Social Aware Multi-modal Pedestrian Crossing Behavior Prediction

Xiaolin Zhai[1,2], Zhengxi Hu[1,2], Dingye Yang[1,2], Lei Zhou[1,2], and Jingtai Liu[1,2](✉)

[1] Institute of Robotics and Automatic Information System, College of Artificial Intelligence, Nankai University, Tianjin, China
{2120210410,hzx,1711502}@mail.nankai.edu.cn, liujt@nankai.edu.cn
[2] Tianjin Key Laboratory of Intelligent Robotics, Nankai University, Tianjin, China

Abstract. With the development of self-driving vehicles, pedestrian behavior prediction plays a vital role in constructing a safe human-robot interactive environment. Previous methods ignored the inherent uncertainty of pedestrian future actions and the temporal correlations of spatial interactions. To solve the aforementioned problems, we propose a novel social aware multi-modal pedestrian crossing behavior prediction network. In this research field, our network innovatively explores the multimodality nature of pedestrian future action prediction and forecasts diverse and plausible futures. Also, to model the social aware context in both the spatial and temporal domain, we construct a spatial-temporal heterogeneous graph, bridging the spatial-temporal gap between the scene and the pedestrian. Experiments show that our model achieves state-of-the-art performance on pedestrian action detection and prediction task. The code is available at https://github.com/zxll0106/Pedestrian_Crossing_Behavior_Prediction.

Keywords: Pedestrian crossing behavior prediction · Video understanding

1 Introduction

Predicting pedestrian behaviors plays a critical role in the human-robot interactive scene. Pedestrians in the urban traffic scenario can extract useful information from the surroundings to infer others' motion patterns and make reasonable decisions. We hope that autonomous systems can acquire the capability to mimic human perception to predict pedestrian behaviors, which is important for creating a safer environment for both robots and pedestrians.

This work is supported in part by National Key Research and Development Project under Grant 2019YFB1310604, in part by National Natural Science Foundation of China under Grant 62173189.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-26316-3_17.

L. Wang et al. (Eds.): ACCV 2022, LNCS 13844, pp. 275–290, 2023.
https://doi.org/10.1007/978-3-031-26316-3_17

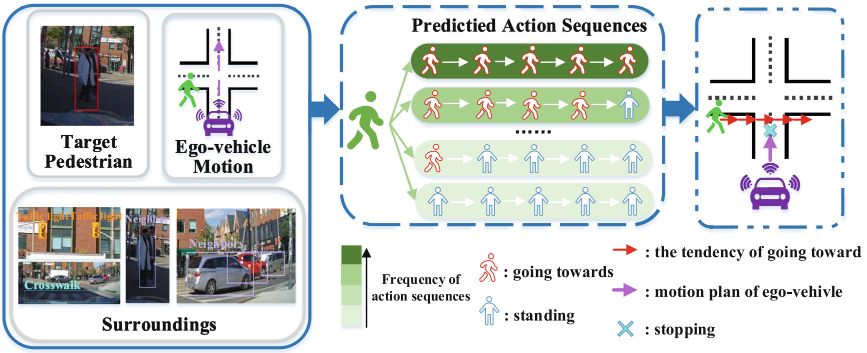

Fig. 1. Our model captures useful information not only from the target pedestrian but also from the social aware context. Interactions with the surrounding traffic objects and the ego-vehicle motion are utilized to enhance the dynamical representation of the target pedestrian. Our model also considers the uncertainty of pedestrian future actions and performs diverse predictions. We choose the most frequently appearing action sequence as the final prediction result. The ego-vehicle will make reactive decisions based on the predicted action sequence of the target pedestrian. Only *standing* and *going towards* are involved in this figure and other action labels are not involved.

Many works [10,13,14,22] explored pedestrian crossing behaviors and achieved significant improvement in the pedestrian behavior understanding. SF-GRU [13] designed the stacked RNNs to gradually fuse multiple inputs to estimate pedestrian intention. MMH-PAP [14] modeled the temporal dynamics of different inputs and designed an attention module to calculate the weights of each input to predict binary crossing action. Yao *et al.* [22] proposed an intention estimation and action detection network. A soft-attention module is designed to capture spatial interaction between different traffic objects. However, they not only neglect the multi-modal nature of pedestrian future actions, but also ignore the temporal continuity of relations between traffic objects.

Different from previous works, we take account into the inherent uncertainty of pedestrian behaviors which is a critical cue for inferring future actions, as shown in Fig. 1. Under the same history states, this uncertainty can cause diverse and plausible futures. For example, when the pedestrian comes to the crossroad, he may cross directly or wait for the red traffic light. Deterministic models are not suitable for capturing a one-to-many mapping and producing probabilistic inference. As a result, we propose Multi-Modal Conditional Generative Module to learn multiple modes of pedestrian future actions. In this module, we introduce multiple latent variables to model the multimodality of pedestrian future and perform diverse action predictions.

Spatial-temporal interactions between traffic objects play a vital role in refining the scene information and enhancing the pedestrian representation. Pedestrians observe the current and past states of other traffic objects to perceive the surrounding environment and make reasonable decisions. To enable the robot with the ability of perception, we propose Social Aware Encoder to extract the pedestrian-specific contextual information. Spatial relations between traffic objects and the temporal continuity of these relations should also be fully considered. In Social Aware Encoder, we construct a spatial-temporal heterogeneous graph to model spatial-temporal interactions between the target pedestrian and heterogeneous traffic objects.

The key contributions of our work are threefold:

- We highlight a new direction in pedestrian behavior understanding, namely the multi-modal pedestrian action prediction. Multi-Modal Conditional Generative Module is proposed to capture the multimodality of pedestrian future actions.
- Our proposed Social Aware Encoder can jointly model spatial-temporal contextual interactions and augment the target pedestrian representation.
- To aggregate the above modules, we propose a social aware multi-modal pedestrian behavior prediction network. Experiment results on the PIE dataset and JAAD dataset show that our network achieves state-of-the-art performance on the action detection and action prediction task. On the intention estimation task, our model improve by 1.1%–5.7% based on different metrics.

2 Related Work

2.1 Pedestrian Intention Estimation

Pedestrian intention estimation plays a critical role in helping the autonomous system perform safer decisions and construct a safe urban traffic environment. Previous works [4,8,20,24] took the destination of the trajectory as pedestrian intentions and lacked a deeper semantic interpretation of pedestrian intentions. PIE [10] extracted pedestrian intention from RGB images, captured temporal dynamics of intention features, and utilized intentions to guide pedestrian motion prediction. ST-DenseNet [17] proposed a real-time network that incorporates pedestrian detection, tracking, and intention estimation. They utilized YOLOv3 [15] to detect pedestrians, used SORT [21] algorithm to track them, and designed a spatial-temporal DenseNet [5] to estimate their intentions. SF-GRU [13] collected visual inputs from pedestrians and their surrounding scenes. Then they gradually concatenated inputs and fused them into the stacked RNNs. Kotseruba *et al.* [6] analyzed the influence of the pedestrian self and the environment on intentions. They evaluate the impact of the gaze, location, orientation, and interaction of pedestrians. And locations of designated crosswalks and curbs in the environment are also utilized to estimate pedestrian intention. They combined the influences of these factors and used a logistic regression classifier to infer pedestrian intentions. FuSSI-Net [9] utilized the pose estimation network to obtain human joint coordinates and designed different strategies to fuse human joint features and visual features. Liu *et al.* [7] estimated intention from the pedestrian-centric and location-centric perspectives. They constructed a pedestrian-centric graph based on pedestrian position relation and appearance relation. Then the pedestrian-centric graph is modified to the location-centric graph to predict whether there are pedestrians crossing in front of the ego-vehicle.

Considering that intentions are forward-looking, intentions will eventually be reflected in actions. They ignored the interaction between intentions and actions. Our network produces the multi-modal future action sequences under the direction of intentions which fully account for relations between pedestrian intentions and actions.

2.2 Pedestrian Behavior Prediction

In the human-robot interactive scene, uncertainties of pedestrian behaviors are an important challenge for autonomous systems. Pedestrian behavior understanding can provide a reasonable inference of pedestrian behaviors which is beneficial for robot navigation.

MMH-PAP [14] utilized LSTM to integrate temporal dynamics of visual features and position features and applied an attention mechanism to generate weighted representations of each input. Graph-SIM [23] proposed a pedestrian action prediction network based on graph neural networks. Given their locations and speeds, they clustered road users and assigned importance weights to relations with the target pedestrian. CIR-Net [2] paid particular attention to the relation between pedestrian actions and trajectories. To couple actions and trajectories, they aligned changes in trajectories with changes in actions. The multi-task network proposed by [22] can estimate pedestrian intention and detect crossing action simultaneously. They utilized the soft attention network [3] to model spatial relations between the target pedestrian and other road components. Previous works ignored the inherent multimodality of pedestrian future actions and their deterministic models are not suitable for modeling a distribution over diverse futures. We take consideration into the uncertainty of pedestrian futures and design Multi-Modal Conditional Generative Module to produce diverse and plausible future action sequences.

3 Method

Our goal is to produce diverse and plausible action predictions of the target pedestrian P. At time t, we summarize P's current state s_t and history states $s_{t-1}, ..., s_{t-H}$ for H history time steps as input $X_t = s_{(t-H:t)}$. There is also additional surrounding information S_t of P, including the position and type of other traffic objects, and the ego-vehicle's motion to which autonomous systems have access. Given X_t and S_t, we aim to predict the target pedestrian's actions $Y_t = \widehat{A}_{(t+1:t+F)}$ for the future F time steps, which is referred to $P(Y_t|X_t, S_t)$.

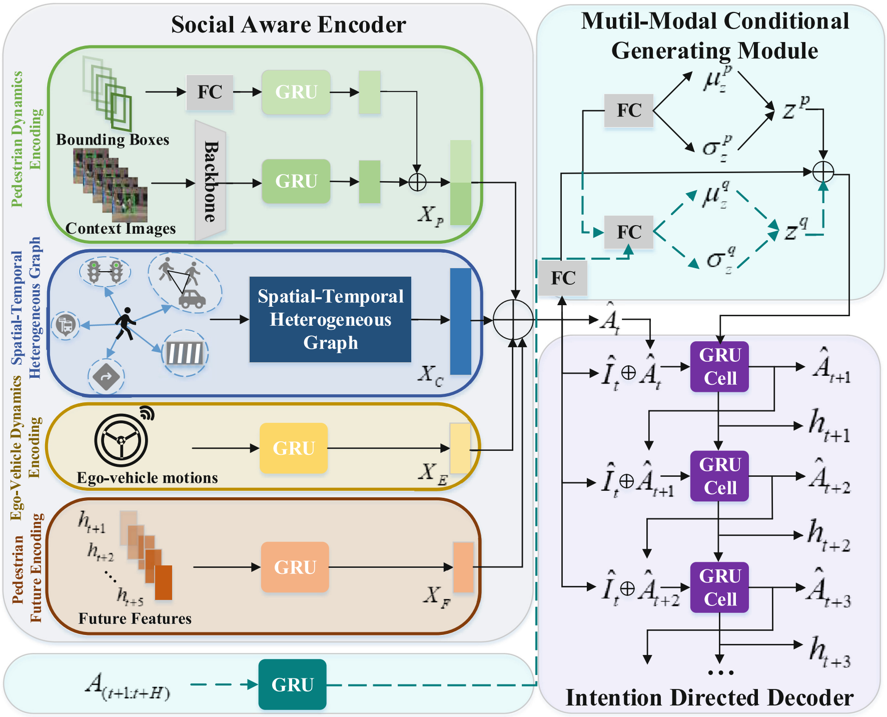

Fig. 2. Overview of our social-aware multi-modal pedestrian behavior prediction network. Our proposed network contains 3 parts, namely Social Aware Encoder, Multi-Modal Conditional Generative Module, and Intention Directed Decoder. Social Aware Encoder is composed of four modules: Pedestrian Dynamics Encoding Module (green), Spatial-Temporal Heterogeneous Graph Module (blue), Ego-Vehicle Dynamics Encoding Module (yellow), and Pedestrian Future Fusion Module (brown). $\bigoplus$ denotes the concatenation operation. (Color figure online)

3.1 Social Aware Encoder

Considering the target pedestrian and the surrounding environment, we proposed Social Aware Encoder which is composed of four modules in parallel, including Pedestrian Dynamics Encoding, Spatial-Temporal Heterogeneous Graph, Ego-vehicle Dynamics Encoding, and Pedestrian Future Fusion, as shown in Fig. 2. We explore the temporal dynamics of the target pedestrian P to enhance the individual representation X_P in the Pedestrian Dynamics Encoding (PDE). Simultaneously, Spatial-Temporal Heterogeneous Graph (STHG) models pairwise relation between P with heterogeneous traffic objects to obtain context encoded feature X_C. Meanwhile, Ego-Vehicle Dynamics Encoding (EVDE) considers decision making of autonomous systems and models the temporal dependency X_E of the ego-vehicle's motion. We aggregate future features of P to capture rich latent information X_F from the future in the Pedestrian Future Fusion (PFF).

We concatenate X_P, X_C, X_E and X_F and apply a fully-connected layer to obtain the social-aware feature X_S. Finally, an action classifier and an intent classifier are applied on X_S to estimate the current intention $\widehat{I}_t$ and the current action $\widehat{A}_t$, respectively.

Pedestrian Dynamics Encoding (PDE). This module models temporal dynamics of the current and history states of the target pedestrian. We note that the current state lacks temporal context, so we merge the information of the history states in the temporal domain. The states $s_{(t-H:t)}$ contain position features $pos_{(t-H:t)}$ and visual features $v_{(t-H:t)}$. Position features are the embedded bounding box of the target pedestrian through a fully-connected layer. Visual features are captured by pretrained CNN backbone network [18] on the cropped image patch which includes the pedestrian and the surrounding environment. These two features have different semantic attributes, so we model their temporal dynamics separately. We input these feature sequences into the corresponding GRU network to enrich them with temporal dynamical evolution clues. Outputs of two GRU networks are concatenated to obtain the Pedestrian Dynamical Feature X_P which aggregates temporal dynamics from multiple inputs.

Spatial-Temporal Heterogeneous Graph (STHG). We propose the Spatial-Temporal Heterogeneous Graph to capture other traffic objects' influence on the target pedestrian. There are various types of traffic objects, so their inherent heterogeneity can not be ignored. For example, traffic objects of the same type have the same semantic attributes and relations with the target person. Hence, we regard traffic objects in the traffic scenario as a heterogeneous multi-instance system. Taking account into the inherent heterogeneity of traffic objects, all traffic objects of the same type are aggregated in the local graph to capture intra-type interaction. Then we model the high-order relation between the target pedestrian and traffic objects of the same type in the global graph.

We construct a local graph $\mathcal{G}^c = (\mathcal{V}^c, \mathcal{E}^c)$ for each type $c \in C$, where the type set C includes 5 types, namely traffic neighbors (pedestrians, cyclists, vehicles, *etc.*), traffic signs, traffic lights, stations, and crosswalks. $\mathcal{V}^c = \{v_1, v_2, ..., v_{N_c}\}$ contains traffic objects of type c at the same time step. $v_i = \{x_{tl}, y_{tl}, x_{br}, y_{br}, c\}$ represents the bounding box coordinates and the type of v_i. In the edge set $\mathcal{E}^c$, we allow traffic objects belonging to the same type to connect with each other. We propose an information propagation mechanism on the local graph $\mathcal{G}^c$ to aggregate contextual information from the neighbor objects:

$$v_c^{(l)} = Avgpool\left(\phi_c(v_1^{(l)}), \dots, \phi_c(v_{N_c}^{(l)})\right) \tag{1}$$

$$v_i^{(l+1)} = concat\left(\phi_c(v_i^{(l)}), v_c^{(l)}\right) \tag{2}$$

where $v_i^{(l)}$ is the feature of object i in the l-th layer of the local graph, $v_c^{(l)}$ is the spatial interaction feature of the type c, and $v_i^{(0)}$ is the initial feature v_i.

ϕ_c is the fully connected layer and embeds the object feature of type c into the high-dimensional spaces. Average pooling is utilized to propagate spatial context information from neighbor objects. We stack multiple layers of the local graph $\mathcal{G}^c$ to fuse features of different objects in the different layers.

In the global graph, we integrate the interaction between the target pedestrian and traffic objects of type c into the global feature X_c. We note that the attention mechanism can assign an adaptive weight to each object and understand underlying relations better. Hence, we utilize the attention mechanism to calculate traffic objects' weights of type c based on relations with target pedestrian:

$$A_i = \frac{\exp(\theta_p(v_P)^T \theta_q(v_i)/\sqrt{d_\theta})}{\sum_{j=1}^{N_c} \exp(\theta_p(v_P)^T \theta_q(v_j)/\sqrt{d_\theta})} \tag{3}$$

$$X_c = \sum_{i=1}^{N_c} A_i \cdot v_i \tag{4}$$

where $v_i \in \mathcal{V}^c$, v_P is the target pedestrian feature, and θ_p and θ_q embed them into the d_θ-dimension space.

Previous works considered only spatial interaction [22]. Different from them, we capture the spatial-temporal dynamics integrally to build the bridge between the spatial and temporal context. We utilize GRU to aggregate contextual information of X_c in the temporal domain to obtain the final context encoded feature.

$$X_C = concat\left(\{GRU_c(X_c), c \in C\}\right) \tag{5}$$

Finally, we obtain the Context Encoded Feature X_C which aggregates spatial-temporal dependency of relations between the target pedestrian and heterogeneous traffic objects.

Ego-Vehicle Dynamics Encoding (EVDE). Interactions between pedestrians and the ego-vehicle contain important latent information. For example, when observing the crossing action of the pedestrian in front of the ego-vehicle, the controller will slow down or make way accordingly. Similarly, the pedestrian observes the motion tendency of the ego-vehicle and then may wait for the ego-vehicle to pass. As a result, we take into account ego-vehicle future motion plans and infer the pedestrian intention and action from them. We consider that the attribute of ego-vehicle motion is different from other road users, so this module utilizes a separate bi-directional GRU to encode the ego-vehicle future motions $M_{E(t+1:t+F)}$, including speed, acceleration, yaw rate, and yaw acceleration. We refer to the output of the GRU as Ego-vehicle Dynamical Feature X_E which integrates the bi-directional long-term dependency of the future motion plan.

Pedestrian Future Fusion (PFF). Considering that future actions can reflect pedestrian intentions, we collect future features of the target pedestrian, namely hidden states h of GRU cells in the Intention-Directed Decoder for the future F time steps. Future features for the future F time steps are denoted as $H_F = h_{(t+1:t+F)}$. We apply a bi-directional GRU to model the temporal dynamics of the target pedestrian's future features. In a bi-directional GRU, information flows in forward and backward two directions, so bi-directional information can be integrated. We obtain the last hidden state of the GRU and refer to it as Pedestrian Future Features X_F.

3.2 Multi-modal Conditional Generating Module

There are multiple uncertainties in action prediction since the prediction is inherently probabilistic. For example, when the pedestrian faces the crosswalk, the pedestrian may cross the road, wait for a red light or wait for other vehicles to pass. Hence, under the same situation, there will be diverse future scenarios where each one contains a reasonable explanation. A deterministic function that projects one input to one output may not have the adequate capacity to represent the diverse latent space. To learn a one-to-many mapping, we adopt a deep conditional generative model, conditional variational auto-encoder [19].

This module contains the prior network $P_\nu(z|X_t)$, and the recognition network $Q_\phi(z|X_t, Y_t)$. ϕ, ν refer to the weights of corresponding networks. The latent variables z play a critical role in modeling the inherent multimodality of the pedestrian future. In the training stage, we utilize a bi-directional GRU to encode the ground truth future actions $A_{(t+1:t+F)}$ obtaining Y_t. $Q_\phi(z|X_t, Y_t)$ takes X_t and Y_t as inputs to predict the mean μ_z^q and covariance σ_z^q and then sample the latent variable z_q from $N(\mu_z^q, \sigma_z^q)$. The goal of Q_ϕ is to learn a distribution from observations and ground truth to the latent variable z_q. With no prior knowledge of ground truth future actions, $P_\nu(z|X_t)$ predicts the mean μ_z^p and covariance σ_z^p based on observations. Similarly, the latent variable z^p is sampled from the distribution $N(\mu_z^p, \sigma_z^p)$. We optimize Kullback-Leibler divergence between $N(\mu_z^q, \sigma_z^q)$ and $N(\mu_z^p, \sigma_z^p)$ to make z_p produced by prior network learn the distribution modeled by recognition network. During the training stage, the latent variables z are sampled from the recognition network and then concatenated with X_S from the encoder. In the testing stage, we sample z from the prior network because we can not get access to ground truth future actions.

3.3 Intention-Directed Decoder

The intention is prospective and action can be described as following along line to attain a certain intention. Under the guidance of intentions, generation network $P_\theta(Y_t|X_t, \widehat{I}_t, z)$ is utilized to model the diversity of future actions. Multiple latent variables z are used to generate future actions with multiple modes. And pedestrian intention $\widehat{I}_t$ provides a direction for the development of future actions. Considering that future actions can be regarded as a temporal sequence, we adopt GRU cells to construct the Intention-Directed Decoder. The input of the decoder is the concatenation of estimated intention $\widehat{I}_t$ and the predicted action at the last prediction time step. Under the guidance of pedestrian intention, the decoder iteratively predicts future actions. And we collect the hidden states $h_{(t+1:t+F)}$ of the GRU cell and pass them to Social Aware Encoder to extract important information in the future.

3.4 Multi-task Loss

The loss function of our model contains Kullback-Leibler divergence (KLD) between Q_ϕ and P_ν, binary cross entropy loss $\mathcal{L}_I$ for intention estimation at time step t, cross entropy loss $\mathcal{L}_A$ for action detection at time step t and action prediction for future time steps from $t+1$ to $t+F$:

$$\begin{aligned}\mathcal{L} = \sum_{t=1}^{T}(&\lambda_1 KLD(Q_\phi(z|X,Y), P_\nu(z|X)) + \lambda_2 \mathcal{L}_I(\widehat{I}_t, I_t) \\ &+\lambda_3 \mathcal{L}_A(\widehat{A}_t, A_t) + \lambda_4 \mathcal{L}_A(\widehat{A}_{(t+1:t+F)}, A_{(t+1:t+F)}))\end{aligned} \tag{6}$$

where T is the total sample length, $\widehat{I}$ and $\widehat{A}$ are the predicted value of intention and action, and I and A are the ground truth of intention and action. $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are utilized to balance multiple tasks.

4 Experiments

We conduct experiments on the PIE and JAAD datasets under the original data setting and the 'time to event' setting to verify the effectiveness of our pedestrian behavior prediction model. Firstly, we introduce experiment settings, including datasets, implementation details, data sampling strategies, and evaluation metrics. Secondly, we perform quantitative experiments to compare our model with the state-of-the-art methods and then demonstrate the effectiveness of our proposed modules through the ablation study. Finally, we present qualitative experiments to visualize the efficiency of our proposed modules.

4.1 Experiment Settings

Datasets. There are two publicly available naturalistic pedestrian behavior datasets, namely Pedestrian Intention Estimation (PIE) [10] and Joint Attention in Autonomous Dataset (JAAD) [11,12].

PIE contains 6 h of driving videos under the first-person perspective which are shot by a monocular dashboard camera. The dataset provides annotations of 1842 pedestrians and traffic objects appearing at the same time. Pedestrian annotations contain the bounding box coordinates, intentions, and actions. Traffic objects consist of cyclists, vehicles, signs, traffic lights, crosswalks, and stations. Annotations of traffic objects contain the bounding box and the type. In addition, the ego-vehicle motion is also captured by the onboard diagnostics sensor. We follow [22] to split 1842 pedestrians into 880 for training, 243 for validation, and 719 for testing.

JAAD provides 300 video clips of 686 pedestrians which range from 5 to 15 s. The dataset also collects the video from the first-person perspective. There are annotations of pedestrians, including the bounding box, intention, and action labels. In addition, the dataset provides contextual tags which contain the traffic light state and the sign type. The ego-vehicle motion state is also accessible. We follow [22] to split 686 pedestrians into 188 for training, 32 for validation, and 126 for testing.

Implementation Details. The image patches in the input contain the pedestrian and the surrounding environment. We follow [10] to expand the bounding box to twice its original size and crop the image based on the enlarged bounding box. The VGG-16 [18] pretrained on ImageNet [16] is used as the backbone network to extract the feature map of the image patch. In Multi-Modal Conditional Generative Module, we sample $K = 20$ latent variables z in the prior network and recognition network. Inspired by [22], we expand 2 action labels (*walking* and *standing*) to 7 semantic action labels (*standing, waiting, going towards the crossing point, crossing, crossed and standing, crossed and waiting,* and *other walking*). And the pedestrian intention consists of two binary labels, namely *crossing* and *not-crossing.* At each frame, we detect the intention and action of the current frame and predict actions for the future $F = 5$ frames. To balance the loss of multiple tasks, we set $\lambda_1, \lambda_2, \lambda_3$, and λ_4 as 1. To train our model, we set the batch size as 64, learning rate as 10^{-5}, and adopt RMSprop optimizer with $\alpha = 0.9, \epsilon = 10^{-7}$. In the testing stage, we consider the inconsistent length of trajectories, so we input one trajectory at each testing time. We take the future action sequence which appears most in K predicted sequences as the final prediction result. We collect action prediction results at each time step of the trajectory to calculate mAP of action prediction.

Data Sampling Strategy. We adopt two data sampling strategies that previous works widely used, and we list the results under the two strategies: (1) the PIE strategy: all of the original data was utilized in the PIE [10]. In the training stage, we sample the trajectories which are truncated to the length T to make batch training feasible, where we set T as 15 and 30. (2) the time to event (TTE) strategy: During training, we follow [13,14] to sample trajectories with 1–2 s before the crossing event.

Evaluation Metrics. Pedestrian actions contain multiple labels, so we utilize mAP to evaluate the results of action prediction and detection. Pedestrian intention estimation is a binary classification problem, so we adopt accuracy, F1 score, precision, and area under curve (AUC) as evaluation metrics.

Table 1. Results of pedestrian action detection and prediction are compared with state-of-the-art methods on the PIE dataset using the PIE sampling strategy. '*' indicates the result produced by the re-implemented model. 'Det' represents pedestrian action detection and 'Pred' represents pedestrian action prediction.

Method	T = 15		T = 30	
	mAP(Det)	mAP(Pred)	mAP(Det)	mAP(Pred)
Yao *et al.* [22]	0.23*	0.22*	0.24	0.23*
Ours	**0.29**	**0.25**	**0.29**	**0.26**

Table 2. Results of pedestrian intention estimation are compared with state-of-the-art methods. 'PIE(PIE)', 'PIE(TTE)', 'JAAD(TTE)' in the first row indicate tested on the PIE dataset using the PIE sampling strategy, on the PIE dataset using the time to event (TTE) sampling strategy, on the JAAD dataset using the TTE sampling strategy, respectively.

Method	PIE(PIE)				PIE(TTE)				JAAD(TTE)			
	Acc	F1	Prec	AUC	Acc	F1	Prec	AUC	Acc	F1	Prec	AUC
I3D [1]	–	–	–	–	0.63	0.42	0.37	0.58	0.79	0.49	0.42	0.71
PIE [10]	0.79	0.87	0.86	0.73	–	–	–	–	–	–	–	–
SF-GRU [13]	–	–	–	–	0.87	0.78	0.74	0.85	0.83	0.59	0.50	0.79
MMH-PAP [14]	–	–	–	–	**0.89**	0.81	0.77	0.88	0.84	0.62	0.54	0.8
Yao *et al.* [22]	0.82	0.88	**0.94**	0.83	0.84	0.90	**0.96**	0.88	0.87	0.70	0.66	**0.92**
Ours	**0.85**	**0.91**	0.92	**0.87**	0.85	**0.91**	0.93	**0.89**	**0.88**	**0.74**	**0.67**	0.91

4.2 Performance Evaluation and Analysis

Quantitative Results on Pedestrian Action Detection and Prediction. To verify the effectiveness of our model on the pedestrian action detection and prediction tasks, we compare our model with state-of-art models on the PIE dataset using the PIE sampling strategy in Table 1. Our model surpasses the previous works by a good margin on pedestrian action detection and prediction under $T = 15$ and $T = 30$ sampling length. Compared with [22], our model captures the spatial-temporal interaction between the target pedestrian with other traffic objects. The uncertainty of action prediction is also modeled in Conditional Multi-Modal Generating Model to sample multi-modal future action sequences. It is clear that our model incorporates the social aware context and the multimodality of pedestrian futures, which significantly improves the effectiveness of the pedestrian detection and prediction task.

Quantitative Results on Pedestrian Intention Estimation. In Table 2, we also conduct pedestrian intention estimation experiments on the PIE and JAAD datasets adopting different sampling strategies. On the PIE dataset under the PIE sampling strategy, our model surpasses the state-of-the-art method [22] and achieves the best 0.85 accuracy, 0.91 F1 score, and 0.87 AUC. F1 score

Table 3. Ablation study of Spatial-Temporal Heterogeneous Graph module on the PIE dataset using the PIE sampling strategy and sampling length $T = 30$. The third row shows the result without traffic objects, namely removing the STHG module in our model. Row 4–8 show the effect of using one type of traffic object alone. The last row is the result with all traffic object types.

Traffic objects					Action		Intention			
Traffic neighbor	Crosswalk	Traffic light	Traffic sign	Station	mAP(Det)	mAP(Pred)	Acc	F1	Prec	AUC
–	–	–	–	–	0.24	0.17	0.78	0.87	0.86	0.72
✓	–	–	–	–	0.25	0.21	0.80	0.88	0.88	0.75
–	✓	–	–	–	0.26	0.21	0.83	0.89	0.89	0.82
–	–	✓	–	–	0.27	0.22	0.84	0.90	0.90	0.82
–	–	–	✓	–	0.25	0.21	0.81	0.88	0.88	0.76
–	–	–	–	✓	0.26	0.21	0.81	0.88	0.88	0.77
✓	✓	✓	✓	✓	**0.29**	**0.26**	**0.85**	**0.91**	**0.92**	**0.87**

Table 4. Ablation study of the local graph in Saptial-Temporal Heterogeneous Graph on the PIE dataset using the PIE sampling strategy and sampling length $T = 30$.

Layers of local graphs	mAP(Det)	mAP(Pred)	Acc	F1	Prec	AUC
L = 1	0.27	0.23	0.84	0.90	0.91	0.86
L = 2	0.28	0.24	**0.85**	**0.91**	0.91	0.87
L = 3	**0.29**	**0.26**	**0.85**	**0.91**	**0.92**	**0.87**
L = 4	0.28	0.24	**0.85**	0.90	**0.92**	0.86

and AUC metrics are also increased by our model on the PIE dataset using the TTE strategy. On the JAAD dataset with the TTE sampling strategy, our model outperforms previous methods by 1.1% to 5.7% on the multiple metrics. Experiment results on intention estimation demonstrate the effectiveness of our multi-task model.

Ablation Study on Spatial-Temporal Heterogeneous Graph. To explore the impact of traffic objects belonging to different types, we conduct the ablation study on the STHG module. Results in Table 3 show that adding any type of traffic object enhances the effectiveness of the STHG module. Among them, crosswalks and traffic lights both make a significant performance boost. We consider that the crosswalk determines where pedestrians will cross, and the traffic light determines when pedestrians can cross the road. Results also demonstrate that our model can mimic human perception to observe the surrounding environment and extract useful information. In addition, we investigate the impact of different layers of the local graph in the STHG module in Table 4. As the number of layers increases, the effectiveness on multiple metrics will improve. And the performance achieves the best when the number of layers is three.

Table 5. Ablation study on Multi-Modal Conditional Generative Module (MMCGM) on the PIE dataset using PIE sampling strategy and sampling length $T = 30$.

MMCGM	Action		Intention			
	mAP(Det)	mAP(Pred)	Acc	F1	Prec	AUC
K = 1	0.28	0.24	**0.85**	**0.91**	0.91	**0.87**
K = 20	**0.29**	**0.26**	**0.85**	**0.91**	**0.92**	**0.87**

Ablation Study on Multi-modal Conditional Generating Module. As shown in Table 5, we conduct the ablation study on Multi-Modal Conditional Generating Module (MMCGM). The performance of the deterministic form ($K = 1$) on the action prediction is degraded. We consider that the deterministic form may not fully account for the inherent stochasticity of the future. We set the number of samples K as 20 to evaluate the effectiveness of the multi-modal action prediction. The result suggests that the multi-modal prediction makes a performance improvement from 0.24 mAP to 0.26 mAP. The introduction of multiple latent variables plays a critical role in approximating the one-to-many distribution to model the diversity of future actions.

4.3 Qualitative Example

Figure 3 shows the visualization of the STHG module and the predicted multi-modal action sequences. In the frame $t - 2 \sim t + 2$, the target pedestrian facing the crosswalk intends to cross the road. Observing green traffic lights for vehicles, he is waiting for crossing. At the time step t, STHG module pays attention to the traffic lights and the crosswalk. In addition, traffic neighbors in the urban scenario also deserve our attention. Traffic neighbor 1 and 2 waiting for crossing are highlighted, since the pedestrian usually observes surrounding neighbor pedestrians to infer the current scene information. STHG module also notes that there are many passing vehicles, which suggests that pedestrians should wait for vehicles to pass. At the bottom of the Fig. 3, we present multi-modal predictions of future action sequences. At the time step $t - 2$, our model (K = 20) outperforms the deterministic form (K = 1). It is clear that our multi-modal model considers the uncertainty of pedestrians and conducts probabilistic predictions. In the probabilistic outputs, we take the future action sequence which appears most as the final prediction result. The first action sequence appears 7 times and is highly likely to happen. We consider that the multi-modal model which can produce probabilistic inference is more preferable than the deterministic form in some application scenarios. At the time step t, both the multi-modal and deterministic model perform well, since the two models can capture the useful spatial-temporal interactions and infer the correct future actions from them.

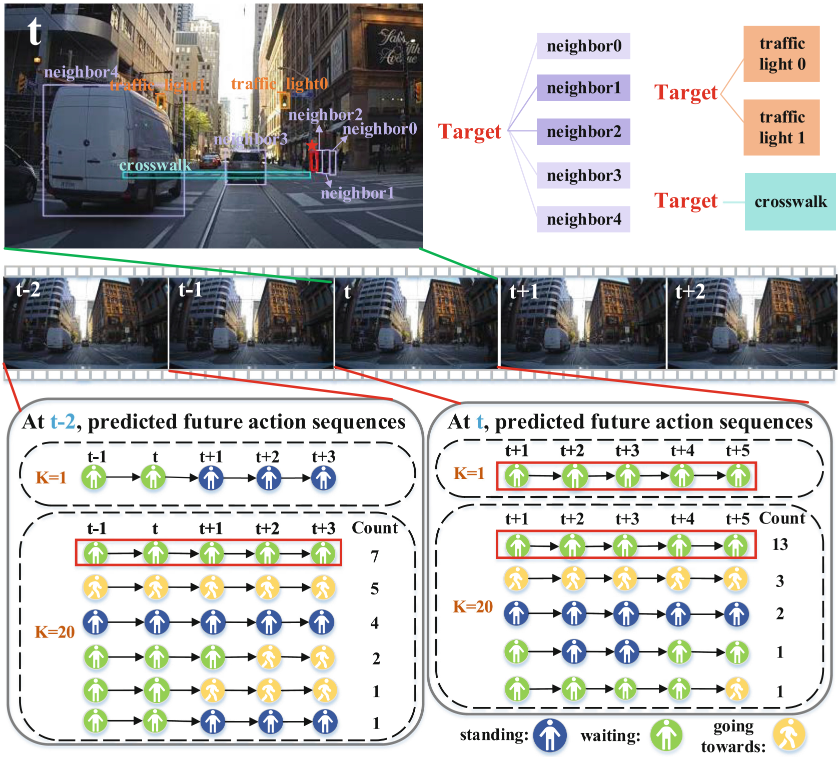

Fig. 3. At the top of the figure, we present the visualization of attention matrices of Spatial-Temporal Heterogeneous Graph Module at the t frame. The bottom of the figure is the visualization of the deterministic output ($K = 1$) and the probabilistic output ($K = 20$). Red bounding boxes indicate the correctly predicted action sequence. 'Count' represents the frequency the action sequence appeared. (Color figure online)

5 Conclusion

In this work, we propose a social aware multi-modal pedestrian behavior prediction network. In the Social Aware Encoder, we capture the spatial-temporal interaction between the target pedestrian and heterogeneous traffic objects. Multi-Modal Conditional Generative Module is designed to model the inherent uncertainties of future action sequences. Experiments demonstrate that our model outperforms previous methods on multiple tasks on the PIE and JAAD datasets using multiple metrics. The comprehensive ablation study and visualization also verify the effectiveness of our proposed modules.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
2. Chen, B., Li, D., He, Y.: Simultaneous prediction of pedestrian trajectory and actions based on context information iterative reasoning. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1007–1014. IEEE (2021)
3. Chen, L., et al.: SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5659–5667 (2017)
4. Gu, J., Sun, C., Zhao, H.: DenseTNT: end-to-end trajectory prediction from dense goal sets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15303–15312 (2021)
5. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 2261–2269. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.243
6. Kotseruba, I., Rasouli, A., Tsotsos, J.K.: Do they want to cross? Understanding pedestrian intention for behavior prediction. In: 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 1688–1693. IEEE (2020)
7. Liu, B., et al.: Spatiotemporal relationship reasoning for pedestrian intent prediction. IEEE Robot. Autom. Lett. **5**(2), 3485–3492 (2020)
8. Mangalam, K., et al.: It is not the journey but the destination: endpoint conditioned trajectory prediction. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 759–776. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_45
9. Piccoli, F., et al.: FuSSi-Net: fusion of spatio-temporal skeletons for intention prediction network. In: 2020 54th Asilomar Conference on Signals, Systems, and Computers, pp. 68–72. IEEE (2020)
10. Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J.K.: 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 (2019)
11. Rasouli, A., Kotseruba, I., Tsotsos, J.K.: Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior. In: ICCVW, pp. 206–213 (2017)
12. Rasouli, A., Kotseruba, I., Tsotsos, J.K.: It's not all about size: on the role of data properties in pedestrian detection. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11129, pp. 210–225. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11009-3_12
13. Rasouli, A., Kotseruba, I., Tsotsos, J.K.: Pedestrian action anticipation using contextual feature fusion in stacked RNNs. arXiv preprint arXiv:2005.06582 (2020)
14. Rasouli, A., Yau, T., Rohani, M., Luo, J.: Multi-modal hybrid architecture for pedestrian action prediction. In: 2022 IEEE Intelligent Vehicles Symposium (IV), pp. 91–97. IEEE (2022)
15. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. CoRR abs/1804.02767 (2018). http://arxiv.org/abs/1804.02767

16. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
17. Saleh, K., Hossny, M., Nahavandi, S.: Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal DenseNet. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 9704–9710. IEEE (2019)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
19. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
20. Wang, C., Wang, Y., Xu, M., Crandall, D.: Stepwise goal-driven networks for trajectory prediction. IEEE Robot. Autom. Lett. **7**(2), 2716–2723 (2022)
21. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing, ICIP 2017, Beijing, China, 17–20 September 2017, pp. 3645–3649. IEEE (2017). https://doi.org/10.1109/ICIP.2017.8296962
22. Yao, Y., Atkins, E., Roberson, M.J., Vasudevan, R., Du, X.: Coupling intent and action for pedestrian crossing behavior prediction. arXiv preprint arXiv:2105.04133 (2021)
23. Yau, T., Malekmohammadi, S., Rasouli, A., Lakner, P., Rohani, M., Luo, J.: Graph-SIM: a graph-based spatiotemporal interaction modelling for pedestrian action prediction. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 8580–8586. IEEE (2021)
24. Zhao, H., et al.: TNT: target-driven trajectory prediction. arXiv preprint arXiv:2008.08294 (2020)