# Multi-modal Pedestrian Trajectory Prediction based on Pedestrian Intention for Intelligent Vehicle

**Youguo He[1], Yizhi Sun[1], Yingfeng Cai[1*], Chaochun Yuan[1], Jie Shen[2], and Liwei Tian[3]**
[1] Automotive Engineering Research Institute, Jiangsu University
Zhenjiang, Jiangsu 212013 China
[e-mail: wanghai1019@163.com]
[2] Department of Computer and Information Science, University of Michigan-Dearborn
MI 48128, USA
[e-mail: shen@umich.edu]
[3] College of Information Engineering, Shenyang University
Shenyang, Liaoning 110044, China
[e-mail: tianliwei@163.com]
* Corresponding author: Yingfeng Cai

*Received February 27, 2023; revised December 6, 2023; revised April 23, 2024; accepted June 6, 2024; published June 30, 2024*

## Abstract

The prediction of pedestrian trajectory is conducive to reducing traffic accidents and protecting pedestrian safety, which is crucial to the task of intelligent driving. The existing methods mainly use the past pedestrian trajectory to predict the future deterministic pedestrian trajectory, ignoring pedestrian intention and trajectory diversity. This paper proposes a multi-modal trajectory prediction model that introduces pedestrian intention. Unlike previous work, our model makes multi-modal goal-conditioned trajectory pedestrian prediction based on the past pedestrian trajectory and pedestrian intention. At the same time, we propose a novel Gate Recurrent Unit (GRU) to process intention information dynamically. Compared with traditional GRU, our GRU adds an intention unit and an intention gate, in which the intention unit is used to dynamically process pedestrian intention, and the intention gate is used to control the intensity of intention information. The experimental results on two first-person traffic datasets (JAAD and PIE) show that our model is superior to the most advanced methods (Improved by 30.4% on $MSE_{0.5s}$ and 9.8% on $MSE_{1.5s}$ for the PIE dataset; Improved by 15.8% on $MSE_{0.5s}$ and 13.5% on $MSE_{1.5s}$ for the JAAD dataset). Our multi-modal trajectory prediction model combines pedestrian intention that varies at each prediction time step and can more comprehensively consider the diversity of pedestrian trajectories. Our method, validated through experiments, proves to be highly effective in pedestrian trajectory prediction tasks, contributing to improving traffic safety and the reliability of intelligent driving systems.

***Keywords:*** Pedestrian trajectory prediction, Pedestrian Intention, *Gate Recurrent Unit* (GRU), Intelligent Driving.

This work has been supported by the National Key Research and Development Program of China(2022YFB2503302), the National Natural Science Foundation of China (52225212, U20A20333, U20A20331, 51875255, 5217120774), Key Research and Development Program of Jiangsu Province (BE2020083-3, BE2019010-2), Project of Faculty of Agricultural Equipment of Jiangsu University (NZXB20210103).

## 1. Introduction

In the past few years, we have witnessed significant progress in assisted driving and autonomous driving systems [1-5]. With the development of image processing technology [6-7], autonomous driving technology has achieved success in the field of pedestrian detection [8-10]. However, compared to pedestrian detection, research on pedestrian trajectory prediction is relatively limited, and this task is highly challenging. Predicting pedestrian movements, especially the trajectory at intersections, is essential to ensure pedestrian safety. By predicting the pedestrian trajectory [11-14], the intelligent vehicle can plan a safe path while driving to avoid path conflict between the intelligent vehicles and pedestrians.

Most of the previous pedestrian trajectory prediction methods assume that the future pedestrian trajectory is deterministic [15-18] and then predict the future trajectory based on the past pedestrian trajectory. However, in the process of walking, the pedestrian trajectory is uncertain. Even if the observation sequence of pedestrians has the same trajectory, pedestrians will also have multiple reasonable trajectories in the future. Early prediction methods ignored the diversity of pedestrian trajectories. Compared with the multi-modal trajectory prediction model, the deterministic trajectory prediction model has a larger error in predicting future trajectory distribution, so the multi-modal method is more suitable for predicting pedestrian trajectory. A recent work [19] adopted a multimodal approach in pedestrian trajectory prediction, but it only involved simple modeling based on the pedestrian's past trajectory. This method takes the pedestrian's past trajectory as input, generates multimodal results through Conditional Variational Autoencoder (CVAE), and then generates multiple trajectories using Recurrent Neural Network (RNN). Methods based on past trajectories have been proven to predict pedestrian trajectories [16,20,21], unfortunately, the past trajectory of the traveler may not reflect his future action goals. For example, students on the roadside may go to the road to check whether the school bus is coming. The trajectory-based models will assume that students will cross the road for this scenario. This problem can be improved by introducing pedestrian intentions. When pedestrians move, they will have their intentions, and the intentions will be used to plan the action route. The pedestrian intentions reflect the goal of pedestrians crossing the road, which is important for predicting the pedestrian trajectory. Another recent work [17] utilized pedestrian intent information in predicting pedestrian trajectories, but it assumed that pedestrian trajectories are deterministic and did not consider the variability of pedestrian trajectories.

Most pedestrian trajectory prediction methods are based on a bird-eye view [21-24]. These works simulate the bird-eye view by projecting self-centered video frames onto the ground. However, many roads are irregular, which will affect the accuracy of the projection, thus affecting the accurate prediction of pedestrian positions. The method [25] proposed to predict pedestrian trajectory in three-dimensional space requires expensive LIDAR equipment to obtain three-dimensional coordinates of the real scene. Methods from the first-person perspective not only better suit the scenarios of autonomous driving but also can reduce costs.

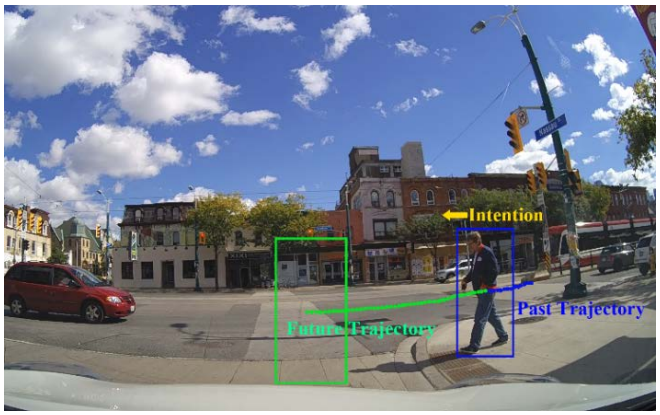



**Fig. 1.** The actual pedestrian trajectory is represented by a dark blue box, and the predicted trajectory is represented by a green box. Our goal is to predict the future trajectory of pedestrian through his past trajectory and intention.

In this paper, we propose a multi-modal goal-conditioned pedestrian trajectory prediction model based on pedestrian intention from the first-person perspective. **Fig.1** is a simple example of our model. Given the pedestrian's past trajectory within the deep blue box, we can predict the trajectory for the next few seconds based on the past trajectory and intention. We simultaneously utilize the pedestrian's past trajectory and intention to predict multiple future trajectories. Our model consists of three main modules: (1) Intention estimation module that estimates and generates intention values for pedestrians. (2) Sample generation module that learns the distribution of future trajectories based on past pedestrian trajectories through a random latent variable in CVAE. This module will also generate multiple samples and estimate the endpoint of the pedestrian's future trajectory based on pedestrian intention. (3) Trajectory generation module that predicts pedestrian future trajectories based on the samples and the endpoint of pedestrian trajectories generated by the sample generation module and the intention values generated by the intention estimation module. In the trajectory generation module, to better utilize pedestrian intention, we propose a novel GRU (Intention GRU) to dynamically analyze pedestrian intention. Pedestrian intention changes constantly due to environmental influences. To address this challenge, our Intention GRU adds an intention unit based on traditional GRU, which can analyze the pedestrian intention at each time step. We also designed an intention gate for Intention GRU to control the extent to which the pedestrian position changes with the change in pedestrian intent. For models based on past outputs, their errors increase over time. To address this issue, we applied a bidirectional trajectory predictor. It consists of Intention GRUs in both forward and backward directions. When predicting pedestrian trajectories, we first estimate the endpoint of the pedestrian's trajectory (not the endpoint of our model) and then combine the predicted trajectories in both directions to predict the final pedestrian trajectory.

In summary, existing pedestrian trajectory prediction methods face two key issues: firstly, they often assume determinism in future trajectories, overlooking the diversity of pedestrian trajectories; secondly, most of them rely solely on past trajectory information, neglecting the pedestrian intention and dynamic characteristics of intention. To address these challenges, we propose a multi-modal goal-conditioned pedestrian trajectory prediction model based on pedestrian intention. Unlike methods that generate deterministic predictions, our approach takes into account the diversity of trajectory outcomes. Additionally, we introduce the "Intention Gated Recurrent Unit" (Intention GRU) to analyze pedestrian intention at each time

step.

For this paper, the main contributions of our work are as follows:

- We propose Intention GRU to learn the pedestrian intention at every time step during prediction so that the pedestrian intention information can dynamically influence the future pedestrian trajectory.
- When conducting multi-modal goal-conditioned trajectory pedestrian prediction, we simultaneously utilize past pedestrian trajectory and pedestrian intention to improve the accuracy of endpoint estimation for pedestrian trajectory, ultimately improving the performance of the trajectory prediction model.
- The experimental results show that our model has reached an advanced level on two first-person datasets.

## 2. Related Works

Pedestrian detection, pedestrian tracking, and pedestrian trajectory prediction are three crucial tasks in image processing [26-28] related to pedestrians, where pedestrian detection and tracking form the foundation for pedestrian trajectory prediction. Pedestrian detection enables computers to locate pedestrians in images or videos, typically implemented using target detection algorithms, such as Convolutional Neural Networks (CNNs) based on deep learning methods. Recent advancements, as demonstrated in [29,30], have achieved state-of-the-art performance in this area. Pedestrian tracking involves continuously tracking the detected positions of pedestrians in consecutive images or videos and associating these positions. Target tracking algorithms, including those based on Kalman filtering, correlation filtering, or deep learning, can be employed for pedestrian tracking. The most advanced methods currently available include [31-33]. After pedestrian detection and tracking, the computer can obtain the position and past trajectory of pedestrians, and then predict their future trajectory. Currently, most methods use recurrent neural networks (RNN) to encode pedestrian past trajectories and then decode them to generate future trajectories. For multi-modal trajectory prediction, the current method is achieved through Conditional Variational Autoencoder CVAE [34,35]. Lee et al. [36] first used CVAE for multi-modal trajectory prediction. They combine CVAE with RNN encoding to generate random prediction hypotheses to deal with multi-modal targets in future predictions. Choi et al. [37] proposed a CVAE model with LSTM, which is an interactive perception model for vehicle motion prediction. The model predicts future movement with multi-modal distribution based on the confidence value of vehicle mobility. Salzmann et al. [38] extended Trajectron and proposed a modular recursive model Trajectron++, combining heterogeneous data of previous trajectory information, generating future condition predictions consistent with dynamic constraints, and generating full probability distributions. Yao et al. [19] divided CVAE target trajectories into parametric distribution and nonparametric distribution, designed nonparametric models (BiTraP-NP) and parametric models (BiTraP-GMM), respectively, and proved that the selection of potential variables affects the diversity of target distribution. Wang et al. [15] designed a SGNet, which includes an encoder module to capture historical information, a stepwise target estimator to predict future continuous targets, and a decoder module to predict future trajectories. SGNet predicts both long-term and short-term targets to better capture historical observations. Zhou et al. [39] proposed a multi-modal trajectory prediction method. In this approach, CVAE is employed to generate multi-modal trajectories, and GAN is used for adversarial training. They utilized spatiotemporal graphs to encode pedestrian social interactions and employed RNN to capture the temporal dependencies of evolving patterns. However, their research is only based

on the past trajectories of pedestrians, ignoring other information about pedestrians, such as their intentions.

In practice, once a pedestrian is detected, the computer can obtain pedestrian data through image processing programs [40,41], so we can extract information that is conducive to predicting the future trajectory from the previously obtained data, such as pedestrian intention. Intention estimation provides an intuitive understanding of pedestrian behavior, which is crucial for predicting pedestrian trajectory. Zhang et al. [42] analyzed pedestrian crossing intentions at red lights by examining changes in key points of the pedestrian's body and facial expressions. They employed four machine learning models to predict pedestrians' intentions to cross the red light. Zhou et al. [43] proposed a method for identifying and predicting pedestrian crossing intentions in obstructed visual fields. They constructed a specialized LSTM to integrate pedestrian posture, speed, interaction status features, and blind-state features to estimate pedestrian intentions. Ahmed et al. [44] introduced a multi-scale pedestrian intent prediction method based on 2D pose estimation and LSTMs. 2D pose estimation analyzes changes in pedestrian joints over time, and LSTMs predict pedestrian intentions based on the spatiotemporal information provided by pose estimation. Moreno et al. [45] introduced a random forest classifier to estimate pedestrian crossing intentions. They utilized features such as pedestrian position, speed, and heading. In estimating pedestrian intentions, they not only analyzed whether pedestrians were crossing but also provided a quantitative confidence level.

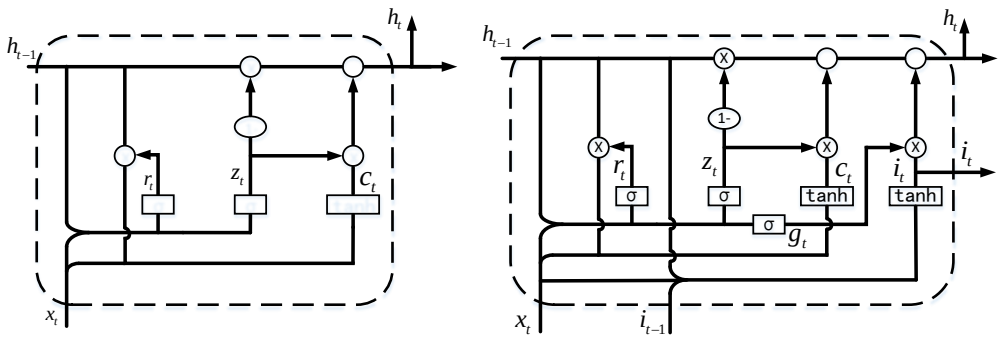
Current research on multimodal trajectory prediction ignored pedestrian intention when generating multi-modal trajectories. Methods regarding pedestrian intentions have not taken into account the characteristics of the temporal changes in pedestrian intentions. We propose a multimodal trajectory prediction model based on pedestrian intention. To estimate the intention of pedestrians at each prediction time step, we designed an Intention GRU based on Gate Recurrent Unit (GRU), which can simultaneously process intention information and pedestrian past trajectory information. Specifically, it added an intention unit and an intention gate based on a standard GRU to learn pedestrian intention at each time step.

## 3. Methodology

We first introduce our proposed Intention GRU in Section 3.1 and subsequently introduce our multi-modal pedestrian trajectory prediction model that applied Intention GRU in Section 3.2.

### 3.1 Intention GRU

The intention information of pedestrians is highly related to the future pedestrian trajectory. To make better use of the intention information, we propose additional intention units and intention gates. The standard GRU and our Intention GRU structure are shown in **Fig. 2.**




(a)GRU (b) Intention GRU

**Fig. 2.** (a) GRU. (b) Intention GRU. Where σ and tanh are the sigmoid activation function and tanh activation function respectively.

The standard GRU includes reset gate $r_t$ and update gate $z_t$, in addition, our Intention GRU also includes an intention unit $i_t$ and an intention gate $g_t$. In the two GRUs, how much historical information about the state at the previous time of reset gate control is written to the current candidate set $c_t$. The update gate controls how much current memory is saved to the current time step, while our intention unit and intention gate are used to process the intention information. In mathematics, we use $g_t \cdot i_t$ to dynamically change the output $h_t$. Our method makes the intention information not simply connect with the past track information of pedestrians but make full use of it. To better understand intention units and intention gates, we will introduce their details as follows:

**Intention Gate.** The future location of a pedestrian is related to his intention, and when he wants to cross the road, his position will change in the future. We propose an intention gate to control the influence of pedestrian intentions on their future location changes. We use a sigmoid activation function to realize this idea $g_t = \sigma(Wx_t + Wh_{t-1} + b_g)$. Once the intention value is obtained, the intention gate will analyze its reliability in each time step to make robust to noisy intention data. In this way, intention units can better utilize intention information.

**Intention Cell.** Intention information has a great impact on its future trajectory. However, pedestrian intentions are not static but change from time to time. It is not reasonable to use a fixed intention value. For example, when a pedestrian encounters a speeding vehicle crossing the road, he will stop crossing. Therefore, we propose an intention unit to deal with changing pedestrian intentions instead of using a fixed intention value. The initial input of the intention unit is the intention value from the intention estimation module. Then the intention state will be updated each time step by $i_t = tanh(Wx_t + Wi_{t-1} + Wh_{t-1} + b_i)$, and it is used as the input of the intention unit of the next GRU neuron.

**Recurrent State.** Mathematically, our Intention GRU is formulated as:

$$z_t = \sigma(Wx_t + Wh_{t-1} + b_z) \quad (1)$$

$$r_t = \sigma(Wx_t + Wh_{t-1} + b_r) \quad (2)$$

$$c_t = tanh[Wx_t + W(r_t * h_{t-1}) + b_c] \quad (3)$$

$$g_t = \sigma(Wx_t + Wh_{t-1} + b_g) \quad (4)$$

$$i_t = tanh(Wx_t + Wi_{t-1} + Wh_{t-1} + b_i) \quad (5)$$

$$h_t = (1 - z_t)h_{t-1} + z_t c_t + g_t i_t \quad (6)$$

Where σ and tanh are the sigmoid activation function and tanh activation function, W and b represent the weight matrix and bias vector, respectively. In the proposed Intention GRU network, the intention unit will update its state at each time step. The intention unit directly changes the output, which is consistent with the intention value that the pedestrian spatial movement depends on crossing the street. In Intention GRU, the intention state is continuously updated and used as input to the next Intention GRU neuron, the output response of each Intention GRU layer will also be used as the input of the next Intention GRU layer to affect future gates and units further. Therefore, Intention GRU can be applied in a pedestrian trajectory prediction method based on pedestrian past trajectory information and intention information cues.

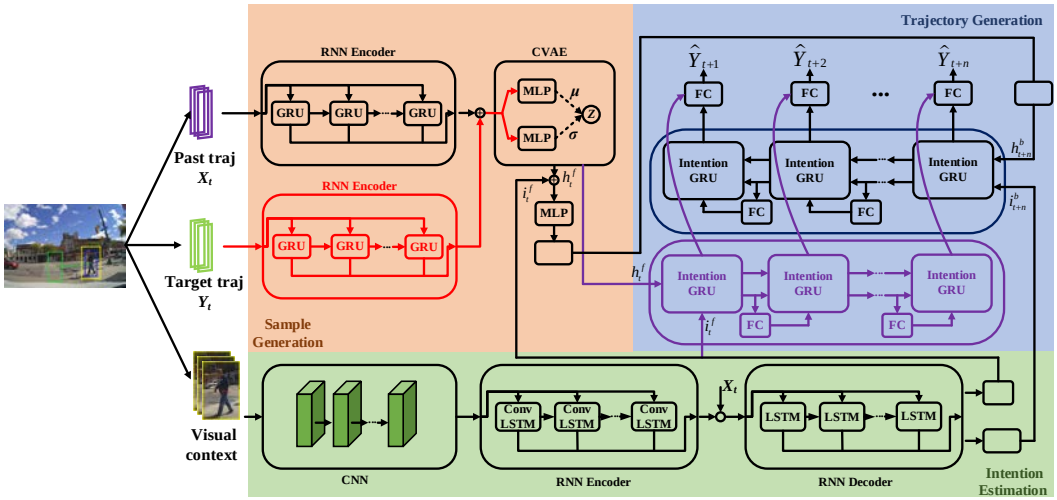



**Fig. 3.** Structure diagram of our model. The red arrows in the diagram show progress only during training.

## 3.2 Intention GRU for Pedestrian Trajectory Prediction

Our model performs the multi-modal trajectory prediction based on pedestrian intention in a first-person view. The pedestrian trajectory in the past m frames is known, and the goal is to predict the pedestrian trajectory in the next n frames. As shown in **Fig. 3**, our model mainly comprises three parts: Intention Estimation Module, Sample Generation Module, and Trajectory Generation Module. Among them, the Intention estimation module is used to estimate and generate intention values for pedestrians; Sample generation module learns the distribution of future trajectories based on past pedestrian trajectories through a random latent variable in CVAE. This module will also generate multiple samples and estimate the endpoint of the pedestrian's future trajectory based on pedestrian intention; Trajectory generation module predicts pedestrian future trajectories based on the samples and the endpoint of pedestrian trajectories generated by the sample generation module and the intention values generated by the intention estimation module. Pedestrian bounding boxes can be represented by the pixel position and size as $X_t = [x, y, w, h]$ in the first-person view. The past pedestrian trajectory can be represented by $\boldsymbol{X_t} = [X_{t-m+1}, X_{t-m+2}, \cdots, X_t]$, and the future pedestrian trajectory can be represented by $\boldsymbol{Y_t} = [Y_{t+1}, Y_{t+2}, \cdots, Y_{t+n}]$. Given the observation trajectory $\boldsymbol{X_t}$ and the pedestrian environment clipping box, we estimate the pedestrian crossing intention to obtain the intention estimation value $i_t$ first. At the same time, the sample generation

module generates multiple sample Z based on the past trajectory, then estimates the endpoint $G_t$ of the pedestrian trajectory by the intention value $i_t$ and sample Z, and finally predicts the pedestrian trajectory $\boldsymbol{Y_t}$ by the Bi-directional decoder composed of two Intention GRUs. In this section, we describe the details of our model (**Fig. 3**) in the following structure: Intention Estimation Module (Sec. 3.2.1), Sample Generation Module (Sec. 3.2.2), and Trajectory Generation Module (Sec. 3.2.3).

### 3.2.1 Intention Estimation Module

It has been shown that the visual information of pedestrian crossing intention is implied in the pedestrian-coded posture and direct local environment, and it implies the future movement of pedestrian, which helps predict the future pedestrian trajectory. Given the local visual context $\boldsymbol{C_{obs}} = [C_{t-m+1}, C_{t-m+2}, \cdots, C_t]$ and the past trajectory of an observation pedestrian $\boldsymbol{X_t} = [X_{t-m+1}, X_{t-m+2}, \cdots, X_t]$, we define a binary classification task $i_t \in \{0,1\}$ to predict whether pedestrians will cross the street, which represents the probability of pedestrians crossing the street, where 1 means pedestrians will cross, and 0 means they will not.

For pedestrian intention estimation, we use an encoder-decoder structure, where ConvLSTM [46] network is the encoder and LSTM is the decoder. First, VGG16 [47] pre-trained on ImageNet [48] is used to extract image features, then input the encoded features into the encoder, and finally input the encoder output of the encoder and past pedestrian trajectory $\boldsymbol{X_t}$ into the decoder, the intention value is then output by the decoder. We generate the intention values $i_t^f$ of pedestrian at time t and $i_{t+n}^b$ of pedestrian at time $t+n$, respectively, and use them to predict pedestrian trajectories in sample generation module and trajectory prediction module.

Algorithm 1 Workflow of Intention Estimation Module

Input: Pedestrian past trajectory $\boldsymbol{X_t}$, and visual context
Output: $i_t^f$, $i_{t+n}^b$: Intention value at time $t$ and time $t+n$
1: for each step $t$ do
2: Compute $h_t^x$ from $\boldsymbol{X_t}$ by linear embedding
3: Get feature from visual context by VGG16
4: Compute $h_t^e$ from feature by ConvLSTM encoder
5: Compute $h_t^i$ from $h_t^e$ and $h_t^x$ by ConvLSTM decoder
5: Compute $i_t^f$ from $h_t^i$ by fully connected layer
6: Estimate $i_{t+n}^b$ from $h_t^i$ by MLP
7: end for
8: return $i_t^f$, $i_{t+n}^b$

### 3.2.2 Sample Generation Module

The prediction of the future is inherently uncertain because there will be multiple credible future trajectories under the same observation sequence. Therefore, learning a deterministic function that directly maps $\boldsymbol{X_t}$ to $\boldsymbol{Y_t}$ will not be sufficient to represent the potential prediction space. To deal with this uncertainty, we adopt a deep generative model, the Conditional Variational Automatic Encoder (CVAE). CVAE can learn the distribution $P(\boldsymbol{Y_t}|\boldsymbol{X_t})$ by introducing a random latent potential variable Z, where $\boldsymbol{X_t}$ is the input and $\boldsymbol{Y_t}$ is the output. Our Sample Generation Module consists of prior network $P_\theta(Z|\boldsymbol{X_t})$ to model latent variable Z from the past pedestrian trajectory $\boldsymbol{X_t}$, recognition network $Q_\varphi(Z|\boldsymbol{X_t}, \boldsymbol{Y_t})$ to capture dependencies between Z and future pedestrian trajectory $\boldsymbol{Y_t}$, and goal generation network $P_\omega(G_t|\boldsymbol{X_t}, i_t^f, Z)$ to generate the estimated target.

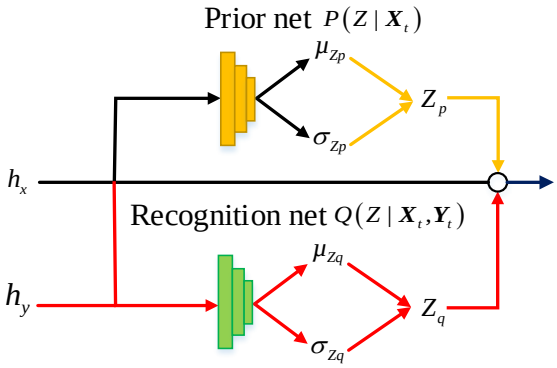



**Fig. 4.** Prior network and recognition network. The red arrows in the diagram show progress only during training.

**Train phase.** The distribution of latent potential variable $Z$ of the CVAE model is Gaussian distribution, $Z \sim Q_\varphi(Z|\boldsymbol{X_t}, \boldsymbol{Y_t}) = N(\mu_Z, \sigma_Z)$. First, the observation trajectory $\boldsymbol{X_t}$ and the ground real goal $\boldsymbol{Y_t}$ are processed by GRU to obtain the hidden state $h_t^x$ and $h_t^y$. Then the recognition network $Q_\varphi(Z|\boldsymbol{X_t}, \boldsymbol{Y_t})$ uses $h_t^x$ and $h_t^y$ to predict the distribution mean $\mu_{Zq}$ and standard deviation $\sigma_{Zq}$ to learn the dependency between the observation and the real target. Prior network $P_\theta(Z|\boldsymbol{X_t})$ only uses $h_t^x$ to predict $\mu_{Zp}$ and $\sigma_{Zp}$. Kullback–Leibler divergence (KLD) loss between $N(\mu_{Zp}, \sigma_{Zp})$ and $N(\mu_{Zq}, \sigma_{Zq})$ is optimized so that the priori network learns the dependency between $\boldsymbol{Y_t}$ and $\boldsymbol{X_t}$. Z is sampled from $N(\mu_{Zq}, \sigma_{Zq})$ and connected with the observation trajectory code $h_t^x$ and the result of intention estimation $i_t^f$ to predict multi-modal goals $\hat{G}_t$ by the goal generation network. We use 3-layer multi-layer perceptrons (MLPs) for prior, recognition and goal generation network.

**Test phase.** At the time of the test, the ground truth future trajectory $\boldsymbol{Y_t}$ cannot be obtained, so we extracted samples from $N(\mu_{Zp}, \sigma_{Zp})$ and connected the observation trajectory encode $h_t^x$ and the result of intention estimation $i_t^f$ to predict the estimated endpoint of trajectory $\hat{G}_t$.

Algorithm 2 Workflow of Sample Generation Module

Input: Pedestrian past trajectory $\boldsymbol{X_t}$, and intention value $i_t^f$
Output: Hidden state $h_t^f$, and estimated endpoint of trajectory $\hat{G}_t$
1: for each step *t* do
2: Compute $x_t$ from $\boldsymbol{X_t}$ by linear embedding
3: Compute $h_t^x$ by GRU
4: CVAE takes $h_t^x$ and sample Z proposals of $h_t^f$
5: Compute $\hat{G}_t$ from $h_t^f$ and $i_t^f$ by MLP
6: end for
7: return $h_t^f$, $\hat{G}_t$

### 3.2.3 Trajectory Generation Module

The input of the Trajectory Generation Module is the estimated endpoint of trajectory $\hat{G}_t$, hidden state $h_t^f$ and intention value $i_t^f$ and $i_{t+n}^b$, the output is the predicted pedestrian trajectory $\hat{Y}_t$. In this module, the estimated endpoint of trajectory $\hat{G}_t$ generates a hidden state $h_{t+n}^b$ through a fully connected layer. We use the intention GRU in both forward and backward directions to construct a bidirectional trajectory predictor to reduce the error in long-term

prediction. The forward trajectory predictor generates $h_{t+1}^f$, $h_{t+2}^f$ ,,, $h_{t+n-1}^f$ and $i_{t+1}^f$, $i_{t+2}^f$ ,,, $i_{t+n-1}^f$ through the hidden state $h_t^f$ and intention value $i_t^f$ as formulas (5) and (6), while the backward trajectory predictor generates $h_{t+n-1}^b$, $h_{t+n-2}^b$ ,,, $h_{t+1}^b$ and $i_{t+n-1}^b$, $i_{t+n-2}^b$ ,,, $i_{t+1}^b$ through the hidden state $h_{t+n}^b$ and intention value $i_{t+n}^b$ as formulas (5) and (6). Then, $h^f$ and $h^b$ at the same time step in both directions are connected to predict future pedestrian trajectories. These steps can be expressed by the formula:

$$h_{t+1}^f = IntentionGRU_f(h_t^f, i_t^f, W_f h_t^f + b_f) \tag{7}$$

$$h_{t+n-1}^b = IntentionGRU_b(h_{t+n}^b, i_{t+n}^b, W_b h_{t+n}^b + b_b) \tag{8}$$

$$\hat{Y}_{t+n-1} = W_f h_{t+n-1}^f + W_b h_{t+n-1}^b + b \tag{9}$$

where, *f*, *b*, W, and b indicate "forward", "backward", weight matrix, and bias vector, respectively.

Algorithm 3 Workflow of Trajectory Generation Module

Input: Intention value $i_t^f$ and $i_{t+n}^b$ , hidden state $h_t^f$ and estimated endpoint of trajectory $\hat{G}_t$
Output: Future trajectory $\hat{\boldsymbol{Y}}_t$
1: for each step *t* do
2: Compute $h_{t+n}^b$ from $\hat{G}_t$ by fully connected layer
3: Compute $i_{t+1}^f$ from $h_t^f$ and $i_t^f$ by forward Intention GRU as (5)
4: Compute $h_{t+1}^f$ from $h_t^f$ and $i_t^f$ by forward Intention GRU as (7)
5: Compute $i_{t+n-1}^b$ from $h_{t+n}^b$ and $i_{t+n}^b$ by backward Intention GRU as (5)
6: Compute $h_{t+n-1}^b$ from $h_{t+n}^b$ and $i_{t+n}^b$by forward Intention GRU as (8)
7: end for
8: concatenate the same time step $h^f$ and $h^b$ to predict trajectory $\hat{\boldsymbol{Y}}_t$ as (9)
9: return $\hat{\boldsymbol{Y}}_t$

### 3.2.4 Loss Functions

Our model is based on residual $\hat{Y}_{t+n}=Y_{t+n}-X_t$ predicts the change of the current position. Compared with the direct prediction of the future position [16] or the integration from the prediction of the future speed [44], the residual prediction can provide less initial loss than the predicted position from the beginning. We use L2 loss between prediction and target to represent CVAE loss [42] and adapt the best-of-many (BoM) [49] method to minimize the distance between the best prediction and target. This method will produce more accurate and diversified predictions and encourage models to capture real changes in data. We use a binary cross-entropy function when training the Intention Estimation Module. So the loss function of the model is the combination of target L2 loss, trajectory L2 loss, KLD loss between prior network and recognition network of CVAE, and Binary Cross Entropy Loss for intention estimation, which can be expressed by the formula:

$$\begin{aligned} L = \min_{i \in N} \left|\left| G_t - X_t - \hat{G}_t^i \right|\right| + \min_{i \in N} \sum_{m=t+1}^{t+n} \left|\left| Y_m - X_t - \hat{Y}_m^i \right|\right| \\ + KLD(Q_\varphi(Z|X_t, Y_t), P_\theta(Z|X_t)) \\ - \frac{1}{N}\sum_{i=1}^{N}(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)) \end{aligned} \tag{10}$$

where $\hat{G}_t$ and $\hat{Y}_m$ are the estimated goals and trajectory path points relative to current position $X_t$. $y_i$ and $\hat{y}_i$ represents the actual pedestrian intention and the estimated pedestrian intention.

## 4. Experiments

### 4.1 Datasets

**Pedestrian Intention Estimation** (PIE) [16]. A dataset suitable for automatic driving tasks for first-person view, including video clips of more than 6 hours of crosswalks in various types, all of which are taken by on-board cameras. This dataset has long traditions and more comprehensive annotations such as semantic content, ego-motion, and neighbor objects. The PIE dataset also provides information such as road boundaries required for traffic visual tasks. There are 1842 pedestrian samples in this dataset, and the proportions of train, test, and validation sets are 50%, 40%, and 10%, respectively. Tracks are sampled at an overlap rate of 0.5.

**Joint Attention for Autonomous Driving** (JAAD) [50]. JAAD contains 2800 pedestrian tracks, all of which were taken by a 30 Hz dash cam. This dataset contains various traffic scenes, videos under light and weather conditions, and provides a basis for studying the behavior of pedestrians and vehicles. The number of samples in this dataset is less than PIE and the track is shorter, so tracks are sampled at an overlap rate of 0.8. We use the same train/test segmentation as [16].

### 4.2 Implementation

**Intention Estimation.** For the encoder, a ConvLSTM with 64 filters and kernel size of 2 × 2 with stride 1 was used; and we used a LSTM with tanh activation,128 hidden units, dropout of 0.4, and recurrent dropout of 0.2 as the encoder. The image features are encoded using VGG16 pre-trained on ImageNet. For visual information, the input is the context of the image around pedestrians, which is clipped to twice the size of the pedestrian bounding box and resized to 224 × 224.

**Trajectory Prediction.** We implement our model in PyTorch and use a single NVIDIA GeForce RTX 3060 Laptop GPU to complete all our experiments. The observation and prediction horizon lengths are set to 0.5s (15 frames) and 1.5s (45 frames), respectively. The hidden size of the model is 256 and set the batch size to 64, learning rate 0.001, an exponential LR schedule [41], and the training is terminated after 50 epochs. We set the number of pedestrian trajectories generated by the model to 20. The sizes of the parameters in the model are shown in **Table 1**.

**Table 1.** The sizes of the parameters in the model. Among them, $IntentionGRU_f$ and $IntentionGRU_b$ represent the outputs of Intention GRU in both positive and negative directions, respectively.

| Parameter | size |
|---|---|
| $\boldsymbol{X_t}$ | 64×15×4 |
| $\boldsymbol{Y_t}$ | 64×45×4 |
| $\boldsymbol{Z}$ | 64×20×32 |
| $\boldsymbol{G}$ | 64×20×4 |
| $i^f$ | 64×20×256 |
| $i^b$ | 64×20×256 |

| | |
|---|---|
| $h^f$ | 64×20×256 |
| $h^b$ | 64×20×256 |
| $IntentionGRU_f$ | 64×45×20×4 |
| $IntentionGRU_b$ | 64×45×20×4 |
| $\widehat{\boldsymbol{Y}}_t$ | 45×20×4 |

### 4.3 Evaluation Metrics

For results of trajectory prediction, following [16,17,21], we report the following metrics: 1) Mean squared error (MSE) over bounding box coordinates, 2) Center mean squared error ($C_{MSE}$), which is the average MSE of the center of the bounding boxes, and 3) center final mean squared error ($CF_{MSE}$), which indicates the $C_{MSE}$ of the last time ($t+n$). For our multi-modal, we report the best-of-20 results (Minimum MSE, $C_{MSE,}$ and $CF_{MSE}$ of 20 random trajectories), following [42,44,51]. For the estimation results of pedestrian goal, we report its MSE and $C_{MSE}$ on the PIE dataset. For all metrics, the lower the value, the smaller the prediction error.

### 4.4 Comparison to State-of-the-Art

#### 4.4.1. Comparison results of trajectory endpoint estimation

The input of the reverse trajectory predictor in the trajectory prediction module is the trajectory endpoint predicted by the sample generation module, so improving the accuracy of the predicted trajectory endpoint has a positive significance in reducing the error of the final trajectory. We conducted a comparative experiment on the PIE dataset to investigate the impact of pedestrian intention information on the estimated trajectory endpoint. **Fig. 5** shows the errors of trajectory endpoint estimation with/without intention during training. **Table 2** shows the experimental results in the test set. We can see that after incorporating intention, the MSE performance of the estimated target improved by 23.9% (from 176 to 134) compared to the model without intention information, and the CMSE performance improved by 17.1% (from 146 to 121) compared to the model without intention information. Therefore, introducing pedestrian intention information helps to estimate the endpoint of pedestrian trajectories.

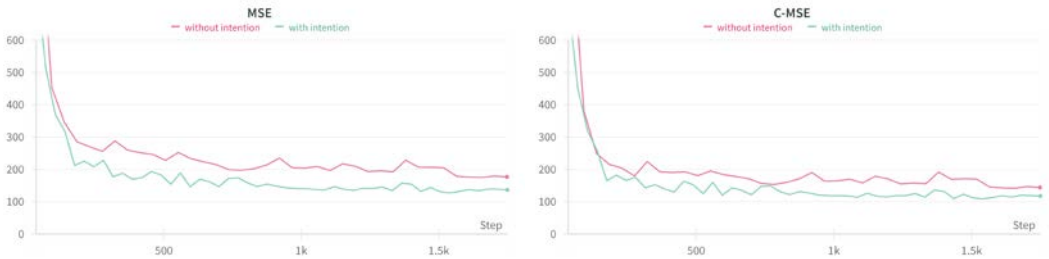



**Fig. 5.** Error of trajectory endpoint estimation with/without intention during training on the PIE. dataset.

**Table 2.** The error pedestrian goal estimation on the PIE dataset.

| Methods | MSE | $C_{MSE}$ |
|---|---|---|
| Without intention | 176 | 146 |
| With intention | 134 | 121 |

**Table 3.** Prediction error on multiple future time steps of different methods on PIE and JAAD datasets.

| Methods | PIE | | | JAAD | | |
|---|---|---|---|---|---|---|
| | MSE | $C_{MSE}$ | $CF_{MSE}$ | MSE | $C_{MSE}$ | $CF_{MSE}$ |
| | 0.5s/1.0s/1.5s | 1.5s | 1.5s | 0.5s/1.0s/1.5s | 1.5s | 1.5s |
| B-LSTM [52] | 101/296/855 | 811 | 3258 | 159/539/1535 | 1447 | 5615 |
| $PIE_{traj}$ [16] | 58/200/636 | 596 | 2477 | 110/399/1248 | 1183 | 4780 |
| Holistic LSTM [17] | 56/167/507 | 466 | 1917 | 105/389/1177 | 1116 | 4493 |
| FOL-X [53] | 47/183/584 | 546 | 2303 | 147/484/1374 | 1290 | 4924 |
| BiTraP-D [19] | 41/161/511 | 481 | 1949 | 93/378/1206 | 1105 | 4565 |
| Model-D | 38/157/503 | 480 | 1993 | 89/364/1189 | 1095 | 4531 |
| BiTraP-GMM[19] | 38/94/222 | 171 | 368 | 153/250/585 | 501 | 998 |
| BiTraP-NP [19] | 23/48/102 | 81 | 261 | 38/94/222 | 177 | 565 |
| **Ours(20)** | **16/40/92** | **67** | **201** | **32/80/192** | **157** | **502** |

### 4.4.2. Comparison results of trajectory prediction

**Table 3** shows the comparison between our model and advanced methods. We also show our deterministic model Model-D, which removes the multi-modal CVAE module compared to the multi-modal method. Among the above methods, only BiTraP-NP and BiTraP-GMM are multi-modal, and others are deterministic. As shown in **Table 3**, our deterministic method has achieved advanced results; for multi-modal methods, our multi-modal results are superior to other methods in all benchmarks. On the PIE dataset, our performance on $MSE_{0.5s}$ is 30% higher than that of BiTraP-NP (from 23 to 16). In addition, our result is better than BiTraP-NP (9.8%) in a long-time prediction on $MSE_{1.5s}$ (from 102 to 92). For the JAAD dataset, we improved the performance on $MSE_{0.5s}$ from 38 to 32 and on $MSE_{1.5s}$ from 222 to 192. The experiment results show that pedestrian intention information helps predict pedestrian trajectory and proves that our model is more capable of short-term and long-term prediction.

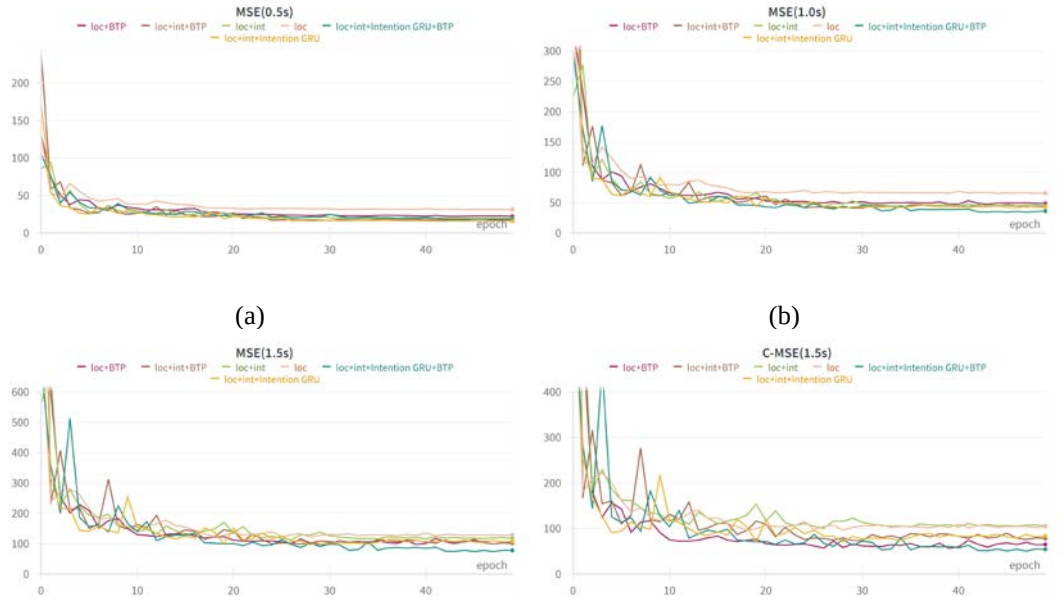
**Table 4.** Results of ablation studies.

| Methods | MSE | | | $C_{MSE}$ | $CF_{MSE}$ |
|---|---|---|---|---|---|
| | 0.5s | 1s | 1.5s | 1.5s | 1.5s |
| *loc* | 30 | 63 | 121 | 105 | 379 |
| *loc* + *int* | 19 | 46 | 117 | 98 | 333 |
| *loc* + *BTP* | 23 | 48 | 102 | 81 | 261 |
| *loc* + *int* + *BTP* | 18 | 45 | 98 | 75 | 231 |
| *loc* + *int* + *Intention GRU* | 19 | 44 | 97 | 77 | 234 |
| *loc* + *int* + *Intention GRU* + *BTP* | 16 | 40 | 92 | 67 | 201 |

## 4.5 Ablation Studies

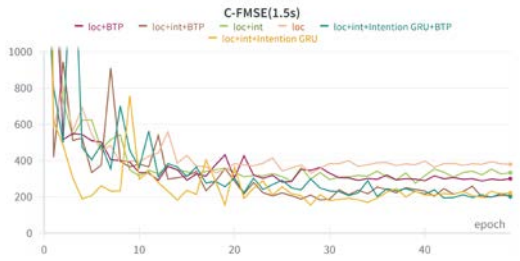
In order to understand how pedestrian intention information, intention GRU, and bidirectional trajectory predictor affect the performance of pedestrian trajectory prediction, we conducted ablation studies on the PIE dataset. We will set the model that only uses pedestrian past trajectories as input and the trajectory generation module is a unidirectional traditional GRU as "*loc*", where "*int*" indicates adding pedestrian intentions to the input. "*BTP*" indicates the use of a bidirectional trajectory predictor in the trajectory generation module, and "*Intention*

*GRU*" indicates the use of Intention GRU in the trajectory generation module. **Fig. 6** shows the changes in various indicators during training using various methods. **Table 4** shows the experimental results of various methods on the test set. We observe that the error of the loc method is the largest in all indicators. After integrating pedestrian intentions, the error decreases, especially in the short term. $MSE_{0.5s}$ decreased from 30 to 19 (a decrease of 36.7%), while in the long term, intention has little effect on improving model performance. This is because the estimated pedestrian intentions are only the thoughts of pedestrians in the short term, and pedestrian intentions may change in the subsequent time. *loc+BTP* reduced $MSE_{1.5s}$, $C_{MSE1.5s}$, and $CF_{MSE1.5s}$ from 121, 105, and 379 to 102, 81, and 261, indicating that using a bidirectional trajectory predictor can indeed reduce errors in long-term prediction. By comparing *loc+int* and *loc+int+Intention GRU*, we can observe that the prediction error is significantly reduced in long-term prediction after using Intention GRU. Specifically, $MSE_{1.5s}$, $C_{MSE1.5s,}$ and $CF_{MSE1.5s}$ decreased from 117, 98 and 333 to 97, 77 and 234, a decrease of 17.1%, 21.4% and 29.7%. This is because our Intention GRU can change the intention value of each prediction step instead of a constant intention value to predict trajectories. Our complete model, *loc+int+Intention GRU+BTP*, achieved the best performance in all indicators. The results of the ablation experiment demonstrate the effectiveness of our method. In summary, adding pedestrian intention information to the input can reduce prediction errors in the short term. Intention GRU can dynamically estimate pedestrian intentions for each prediction step, thereby improving the impact of pedestrian intentions on long-term prediction. The bidirectional trajectory predictor can improve the performance of the model in long-term prediction.




(a) (b)

(c) (d)



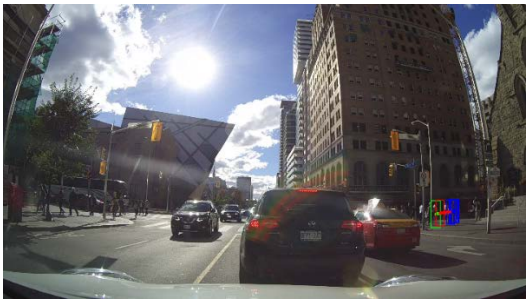

(e)

**Fig. 6.** The error changes of various methods during training.

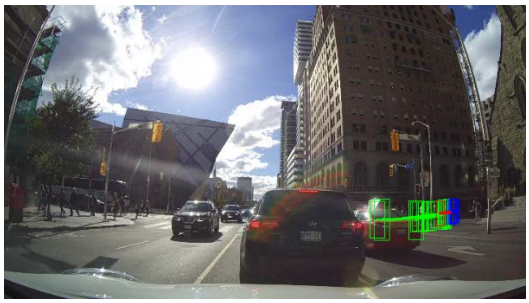
### 4.6 Visualization

In order to express our experimental results intuitively, we show several visual predictions of our model on the PIE dataset in **Fig. 7**. We use a box to represent the position of pedestrians 1.5 seconds later, and lines of points represent the pedestrian trajectory within 1.5 seconds. The dark blue represents the observation pedestrian trajectory, the red represents the ground truth future, and the green represents the predicted trajectory.

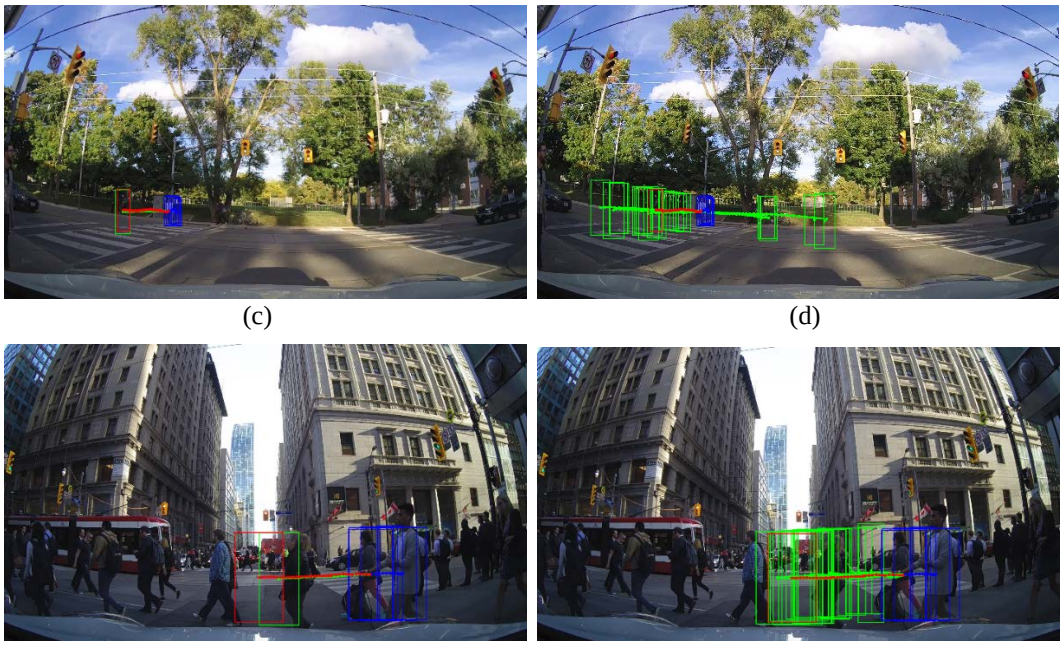
As we can see in **Fig. 7(a)** and **Fig. 7(b)**, a pedestrian has an intention to cross the street in the next few seconds. The deterministic method cannot directly reflect the crossing behavior, while the multi-modal method can predict a variety of possible trajectories, including two trajectories crossing the street, and the distribution of trajectories also indicates that the pedestrian will cross the street. In **Fig. 7(c)**, **Fig. 7(d)**, **Fig. 7(e)**, and **Fig.7(f)** the pedestrian bounding boxes predicted by the multimodal approach are mostly in proximity to the true bounding boxes. Although our multimodal method predicts multiple trajectories, the predicted trajectories have a higher distribution near the actual trajectories. This demonstrates the effectiveness of the trajectories predicted by the multimodal approach. The multi-modal method can predict multiple trajectories and clearly show the future pedestrian trajectory distribution. Therefore, the multi-modal method can obtain more reasonable trajectories than the deterministic method so that intelligent vehicles can better predict risks and ensure pedestrian safety. Our multi-modal method predicts multiple trajectories, which is consistent with pedestrian having multiple reasonable trajectories in the future.



(a)



(b)



(c) (d)

(e) (f)

**Fig. 7.** Our visualization results are shown in the figure. The left column are the deterministic results, and the right column are the multi-modal results. The dark blue, red and green represent the pedestrian past trajectories, the future real trajectories, and the prediction trajectories, respectively.

## 5. Conclusion

In this paper, we propose a multi-modal goal-conditioned pedestrian trajectory prediction model that utilizes past pedestrian trajectory and dynamic pedestrian intentions to generate multi-modal pedestrian trajectories. We introduce the Intention GRU, which is a dynamic learning mechanism that can capture pedestrian intentions throughout the entire prediction process, enabling the model to identify pedestrian intention at each time step. At the same time, we use two directional Intention GRUs to form a bidirectional trajectory predictor to reduce the model's error in long-term prediction. Our experimental results on two first-perspective datasets show that combining pedestrian intention information significantly improves the performance of the model in short-term prediction, while reducing errors in estimating pedestrian trajectory endpoints, thereby improving overall model performance. Intention GRU allows the model to achieve state-of-the-art performance in both short-term and long-term predictions. Our method can improve the accuracy of pedestrian trajectory prediction in intelligent driving systems and ultimately contribute to the broader goal of improving overall traffic safety. However, our work has not fully exploited pedestrian information. Future work could explore the integration of additional pedestrian cues, such as facial expressions and walking postures, to better predict pedestrian behavior and improve the accuracy of trajectory prediction. Additionally, the model can be extended to incorporate environmental factors, such as traffic signals, providing a more comprehensive approach to trajectory prediction.

## References


[1] X. Song, P. Wang, D. Zhou, R. Zhu, C. Guan, Y. Dai, H. Su, H. Li, and R. Yang, "ApolloCar3D: A Large 3D Car Instance Understanding Benchmark for Autonomous Driving," in *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5447-5457, 2019. Article (CrossRef Link)

[2] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol.32, no.11, pp.1231-1237, 2013. Article (CrossRef Link)

[3] Y. He, Y. Yang, Y. Cai, C. Yuan, J. Shen, and L. Tian. "Predicting pedestrian tracks around moving vehicles based on conditional variational transformer," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol.5, no.3, pp.1-14, 2023. Article (CrossRef Link)

[4] H. Wang, Y. Xu, Z. Wang, Y. Cai, L. Chen and Y. Li, "CenterNet-Auto: A Multi-object Visual Detection Algorithm for Autonomous Driving Scenes Based on Improved CenterNet," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol.7, no.3, pp.742-752, 2023. Article (CrossRef Link)

[5] L. Chen, Q. Zhou, Y. Cai, H. Wang, and Y. Li, "CAE-GAN: A hybrid model for vehicle trajectory prediction," *IET Intelligent Transport Systems*, vol.16, no.12, pp.1682-1696, 2022. Article (CrossRef Link)

[6] J. Zhang, Y. Lv, J. Tao, F. Huang and J. Zhang, "A Robust Real-Time Anchor-Free Traffic Sign Detector With One-Level Feature," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol.8, no.2, pp.1437-1451, 2024. Article (CrossRef Link)

[7] Y. Chen, R. Xia, K. Zou, and K. Yang, "FFTI: Image inpainting algorithm via features fusion and two-steps inpainting," *Journal of Visual Communication and Image Representation*, vol.91, 103776, 2023. Article (CrossRef Link)

[8] I. Hasan, S. C. Liao, J. P. Li, S. U. Akram and L. Shao, "Generalizable Pedestrian Detection: The Elephant In The Room," in *Proc. of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.11323-11332, 2021. Article (CrossRef Link)

[9] M. Liu, J. Jiang, C. Zhu, and X-C. Yin, "VLPD: Context-Aware Pedestrian Detection via Vision-Language Semantic Self-Supervision," in *Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6662-6671, 2023. Article (CrossRef Link)

[10] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, R. Jin, and X. Sun, "Beyond Appearance: a Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks," in *Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.15050-15061, 2023. Article (CrossRef Link)

[11] B. I. Sighencea, R. I. Stanciu, and C. D. Căleanu, "A Review of Deep Learning-Based Methods for Pedestrian Trajectory Prediction," *Sensors*, vol.21, no.22, 7543, 2021. Article (CrossRef Link)

[12] A. Kalatian and B. Farooq, "A context-aware pedestrian trajectory prediction framework for automated vehicles," *Transportation Research Part C: Emerging Technologies*, vol.134, 103453, 2022. Article (CrossRef Link)

[13] C. Anderson, R. Vasudevan, and M. Johnson-Roberson, "Off the Beaten Sidewalk: Pedestrian Prediction in Shared Spaces for Autonomous Vehicles," *IEEE Robotics and Automation Letters*, vol.5, no.4, pp.6892-6899, 2020. Article (CrossRef Link)

[14] Y. Li, X-Y. Lu, J. Wang, and K. Li, "Pedestrian Trajectory Prediction Combining Probabilistic Reasoning and Sequence Learning," *IEEE Transactions on Intelligent Vehicles*, vol.5, no.3, pp.461-474, 2020. Article (CrossRef Link)

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

[16] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.6261-6270, 2019. Article (CrossRef Link)
[17] R. Quan, L. Zhu, Y. Wu, and Y. Yang, "Holistic LSTM for Pedestrian Trajectory Prediction," *IEEE Transactions on Image Processing*, vol.30, pp.3229-3239, 2021. Article (CrossRef Link)
[18] Y. Cai, L. Dai, H. Wang, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, "Pedestrian Motion Trajectory Prediction in Intelligent Driving from Far Shot First-Person Perspective Video," *IEEE Transactions on Intelligent Transportation Systems*, vol.23, no.6, pp.5298-5313, 2022. Article (CrossRef Link)
[19] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, "BiTraP: Bi-Directional Pedestrian Trajectory Prediction with Multi-Modal Goal Estimation," *IEEE Robotics and Automation Letters*, vol.6, no.2, pp.1463-1470, 2021. Article (CrossRef Link)
[20] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.961-971, 2016. Article (CrossRef Link)
[21] S. Zamboni, Z. T. Kefato, S. Girdzijauskas, C. Norén, and L. D. Col, "Pedestrian trajectory prediction with convolutional neural networks," *Pattern Recognition*, vol.121, 108252, 2022. Article (CrossRef Link)
[22] L. Huang, J. Zhuang, X. Cheng, R. Xu, and H. Ma, "STI-GAN: Multimodal Pedestrian Trajectory Prediction Using Spatiotemporal Interactions and a Generative Adversarial Network," *IEEE Access*, vol.9, pp.50846-50856, 2021. Article (CrossRef Link)
[23] J. Yue, D. Manocha, and H. Wang, "Human Trajectory Prediction via Neural Social Physics," in *Proc. of 17th European Conference on Computer Vision*, pp.376-394, 2022. Article (CrossRef Link)
[24] J. Lian, X. Wang, L. Li, Y. Zhou, and B. Zhou, "Pedestrian Trajectory Prediction Based on Human-vehicle Interaction," *China Journal of Highway and Transport*, vol.34, no.5, pp.215-223, 2021. Article (CrossRef Link)
[25] S. Zhou, H. Xu, G. Zhang, T. Ma, and Y. Yang, "Deep learning-based pedestrian trajectory prediction and risk assessment at signalized intersections using trajectory data captured through roadside LiDAR," *Journal of Intelligent Transportation Systems*, pp.1-13, 2023. Article (CrossRef Link)
[26] Y. Li, F. Feng, Y. Cai, Z. Li, and M. A. Sotelo, "Localization for Intelligent Vehicles in Underground Car Parks Based on Semantic Information," *IEEE Transactions on Intelligent Transportation Systems*, vol.25, no.2, pp.1317-1332, 2024. Article (CrossRef Link)
[27] Y. Cai, Z. Lu, H. Wang, L. Chen, and Y. Li, "A Lightweight Feature Map Creation Method for Intelligent Vehicle Localization in Urban Road Environments," *IEEE Transactions on Instrumentation and Measurement*, vol.71, pp.1-15, 2022. Article (CrossRef Link)
[28] J. Zhang, X. Liu, X. Zhang, Z. Xi, and S. Wang, "Automatic Detection Method of Sewer Pipe Defects Using Deep Learning Techniques," *Applied Sciences*, vol.13, no.7, 4589, 2023. Article (CrossRef Link)
[29] A. H. Khan, M. Munir, L. van Elst and A. Dengel, "F2DNet: Fast Focal Detection Network for Pedestrian Detection," in *Proc. of 2022 26th International Conference on Pattern Recognition*, pp. 4658-4664, 2022. Article (CrossRef Link)
[30] A. H. Khan, M. S. Nawaz and A. Dengel, "Localized Semantic Feature Mixers for Efficient Pedestrian Detection in Autonomous Driving," in *Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5476-5485, 2023. Article (CrossRef Link)
[31] Y. Cui, C. Jiang, L. Wang, and G. Wu, "MixFormer: End-to-End Tracking with Iterative Mixed Attention," in *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.13598-13608, 2022. Article (CrossRef Link)
[32] X. Wei, Y. Bai, Y. Zheng, D. Shi, and Y. Gong, "Autoregressive Visual Tracking," in *Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.9697-9706, 2023. Article (CrossRef Link)

[33] Q. Wu, T. Yang, Z. Liu, B. Wu, and Y. Shan, A. B. Chan "DropMAE: Masked Autoencoders with Spatial-Attention Dropout for Tracking Tasks," in *Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.14561-14571, 2023. Article (CrossRef Link)
[34] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised Learning with Deep Generative Models," in *Proc. of NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol.2, pp.3581-3589, 2014. Article (CrossRef Link)
[35] K. Sohn, X. Yan, and H. Lee, "Learning structured output representation using deep conditional generative models," in *Proc. of NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, vol.2, 2015. Article (CrossRef Link)
[36] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, "DESIRE: Distant future prediction in dynamic scenes with interacting agents," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2165-2174, 2017. Article (CrossRef Link)
[37] C. Choi, S. Malla, A. Patil, and J. H. Choi, "DROGON: A Trajectory Prediction Model based on Intention-Conditioned Behavior Reasoning," in *Proc. of* 2020 Conference on Robot Learning, pp.49-63, 2021. Article (CrossRef Link)
[38] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Proc. of Computer Vision–ECCV 2020: 16th European Conference, Glasgow, Proceedings, Part XVIII*, pp.683-700, 2020. Article (CrossRef Link)
[39] Z. Zhou, G. Huang, Z. Su, Y. Li, and W. Hua, "Dynamic Attention-Based CVAE-GAN for Pedestrian Trajectory Prediction," *IEEE Robotics and Automation Letters*, vol.8, no.2, pp.704-711, 2023. Article (CrossRef Link)
[40] Y. Chen, R. Xia, K. Yang, and K. Zou, "MICU: Image super-resolution via multi-level information compensation and U-net," *Expert Systems with Applications*, vol.245, 123111, 2024. Article (CrossRef Link)
[41] Y. Chen, R. Xia, K. Yang, and K. Zou, "DNNAM: Image inpainting algorithm via deep neural networks and attention mechanism," *Applied Soft Computing*, vol.154, 111392, 2024. Article (CrossRef Link)
[42] S. Zhang, M. Abdel-Aty, Y. Wu, and O. Zheng, "Pedestrian Crossing Intention Prediction at Red-Light Using Pose Estimation," *IEEE Transactions on Intelligent Transportation Systems*, vol.23, no.3, pp.2331-2339, 2022, Article (CrossRef Link)
[43] J. Zhou, X. Bai, and W. Hu, "Recognition and Prediction of Pedestrian Hazardous Crossing Intentions in Visual Field Obstruction Areas Based on IPVO-LSTM," *Applied Sciences*, vol.13, no.5, 2999, 2023. Article (CrossRef Link)
[44] S. Ahmed, A. A. Bazi, C. Saha, S. Rajbhandari, and M. N. Huda, "Multi-scale pedestrian intent prediction using 3D joint information as spatio-temporal representation," *Expert Systems With Applications*, vol.225, 120077, 2023. Article (CrossRef Link)
[45] E. Moreno, P. Denny, E. Ward, J. Horgan, C. Eising, E. Jones, M. Glavin, and A. Parsi, D. Mullins, B. Deegan, "Pedestrian Crossing Intention Forecasting at Unsignalized Intersections Using Naturalistic Trajectories," *Sensors*, vol.23, no.5, 2773, 2023. Article (CrossRef Link)
[46] X. Shi, Z. Chen, H. Wang, D-Y. Yeung, W-K. Wong, and W-C. Woo, "Convolutional LSTM Network: a machine learning approach for precipitation nowcasting," *in Proc. of NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, vol. 1, pp.802-810, 2015. Article (CrossRef Link)
[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of International Conference on Learning Representations*, pp.1-14, 2015. Article (CrossRef Link)
[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol.115, pp. 211-252, 2015. Article (CrossRef Link)

[49] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.2255-2264, 2018. Article (CrossRef Link)
[50] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Joint Attention in Autonomous Driving (JAAD)," *Computer Science, Engineering, Psychology*, arXiv, pp.1-10, 2016. Article (CrossRef Link)
[51] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1349-1358, 2019. Article (CrossRef Link)
[52] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-Term On-board Prediction of People in Traffic Scenes under Uncertainty," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4194-4202, 2018. Article (CrossRef Link)
[53] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised Traffic Accident Detection in First-Person Videos," in *Proc. of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp.273-280, 2019. Article (CrossRef Link)

**Youguo He** was received his B.Sc. degree from Department of Electronic Engineering, Liaoning University of Technology, Jinzhou, China in 2000, and his M.Sc. and Ph.D. degrees from the School of Information Science and Engineering, Northeastern University, Shenyang, China in 2005 and 2008, respectively.

In 2008, he joined the Faculty of Information Engineering, Shenyang University as a lecturer. In 2015, he joined the Automotive Engineering Research Institute in Jiangsu University. Associate professor title. His research interests are in intelligent automotive and vehicle control system.

**Yizhi Sun** was received his B.Sc. degree from School of Mechanical Engineering, Anhui University of Science and Technology, Huainan, China in 2021. He is currently pursuing the M.Sc. degree in Automotive Engineering Research Institute, Jiangsu University, Zhenjiang, China. His research interests include computer vision, deep learning, and intelligent vehicles.

**Yingfeng Cai** was received her B.S., M.S. and Ph.D. degree all from the School of Instrument Science and Engineering, Southeast University, Nanjing, China, respectively.

In 2013, she joined the Automotive Engineering Research Institute in Jiangsu University. Professor title. Her research interests are in intelligent transportation system and intelligent automotive.

**Chaochun Yuan** was received his B.Sc. degree from Department of Packaging Engineering, Jiangsu Polytechnic University, Zhenjiang, China in 2000, and Ph.D. degree from the School of Agricultural Mechanization Engineering, Jiangsu University, Zhenjiang, China in 2007.

In 2008, he joined the Automotive Engineering Research Institute in Jiangsu University. Professor title. His research interests are in intelligent automotive and vehicle active safety.

**Jie Shen** is the editor-in-chief of International Journal of Modelling and Simulation, which is an EI-indexed, peer-reviewed research journal in the field of modelling and simulation. He also served as an editorial board member for two international journals; an organizer for 8 international conferences; an associate editor of 2 international conference proceedings; a program committee member for 20 international conferences; a session chair for 13 international or national conferences; a board member for 3 international- or national-level technical committees; and a member for various committees at department and campus levels within the University of Michigan - Dearborn.

**Liwei Tian** received the Ph.D. degree from the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China. He is currently a Professor with Shenyang University, Shenyang. His current research interests include artificial intelligence, evolutionary computation, swarm intelligence, and their applications in design and optimization of intelligent transportation systems.