
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Munir, Farzeen; Kucner, Tomasz Piotr

Context-aware multi-task learning for pedestrian intent and trajectory prediction

Published in:
Transportation Research Part C: Emerging Technologies

DOI:
[10.1016/j.trc.2025.105203](https://doi.org/10.1016/j.trc.2025.105203)

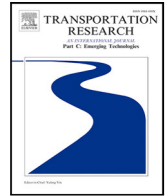
Published: 01/09/2025

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Munir, F., & Kucner, T. P. (2025). Context-aware multi-task learning for pedestrian intent and trajectory prediction. *Transportation Research Part C: Emerging Technologies*, 178, Article 105203.
<https://doi.org/10.1016/j.trc.2025.105203>

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.



Context-aware multi-task learning for pedestrian intent and trajectory prediction

Farzeen Munir ^{ID}*, Tomasz Piotr Kucner ^{ID}

Department of Electrical Engineering and Automation, Aalto University, Espoo, Finland
Finnish Center for Artificial Intelligence, Finland

ARTICLE INFO

Keywords:

Pedestrian trajectory prediction
Intention prediction
Autonomous vehicle
Deep learning

ABSTRACT

The advancement of socially-aware autonomous vehicles hinges on precise modeling of human behavior. Within this broad paradigm, the specific challenge lies in accurately predicting pedestrian's trajectory and intention. Traditional methodologies have leaned heavily on historical trajectory data, frequently overlooking vital contextual cues such as pedestrian-specific traits and environmental factors. Furthermore, there is a notable knowledge gap as trajectory and intention prediction have largely been approached as separate problems, despite their mutual dependence. To bridge this gap, we introduce PTINet (Pedestrian Trajectory and Intention Prediction Network), which jointly learns the trajectory and intention prediction by combining past trajectory observations, local contextual features (individual pedestrian behaviors), and global features (signs, markings etc.). The efficacy of our approach is evaluated on widely used public datasets: JAAD, PIE and TITAN, where it has demonstrated superior performance over existing state-of-the-art models in trajectory and intention prediction. The results from our experiments and ablation studies robustly validate PTINet's effectiveness in jointly exploring intention and trajectory prediction for pedestrian behavior modeling. The experimental evaluation indicates the advantage of using global and local contextual features for pedestrian trajectory and intention prediction. The effectiveness of PTINet in predicting pedestrian behavior paves the way for the development of automated systems capable of seamlessly interacting with pedestrians in urban settings <https://github.com/munirfarzeen/PTINet>.

1. Introduction

In the rapidly evolving field of *Autonomous Vehicles* (AV), social awareness has become a pivotal research focus (Maurer et al., 2016). Incorporating social awareness adds a layer of safety, predictability, and public trust to the operation of autonomous vehicles. However, the development of socially-aware autonomous vehicles presents significant challenge, primarily due to the complexity of modeling human behavior. One of the pertinent reason in modeling human behavior is the unpredictability of actions. To illustrate this unpredictable behavior, consider a scenario as illustrated in Fig. 1, where a young teenager stands in the middle of the road. The red points illustrate the future trajectory as the teenager decides to run to the other side of the road. Human drivers, from their past experiences, would anticipate erratic behavior and make efforts to discern the intentions based on facial expressions, gestures, age, and body language. However, it is challenging for AV to accurately perceive such behavior. This limits their reaction time, significantly increasing the risk of not stopping in time and potentially leading to accidents. Therefore, predicting pedestrian behavior efficiently and accurately is a crucial task for safe, safe human-AV interactions.

* Corresponding author at: Department of Electrical Engineering and Automation, Aalto University, Espoo, Finland.
E-mail address: farzeen.munir@aalto.fi (F. Munir).

<https://doi.org/10.1016/j.trc.2025.105203>

Received 4 December 2024; Received in revised form 3 April 2025; Accepted 21 May 2025

Available online 12 June 2025

0968-090X/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

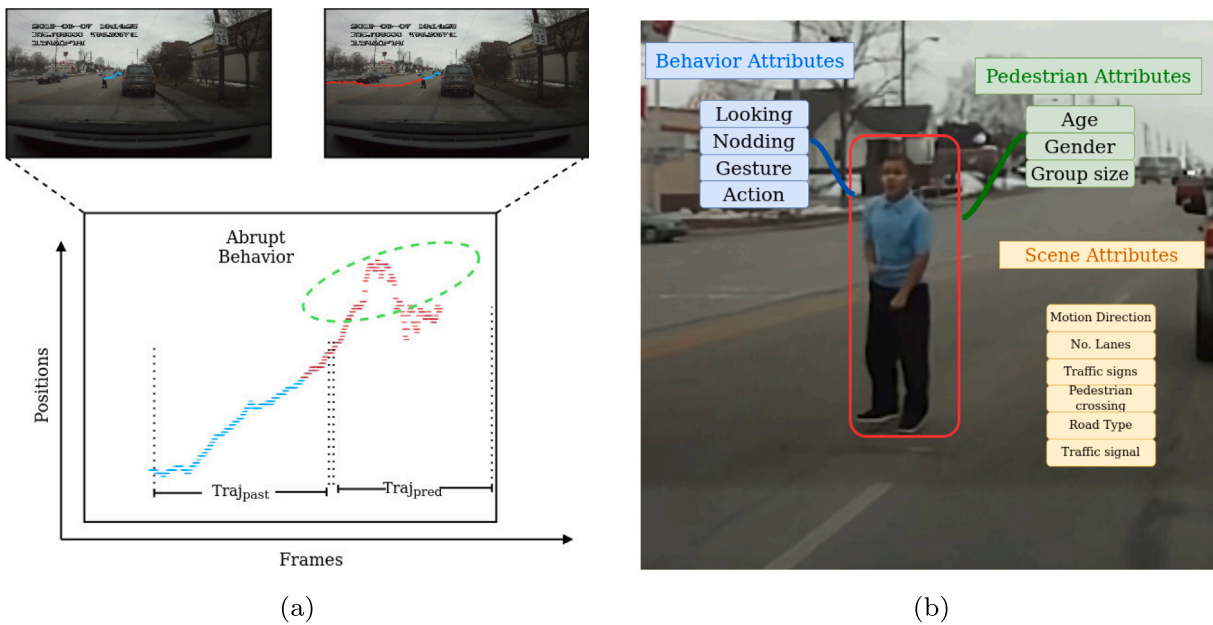


Fig. 1. The figure illustrates pedestrian behavior in urban scenarios during busy road crossings. (a) The trajectory plot shows blue points representing past trajectories and red points representing future trajectories, emphasizing sudden and erratic movements. Note that we only plotted the y-coordinate of the bounding box center to demonstrate the variation in the pedestrian's trajectory over time, with the x-axis representing the frame number (time). (b) highlights pedestrian attributes, behavior attributes, and scene attributes.

Predicting pedestrian behavior presents a significant challenge due to the lack of access to their complete internal state, necessitating the use of external cues. Predicting pedestrian behaviors is crucial for safety, especially when autonomous vehicles navigate in shared urban spaces (Fang and López, 2018). The pedestrian's behavior is affected by two factors as depicted in Fig. 1. The primary factor is the pedestrian's historical trajectory, which encapsulates their latent intent. The secondary factor concerns the environmental context, delineated by accessible and restricted areas (Sharma et al., 2022). A proficient model for predicting pedestrian behavior necessitates the integration of these two critical factors. In research methodologies, modeling pedestrian behaviors involves two approaches (Rudenko et al., 2020; Herman et al., 2021; Kotseruba et al., 2021) that are categorized into (i) **Intention Prediction** and (ii) **Trajectory Prediction**. In this work, we posed the research question that **predicting the former without considering the latter often limits understanding the pedestrian behavior in context to AV**.

Formally, we can define intention prediction as the process of anticipating the pedestrian's next move. In the case of urban settings, intention prediction has been mostly studied in literature by modeling factors, such as predicting intentions by analyzing historical data on position and contextual features (i.e., gait Mínguez et al., 2018, activity Lorenzo et al., 2020, and gestures Gesnoin, 2022). While these elements are instrumental in deciphering pedestrian intentions, they predominantly model intentions from the individual's perspective without adequately accounting for environmental influences. In the literature, environmental factors and local attributes are also studied in addition to the above pedestrian-centric factors to model pedestrian intention prediction (Crosato et al., 2023; Millard-Ball, 2018). The core research challenge lies in integrating contextual and environmental factors into a cohesive framework for predicting pedestrian intentions. As stated earlier, intention prediction is interlinked with trajectory prediction, formulating the former without incorporating the latter would not fully capture pedestrian's behavior, potentially resulting in unsafe human-AV interactions. Integrating both pedestrian's intention and trajectory predictions enhances the ability of AV systems to anticipate pedestrian movements more accurately, thereby directly improving safety mechanisms and reducing the likelihood of accidents. In existing studies (Salzmann et al., 2020; Mohamed et al., 2022; Kothari et al., 2021), pedestrian trajectory prediction often relies only on past movements, ignoring its interdependence with intention prediction. Moreover, these studies usually overlook important contextual and environmental factors that are key to understanding human behavior. Recent studies (Gupta et al., 2018; Kothari et al., 2021; Yuan et al., 2021; Rudenko et al., 2020) have developed robust algorithms for human trajectory prediction, focusing on pedestrian interactions with surroundings and past trajectory data. Others have incorporated scene information through scene graphs (Salzmann et al., 2020), obstacle maps (Zhang et al., 2023a), and heat maps (Mangalam et al., 2021) to predict feasible trajectories. However, they are deficient in accurately predicting pedestrian behavior because they do not consider specific AV features, pedestrian attributes, or traffic conditions. This limitation is particularly challenging in dynamic urban settings where various factors can cause pedestrian's behavior to change unpredictably.

Based on our underlying research question, in this work, we present a comprehensive framework for *Pedestrian Trajectory and Intention prediction Network* (PTINet) that considers pedestrian past trajectory, *Local Contextual Features* (LCF), and *Global Features* (GF) to predict both trajectory and intention simultaneously. Our proposed framework (PTINet) integrates past trajectories and

visual data gathered by the ego vehicle's field-of-view camera, contrary to bird-eye-view perspective of data. The LCF capture attributes specific to pedestrians, including their behavior and the surrounding scene characteristics. These features are represented as vectors, encompassing both pedestrian attributes like age and gender and their behaviors, such as gestures, gaze direction, movement, and nodding. Additionally, they include traffic-related information, for instance, pedestrian crossings, road types, traffic signs, number of lanes, and traffic signals. LCF enable the model to understand and represent pedestrian behavior, capturing their immediate interactions, which are essential for accurate trajectory and intention prediction. The GF which consists of image data and optical flow derived from consecutive frames, are integrated into the model. Introducing image and optical flow data is particularly advantageous as it endows the framework with a more comprehensive understanding of the environment. Image data offers rich visual information, while optical flow enables the model to discern the temporal evolution of visual cues. The synergistic integration of past trajectories, GF and LCF is instrumental in identifying complex spatial-temporal patterns, ultimately enhancing the robustness and accuracy of intention and trajectory predictions within the proposed multi-task framework.

The main contributions of our work are:

1. We have developed a novel multi-task framework, PTINet, that integrates LCF, represented by pedestrian-specific attributes, and GF, embodied by image data and optical flow. This approach enables accurate prediction of pedestrian behaviors by analyzing complex spatial-temporal patterns.
2. To learn the spatial and temporal representation, we integrate C-LSTM, LSTM-VAE, and MLP in a unified encoder network, followed by an LSTM-based intention and trajectory prediction decoder.
3. Our experimental analysis and ablation studies show the efficacy of the proposed PTINet on widely used benchmark datasets, outperforming the state-of-the-art methods.

2. Related work

2.1. Intention prediction

Intention prediction is crucial for facilitating interactions between AV and pedestrians (Liu et al., 2020; Chen, 2021; Fang and López, 2018; Rasouli et al., 2017; Yau et al., 2021; Lorenzo et al., 2020; Song et al., 2022), involving the prediction of pedestrians' future actions, such as the likelihood of crossing a road. This capability is vital for allowing AV to make timely safety-related decisions. Early approaches primarily focused on learning feature representations from static driving scenes (Kotseruba et al., 2016), later incorporating pedestrian pose estimation to enhance intention prediction (Fang and López, 2018). More recent studies have employed transformer-based architectures to extract temporal correlations from video data while modeling pedestrian uncertainty (Zhang et al., 2023b). Similarly, Achaji et al. (2022) proposed a framework leveraging multiple Transformer variants and demonstrated that bounding box information alone can effectively capture key spatial features for pedestrian intent prediction. In addition, Marchetti et al. (2024) introduced a semantic cross-modal attention mechanism to fuse visual, motion, and contextual cues, enabling robust prediction of pedestrian crossing behavior under STOP and GO scenarios. Furthermore, Song et al. (2022) presented a traffic-aware scene graph model that captures structured relationships among pedestrians, vehicles, and road elements, highlighting the importance of contextual understanding. Schörkhuber et al. (2022) further evaluated the impact of various input modalities—such as human poses, bounding boxes, ego-vehicle speed, and image-based features—on improving pedestrian crossing prediction performance. Our proposed method builds upon these advances by utilizing visual, motion, and contextual cues within a multitask learning framework to jointly predict pedestrian intention and trajectory. Additionally, we incorporate global features extracted from both images and optical flow—an aspect that has not been extensively explored in prior studies.

The most relevant state-of-the-art studies in pedestrian intent prediction closely aligned with our work include PIE-intent (Rasouli et al., 2019), FF-STA (Yang et al., 2022), TAMformer (Osman et al., 2023), PedFormer (Rasouli and Kotseruba, 2023), and BiPed (Rasouli et al., 2021). These studies are selected based on input modality, feature extraction methods, evaluation measures, and the benchmark datasets employed, that aligns with our pedestrian intention prediction settings. While serving as solid baseline approaches, these methods fall short in incorporating global context from the ego-vehicle perspective. For instance, FF-STA (Yang et al., 2022), PedFormer (Rasouli and Kotseruba, 2023), and BiPed (Rasouli et al., 2021) segment the environment to model global context, potentially overlooking environmental dynamics. Conversely, PIE-intent (Rasouli et al., 2019) and TAMformer (Osman et al., 2023) heavily rely on local environmental context. In this work, we have selected these state-of-the-art methods that reflect quantitative and qualitative comparison with our proposed method in terms of modeling local contextual and global feature for predicting the pedestrian intentions. In contrast to PTINet, some other works for instance, PIE-intent employs a convolutional LSTM network to encode past visual data, combined with bounding box information to predict a pedestrian's intention (Rasouli et al., 2019). Similarly, TAMformer uses features similar to PIE-intent but utilizes a transformer-based architecture for intention prediction (Osman et al., 2023). FF-STA extracts pedestrian appearance and context features using two separate CNNs and pre-computed pose data (Yang et al., 2022).

The two state-of-the-art methods PedFormer (Rasouli and Kotseruba, 2023) and BiPed (Rasouli et al., 2021), are closely aligned with our work that use LSTM-based network to predict both intention and trajectory using local images, past trajectories, and semantic segmentation maps. BiPed (Rasouli et al., 2021) and PedFormer (Rasouli and Kotseruba, 2023) rely on semantic segmentation maps to incorporate global context into their pedestrian behavior prediction models. These maps assign categorical labels (e.g., "road", "sidewalk", "vehicle") to each pixel in the input image at every timestep, providing a static, high-level representation of the scene. While their approach captures coarse environmental structure, it inherently struggles to model the

subtle, dynamic spatial–temporal interactions that are critical for understanding complex pedestrian behaviors in urban settings. For instance, consider a scenario where a pedestrian is standing near a crosswalk, glancing toward an approaching vehicle while shifting their body slightly forward. A segmentation map might accurately label the crosswalk, road, and vehicle, but it misses critical dynamic cues such as temporal motion patterns, fine-grained contextual interactions and pedestrian-specific dynamics. BiPed (Rasouli et al., 2021), for example, uses semantic segmentation at each timestep to model global context, yet its reliance on these maps limits its ability to capture the continuous flow of motion or subtle behavioral shifts. Similarly, PedFormer (Rasouli and Kotseruba, 2023) integrates segmentation maps alongside trajectory data, but the static nature of these maps restricts its sensitivity to real-time environmental dynamics, such as the optical flow of moving objects or pedestrians adjusting their pace in response to traffic. In contrast, PTINet integrates Global Features (GF) using image data and optical flow, alongside Local Context Features (LCF), to provide a more comprehensive and dynamic representation of the scene. Unlike segmentation maps, our approach captures both spatial and temporal features that are critical for modeling complex pedestrian behaviors. Additionally, our ablation study quantifies the impact of removing optical flow and image data, showing a significant performance drop without image data (due to loss of spatial context) and a milder drop without optical flow (due to reduced temporal sensitivity). This underscores the complementary roles of (GF) and (LCF) in PTINet, which go beyond the capabilities of segmentation-based methods.

2.2. Trajectory prediction

Pedestrian trajectory prediction involves forecasting the future position of pedestrians based on their current and past locations, behaviors, and the surrounding environment. Trajectory prediction algorithms often rely on *Bird-Eye-View* (BEV) data and operate from a top-down perspective, which simplifies the calculation of relative distances between objects (Alahi et al., 2016; Sadeghian et al., 2019; Salzmann et al., 2020). For instance, Social LSTM uses a specialized pooling module to consider the influence of other agents (Alahi et al., 2016). Other approaches like adversarial networks (Gupta et al., 2018) and MID algorithm (Mohamed et al., 2022) also focus on modeling interactions. Trajectron++ incorporates semantic maps and dynamic constraints (Salzmann et al., 2020), while Gu et al. (2022) employs a transformer-based model to capture temporal dependencies. Despite their advances, these methods generally rely on past trajectory data, limiting their accuracy in predicting complex human behavior, especially in the context of autonomous vehicles.

In contrast to BEV approaches, some algorithms use a first-person perspective, adding complexity due to the ego vehicle's motion (Rasouli et al., 2019; Cao and Fu, 2020). These methods mainly aim to predict pedestrian behavior by predicting trajectory. The trajectory prediction algorithm in this context uses diverse inputs like bounding boxes, ego-vehicle distance (Herman et al., 2021), and contextual information (Sui et al., 2021; Yau et al., 2021). Visual features and behavioral cues such as orientation and awareness level are also considered (Rasouli et al., 2017; Cao and Fu, 2020; Kooij et al., 2019). Despite integrating various features, the trajectory prediction algorithms show limited trajectory accuracy improvement on datasets like PIE (Rasouli et al., 2019; Cao and Fu, 2020). In contrast to the methods mentioned above (Rasouli et al., 2019; Cao and Fu, 2020), our approach incorporates LCF, such as gesture, walk direction, and head nodding of pedestrians, as well as their attributes to enhance the prediction of future trajectories in the image plane. Additionally, we integrate GF using image features and motion information from optical flow to improve overall scene understanding.

We propose that intention and trajectory prediction are interconnected aspects essential for accurately modeling pedestrian behavior from the perspective of an ego-vehicle. Addressing one without the other could result in an incomplete representation of pedestrian behavior, as both elements are crucial to understanding and predicting their actions in traffic scenarios.

3. Methodology

3.1. Problem formulation

This study proposes a multi-task learning framework (PTINet) to predict pedestrian trajectory and intention concurrently. In addition to pedestrian-centric features like key points, head orientations, and past trajectories, our approach extends its scope to include a more comprehensive set of features, especially GF and LCF, as shown by Fig. 2. By incorporating these additional features, our goal is to capture the intricacies of human behavior more comprehensively, ultimately enhancing predictions for both trajectory and intentions. The formulation of our framework is outlined as follows.

Given a video sequence \mathcal{V} of an urban scenario, we define a sequence of observed video frames as $\mathcal{V} = \{f_1, f_2, \dots, f_t\}$ where t represents discrete time steps corresponding to individual image frames (f_t). Our approach aims to estimate the probability of a pedestrian's intention to cross the street, represented as $I \in [0, 1]$, while concurrently predicting the pedestrian's future trajectory. The trajectory of the pedestrian is characterized by a sequence of bounding boxes $b_t = \{x_t, y_t, w_t, h_t\}$ where (x_t, y_t) is center coordinates, w_t is the width, and h_t is height in the t th image frame. At a given time step t , our framework predict the future trajectory, $\Omega_{f_t} = \{b_i^{t+1}, \dots, b_i^{t+n}\}$ and the future intention to cross, $\Psi_{f_t} = \{I_i^{t+1}, \dots, I_i^{t+n}\}$ for a pedestrian i over a prediction horizon of n time steps. This prediction is based on the pedestrian's past trajectory Ω_p , LCF, and GF observed over a horizon m . The pedestrian past trajectory Ω_p encompasses both positions $Pos_p = \{b_i^{t-m+1}, \dots, b_i^t\}$ and velocity $Vel_p = \{bv_i^{t-m+1}, \dots, bv_i^t\}$. The velocity Vel_p at t is then estimated as the change in position from the previous frame $t - 1$.

The LCF within this framework are categorized into pedestrian attributes, behavior, and scene attributes.

1. Pedestrian Attributes LCF_p : These attributes are denoted as $LCF_p = \{ap_1, \dots, ap_i\}$, where each ap_i is a vector representing demographic aspects such as age and gender, and group size for each pedestrian.

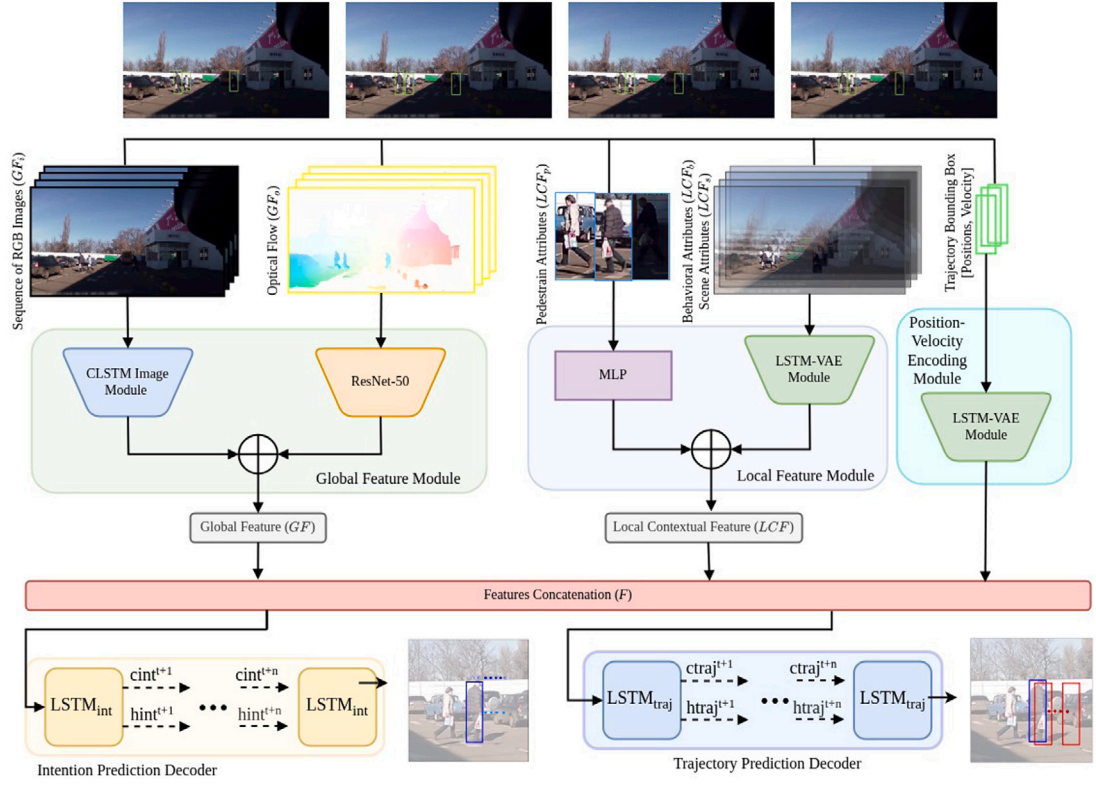


Fig. 2. The figure illustrates a context-aware multi-task learning framework for the prediction of pedestrian trajectories and intentions. The architecture comprises a Global Feature Module, which processes image and optical flow data utilizing a C-LSTM module and Resnet50, respectively. Concurrently, the Local Contextual Module takes in local contextual features and employs a combination of MLP and LSTM-VAE blocks for feature extraction. The Position-Velocity Encoding Module encodes past pedestrian trajectories. The outputs from these distinct modules are concatenated and fed into separate trajectory and intention decoders, facilitating subsequent predictions.

2. Behavior Attributes LCF_b : These attributes are articulated as $LCF_b = \{ab_i^{t-m+1}, \dots, ab_i^t\}$, each ab_i^t is a binary vector that consists of a range of non-verbal behavioral cues, such as looking, nodding, gesturing, and actions.
3. Scene Attributes LCF_s : These attributes, represented as $LCF_s = \{as_i^{t-m+1}, \dots, as_i^t\}$, comprise multidimensional vectors that intricately detail the environmental and infrastructural elements of the pedestrian's surroundings. Each vector as_i^t contains information on motion direction, number of lanes, traffic signs, pedestrian crossings, road types, and traffic signals.

Lastly, the framework incorporates GF , which include image data, denoted as GF_{img} , and optical flow, represented as GF_o . The image data is expressed as $GF_{img} = \{img^{t-m+1}, \dots, img^t\}$ capturing the visual context from a series of frames, while the optical flow is detailed as $GF_o = \{of^{t-m+1}, \dots, of^t\}$, quantifying the motion between these frames. The integration of optical flow is particularly important, as it enables the model to account for and adapt to the dynamic aspects of the environment. Optical flow offers an in-depth perspective on temporal variations and movements within the scene by analyzing motion patterns across sequential frames, which enhances the prediction of pedestrian behavior.

3.2. Architecture

The framework depicted in Fig. 2 illustrates an integrated approach to predicting pedestrian trajectory and intention, using a combination of sequential image data, optical flow, and dynamic pedestrian attributes. The methodology adopts an encoder–decoder architecture, with each encoder module responsible for encoding the pedestrian past trajectory, LCF, and GF respectively.

3.2.1. Position-velocity encoding module

A Long Short-Term Memory Variational Autoencoder (LSTM-VAE) block, as illustrated in Fig. 4, is employed for encoding the pedestrian trajectory consisting of pedestrian position Pos_p and velocity Vel_p (Hsu et al., 2017). LSTM-VAE for trajectory encoding is an optimal choice, as it effectively captures long-term dependencies and leverages the sequential data handling capability essential for maintaining temporal coherence in trajectory forecasting. Furthermore, the LSTM's ability to handle sequential data, combined with the generative modeling capabilities of Variational Autoencoders (VAE), provides a comprehensive approach for accurately capturing the probabilistic nature of pedestrian movements and intentions. This block serves as a sequence-to-sequence autoencoder,

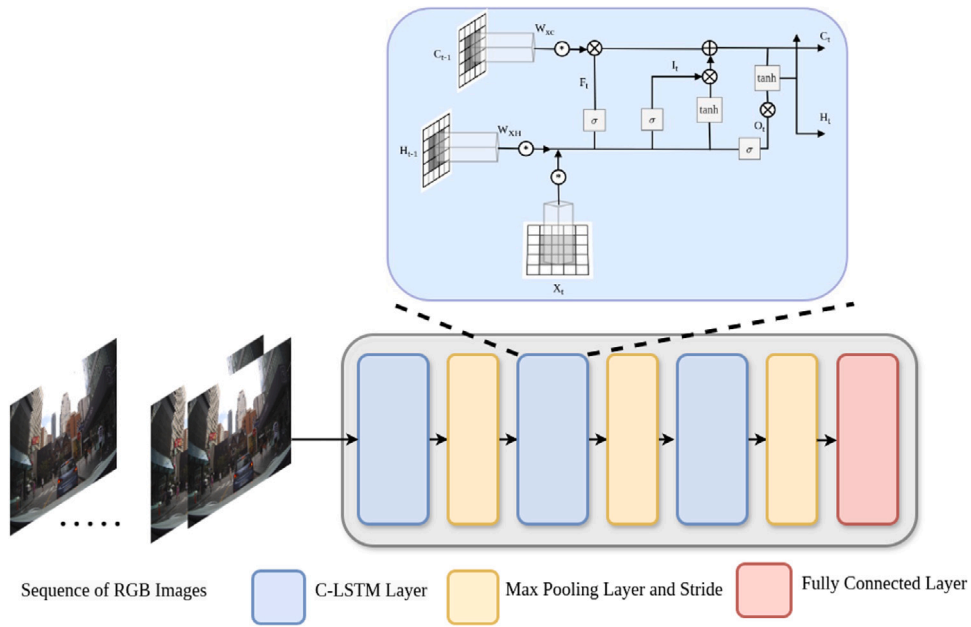


Fig. 3. The C-LSTM module is shown, designed to process input images and generate GF. The detailed framework of the module comprises three layers of C-LSTM, each followed by max pooling. The last max pooling layer is followed by a fully connected layer.

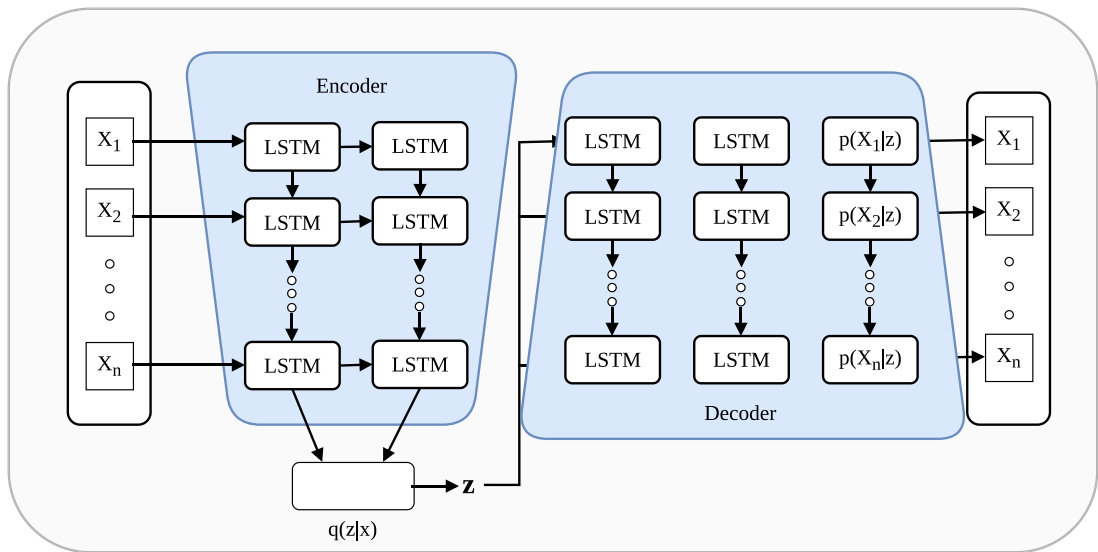


Fig. 4. Illustrates the architecture of the LSTM-VAE module employed in PTINet, which is utilized for learning the LCF and for capturing the temporal representation of past trajectories.

encoding a sequence of input vectors into a latent space and then decoding from a sampled latent variable back to an input sequence. VAE are utilized to learn the generative process of pedestrian trajectories, while the LSTM-VAE block models the temporal relationships. For the LSTM-VAE encoder, both the conditional distribution $p_{\theta}(x|z)$ and the approximate posterior distribution $q_{\phi}(z|x)$ are modeled as diagonal Gaussian distributions, as given by Eqs. (1) and (2), respectively.

$$p_{\theta}(x|z) = \mathcal{N}(z : s_{\mu_x}(x; \theta), \exp(s_{\log\sigma_x^2}(x; \theta))) \tag{1}$$

$$q_{\phi}(z|x) = \mathcal{N}(z : g_{\mu_x}(z; \phi), \exp(g_{\log\sigma_x^2}(z; \phi))) \tag{2}$$

Here, $s_{\mu_x}(x; \theta)$ and $g_{\mu_x}(z; \phi)$ represent mean and $s_{\log\sigma_x^2}(x; \theta)$ and $g_{\log\sigma_x^2}(z; \phi)$ represent the log variance, and are estimated by a neural network. The prior is set as a centered isotropic multivariate Gaussian $p_{\theta}(z) = \mathcal{N}(z; 0, I)$ with 64 dimensions. The LSTM-VAE

encoder consists of a two-layer LSTM with 512 hidden units to process the feature vector. Its outputs are merged and sent to a Gaussian layer to estimate the latent variable (z) mean and log variance. The reparameterization trick is applied to rewrite the latent variable as $z = s_{\mu_z}(x; \theta) + \sqrt{\exp(s_{\log\sigma_z^2}(x; \theta))} \odot \epsilon$ where, \odot denotes element-wise multiplication, and ϵ is sampled from $\mathcal{N}(z; 0, I)$. The decoder of LSTM-VAE features a two-layer LSTM with 512 hidden units and takes the sampled latent variable to generate a sequence. Each generated output is then used as input for a Gaussian parameter layer, which predicts the mean and log variance for a single timestep of the input feature.

3.2.2. Global feature module

This encoder integrates comprehensive global scene dynamics through image data and optical flow, capturing the dynamic changes and interactions that influence pedestrian movement. The image sequence undergoes processing via a *Convolutional Long Short-Term Memory Network* (C-LSTM) block, as depicted in Fig. 3. This block comprises three C-LSTM cells (Shi et al., 2015). Each C-LSTM cell is followed by a max-pooling layer, with the exception of the last cell, which is succeeded by a fully connected layer instead. Convolutional kernels for each layer have a size of 5×5 and a stride of 2×2 , with 32 filters. This block is especially well-suited for image sequences, as it is designed to learn both spatial and temporal dependencies concurrently. C-LSTM cells maintain a continuously updated hidden state as they process the input sequence, enabling them to model non-linear temporal transitions effectively. Additionally, optical flow data are encoded using a ResNet-50 backbone, and the features extracted are amalgamated to form the GF.

3.2.3. Local contextual feature

It processes attributes directly related to pedestrians, such as demographic information, behavioral cues, and immediate environmental context, contributing to a holistic understanding of pedestrian behavior. Given the heterogeneous nature of the data, each attribute category is encoded distinctly to preserve its unique characteristics. Time-invariant pedestrian attributes are encoded utilizing a 64-layer Multi-Layer Perceptron (MLP) network optimized for static data representation. Conversely, pedestrian behavior attributes and scene attributes, which exhibit temporal variability, are processed using an LSTM-VAE block identical to the one previously described. This LSTM-VAE block models temporal dependencies and encodes them into a latent space. Such a space is probabilistically formulated to reflect the complex nature of pedestrian behavior and environmental factors, thus yielding a dense and informative representation of the active dynamics.

The encoded features \mathbf{F} , as shown in Fig. 2 from these modules (LCF and GF) are then fed into the decoder, which consists of the Trajectory Prediction Decoder and the Intention Prediction Decoder, which leverage the temporal and spatial context to predict future pedestrian trajectories and intentions.

Trajectory prediction decoder

The trajectory decoder is designed to forecast pedestrian trajectories over a given timestep n . We opted for the LSTM because its inherent capacity to maintain long-term dependencies makes it ideally suited for the temporal precision required in trajectory forecasting. An initial hidden state h_{traj}^t is supplied to the trajectory decoder, $LSTM_{traj}$, which is the final concatenated feature vector \mathbf{F} from the encoder module, as shown in Fig. 2. This decoder takes the last observed position Pos_b^t as its input and subsequently produces the next predicted position for the bounding box, expressed as $Pos_b^{t+1} = (x^{t+1}, y^{t+1}, w^{t+1}, h^{t+1})$. The initial prediction is formulated through Eq. (3):

$$h_{traj}^{t+1} = LSTM_{traj}(h_{traj}^t, Pos_b^t, \mathcal{W}_{traj}) \quad (3)$$

The predicted hidden state h_{traj}^{t+1} is then passed through a fully connected layer to calculate the output velocity, as described by Eq. (4):

$$Pos_b^{t+1} = \mathcal{W}_o h_{traj}^{t+1} + bias_o \quad (4)$$

Here, \mathcal{W}_{traj} represents the weight matrix of the trajectory decoder, \mathcal{W}_o is the weight matrix for the output layer, and $bias_o$ signifies its associated bias vector. Subsequent trajectory predictions are computed iteratively for a horizon n . In each iteration, the hidden state is updated, and the most recently predicted trajectory is provided as input to the decoder.

Pedestrian intention decoder

Like the trajectory decoder, the intention decoder also employs LSTM network to process the encoded features from the previous modules, generating future intention predictions. The intention decoder is initiated with a combined feature set $hint^t = \mathbf{F}$ as its initial hidden state. It also takes the last observed position of the bounding box, denoted as $Pos_b^t = (x^t, y^t, w^t, h^t)$, as input. The decoder then outputs the subsequent predicted state of the pedestrian, I^{t+1} , as specified by Eq. (5).

$$\begin{aligned} hint^{t+1} &= LSTM_{int}(hint^t, Pos_b^t, \mathcal{W}_{int}), \\ I^{t+1} &= \mathcal{W}_{oi} hint^{t+1} + bias_{oi} \end{aligned} \quad (5)$$

In this context, $LSTM_{int}$ represents the intention decoder, \mathcal{W}_{int} is its weight matrix, \mathcal{W}_{oi} is the weight matrix of the output layer, and $bias_{oi}$ is the associated bias vector. Subsequent pedestrian intentions for future timesteps n are computed iteratively, with the hidden state being updated in each iteration. Finally, the output intentions are subjected to a softmax activation layer to calculate the probabilities associated with each potential outcome.

3.3. Loss functions

The proposed method loss includes two components: trajectory bounding box prediction loss ($\mathcal{L}_{\text{traj}}$), and intention prediction loss (\mathcal{L}_{int}). The ($\mathcal{L}_{\text{traj}}$) is formulated as a combination of the reconstruction loss and the Kullback–Leibler (KL) divergence, which encourages the learned latent space to adhere to a predefined Gaussian distribution. Specifically, the reconstruction loss quantifies the discrepancy between the predicted bounding boxes and the ground truth, facilitating the model's ability to predict future states accurately. The KL divergence serves as a regularization term, ensuring that the distribution of the latent variables does not deviate significantly from the prior distribution. Mathematically, the trajectory bounding box prediction loss ($\mathcal{L}_{\text{traj}}$) is expressed as:

$$\mathcal{L}_{\text{traj}} = \sum_{i=1}^T \left[\beta \cdot D_{KL}(q_{\phi}(z_t|x_t) \parallel p_{\theta}(z_t)) + \text{RMSE}(b_t, \hat{b}_t) \right] \quad (6)$$

where $D_{KL}(q_{\phi}(z_t|x_t) \parallel p_{\theta}(z_t))$ denotes the KL divergence between the approximate posterior distribution $q_{\phi}(z_t|x_t)$, parameterized by ϕ , and the prior distribution $p_{\theta}(z_t)$, parameterized by θ . The root mean squared error (RMSE) between the actual trajectory points b_t and the predicted trajectory points \hat{b}_t is used to measure the reconstruction error over timesteps n for N training samples, as shown in Eq. (7). The parameter β balances the influence of the KL divergence, allowing for control over the degree of regularization imposed on the latent space.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^n \|b_{t,i}^j - \hat{b}_{t,i}^j\|} \quad (7)$$

For the task of intention prediction loss \mathcal{L}_{int} , the binary cross-entropy (BCE) is employed, given by Eq. (8). This choice is particularly well-suited for problems where the output can be categorized into one of two classes, such as predicting a pedestrian's intent to cross the street or not. The BCE loss function measures the divergence between the predicted probabilities, denoted as \hat{I} , and the actual ground truth labels, I , which are either 0 or 1.

$$\mathcal{L}_{\text{int}} = -\frac{1}{N} \sum_{i=1}^N (I_i \log(\hat{I}_i) + (1 - I_i) \log(1 - \hat{I}_i)) \quad (8)$$

The complete loss function for the proposed method is the sum of the trajectory bounding box prediction loss and the intention prediction loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{traj}} \cdot \mathcal{L}_{\text{traj}} + \lambda_{\text{int}} \cdot \mathcal{L}_{\text{int}}, \quad (9)$$

where, λ_{traj} , and λ_{int} are weighting parameters that balance the contributions of the trajectory bounding box prediction loss and the intention prediction loss, respectively. In our experiments, setting $\lambda_{\text{traj}} = 1$ and $\lambda_{\text{int}} = 1$, gives the better results. This composite loss function is instrumental in concurrently optimizing pedestrian trajectory and intention prediction, which is vital for navigating dynamic and intricate environments.

4. Experimentation and results

4.1. Datasets

The effectiveness of the method is evaluated using two specialized datasets for pedestrian behavior prediction in moving vehicles: the Pedestrian Intent Estimation (PIE) dataset (Rasouli et al., 2019) and the Joint Attention in Autonomous Driving (JAAD) dataset (Kotseruba et al., 2016). The JAAD dataset consists of 346 high-resolution clips from 240 h of driving footage, annotated at a 30 Hz frame rate and focusing on 686 pedestrians with pedestrian behavior annotations. These pedestrians are further divided into training, validation, and testing subsets, containing 188, 32, and 126 individuals, respectively. It provides comprehensive annotations, including pedestrian behaviors, poses, and scene-specific details like traffic signs. The PIE dataset is captured at a resolution of 1920×1080 pixels and a frame rate of 30 fps. It comprises over six hours of driving footage and features 1842 annotated pedestrians. These are allocated across training, validation, and test sets, with counts of 880, 243, and 719 individuals, respectively. The PIE dataset includes not only annotations specific to pedestrians but also spatial metadata for other significant elements in the scene, such as traffic infrastructure and interacting vehicles. In both datasets, We have adopted the standard split as provided in the dataset.

To evaluate the generalization efficacy of our proposed approach, we conducted a case study on the TITAN dataset (Malla et al., 2020). This dataset consists of 700 video sequences captured via the front-view camera of a vehicle and provides bounding box annotations for 8592 distinct pedestrians, supplemented with contextual labels that characterize pedestrian attributes and behavioral patterns. It is imperative to note, that the dataset does not incorporate annotations for scene attributes, marking a limitation in the context of environmental feature analysis. In this case, the LCF consists of only pedestrian attributes and behavior. The dataset standard split as specified in Malla et al. (2020) is used in our experimental design, allocating 400 video sequences for training, 200 for validation, and the remaining 100 for testing purposes.



Fig. 5. The figure presents the qualitative results of the proposed framework on the JAAD, PIE, and Titan datasets. Red bounding boxes indicate predictions at the current timestamp, while white bounding boxes represent ground truth values. Dotted lines illustrate predicted trajectories over a 0.5 s time horizon, with blue indicating ground truth and red showing predicted values. The bar graph displays the pedestrian's intentions, providing a comprehensive view of the model's performance.

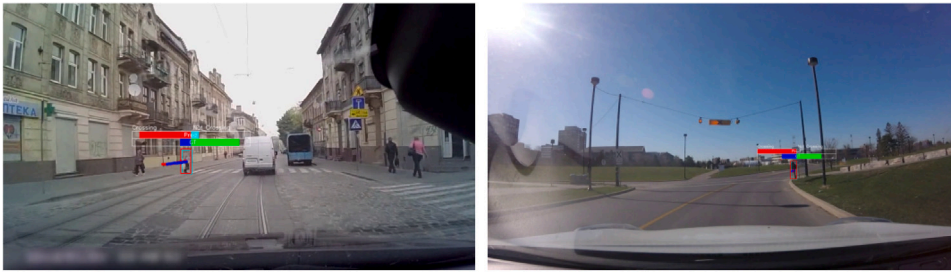
4.2. Training details

The PTNet framework is trained on a GPU server utilizing the PyTorch library, and the network undergoes end-to-end training from scratch. An input horizon of $m = 16$ timesteps, corresponding to 0.5 s, is considered, along with output horizons n of 0.5, 1, and 1.5 s. Image data is resized to dimensions [240, 420], and no other preprocessing or filtering is applied to the input images. The optical flow is estimated using the PyTorch toolkit MMflow (Contributors, 2021), which encompasses a variety of state-of-the-art methodologies. After extensive experimentation and comparative analysis, we selected the Recurrent All Pairs Field Transforms for Optical Flow (RAFT) (Teed and Deng, 2020) method due to its superior performance in capturing detailed motion patterns. The optical flow was computed between consecutive images. The optical flow was also resized to dimensions [240, 420], and no other preprocessing was performed. Pedestrian attributes, scene attributes, and pedestrian behavior data are employed in a categorical format. Training optimization is performed using the Adam optimizer, following the learning rate schedule as specified by $l_r = l_r^{int} \times (\frac{1-epoch}{max-epoch})^p$. Initial learning rate parameters l_r^{int} are set at 0.0001, with epsilon and weight decay values configured at 1^{-9} and 1^{-4} , respectively. The power p during the training phase is set at 0.9. Training proceeds for 200 epochs with a batch size of 4.

4.3. Evaluation metrics

To thoroughly evaluate the proposed methods, two distinct sets of metrics are applied, each tailored to specific aspects of the predictions. For trajectory prediction, *Average Displacement Error* (ADE) and *Final Displacement Error* (FDE) are employed, both calculated based on the vertices of the bounding box position. ADE measures the average Euclidean distance between the predicted and actual bounding box coordinates over a sequence of n time steps. FDE, on the other hand, focuses only on the position in the final time step. All metrics are reported in pixels.

For intention prediction, the F1 score and accuracy serve as the evaluation metrics, gauging the network's ability to identify pedestrian intentions correctly. The F1 score is calculated as the harmonic mean of precision and recall. The accuracy is a measure of the number of correct predictions out of the total number of instances. These metrics together offer a comprehensive evaluation of both the trajectory and intention prediction aspects of the network.



(a) JAAD Dataset



(b) PIE Dataset



(c) TITAN Dataset

Fig. 6. Failure cases.

Table 1

Quantitative evaluation of the proposed method and state-of-the-art approaches on JAAD and PIE datasets for pedestrian intention prediction, focusing on F1-score and accuracy metrics.

Methods	JAAD		PIE	
	F1-score	Accuracy	F1-score	Accuracy
PCPA (Osman et al., 2023)	0.67	0.56	0.77	0.86
R-LSTM (Osman et al., 2023)	0.74	0.65	0.52	0.76
RU-LSTM (Osman et al., 2023)	0.78	0.69	0.77	0.87
PIE-Intent (Rasouli et al., 2019)	–	–	0.87	0.79
TAMformer (Osman et al., 2023)	0.8	0.73	0.79	0.88
FF-STA (Yang et al., 2022)	0.74	0.62	–	–
BiPed (Rasouli et al., 2021)	0.6	0.83	0.85	0.91
PedFormer (Rasouli and Kotseruba, 2023)	0.54	0.93	0.87	0.93
PTINet	0.92	0.96	0.96	0.98

4.4. Results

This section presents the evaluation results of the proposed context-aware multi-task learning framework on two publicly available datasets: JAAD and PIE. Fig. 5 provides qualitative data to elucidate the performance of our proposed framework on the JAAD and PIE datasets. The bounding box in the figure indicates the current location of the pedestrian, while dotted lines represent predicted future trajectories. Bars in the figure indicate the pedestrian's intention, whether it is to cross or not to cross, over the considered time horizon. Fig. 6 shows some of the cases in which our algorithm fails to predict correctly. The Table 2, shows the quantitative evaluation of the PTINet with other state-of-art-methods. The results demonstrate that our framework outperforms

Table 2

Quantitative evaluation of the proposed method and state-of-the-art approaches on JAAD and PIE datasets for pedestrian trajectory prediction, focusing on ADE and FDE metrics across time horizons of 0.5 s, 1 s, and 1.5 s.

Methods	JAAD						PIE					
	ADE			FDE			ADE			FDE		
	0.5 s	1.0 s	1.5 s	0.5 s	1.0 s	1.5 s	0.5 s	1.0 s	1.5 s	0.5 s	1.0 s	1.5 s
Linear (Rasouli et al., 2019)	233	857	2303	–	–	6111	123	477	1365	–	–	3983
LSTM (Rasouli et al., 2019)	289	569	1558	–	–	5766	172	330	911	–	–	3352
B-LSTM (Rasouli et al., 2019)	159	539	1535	–	–	5615	101	296	855	–	–	3259
PIE-intent (Rasouli et al., 2019)	110	399	1248	–	–	4780	58	200	636	–	–	2477
Bi-Trap-D (Yao et al., 2021)	93	378	1206	–	–	4565	41	161	511	–	–	1949
Bi-Trap-NP (Yao et al., 2021)	38	94	222	–	–	565	23	48	102	–	–	261
Bi-Trap-GMM (Yao et al., 2021)	153	250	585	–	–	998	38	90	209	–	–	368
SGNet (Wang et al., 2022)	82	328	1049	–	–	4076	34	133	442	–	–	1761
BiPed (Rasouli et al., 2021)	–	27.98	–	–	55.07	–	–	19.62	–	–	39.12	–
PedFormer (Rasouli and Kotseruba, 2023)	–	24.56	–	–	48.82	–	–	15.27	–	–	32.79	–
PTINet	11.25	22.26	42.05	20.60	46.30	98.23	4.24	9.49	16.94	9.01	23.15	49.03

Table 3

Comparison of model complexity and inference time.

Methods	# of Parameters (M)	Inference time (s)
PIE-Intent	0.62	11.3
Bi-Trap-NP	1.54	0.008
Ours	49.37	0.103

state-of-the-art algorithms across varying time horizons (0.5 s, 1 s, 1.5 s) for both ADE and FDE on the JAAD and PIE datasets. Specifically, our model PTINet showcases the lowest ADE and FDE values in all examined time frames. In our evaluation, we compare our method with simple and complex models and those using multi-task learning frameworks and incorporate social attributes. In the JAAD dataset, the proposed method achieves better ADE and FDE scores at time horizon (0.5 s, 1 s, 1.5 s), outperforming the simple Linear model by a margin of (95.1%, 97.4%, 98.2%) for time horizons (0.5 s, 1 s, 1.5 s), respectively, on ADE. Similarly, in the case of FDE, the proposed method outperforms the Linear model by a margin of 98.4% for a 1.5 s time horizon, showing that the Linear model fails to capture the complexities of pedestrian behavior. In addition, methods such as SGNet and Bi-Trap-D, which mainly depend on trajectory data for their predictions and lack the incorporation of social behavior or pedestrian-centric features, outperform the Linear model. However, compared to our proposed method, they show higher values for ADE and FDE. Similarly, when compared with the proposed method, methods like PedFormer and BiPed, which use semantic segmentation, ego-motion, and trajectory data, have high ADE and FDE scores. Specifically, we outperform BiPed and PedFormer by improving ADE by roughly 20.44% and 9.36% and FDE by 15.93% and 5.16%, respectively for 1 s time horizon. In the PIE dataset, the proposed method also outperforms the state-of-the-art methods, as shown in Table 2. The proposed method obtains the ADE score of 4.26, 9.49, and 16.94 for time horizons (0.5 s, 1 s, 1.5 s) respectively, whereas it achieves the FDE scores of 9.01, 23.15, and 49.025 for the specified time horizons, outperforming the ADE and FDE scores of state-of-the-art methods.

The Table 1 presents a quantitative evaluation of intention prediction algorithms on the JAAD and PIE datasets. The results indicate that our proposed framework achieves an F1-score of 0.92 and an accuracy of 0.96 on the JAAD dataset and an F1-score of 0.965 and an accuracy of 0.98 on the PIE dataset. In the JAAD dataset, TAMformer also shows promising results with an F1-score of 0.8 and an accuracy of 0.73. The model incorporates bounding boxes, poses, and local context and develops a transformer-based framework. Compared to TAMformer, our approach exhibits improvements of approximately 15% in F1-score and 31.5% in accuracy. PedFormer is another noteworthy algorithm. While it attains a high accuracy of 0.93 on the JAAD dataset, its F1-score is 0.54. This suggests that PedFormer may excel in some aspects of prediction, such as correctly identifying true positives and negatives, but may face challenges in minimizing false positives and negatives, which is a crucial factor for a balanced F1 score. For the PIE dataset, PedFormer and BiPed show robust performances. PedFormer has an F1-score of 0.87 and an accuracy of 0.93, while BiPed scores an F1-score of 0.85 and an accuracy of 0.91. Both algorithms benefit from multi-task learning and the mutual reinforcement of trajectory and intention prediction.

It is worth noting here that both JAAD and PIE datasets are collected in urban settings; however, JAAD is collected using three different cameras under diverse weather conditions—including snowy weather and strong sun glare—with most videos recorded during daytime and only a few at night. Its video clips are short (5 to 15 s) and contain fewer examples, which introduces variability in camera viewpoints and environmental challenges that can complicate scene interpretation. These factors inherently limit the variability and richness of pedestrian motion patterns that models can learn from, thus negatively impacting prediction accuracy in both trajectory and intention prediction. In contrast, PIE was recorded using single camera, under clear weather conditions during a single day, resulting in longer pedestrian tracks (10 min), and more number of examples. These characteristics simplify trajectory prediction tasks even for simpler methods such as Linear interpolation, which explains why even these simpler baseline models achieve better performance on PIE compared to JAAD. The deterioration of linear model performance on JAAD for longer-term predictions further underscores the complexity of human motion patterns present in this dataset, which cannot be adequately

Table 4

Quantitative evaluation of the proposed method and state-of-the-art approaches on TITAN datasets for pedestrian trajectory prediction, focusing on ADE and FDE metrics across time horizons of 0.5 s, 1 s, and 1.5 s.

Methods	TITAN						F1-score	Accuracy
	ADE			FDE				
	0.5 s	1.0 s	1.5 s	0.5 s	1.0 s	1.5 s		
Bitrap (Halawa et al., 2022)	194	352	658	–	–	989	–	–
ABC+ (Halawa et al., 2022)	165	302	575	–	–	843	–	–
PTINet	16.97	37.78	56.41	28.79	71.59	115.92	0.96	0.97

captured by simple linear interpolation methods. Consequently, both our proposed method and prior approaches consistently achieve better results on PIE due to its richer contextual information, longer pedestrian tracks, fewer occlusions, and more consistent environmental conditions.

The results suggest that including pedestrian past trajectory, LCF and GF provides a holistic and more comprehensive understanding of pedestrian behavior. Additionally, the utilization of multi-task learning appears to offer mutual benefits for both trajectory and intention prediction tasks. Table 3 shows the comparison of the number of parameters and inference speed of our proposed method along with that of state-of-the-art methods for which source code is publicly available.

4.5. Evaluation on TITAN dataset

The Table 4 presents a quantitative analysis of pedestrian trajectory prediction on the TITAN dataset, comparing our method with the latest state-of-the-art approaches. Our method shows a significant improvement in accuracy. The ADE is 18.4929 for a 0.5 s prediction and 57.20 at 1.5 s. The FDE is 29.6652 at 0.5 s and 116.2543 at 1.5 s, as compared to ABC+ (Action-based contrastive learning) (Halawa et al., 2022), which shows much higher ADE and FDE score across all time horizons. These results highlight our model's strong performance in making accurate predictions over time. Additionally, the F1-score and accuracy of our method are high, at 0.95 and 0.97, respectively, indicating the reliability of our model in predicting pedestrian trajectories and intentions accurately. The comparative analysis highlights the PTINet performance in terms of trajectory accuracy over multiple time horizons and the reliable prediction of pedestrian intentions.

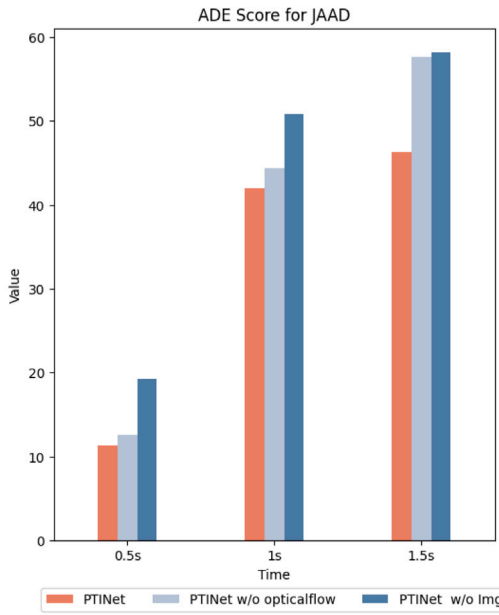
5. Ablation study

5.1. Effect of optical flow on ptinet

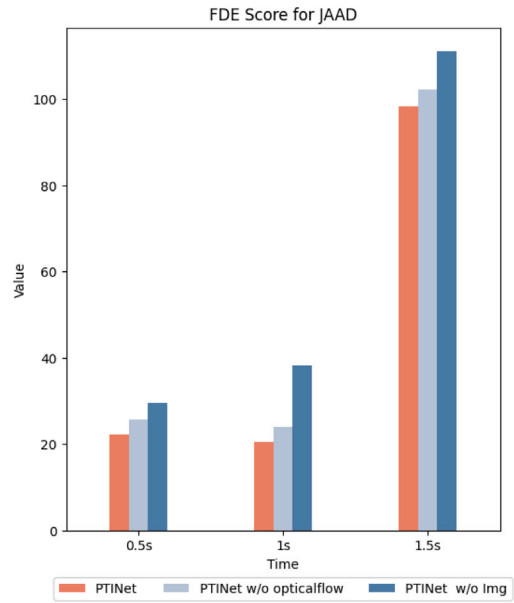
In our experiments, an ablation study was conducted to determine the influence of GF on the performance of the PTINet. This evaluation involved analyzing the PTINet performance in the absence of optical flow and, separately, without image data. The Figs. 7–9 illustrate the impact on ADE and FDE for trajectory prediction across the JAAD, PIE, and TITAN datasets and Fig. 10 shows effect on the F1-score and accuracy for intention prediction. The outcomes of this investigation show the significance of each feature towards enhancing prediction accuracy. The removal of optical flow resulted in a minor reduction in accuracy, highlighting its role in capturing the dynamic elements within scenes. While optical flow is beneficial for understanding movement patterns, its absence can be somewhat mitigated by other features within the model. On the other hand, the exclusion of image data led to a significant decline in the model's performance. This is attributed to the critical role that image data plays in providing comprehensive contextual insights into the environment, which are indispensable for accurately forecasting pedestrian trajectories and intentions. Image data delivers crucial spatial and contextual cues necessary for decoding complex scenarios and anticipating future movements with precision, whereas optical flow contributes important but comparatively less essential information regarding temporal dynamics. This analysis emphasizes the integral role of both optical flow and image data in the predictive accuracy of PTINet, with image data being especially crucial for maintaining model robustness and precision.

5.2. Effect of local contextual feature on ptinet

To further examine the contribution of individual components within our proposed LCF module, we conducted additional ablation studies to analyze the impact of LCF on the overall effectiveness of our method. The results of these experiments are summarized in Tables 5 and 6 for trajectory and intention prediction, respectively, evaluated across the JAAD, PIE, and TITAN datasets. Removing the entire Local Feature Module (GF w/o LCF) resulted in a substantial degradation in prediction performance across all datasets, highlighting the critical role of local contextual information in trajectory prediction tasks. Furthermore, excluding pedestrian attributes (LCF_p), such as age and gender, led to a notable decline in performance, indicating that these attributes significantly contribute to accurate trajectory forecasting. Similarly, omitting behavioral attributes (LCF_b), which capture non-verbal pedestrian cues like looking and nodding, negatively impacted prediction accuracy, although to a slightly lesser extent compared to pedestrian attributes. Additionally, removing scene attributes (LCF_s), encompassing environmental context such as lane numbers and traffic signs, also resulted in decreased performance, though with a comparatively smaller effect than observed with the exclusion of pedestrian or behavioral attributes. Collectively, these findings demonstrate that each component within our LCF provides valuable complementary information that is effectively integrated by the two LSTM networks ($LSTM_{int}$ and $LSTM_{iraj}$), thereby enhancing overall trajectory prediction accuracy across diverse datasets and prediction horizons.

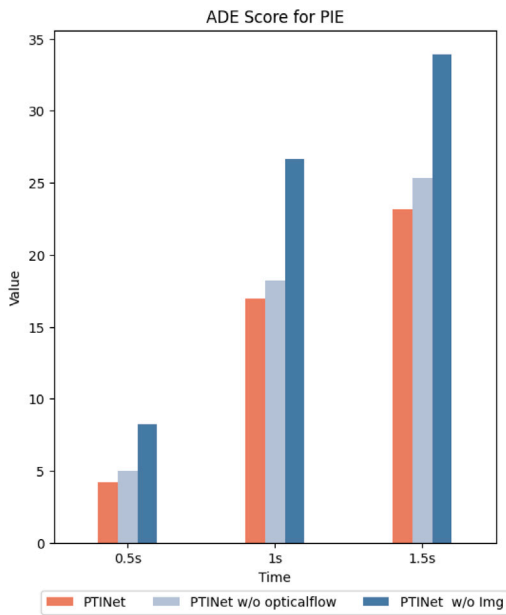


(a)

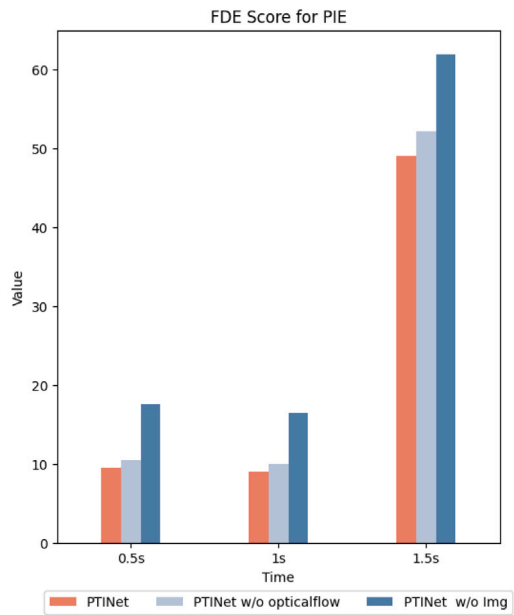


(b)

Fig. 7. ADE (a) and FDE (b) scores for the JAAD dataset, showing performance comparisons of PTINet, PTINet without image data, and PTINet without optical flow.



(a)



(b)

Fig. 8. ADE (a) and FDE (b) scores for the PIE dataset, showing performance comparisons of PTINet, PTINet without image data, and PTINet without optical flow.

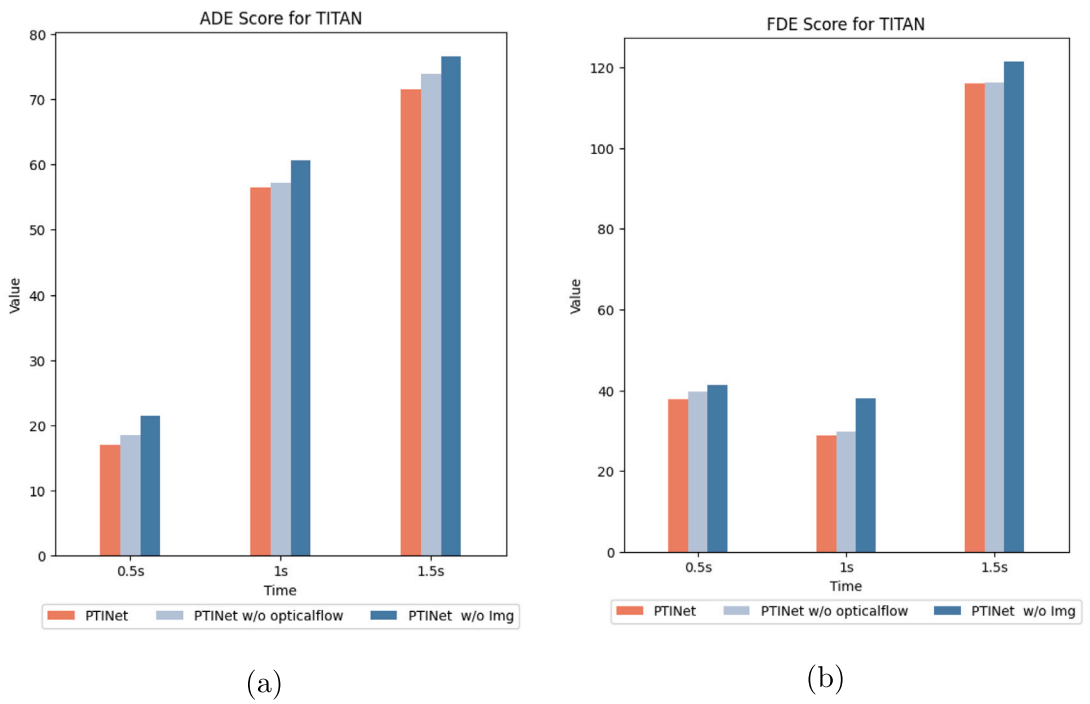


Fig. 9. ADE (a) and FDE (b) scores for the TITAN dataset, showing performance comparisons of PTINet, PTINet without image data, and PTINet without optical flow.

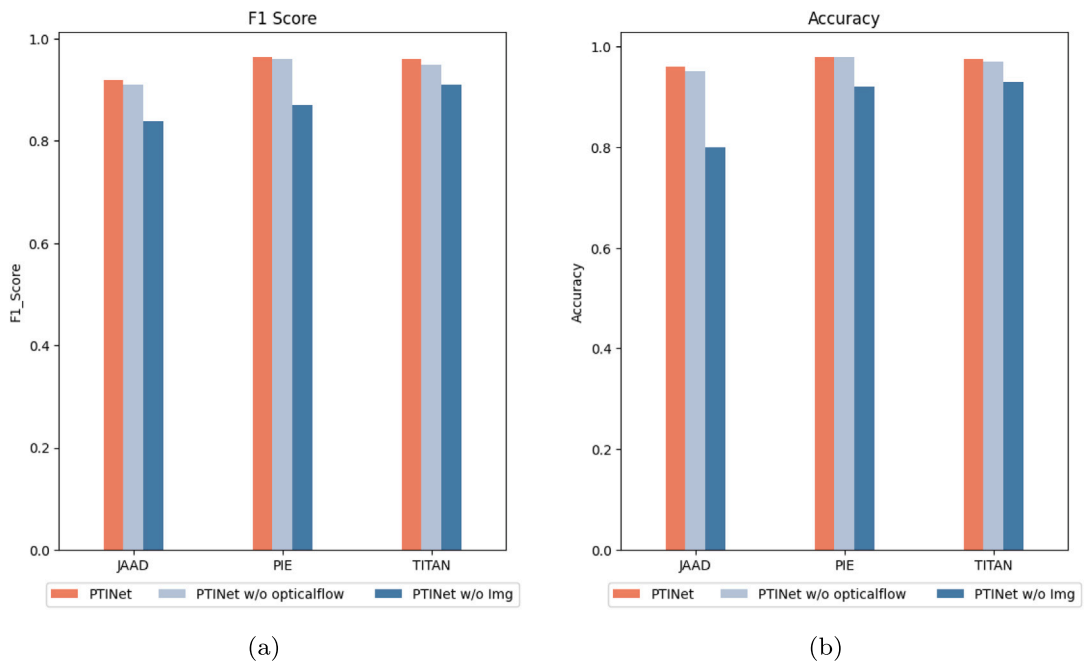


Fig. 10. The bar plots show the comparison of F1 score and accuracy for intention prediction for PTINet, PTINet without image data, and PTINet without optical flow on the JAAD, PIE, and TITAN datasets.

Table 5

Ablation study results evaluating the impact of different Local Feature Module components on model performance across JAAD, PIE, and TITAN datasets using ADE and FDE metrics at prediction horizons of 0.5 s, 1 s, and 1.5 s.

Methods	JAAD						PIE						TITAN					
	ADE			FDE			ADE			FDE			ADE			FDE		
	0.5 s	1 s	1.5 s	0.5 s	1 s	1.5 s	0.5 s	1 s	1.5 s	0.5 s	1 s	1.5 s	0.5 s	1 s	1.5 s	0.5 s	1 s	1.5 s
GF w/o LCF	14.27	28.56	51.43	23.89	57.64	103.15	9.78	15.35	24.42	16.65	33.21	68.47	21.34	49.59	71.24	39.11	87.34	135.86
GF & LCF w/o LCF _p	12.85	25.65	46.00	21.49	55.11	100.43	5.75	10.32	17.95	10.18	24.87	51.65	18.32	38.88	58.01	29.50	72.62	117.32
GF & LCF w/o LCF _b	13.12	26.01	46.50	22.11	56.87	101.32	6.15	10.51	18.13	11.32	25.21	52.25	19.21	39.76	58.31	30.03	73.14	118.76
GF & LCF w/o LCF _s	12.50	25.00	45.20	21.13	54.21	99.51	5.10	9.89	17.54	10.53	24.05	50.32	17.53	38.45	57.88	29.23	70.66	116.63
PTINet	11.25	22.26	45.05	20.60	46.30	98.23	4.24	9.49	16.94	9.012	23.15	49.02	16.97	37.78	56.41	28.79	71.59	115.92

Table 6

blation study results assessing the impact of different Local Feature Module components on intention prediction performance across the JAAD, PIE, and TITAN datasets using F1-score and accuracy metrics.

Methods	JAAD		PIE		TITAN	
	F1	ACC	F1	ACC	F1	ACC
Ours (GF w/o LCF)	0.83	0.94	0.88	0.92	0.87	0.89
Ours (GF & LCF w/o LCF _p)	0.85	0.91	0.89	0.96	0.92	0.94
Ours (GF & LCF w/o LCF _b)	0.86	0.88	0.91	0.94	0.91	0.96
Ours (GF & LCF w/o LCF _s)	0.89	0.93	0.94	0.97	0.95	0.97
PTINet	0.92	0.96	0.965	0.98	0.97	0.98

5.3. The impact of multi-task learning on pedestrian trajectory and intention prediction

We conducted an ablation study to investigate the performance of solving individual tasks compared to using a multi-task approach. Specifically, we performed two experiments: (1) disabling the intention prediction loss and focusing solely on trajectory prediction, and (2) disabling the trajectory prediction loss and focusing exclusively on intention prediction. In both cases, we retained our backbone model to encode local and global information. The results of our ablation study reveal that single-task models suffer notable performance degradation compared to PTINet’s multi-task approach. Fig. 11(a) shows the comparison for intention prediction, single-task performance shows a significant decline across all datasets. On JAAD, the F1 score drops by 10.87% and accuracy decreases by 9.38%. Similarly, on PIE, the F1 score and accuracy decrease by 5.98% and 5.1%, respectively. The TITAN dataset follows this trend, with a 5.15% reduction in the F1 score and a 3.06% drop in accuracy. This decline highlights the limitations of learning intention prediction in isolation, suggesting that integrating trajectory information improves the model’s ability to capture subtle pedestrian cues. Fig. 11(b) and Fig. 12 shows the comparison for trajectory prediction, single-task models exhibit increased errors across both ADE and FDE metrics. On JAAD, ADE increases by 14.2%, 3.6%, and 9.4% for 0.5 s, 1 s, and 1.5 s, respectively, while FDE rises by 4.3%, 3.4%, and 2.2% over the same horizons. On PIE, ADE increases sharply by 18.9%, 8.7%, and 5.95%, with FDE rising by 13.0%, 7.4%, and 5.35%. For TITAN, ADE increases by 7.9%, 2.9%, and 2.8%, accompanied by FDE increases of 2.4%, 1.42%, and 1.2% at 0.5 s, 1 s, and 1.5 s. These results emphasize that single-task trajectory models, lacking intent awareness, struggle to maintain accuracy — especially over longer prediction horizons where uncertainty accumulates. The consistent degradation in performance across both tasks supports our hypothesis that multi-task learning, leveraging the interplay between intention and trajectory prediction, yields more accurate and reliable pedestrian behavior modeling in dynamic urban environments.

6. Comparison with SOTA pedestrian trajectory prediction algorithms

To evaluate the performance of state-of-the-art (SOTA) pedestrian trajectory prediction algorithms on ego-centric datasets like JAAD and PIE, we conducted a sanity check using four prominent models: Social GAN, Trajectron++, MID, and Social Implicit. These models were selected based on their established performance on benchmark datasets such as ETH/UCY, a pedestrian trajectory prediction dataset from BEV perspective. As discussed in the related work section, these algorithms predict pedestrian trajectories using historical data but lack contextual information critical for dynamic environments. For pedestrian trajectory prediction in dynamic environments, contextual information is critical. Ego-centric datasets like JAAD and PIE include rich cues such as pedestrian behavior (e.g., gestures, gaze) and environmental factors (e.g., traffic signals), which are essential for accurate predictions. Our motivation for including this analysis is twofold: first, to highlight the limitations of existing BEV-based approaches when applied to ego-centric settings, and second, to establish a clear baseline for evaluating the effectiveness of our proposed model, PTINet, which explicitly addresses these challenges through its multi-task design and integration of local and global contextual information.

To address this, evaluations were conducted on the JAAD and PIE datasets using publicly available source code. The Table 7 showcases the trajectory prediction results over a time horizon of 0.5 s for both the JAAD and PIE datasets. Social Implicit (SI) and Social GAN emerged as top performers. Specifically, SI registered the lowest ADE and FDE values of 27.77 and 50.03, respectively, on the JAAD dataset, and 15.58 and 29.94 on the PIE dataset. Conversely, Trajectron++ and MID lagged in performance; Trajectron++ recorded an ADE of 206.66 and an FDE of 245.51 on JAAD, and an ADE of 115.32 and an FDE of 180.47 on PIE. Both SI and Social GAN model human behavior and incorporate important spatio-temporal variables, such as changes in human speed, the presence of

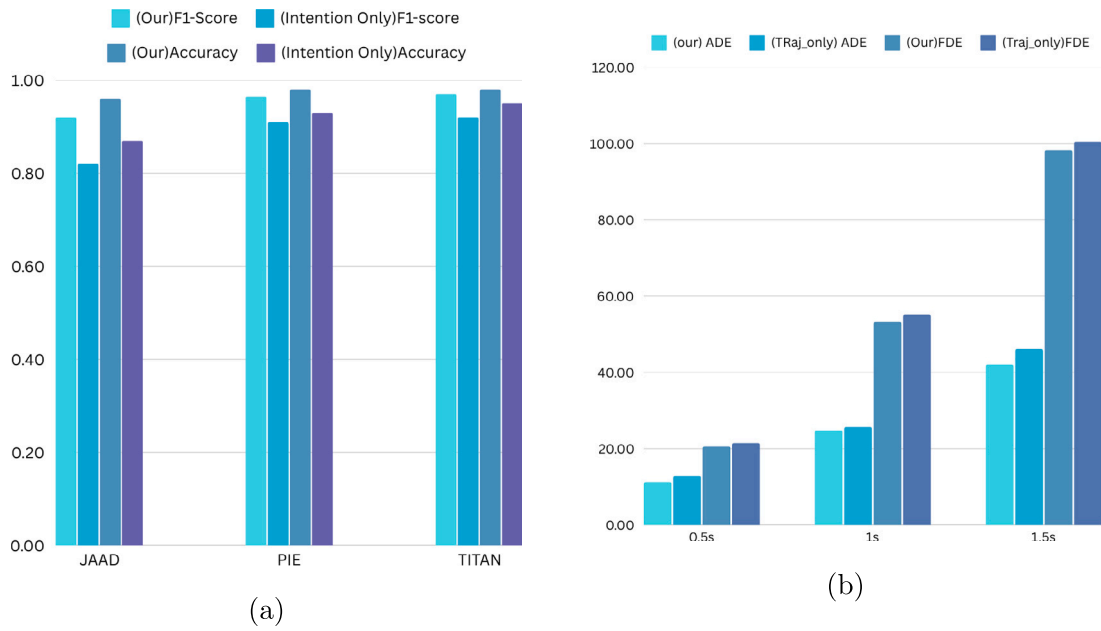


Fig. 11. (a) Comparison of F1 Score and Accuracy for Exclusive Intention Prediction. (b) Comparison of ADE and FDE for Exclusive Trajectory Prediction on JAAD dataset.

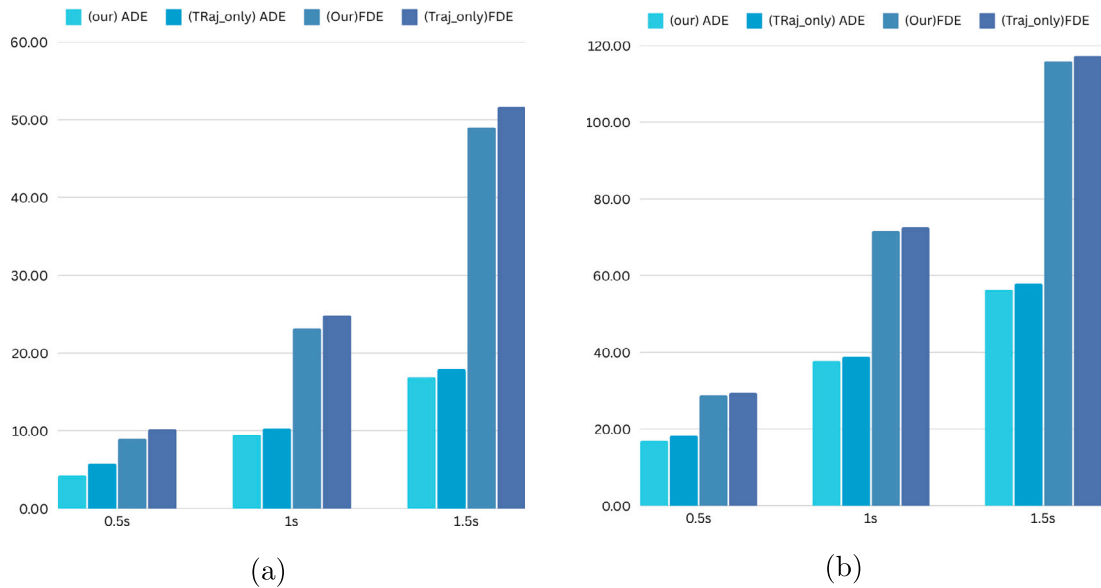


Fig. 12. Comparison of ADE and FDE for Exclusive Trajectory Prediction on (a) PIE dataset. (b) TITAN Dataset.

nearly pedestrians, and average walking speed. These factors could significantly affect the prediction of pedestrian trajectories. Conversely, Trajectron++ and MID primarily employ scene graph representations to model the spatio-temporal interactions of pedestrians with their environment. The lack of human behavioral considerations in the design of Trajectron++ and MID leads to their inferior performance on the JAAD and PIE datasets. However, it is worth noting that these results still lag behind those of Bifold and Pedformer, which utilize pedestrian-centric features to predict pedestrian trajectories.

The observed performance patterns on the JAAD and PIE datasets highlight the pressing need for the development of algorithms that focus specifically on pedestrian behavior in urban scenarios. Such algorithms would be essential for enhancing the safety measures implemented in autonomous vehicles.

Table 7
Comparison of pedestrian trajectory prediction results for state-of-the-art methods on JAAD and PIE datasets.

Methods	JAAD		PIE	
	ADE@0.5	FDE@0.5	ADE@0.5	FDE@0.5
Social GAN (Gupta et al., 2018)	29.25	52.86	17.25	31.15
Trajectron ++ (Salzmann et al., 2020)	206.66	245.51	115.32	180.47
Social Implicit (Mohamed et al., 2022)	27.77	50.03	15.58	29.94
MID (Gu et al., 2022)	151.87	180.45	95.48	121.62
PTINet	11.25	20.60	4.24	9.01

7. Conclusion

In this work, we introduce PTINet, a novel multi-task framework designed for pedestrian trajectory and intention prediction, aimed at comprehensively modeling pedestrian behavior in urban settings. Our approach integrates Local Contextual Features (LCF)—such as pedestrian attributes (e.g., age, gaze direction) and behavioral cues with Global Features (GF), including image data and optical flow, to achieve a holistic understanding of pedestrian dynamics. Extensive experimental evaluations and ablation studies validate our hypothesis that jointly modeling trajectory and intention prediction outperforms methods that treat these tasks in isolation. Specifically, our ablation studies (Section 6) reveal that models relying solely on trajectory data overlook critical human behaviors, whereas incorporating contextual information and pedestrian-specific cues markedly improves prediction accuracy. This comprehensive feature integration within a multi-task framework significantly boosts PTINet’s performance, as demonstrated by lower Average Displacement Error (ADE) and Final Displacement Error (FDE) scores on benchmark datasets like JAAD, PIE, and TITAN, surpassing state-of-the-art methods. These findings highlight the power of a context-aware, multi-task learning strategy for enhancing pedestrian behavior prediction in complex urban environments.

Despite its strong predictive capabilities, PTINet faces challenges, particularly in the explainability of its outputs. The reliance on LSTMs and feature concatenation obscures how specific inputs contribute to individual predictions, limiting transparency. To address this, future research could incorporate attention mechanisms or text-based tokens, drawing from advances in explainable AI for autonomous systems, to make the model’s decision-making process more interpretable. Moreover, while LSTMs have proven effective for our datasets, exploring advanced sequence modeling architectures like Mamba presents an exciting opportunity for further improvement.

CRedit authorship contribution statement

Farzeen Munir: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tomasz Piotr Kucner:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization.

Acknowledgment

This project has received funding from Finnish Center for Artificial Intelligence. We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC IT Center for Science, Finland, and the LUMI consortium through a EuroHPC Regular Access call. We thank CSC for providing computational resources.

References

- Achaji, L., Moreau, J., Fouquerey, T., Aioun, F., Charpillat, F., 2022. Is attention to bounding boxes all you need for pedestrian action prediction? In: 2022 IEEE Intelligent Vehicles Symposium. IV, IEEE, pp. 895–902.
- Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S., 2016. Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–971.
- Cao, D., Fu, Y., 2020. Using graph convolutional networks skeleton-based pedestrian intention estimation models for trajectory prediction. In: Journal of Physics: Conference Series. Vol. 1621, IOP Publishing, 012047.
- Chen, C., 2021. Modeling Spatiotemporal Pedestrian-Environment Interactions for Predicting Pedestrian Crossing Intention from the Ego-View (Ph.D. thesis). Purdue University.
- Contributors, M., 2021. MMFlow: OpenMMLab optical flow toolbox and benchmark. <https://github.com/open-mmlab/mmlflow>.
- Crosato, L., Tian, K., Shum, H.P., Ho, E.S., Wang, Y., Wei, C., 2023. Social interaction-aware dynamical models and decision-making for autonomous vehicles. Adv. Intell. Syst. 2300575.
- Fang, Z., López, A.M., 2018. Is the pedestrian going to cross? answering by 2d pose estimation. In: 2018 IEEE Intelligent Vehicles Symposium. IV, IEEE, pp. 1271–1276.
- Gesnouin, J., 2022. Analysis of Pedestrian Movements and Gestures Using an On-Board Camera to Predict Their Intentions (Ph.D. thesis). Université Paris sciences et lettres.
- Gu, T., Chen, G., Li, J., Lin, C., Rao, Y., Zhou, J., Lu, J., 2022. Stochastic trajectory prediction via motion indeterminacy diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17113–17122.
- Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A., 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2255–2264.

- Halawa, M., Hellwich, O., Bideau, P., 2022. Action-based contrastive learning for trajectory prediction. In: European Conference on Computer Vision. Springer, pp. 143–159.
- Herman, M., Wagner, J., Prabhakaran, V., Möser, N., Ziesche, H., Ahmed, W., Bürkle, L., Kloppenburg, E., Gläser, C., 2021. Pedestrian behavior prediction for automated driving: Requirements, metrics, and relevant features. *IEEE Trans. Intell. Transp. Syst.* 23 (9), 14922–14937.
- Hsu, W.-N., Zhang, Y., Glass, J., 2017. Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop. ASRU, IEEE, pp. 16–23.
- Kooij, J.F., Flohr, F., Pool, E.A., Gavrilu, D.M., 2019. Context-based path prediction for targets with switching dynamics. *Int. J. Comput. Vis.* 127 (3), 239–262.
- Kothari, P., Kreiss, S., Alahi, A., 2021. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Trans. Intell. Transp. Syst.* 23 (7), 7386–7400.
- Kotseruba, I., Rasouli, A., Tsotsos, J.K., 2016. Joint attention in autonomous driving (JAAD). *arXiv preprint arXiv:1609.04741*.
- Kotseruba, I., Rasouli, A., Tsotsos, J.K., 2021. Benchmark for evaluating pedestrian action prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1258–1268.
- Liu, B., Adeli, E., Cao, Z., Lee, K.-H., Sheno, A., Gaidon, A., Niebles, J.C., 2020. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robot. Autom. Lett.* 5 (2), 3485–3492.
- Lorenzo, J., Parra, I., Wirth, F., Stiller, C., Llorca, D.F., Sotelo, M.A., 2020. Rnn-based pedestrian crossing prediction using activity and pose-related features. In: 2020 IEEE Intelligent Vehicles Symposium. IV, IEEE, pp. 1801–1806.
- Malla, S., Dariush, B., Choi, C., 2020. Titan: Future forecast using action priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11186–11196.
- Mangalam, K., An, Y., Girase, H., Malik, J., 2021. From goals, waypoints & paths to long term human trajectory forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15233–15242.
- Marchetti, F., Mordan, T., Becattini, F., Seidenari, L., Del Bimbo, A., Alahi, A., 2024. CrossFeat: Semantic cross-modal attention for pedestrian behavior forecasting. *IEEE Trans. Intell. Veh.*
- Maurer, M., Gerdes, J.C., Lenz, B., Winner, H., 2016. *Autonomous Driving: Technical, Legal and Social Aspects*. Springer Nature.
- Millard-Ball, A., 2018. Pedestrians, autonomous vehicles, and cities. *J. Plan. Educ. Res.* 38 (1), 6–12.
- Mínguez, R.Q., Alonso, I.P., Fernández-Llorca, D., Sotelo, M.A., 2018. Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition. *IEEE Trans. Intell. Transp. Syst.* 20 (5), 1803–1814.
- Mohamed, A., Zhu, D., Vu, W., Elhoseiny, M., Claudel, C., 2022. Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In: European Conference on Computer Vision. Springer, pp. 463–479.
- Osman, N., Camporese, G., Ballan, L., 2023. TAMformer: Multi-modal transformer with learned attention mask for early intent prediction. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1–5.
- Rasouli, A., Kotseruba, I., 2023. PedFormer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning. In: 2023 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 9844–9851.
- Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J.K., 2019. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6262–6271.
- Rasouli, A., Kotseruba, I., Tsotsos, J.K., 2017. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 206–213.
- Rasouli, A., Rohani, M., Luo, J., 2021. Bifold and semantic reasoning for pedestrian behavior prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15600–15610.
- Rudenko, A., Palmieri, L., Herman, M., Kitani, K.M., Gavrilu, D.M., Arras, K.O., 2020. Human motion trajectory prediction: A survey. *Int. J. Robot. Res.* 39 (8), 895–935.
- Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S., 2019. 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. pp. 1349–1358.
- Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M., 2020. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. Springer, pp. 683–700.
- Schörkhuber, D., Pröll, M., Gelautz, M., 2022. Feature selection and multi-task learning for pedestrian crossing prediction. In: 2022 16th International Conference on Signal-Image Technology & Internet-Based Systems. SITIS, IEEE, pp. 439–444.
- Sharma, N., Dhiman, C., Indu, S., 2022. Pedestrian intention prediction for autonomous vehicles: A comprehensive survey. *Neurocomputing*.
- Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* 28.
- Song, X., Kang, M., Zhou, S., Wang, J., Mao, Y., Zheng, N., 2022. Pedestrian intention prediction based on traffic-aware scene graph model. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 9851–9858.
- Sui, Z., Zhou, Y., Zhao, X., Chen, A., Ni, Y., 2021. Joint intention and trajectory prediction based on transformer. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 7082–7088.
- Teed, Z., Deng, J., 2020. RAFT: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision. Springer, pp. 402–419.
- Wang, C., Wang, Y., Xu, M., Crandall, D.J., 2022. Stepwise goal-driven networks for trajectory prediction. *IEEE Robot. Autom. Lett.* 7 (2), 2716–2723.
- Yang, D., Zhang, H., Yurtsever, E., Redmill, K.A., Özgüner, Ü., 2022. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Trans. Intell. Veh.* 7 (2), 221–230.
- Yao, Y., Atkins, E., Johnson-Roberson, M., Vasudevan, R., Du, X., 2021. Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation. *IEEE Robot. Autom. Lett.* 6 (2), 1463–1470.
- Yau, T., Malekmohammadi, S., Rasouli, A., Lakner, P., Rohani, M., Luo, J., 2021. Graph-sim: A graph-based spatiotemporal interaction modelling for pedestrian action prediction. In: 2021 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 8580–8586.
- Yuan, Y., Weng, X., Ou, Y., Kitani, K.M., 2021. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9813–9823.
- Zhang, W., Chai, Q., Zhang, Q., Wu, C., 2023a. Obstacle-transformer: A trajectory prediction network based on surrounding trajectories. *IET Cyber-Syst. Robot.* 5 (1), e12066.
- Zhang, Z., Tian, R., Ding, Z., 2023b. TrEP: Transformer-based evidential prediction for pedestrian intention with uncertainty. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37.