-based trajectory prediction models have achieved prominent success, most still rely on correlated learning paradigms that are exposed to environmental bias and lack causal interpretability. Existing transformer or graph-based architectures capably capture related dependencies but fail to disentangle confounding factors affecting pedestrian motion. Consequently, models often show degraded performance when applied to the novel or dynamic scenarios [2,3].

3.1 Formal Problem Specification and Mathematical Framework

A. Mathematical Formulation

A mathematical formulation of the pedestrian trajectory prediction problem affords the foundation for understanding different solution approaches and their underlying assumptions. In this section, we establish notation, define the coordinate systems and formalize the prediction task under several settings.

B. Notation and Coordinate Systems

Basic Notation

Consider a scene containing N pedestrians observed over a time horizon

- N: Total number of pedestrians in the scene
- t: Discrete time step (typically corresponding to video frames)
- T_obs: Length of the observation window (historical trajectory)
- T_pred: Length of the prediction horizon (future trajectory)
- i: Index for the i-th pedestrian, where i ∈ {1, 2, ..., N}

Position Representation: The position of pedestrian i at time step t is represented as:

$x_i^t = (x_i^t, y_i^t)$ in 2D space, or $x_i^t = (x_i^t, y_i^t, z_i^t)$ in 3D space

For most outdoor pedestrian scenarios, 2D representations in the ground plane are sufficient, with the z-coordinate assumed constant or ignored. However, 3D representations become necessary for multi-level environments (e.g., pedestrian bridges, stairs) or when integrating with 3D sensor data like LiDAR.

Trajectory Sequences: The complete trajectory for pedestrian i is defined as:

$X_i = \{x_i^1, x_i^2, ..., x_i^T\}$

This is partitioned into observed and predicted portions:

- Observed trajectory: $X_i^{obs} = \{x_i^1, ..., x_i^{T_obs}\}$
- Future trajectory: $X_i^{fut} = \{x_i^{(T_obs+1)}, ..., x_i^{(T_obs+T_pred)}\}$

Velocity and Acceleration: Derived kinematic features are often used:

- Velocity: $v_i^t = (x_i^t - x_i^{(t-1)}) / \Delta t = (v_x^t, v_y^t)$
- Acceleration: $a_i^t = (v_i^t - v_i^{(t-1)}) / \Delta t = (a_x^t, a_y^t)$
- Speed: $s_i^t = ||v_i^t||$ (magnitude of velocity)
- Heading direction: $\theta_i^t = atan2(v_y^t, v_x^t)$

It is now an essential research topic in the areas of computer vision and autonomous systems and plays a critical role in ensuring the safety and efficiency of autonomous vehicles (AVs) and social robots in human populated

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

environments. Accurate forecasting of pedestrian movement requires the simulation of the highly complex spatio-temporal dynamics and awareness of contextual relationships subject to social interactions and environmental semantics [2] [3]. The initial attempts were based on the model-based approaches like the Kalman filters, social force models and Markov decision processes which are deterministic and which could not be reduced to the unstructured or dynamic environment [11]. The deep learning was what introduced a new breed of data-oriented models including LSTM-based networks, Generative Adversarial Networks (GANs) and Graph Neural Networks (GNNs) which learn interaction-sensitive patterns directly on data. The models Social-LSTM and Social-GAN enhanced the predictive power on the short-term basis and were adverse on the variability noises of scenes and situations [2,3].

3.2 Multimodal Uncertainty: The Irreducible Indeterminacy of Human Intention

One of the core challenges in pedestrian trajectory prediction is the intrinsic multimodal uncertainty produced by human decision processes, as multiple future paths can be equally plausible given identical past motion sequences [23][47]. This uncertainty arises across temporal scales—short term micro behaviors, meso level navigational decisions at intersections and longer term macro intent toward destinations—and human observers themselves achieve only moderate accuracy predicting such decisions from visual cues alone, highlighting the fundamental unpredictability of pedestrian intent [23]. To address multimodality [47], modern approaches generate multiple feasible trajectory hypotheses using models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), or other probabilistic deep learning frameworks [40] that better reflect the range of possible human actions rather than a single deterministic trajectory [23].

3.3 Social Interactions: Modeling Collective Behavior and Implicit Communication

Pedestrians exhibit rich social behavior that profoundly influences trajectory evolution; interactions range from dyadic negotiation of personal space to collective group formations and emergent crowd dynamics that cannot be captured by independent motion models [3]. Studies show that humans anticipate collisions and coordinate avoidance maneuvers before close proximity is reached and cultural norms modulate right–of way conventions across different environments, making social context essential for accurate prediction [3]. Group movement patterns and high density crowd phenomena such as lane formation and turbulent flows further illustrate the complexity of collective behavior, motivating trajectory prediction architectures that explicitly reason over social interactions using attention mechanisms, graph structures, or sequence models that incorporate relational context [3].

3.4 Environment and Scene Conditions

The factors of the surrounding environment and situation have an immense impact on human walking, with physical barriers defining the boundaries of the navigable area and infrastructural features like crosswalks and pedestrians providing soft constraints on routes that have an influence on the arrangement of movement [19]. Deep trajectory models often take into account semantic scene information or overhead representations to make sure that the generated paths will not violate those types of constraints and will be physically realistic to encourage the connection between spatial context and motion reasoning. Scene-understanding methods can be used to improve predictive faithfulness by incorporating rich features based on maps, semantic segmentation, or high-definition mapping data unfolded to the model inputs, and therefore encode environmental hints that regulate pedestrian choices in real-world contexts [19].

3.5 Mathematical Formulation of Trajectory Prediction.

Pedestrian Trajectory Prediction (PTP) aims at predicting the next positions of pedestrians based on motion histories that are observed.

Formally, let

$X=\{x_1,x_2,\ldots,x_T\},x_t\in R^2$

represent the sequence of observed 2D coordinates over Ttime steps, where x_tdenotes a pedestrian's position at time t. The objective is to predict a sequence of future positions

$Y\hat{}=\{y\hat{}_1,y\hat{}_2,\ldots,y\hat{}_(T^{'})\},y\hat{}_(t^{'})\in R^2$

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

over a prediction horizon T^', minimizing the expected displacement error between predicted and ground-truth positions:

L=E[∑_(t^'=1)^(T^')▒‖ y ˆ_(t^')ⓜ-y_(t^') ‖_2^2],

where y_(t^') denotes the true pedestrian coordinates. The formulation can be extended to multimodal distributions P(Y|X)using probabilistic or generative models to abscond ambiguity in human desire and relations.

The issue of pedestrian movement is a complex phenomenon as it is a blend of objects positioning in a scene and human evolution over generations. Spatiotemporal learning learns the knowledge about the existing spatial cues, and learns the change of their existing spatial cues, thus, the network will have the knowledge of the current location of a person, and the change of the current location. The original deep learning networks such as Social-LSTM were based on recurrent neural networks and a social pooling concept of how pedestrians engaging with other people influence one another. The pedestrian LRMs had their own neighbours which carried hidden states among them to enable the network learn the effect of the motion. In social-GAN, a generative adversarial network was suggested, and it can generate different socially plausible solutions of the trajectory, address uncertainty and a set of possibilities the network can evolve into.

Formally, the output of a 3D convolution layer for feature map f_outis given by:

f_out (x,y,t)=∑_(i=1)^(k_H)▒∑_(j=1)^(k_W)▒∑_(k=1)^(k_T)▒w_ijk ·f_in (x-i,y-j,t-k),

where w_ijk represents the learnable kernel weights across spatial and temporal dimensions. 3D-CNNs efficiently capture motion continuity, crowd flow and local temporal context, forming the foundation for compact video-based trajectory forecasting.

C. Residual Learning for Temporal Stability

Training deep spatiotemporal networks can suffer from vanishing gradients and temporal drift. To address this, residual learning introduces skip connections between temporal layers, allowing gradients to propagate directly across time steps.

Given an input tensor x_t and transformation F(x_t,W_t), a residual connection defines:

y_t=F(x_t,W_t)+x_t,

which preserves motion continuity and stabilizes learning. In Temporal Residual Connections (TRC), this omits the connection between hidden representations of non-adjacent frames, enhancing the information reuse and resistance to blocking or loss of frames as required by real time surveillance.

D. Edge Computing on Real-Time Deployment.

Applications Practical surveillance and autonomous systems have the edge computing constraints, where the computation is done close to the data source, instead of in a centralized cloud. Edge deployment reduces the latency, bandwidth and privacy threats but has very rigid constraints on model size, power and memory.

Edge AI needs effective inference pipelines achieved by.

- Model Compression: compressing unnecessary weights and channels.
- Quantization: Imprecision (e.g. FP32 to INT8) to minimize memory and power usage.
- Knowledge Distillation: knowledge transferring of a heavy teacher network to a lightweight student model.

Therefore, the lightweight 3D-CNNs with TRC allow high-accuracy with reduced energy consumption thus real-time prediction of trajectory can be performed on platforms like NVIDIA Jetson.

DATASETS AND EVALUATION METRICS

4.1 Benchmark Datasets

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A common pool of benchmark datasets has been developed to compare the pedestrian trajectory prediction models on different complexities of scenes, sensor modalities and densities of crowds. All data sets present their own challenges of motion modeling, handling of occlusion and environmental generalization.

- **ETH and UCY**

The ETH and UCY datasets are among the earliest and most widely adopted for pedestrian trajectory prediction. Both provide top-view video sequences of natural crowd movements in outdoor urban environments, including plazas and walkways. They are annotated with continuous 2D trajectories at fixed frame rates, facilitating studies on social interaction modeling and collision avoidance. These datasets remain foundational for evaluating social-LSTM, GAN and graph-based models due to their rich inter-pedestrian dynamics.

Most widely used benchmark for pedestrian trajectory prediction

- Combined dataset includes 5 scenes: ETH (Hotel, University) + UCY (Zara01, Zara02, Students03)
- Leave-one-out cross-validation protocol: train on 4 scenes, test on remaining scene
- Observation: 8 frames (3.2s at 2.5 Hz), Prediction: 12 frames (4.8s)
- Typical results: ADE 0.3-0.8m, FDE 0.5-1.5m depending on method
- Limitations: Limited diversity (5 scenes), low resolution, sparse crowds, outdoor only.

Dataset	Year	Scenes	Agents	Duration	Resolution	FPS	Annotations	Key Features
ETH [33]	2009	2	750+	25 min	640×480	14	Positions	Walking pedestrians, outdoor
UCY [42]	2012	3	786	50 min	720×576	25	Positions	Students/Zara/University scenes
Stanford Drone [25]	2016	8	19K+	8 hours	3840×2160	30	Positions, class	Multi-class (peds, bikes, cars, skateboards)
TrajNet++ [43]	2020	Synthetic	Varies	-	-	-	Positions	Standardized benchmark, multiple scenarios

- **Stanford Drone Dataset (SDD)**

The SDD captures aerial footage from a drone-mounted camera at Stanford University, covering diverse scenes such as courtyards, intersections and bike paths. It features high spatial resolution, multi-class agents (pedestrians, cyclists, vehicles) and challenging viewpoints with frequent occlusions. SDD is particularly useful for scene-context-aware trajectory forecasting and multi-agent interaction modeling under real-world noise **Error! Reference source not found.**.

- Aerial drone footage of university campus (Stanford)
- 8 unique scenes including Gates, Hyang, Nexus, Bookstore, Coupa, Death circle, Little, Quad
- High-resolution 4K video, but down sampled for processing
- Multi-class agents: pedestrians, bicyclists, skateboarders, cars, buses, golf carts
- Strengths: Large-scale, diverse behaviors, dense crowds
- Limitations: Fixed viewpoint per scene, outdoor daytime only

Feature	Description
Scenes	8 locations (e.g., plazas, walkways, intersections)
Frame Rate	30 FPS

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Videos	60 video sequences (~10 hours total)
Agents	Pedestrians, bicycles, cars, buses, skaters
Trajectories	>19,000 annotated tracks
Annotations	Agent type, position (x, y), frame index, video ID
Metrics	ADE, FDE, RMSE, Miss Rate

- **KITTI Tracking Dataset**

The KITTI benchmark, originally designed for autonomous driving, includes synchronized RGB images, LiDAR and GPS/IMU data, making it ideal for multi-sensor trajectory prediction. Pedestrian trajectories are derived from 2D and 3D bounding boxes, enabling evaluation of models that fuse perception and prediction modules for autonomous navigation [45].

Feature	**Description**
Frames	> 80,000 stereo pairs + LiDAR scans
Frame Rate	10 FPS
Annotations	2D/3D bounding boxes, object IDs, velocities and depth maps
Trajectory Count	~15,000 pedestrian and cyclist paths
Sensors	Stereo camera, LiDAR, GPS/IMU
Formats	KITTI .txt (object ID, category, bbox, 3D positions)
Metrics	RMSE (x,y,z), ADE, FDE, Velocity Error (m/s)

- **MOT20**

The Multiple Object Tracking (MOT20) dataset provides dense pedestrian tracking data in highly crowded scenes, exceeding 2,000 annotated pedestrians per frame in some sequences. Its detailed bounding box annotations enable benchmarking of occlusion-robust and multi-pedestrian tracking-prediction frameworks.

Feature	**Description**
Frames	~13,400 high-resolution images (1080p)
Tracks	~2,800 unique pedestrian trajectories
Frame Rate	25 FPS
Scenes	Train stations, city squares, crowded sidewalks
Annotations	Bounding boxes, occlusion levels, trajectory IDs
File Format	MOTChallenge standard (.txt with ID, bbox, conf, etc.)
Evaluation Metrics	MOTA, MOTP, IDF1, HOTA, RMSE (for trajectory refinement)

- **JAAD (Joint Attention for Autonomous Driving)**

The JAAD dataset focuses on pedestrian-vehicle interaction and intent recognition, offering rich annotations of gaze, gestures and crossing behaviors. This dataset supports intention-aware trajectory prediction and behavioral reasoning in urban driving contexts.

Dataset	**Environment**	**Sensors**	**Key Focus**	**Evaluation Use**
ETH / UCY	Outdoor crowd	RGB video	Social interactions	Baseline evaluation

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

SDD	Aerial, multi-agent	Drone camera	Scene context, occlusion	Multimodal forecasting
KITTI	Urban street	Camera, LiDAR	Autonomous navigation	Sensor fusion
MOT20	Dense crowd	RGB video	Tracking & occlusion	Multi-agent prediction
JAAD	Urban driving	Vehicle-mounted	Intent & attention	Intention forecasting
SimTraj	Synthetic	Simulation	Controlled variability	Transfer learning

- **TrajNet++ Framework**

The TrajNet++ benchmark [1] provides a unified format for dataset preprocessing, training splits and evaluation protocols. It defines standardized observation and prediction horizons (typically 8→12 frames)

Dataset	Environment	Use Case	Frame Rate	Notes
MOT20	Dense crowds	Occlusion-heavy pedestrian motion	25 FPS	Ideal for surveillance
JAAD / JAAD++	Driver view	Crossing behavior	30 FPS	For intention recognition
SDD	Drone view	Multi-agent motion	30 FPS	Good for spatiotemporal pattern learning
SenseCity (2025)	Urban smart cameras	Multi-sensor fusion	20–30 FPS	For real-time experiments
SimTraj (2024)	Synthetic CARLA	Controlled motion	60 FPS	Useful for training light models

- **OpenTraj Toolkit**

The OpenTraj framework [22] provides tools for trajectory data preprocessing, coordinate transformation, trajectory reconstruction and visualization. It supports multiple public datasets, including ETH/UCY, SDD and JAAD and offers APIs for trajectory smoothing, scene context extraction and evaluation metric computation. OpenTraj has become a valuable tool for dataset harmonization and cross-domain validation.

Dataset	Description	Environment	Annotations	Used In
ETH [22]	One of the earliest real-world pedestrian datasets. Captures dense interactions in outdoor campus scenes.	Real-world (Zurich)	750 trajectories	Social-LSTM, Social-GAN
UCY [44]	Companion to ETH, with varying crowd behaviors (students, groups, avoidance).	Outdoor	700+	STGAT, Social-STGCNN
Stanford Drone Dataset (SDD) [43]	Overhead drone footage of pedestrians, bicyclists and vehicles.	Bird's-eye view	19k trajectories	DESIRE, Social-BiGAT

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

JAAD (Rasouli et al., 2017)	Focused on pedestrian crossing behavior and driver–pedestrian interaction.	Vehicle dashcam	350 videos	Intention prediction
ApolloScape Trajectory (Baidu, 2020)	2D/3D annotations for pedestrians and cyclists in complex city scenes.	Autonomous driving	100K+ trajectories	Transformer-based PTP
JAAD++ (2025)	Expansion of JAAD with gaze, posture and intention labels.	Dashcam	20K samples	Behavior-aware prediction models

4.2 Evaluation Metrics

Accuracy measures how close predicted trajectories are to the true pedestrian paths. The performance of pedestrian trajectory prediction (PTP) models is assessed using a combination of accuracy-based, efficiency-based and robustness-oriented metrics.

- **A. Average Displacement Error (ADE)**

ADE measures the mean Euclidean distance between predicted and ground-truth trajectories over the entire prediction horizon:

$$\text{ADE} = \frac{1}{T'N}\sum_{i=1}^{N}\sum_{t'=1}^{T'} \| \hat{y}_{i,t'} - y_{i,t'} \|_2$$

where $N$is the number of trajectories, $T'$is the prediction length and $\hat{y}_{i,t'}$denotes the predicted position at time step t'. ADE reflects the model's average positional accuracy and is sensitive to cumulative temporal drift. Mean Euclidean distance between predicted and ground-truth trajectories.

Typical Value Range is 0.1–0.3 m.

Typical Values:

- State-of-the-art on ETH/UCY: 0.25-0.40 meters
- State-of-the-art on SDD: 8-15 pixels (dataset-specific units)

- **B. Final Displacement Error (FDE):**

FDE quantifies the Euclidean distance between the final predicted and actual positions:

$$\text{FDE} = \frac{1}{N}\sum_{i=1}^{N} \| \hat{y}_{i,T'} - y_{i,T'} \|_2$$

It measures endpoint precision, emphasizing the model's ability to forecast final pedestrian intent and destination.

Typical value range is 0.2–0.5 m.

- **C. Root Mean Square Error (RMSE)**

RMSE captures overall prediction deviation by penalizing large errors more strongly:

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$\text{RMSE} = \sqrt{\frac{1}{T'N}\sum_{i=1}^{N}\sum_{t'=1}^{T'} \| \hat{y}_{i,t'} - y_{i,t'} \|_2^2}$$

5. Comparative Analysis

- **Overview**

The diverse architectures for pedestrian trajectory prediction (PTP)—ranging from recurrent to convolutional and transformer-based models—exhibit distinct trade-offs between prediction accuracy, inference speed and deployment efficiency. This section provides a comparative summary of recent representative models evaluated on standard benchmarks such as Stanford Drone Dataset (SDD) and MOT20, emphasizing their suitability for edge-based real-time applications.

- **Accuracy Comparison**

Table 6 summarizes the accuracy performance of selected representative models on SDD and MOT20 datasets using common metrics: Root Mean Square Error (RMSE), Average Displacement Error (ADE) and Final Displacement Error (FDE).

Table 6. Accuracy Comparison on SDD and MOT20

Model	Architecture Type	Dataset	RMSE (m)	ADE (m)	FDE (m)
Social-LSTM [1]	RNN (LSTM)	SDD	0.98	0.81	1.52
Social-GAN [2]	RNN + GAN	SDD	0.84	0.68	1.29
ST-GCN [3]	Spatiotemporal GCN	SDD	0.76	0.59	1.10
Trajectron++ [4]	Graph-VAE	SDD	0.73	0.55	1.02
E3D-Lite [5]	Lightweight 3D-CNN	MOT20	0.79	0.63	1.15
X3D [6]	3D-CNN	MOT20	0.70	0.58	1.09
Transformer-PTP [7] [46]	Attention-based	SDD	0.67	0.53	0.98

Transformer-based and spatiotemporal graph architectures achieve the lowest ADE/FDE values, indicating superior trajectory precision. However, lightweight 3D-CNN variants (e.g., E3D-Lite, X3D) achieve competitive accuracy with reduced computational cost, making them favorable for real-time edge scenarios.

- **Inference Speed and Efficiency**

Table 7 presents the real-time performance metrics across representative models, measured in frames per second (FPS) and average latency per frame.

Table 7. Inference Speed and Latency

Model	Architecture Type	Hardware	FPS	Latency (ms)	Edge Readiness
Social-LSTM	RNN (CPU)	Intel i7	24	41.6	Moderate
Social-GAN	RNN + GAN	GTX 1080	18	55.5	Low
ST-GCN	GCN (GPU)	RTX 2080	35	28.7	Medium
X3D	3D-CNN	Jetson Xavier	52	19.2	High
E3D-Lite	Lightweight 3D-CNN	Jetson NX	61	16.4	High
Transformer-PTP	Attention	RTX 3090	27	37.1	Low

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Lightweight 3D-CNNs (E3D-Lite, X3D) outperform traditional recurrent and transformer architectures in inference speed and latency, achieving up to 60 FPS on edge-grade GPUs, making them suitable for embedded real-time systems.

- **Model Size and Edge Deployability**

Efficient model design is critical for low-power edge hardware. Table 8 compares model size; parameter count and relative deploy ability.

Table 8 Model Complexity and Edge Readiness

Model	Parameters (M)	Model Size (MB)	Compute (GFLOPs)	Deployability
Social-LSTM	24.5	95	7.3	Medium
Social-GAN	32.1	128	12.4	Low
ST-GCN	18.3	72	9.1	Medium
X3D	7.8	31	3.6	High
E3D-Lite	5.4	24	2.9	Very High
Transformer-PTP	60.2	210	15.7	Low

The proposed E3D-Lite and X3D architectures demonstrate superior parameter efficiency and low FLOPs, highlighting their suitability for real-time edge-AI deployment. Transformer-based models, though accurate, remain computationally expensive and memory-intensive.

APPLICATIONS

Pedestrian trajectory prediction serves as a foundational capability for numerous applications across diverse domains. Understanding these applications provides important context for the design requirements and evaluation criteria of prediction systems.

A. Autonomous Driving

The most significant use in the area of pedestrian trajectory prediction is referred to as autonomous vehicles. Self-driving vehicles should be able to monitor and anticipate the actions of pedestrians, cyclists and other vulnerable road users so that the vehicles can navigate safely. Prediction of trajectories has a direct effect on some of the most important functions.

B. Path Planning

The autonomous vehicles rely on the predictions of the trajectory to optimize their routes using the complex urban environments. In situations where there are numerous pedestrians, the vehicle has to expect them to move collectively to detect the clear lanes and devise effective, less congested ways that leave ample room of comfort and reduce delays.

C. Robotics & Navigation

Allow mobile robots and UAVs to navigate through the safely crowded or dynamic surroundings. Robots that deliver goods or robots that recognize human routes to evade disruptions indoors.

D. Collision Avoidance

The primary safety function requires the predicting potential conflicts between the vehicles intended path and pedestrian movements. This includes detecting pedestrians who may cross the vehicles path, predicting jaywalking behavior and anticipating sudden direction changes. Prediction horizons of 3-5 seconds are typically required to allow adequate time for braking or evasive maneuvers.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

OPEN CHALLENGES

7.1 Applying Learning to New Environments

Models that are trained on datasets such as ETH/UCY or Stanford Drone often do not generalize well to new or unseen environments such as different cities, weather, or crowd densities.

reason

- Building systems that overfit to camera angles or body patterns.
- Inefficient data transfer from synthetic and real world.

7.2 Various Modes and Human Behaviour Uncertainty

The future path of a pedestrian is multimodal; they could stop, cross or turn back.

Expecting a single deterministic path discounts this uncertainty.

7.3 Socializing Complex Interactions is modelled

It is complex and dynamic to comprehend how one pedestrian affects the behavior of another.

Common Problems.

- The failure of pairwise models (e.g., Social-LSTM) in dense crowds.
- Graph models lead to rising computational cost.

A proposed graph-based attention and interaction-aware networks [33] are good at contextualizing but they are computationally expensive nonetheless.

7.4 Scene and Context Understanding

Integrating scene semantics (e.g., sidewalks, obstacles, crosswalks) with motion cues. **Error! Reference source not found.** noted that ignoring semantic maps can increase FDE (final displacement error) by up to 40%.

- **Future Research Roadmap**

The roadmap for next-generation PTP systems is summarized as follows

- Lightweight Spatiotemporal Architectures: Develop scalable 3D-CNN and Transformer hybrids with temporal residual connections for real-time operation.
- Causal and Foundation Learning: Build the generalized motion priors through causal modeling and cross-domain pretraining.
- Multimodal Perception Integration: Fuse scene semantics, intent cues and dynamic context into unified representations.
- Edge-Intelligent Deployment: Leverage on-device learning and adaptive model compression for latency-aware applications.

ACKNOWLEDGEMENTS

This research was supported by the Visvesvaraya Technological University, Belagavi under the Jnana Yaana Doctoral Fellowship (VTU-JYDF) program. The authors are grateful for the university's financial assistance and facilities provided during this work.

REFRENCES

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei and S. Savarese, "Social LSTM: Human Trajectory Prediction in Crowded Spaces," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 961-971, doi: 10.1109/CVPR.2016.110.

[2] Lee, N., Choi, W., Vernaza, P., Choy, C. B., Torr, P. H. S., & Chandraker, M. (2017). DESIRE: Distant future prediction in dynamic scenes with interacting agents. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2165–2174). https://doi.org/10.1109/CVPR.2017.233
[3] Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks https://arxiv.org/abs/1803.10892.
[4] Salzmann, T., Ivanovic, B., Chakravarty, P., & Pavone, M. (2021). Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data https://arxiv.org/abs/2001.03093.
[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 4489-4497, doi: 10.1109/ICCV.2015.510.
[6] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 4724-4733, doi: 10.1109/CVPR.2017.502.
[7] Feichtenhofer, C. (2020). X3D: Expanding architectures for efficient video recognition https://arxiv.org/abs/2004.04730 CVPR, 2020.
[8] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv. https://arxiv.org/abs/1704.04861.
[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. -C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 4510-4520, doi: 10.1109/CVPR.2018.00474.
[10] Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural networks. In Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15), Vol. 1. MIT Press, Cambridge, MA, USA, 1135–1143.
[11] Helbing, D., & Johansson, A. (2009). Social force model of pedestrian dynamics. In R. A. Meyers (Ed.), Encyclopedia of complexity and systems science (pp. 6476–6495). Springer. https://doi.org/10.1007/978-0-387-30440-3_382
[12] Kumar, R., Singh, S. P., & Gill, S. S. (2023). Smart cities and edge computing: Architecture, applications, and challenges. Journal of Cloud Computing, 12(1), 1–24. https://doi.org/10.1186/s13677-023-00569-6
[13] Lefèvre, S., Vasquez, D., & Laugier, C. (2011). A survey on motion prediction and risk assessment for intelligent vehicles. Robotics and Autonomous Systems, 59(9), 710–723. https://doi.org/10.1016/j.robot.2011.03.008.
[14] Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39. https://doi.org/10.1109/MC.2017.9
[15] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198
[16] Still, G. K. (2014). Introduction to crowd science. CRC Press.
[17] United Nations. (2024). World urbanization prospects: The 2024 revision. UN Department of Economic and Social Affairs. https://population.un.org
[18] World Health Organization. (2023). Global status report on road safety 2023. https://www.who.int
[19] Jiang, J., Yan, K., Xia, X., & Yang, B. (2025). A survey of deep learning based pedestrian trajectory prediction: Challenges and solutions. Sensors, 25(3), 957. https://doi.org/10.3390/s25030957.
[20] Gu, X., Li, C., Gao, L., & Niu, X. (2025). A review of pedestrian trajectory prediction methods based on deep learning technology. Sensors, 25(23), 7360. https://doi.org/10.3390/s25237360.
[21] Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, S. H., & Savarese, S. (2018). SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints. arXiv. https://arxiv.org/abs/1806.01482
[22] Melo Castillo, A. N., Salinas Maldonado, C., & Sotelo, M. A. (2025). Towards explainable pedestrian behavior prediction: A neuro-symbolic framework for autonomous driving. Applied Sciences, 15(6283). https://doi.org/10.3390/app15116283.

[23] Jinrui Geng, Yong Lu, Ruishi Liang, Jianlin Li, and Hannan Shen. 2025. Spatio-Temporal Graph Convolutional Networks Pedestrian Trajectory Prediction. In Neural Information Processing: 31st International Conference, ICONIP 2024, Auckland, New Zealand, December 2–6, 2024, Proceedings, Part V. Springer-Verlag, Berlin, Heidelberg, 271–285. https://doi.org/10.1007/978-981-96-6588-4_19.

[24] Cui, Z., Peng, W., Zhang, Y., Duan, Y., & Tao, X. (2024). Spatio-temporal-interaction graph neural networks for multi-agent trajectory prediction. Journal of Physics: Conference Series, 2833(1), 012010. https://doi.org/10.1088/1742-6596/2833/1/012010.

[25] Bansal, A., Agarwal, A., Lalit, M., Seeja, K.R. (2023). Survey of Pedestrian Trajectory Prediction Techniques Using Surveillance Videos. In: Chakraborty, B., Biswas, A., Chakrabarti, A. (eds) Advances in Data Science and Computing Technologies. ADSC 2022. Lecture Notes in Electrical Engineering, vol 1056. Springer, Singapore. https://doi.org/10.1007/978-981-99-3656-4_64.

[26] Nikhil, N., & Morris, B. T. (2018). Convolutional neural network for trajectory prediction. arXiv. https://arxiv.org/abs/1809.00696.

[27] Zhang, P., Ouyang, W., Zhang, P., Xue, J., & Zheng, N. (2019). SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction. arXiv. https://arxiv.org/abs/1903.02793.

[28] Pan, B., Cai, H., Huang, D.-A., Lee, K.-H., Gaidon, A., Adeli, E., & Niebles, J. C. (2020). Spatio-temporal graph for video captioning with knowledge distillation. arXiv. https://arxiv.org/abs/2003.13942.

[29] Cao, D., Li, J., Ma, H., & Tomizuka, M. (2021). Spectral Temporal Graph Neural Network for Trajectory Prediction. 2021 IEEE International Conference on Robotics and Automation (ICRA), 1839-1845.

[30] Yu, C., Ma, X., Ren, J., Zhao, H., & Yi, S. (2020). Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), Computer Vision – ECCV 2020 (pp. 507–523). Springer. https://doi.org/10.1007/978-3-030-58610-2_30.

[31] R. Korbmacher and A. Tordeux, "Review of Pedestrian Trajectory Prediction Methods: Comparing Deep Learning and Knowledge-Based Approaches," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 12, pp. 24126-24144, Dec. 2022, doi: 10.1109/TITS.2022.3205676.

[32] Bae, I., Oh, J., & Jeon, H.-G. (2023). Eigen Trajectory: Low-rank descriptors for multi-modal trajectory forecasting. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10017–10029 https://arxiv.org/abs/2307.09306.

[33] Saleh, K. (2022). Pedestrian trajectory prediction for real-time autonomous systems via context-augmented transformer networks. Sensors, 22(7495). https://doi.org/10.3390/s22197495.

[34] Han, X., & Xu, H. (2025). Causal intervention and counterfactual reasoning for multimodal pedestrian trajectory prediction. Journal of Imaging, 11(379). https://doi.org/10.3390/jimaging11110379.

[35] Neha Sharma, Chhavi Dhiman, S. Indu, Pedestrian Intention Prediction for Autonomous Vehicles: A Comprehensive Survey, Neurocomputing, Volume 508, 2022, Pages 120-152, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2022.07.085.

[36] Kothari, P., Kreiss, S., & Alahi, A. (2021). Human Trajectory Forecasting in Crowds: A Deep Learning Perspective. arXiv. https://arxiv.org/abs/2007.03639.

[37] Lerner, A., Chrysanthou, Y. and Lischinski, D. (2007), Crowds by Example. Computer Graphics Forum, 26: 655-664. https://doi.org/10.1111/j.1467-8659.2007.01089.x.

[38] Pellegrini, Stefano et al. "You'll never walk alone: Modeling social behavior for multi-target tracking." 2009 IEEE 12th International Conference on Computer Vision (2009): 261-268.

[39] Giuliari, F., Hasan, I., Cristani, M., & Galasso, F. (2020). Transformer Networks for Trajectory Forecasting. arXiv. https://arxiv.org/abs/2003.08111.

[40] Honghui Wang, Weiming Zhi, Gustavo Batista, Rohitash Chandra, Pedestrian trajectory prediction using goal-driven and dynamics-based deep learning framework, Expert Systems with Applications, Volume 271, 2025,126557, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2025.126557.

[41] Yang, Y., Wang, X., Song, M., Yuan, J., & Tao, D. (2021). SPAGAN: Shortest Path Graph Attention Network. arXiv. https://arxiv.org/abs/2101.03464.

[42] Nayakanti, N., Al-Rfou, R., Zhou, A., Goel, K., Refaat, K. S., & Sapp, B. (2022). Way former: Motion Forecasting via Simple & Efficient Attention Networks. arXiv. https://arxiv.org/abs/2207.05844.

[43] Shi, S., Jiang, L., Dai, D., & Schiele, B. (2023). Motion Transformer with Global Intention Localization and Local Movement Refinement. arXiv. https://arxiv.org/abs/2209.13508.

[44] Feng, L., Bahari, M., Amor, K. M. B., Zablocki, É., Cord, M., & Alahi, A. (2024). Uni Traj: A Unified Framework for Scalable Vehicle Trajectory Prediction. arXiv. https://arxiv.org/abs/2403.15098.

[45] Zuo, X., Mukai, M., & Kamal, M. (2025). Model predictive control of an autonomous wheelchair considering pedestrian reaction. SICE Journal of Control, Measurement, and System Integration.

[46] Zhou, Zeyu & Wang, Shanqing & Huang, Anmin & Lou, Jin & wei, tang & Navarro-Alarcon, David. (2025). KITNet: A Region Attention Activated Trajectory Predictor with Hierarchical Graph Neural Network in Dynamic-Mutant Multi-Agent System. IEEE Transactions on Intelligent Transportation Systems. PP. 1-17. 10.1109/TITS.2025.3635237.

[47] Wang, Zan and Wang, Zhiqiang and Niu, Fangqu and Wang, Xuanhui, SetTraj: Reformulating Multimodal Pedestrian Trajectory Prediction as End-to-End Dynamically Adaptive Set Prediction. Available at SSRN: https://ssrn.com/abstract=5854926.

[48] Zhenzhen He, Xiaorong Gan, Ziyang Chen, Yabo Wu, Yongjun Zhang, Wenting Li, Decoupled pedestrian trajectory prediction network with near-Aware attention, Knowledge-Based Systems, Volume 333, 2025, 114913, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2025.114913.

[49] Patel, V. A., Guo, Y., Park, L., & Obst, O. (2026). TAG: Temporal Attention Graph for heterogeneous traffic trajectory prediction. In T. T. Quan, H.-A. Pham, N. T. Tran, & C. Sombattheera (Eds.), Multi-disciplinary Trends in Artificial Intelligence: 18th International Conference, MIWAI 2025, Ho Chi Minh City, Vietnam, December 3-5, 2025, Proceedings, Part III (pp. 226-238). (Lecture Notes in Computer Science; Vol. 16355 LNAI). Springer. https://doi.org/10.1007/978-981-95-4963-4_19.

[50] Robicquet, Alexandre & Sadeghian, Amir & Alahi, Alexandre & Savarese, Silvio. (2016). Learning Social Etiquette: Human Trajectory Understanding in Crowded Scenes. 9912. 549-565. 10.1007/978-3-319-46484-8_33.

[51] Geiger, Andreas & Lenz, Philip & Urtasun, Raquel. (2012). Are we ready for autonomous driving? The KITTI Vision Benchmark Suite. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 3354-3361. 10.1109/CVPR.2012.6248074.