J. Shanghai Jiao Tong Univ. (Sci.), 2023, 28(1): 20-27
https://doi.org/10.1007/s12204-023-2565-3

Action-Aware Encoder-Decoder Network for Pedestrian Trajectory Prediction

*FU Jiawei** (傅家威), *ZHAO Xu* (赵　旭)
(Department of Automation, School of Electronic Information and Electrical Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Accurate pedestrian trajectory predictions are critical in self-driving systems, as they are fundamental to the response- and decision-making of ego vehicles. In this study, we focus on the problem of predicting the future trajectory of pedestrians from a first-person perspective. Most existing trajectory prediction methods from the first-person view copy the bird's-eye view, neglecting the differences between the two. To this end, we clarify the differences between the two views and highlight the importance of action-aware trajectory prediction in the first-person view. We propose a new action-aware network based on an encoder-decoder framework with an action prediction and a goal estimation branch at the end of the encoder. In the decoder part, bidirectional long short-term memory (Bi-LSTM) blocks are adopted to generate the ultimate prediction of pedestrians' future trajectories. Our method was evaluated on a public dataset and achieved a competitive performance, compared with other approaches. An ablation study demonstrates the effectiveness of the action prediction branch.

Key words: pedestrian trajectory prediction, first-person view, action prediction, encoder-decoder, bidirectional long short-term memory (Bi-LSTM)
CLC number: TP 391.4　　**Document code:** A

0 Introduction

Predicting the future behavior and trajectory of agents (individuals, vehicles, etc.) is critical in many applications[1], particularly in fields such as automotive systems and robotics. Owing to the continuous progress in autonomous driving systems and computer vision[2-4], trajectory prediction has drawn considerable attention from researchers. In the autonomous driving field, understanding and predicting future trajectories is a crucial component of perception, which is the basis of path planning and decision-making[5-6].

Pedestrian trajectory prediction typically involves two types of datasets: first-person and bird's-eye view datasets. For first-person view datasets, such as the pedestrian intention estimation (PIE) dataset[7] and joint attention in autonomous driving (JAAD) dataset[8], scenes were captured using an on-board camera moving along with the vehicle. For bird's-eye view datasets, such as Eidgenössische Technische Hochschule Zürich (ETH) pedestrian dataset[9] and University of Cyprus (UCY) multi-person trajectory dataset[10], scenes were captured in public spaces from overhead.

Most existing pedestrian trajectory prediction methods focus on bird's-eye view datasets[11-14]. However, in practice, only the first-person view can be applied to autonomous driving systems. In this study, we focus on the problem of predicting the future trajectory of pedestrians in the first-person view. Currently, some research on trajectory prediction in the first-person view ignores that, in addition to the motion of pedestrians, the constantly changing viewpoint of the camera also affects the 2D position of pedestrians[15-16]. Even for standing pedestrians, their bounding boxes appear at different positions in different frames owing to the motion of the ego vehicle. In general, the source of the shift of bounding boxes is composed of two elements: pedestrian movement and ego-vehicle motion. Evidently, the trajectories of standing pedestrians are more affected by the ego-vehicle's motion, and therefore, the trajectories for pedestrians walking are more affected by their own movement. Therefore, the performance of trajectory prediction may be improved if we predict the action (standing/walking) first, and use it as a weight between those two facts for the bounding box shift. Moreover, recent research has shown that trajectory prediction is improved if the goal, which is the bounding box in the last frame we must predict, is predicted first[17-19].

To this end, we propose an action-aware network for bridge action and trajectory prediction. The

Received: 2022-02-28　**Accepted:** 2022-04-19

***E-mail:** qqniaolong@sjtu.edu.cn

entire network is based on an encoder-decoder architecture. The input of the entire network includes three components: bounding boxes, ego-vehicle motion, and pedestrian pose estimation. The long short-term memory (LSTM) blocks encode the inputs into a vector[20], along with the action and goal prediction. For the decoder part, bidirectional long short-term memory (Bi-LSTM) is used to predict the future trajectory with the help of goal prediction, where the action prediction functions generate weights for the ego-vehicle motion, which is the input of the Bi-LSTM.

In summary, the contributions of this study can be summarized as follows: we propose a framework to realize pedestrian trajectory and action prediction, and we show that action prediction can help offer a weight between pedestrian movement and ego-vehicle motion, and our method achieves a competitive performance compared with other algorithms on the PIE dataset.

1 Related Work

1.1 Pedestrian Trajectory Prediction for Bird's-Eye View

Various studies are being conducted on the pedestrian trajectory prediction problem using static cameras (bird's-eye view). These methods tend to focus on pedestrian-to-pedestrian and pedestrian-to-vehicle interaction. Among these methods, social LSTM is a typical framework for bird's-eye view trajectory prediction[11], combining local and global information obtained by LSTM blocks with a pooling mechanism. Other models utilize generative adversarial networks (GANs) to learn the latent distribution of pedestrian trajectories based on past observations[21-23]. However, these methods fail on first-person view datasets, by neglecting the movement of the camera, that is, the ego-vehicle motion.

1.2 Pedestrian Trajectory Prediction for First-Person View

The first-person view has richer dynamic visual information of the scene than that of bird's-eye view and hence, is more practical for autonomous driving systems[24]. However, relatively less research has been conducted on the first-person view. Yao et al.[15] used a conditional variational auto-encoder (CVAE)[25-27] to learn future trajectory distributions using a stochastic latent variable with the help of a Gaussian mixture model (GMM)[28]. Their study was conducted on both bird's-eye and first-person view datasets without exploring their differences.

Recently, some studies have noticed the effect of ego-vehicle motion, attempting to disentangle the two sources of the shift of bounding boxes, that is, the actual movement of pedestrians and ego-vehicle motion. Quan et al.[29] designed a holistic LSTM block to encode the motion of a pedestrian and ego-vehicle, which is estimated from the optical flow. Neumann and Vedaldi[30] introduced a self-supervised camera-pose prediction network to predict the entire future frame, which infers the ego-vehicle motion. The pedestrian trajectory is then projected onto the predicted frame, in which the two sources are thoroughly decoupled. However, these methods attempt to solve this problem either by introducing the optical flow of the video or feeding the entire picture to the network, which considerably increases the computational complexity.

Some work has been conducted to attempt to solve this problem in 3D space using light detection and ranging (LIDAR) equipment[31-33]. The main drawback of LIDAR-based methods is their high computational complexity and low resolution for distant objects.

1.3 Goal Estimation for Trajectory Prediction

Goal estimation has proven its effectiveness in trajectory prediction. Rehder and Kloeden[18] used the estimated goal distribution as prior information in a prediction procedure based on a particle-filter method. Deo and Trivedi[34] utilized inverse reinforcement learning (IRL) to estimate the goal states that are sent to the decoder along with past trajectory encodings to generate the final prediction. Yao et al.[15] designed a bidirectional trajectory network in which goals are first predicted and then propagated back.

Following this idea, we adopted goal estimation as a branch at the end of the encoder to generate more accurate trajectories.

1.4 Action Prediction for Trajectory Prediction

The definition of an action in trajectory prediction is vague[7], including behaviors and statuses such as walking, standing, crossing the road, and not crossing the road. In terms of the impact of ego-vehicle motion and pedestrians moving on the bounding box shift, the movement of the bounding box for standing pedestrians is more easily affected by ego-vehicle motion, whereas walking people are influenced more by their own movement. Therefore, we pay more attention to the walking status of pedestrians, and we refer to the "standing/walking" behavior as "action." Fang and López[35] showed that 2D pose estimation can help recognize the behaviors of pedestrians and determine pedestrian intentions, such as crossing roads and stopping before crossing the road. Enlightened by this finding, we used 2D pose estimation in the input module to estimate pedestrian actions. We chose OpenPose as our 2D pose-estimation generator because of its simplicity and satisfactory performance[36].

2 Methodology

The target of this problem is to forecast the future pedestrian trajectory of a specific pedestrian based on observed frames and information about the ego vehicle.

We first present the formulations of this problem and then introduce the network. Figure 1 illustrates the overall architecture.

2.1 Formulations

Given a short video cut of a specific labeled pedestrian in the past t frames, trajectory prediction aims to predict the bounding box of this specific pedestrian in the subsequent τ frames. The bounding box of the ith pedestrian at the time step j can be specified by the top-left (TL) and bottom-right (BR) coordinates of a pixel:

$$b_{i,j} = \{(x_{\mathrm{TL}}, y_{\mathrm{TL}}), (x_{\mathrm{BR}}, y_{\mathrm{BR}})\}. \tag{1}$$

The observation of the bounding boxes of the ith pedestrian for t time steps can be expressed as

$$O_i^{\mathrm{B}} = \{b_{i,1}, b_{i,2}, \cdots, b_{i,t}\}. \tag{2}$$

The future sequences of the bounding boxes of the ith pedestrian from the time step t to the time step $t+\tau$ can be expressed as

$$P_i^{\mathrm{B}} = \{b_{i,t+1}, b_{i,t+2}, \cdots, b_{i,t+\tau}\}. \tag{3}$$

The motion information of the ego vehicle includes its speed, roll, yaw, pitch, and heading angle, which is denoted as m_j^{V} at the time step j, and the observation of motion information for t time steps can be expressed as

$$O_{\mathrm{VM}} = \{m_1^{\mathrm{V}}, m_2^{\mathrm{V}}, \cdots, m_t^{\mathrm{V}}\}. \tag{4}$$

The pedestrian pose information is a 36-dimensional vector for every pedestrian i in each frame, which is denoted as $p_{i,j}$ at the time step j, and the observation of pedestrian pose of the ith pedestrian for t time steps can be expressed as

$$O_i^{\mathrm{P}} = \{p_{i,1}, p_{i,2}, \cdots, p_{i,t}\}. \tag{5}$$

We denote the action at the time step j of the ith pedestrian as $a_{i,j} \in \{0, 1\}$: 0 for walking, and 1 for standing. The action sequence of the prediction time of the ith pedestrian can be denoted as

$$P_i^{\mathrm{A}} = \{a_{i,t+1}, a_{i,t+2}, \cdots, a_{i,t+\tau}\}. \tag{6}$$

2.2 Input Module and Encoder

As shown in Fig. 2, the input of our network includes three components: the observation bounding boxes, ego-vehicle's motion, and pedestrian pose estimation. The first two components are labeled in the dataset, and the pose estimation is generated by a pretrained OpenPose network. First, we used a multilayer perceptron (MLP) to resize the inputs to a shape of (t, 128) separately, which were added together and then fed to

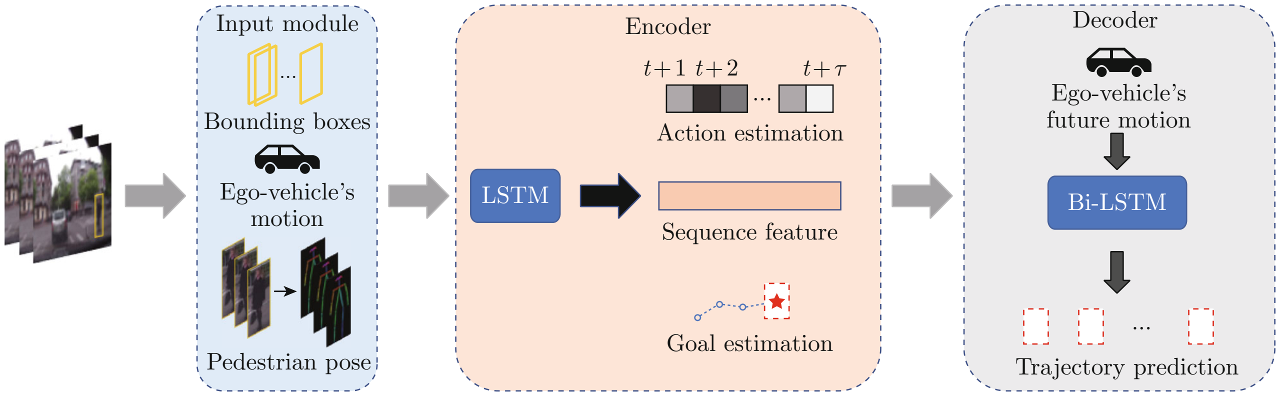


Fig. 1 Architecture overview of action-aware network

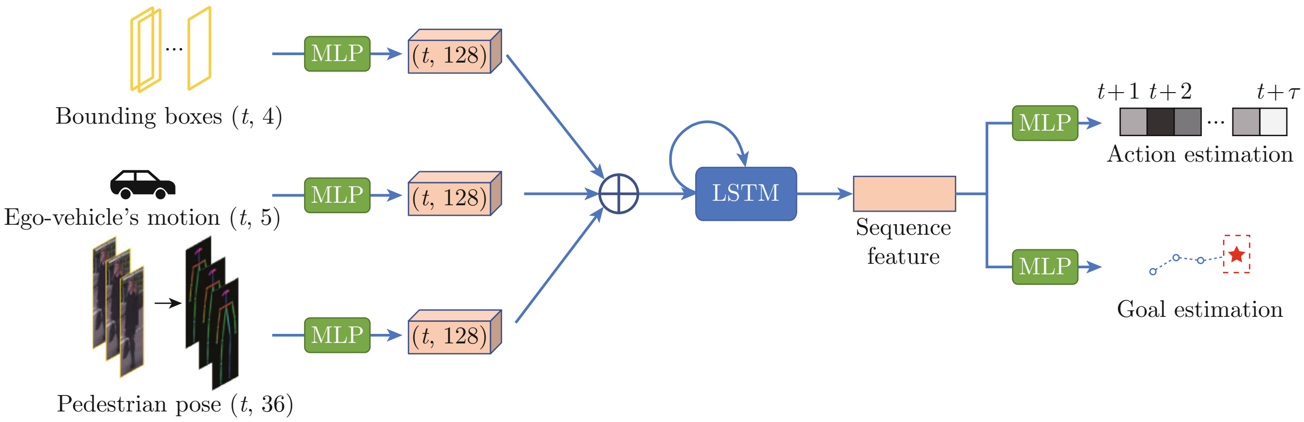


Fig. 2 Input module and network encoder

the LSTM blocks to generate the output of the encoder. The sequence feature of the ith pedestrian can be expressed as

$$f_i^{\text{seq}} = M(S(O_i^{\text{B}}) + S(O_{\text{VM}}) + S(O_i^{\text{P}})), \tag{7}$$

where $M(\cdot)$ is a 128-hidden-unit LSTM block function, and $S(\cdot)$ is a three-layer MLP function.

The feature is then sent to two branches by the MLP to predict the action of the pedestrian and goal, which is the bounding box of the last frame that needs to be predicted. The procedure can be described as follows:

$$\hat{P}_i^{\text{A}} = S(f_i^{\text{seq}}), \tag{8}$$

$$\hat{b}_{i,t+\tau} = S(f_i^{\text{seq}}), \tag{9}$$

where $\hat{P}_i^{\text{A}}$ denotes the action prediction for the ith pedestrian, and $\hat{b}_{i,t+\tau}$ denotes the goal estimation at the time step $t+\tau$.

Note that the action prediction is a vector with a length τ, where each element represents the action prediction for every time step. The element in the action prediction is a double number ranging from 0 to 1, where 0 stands for "walking," and 1 for "standing." For example, when the estimation is 0.95 for a specific pedestrian in one frame, the pedestrian tends to be "standing" based on our prediction.

2.3 Decoder

The main part of the decoder is the Bi-LSTM. The output of the encoder is the beginning of the forward LSTM, and the goal estimation is the beginning of the backward LSTM. The forward and backward LSTMs exchange information at the same time step. The input to the Bi-LSTM block originates from the motion of the ego vehicle at the corresponding time step. As discussed, the action can measure the ratio of ego-vehicle movement in the shift of bounding boxes. The action prediction is fed to a mask function before it is used as the weight of the ego-vehicle motion. Figure 3 shows the decoder procedure.

In general, the decoder procedure is expressed as

$$\hat{P}_i^{\text{B}} = M^{\text{Bi}}(O_{\text{enc}}, \hat{b}_{i,t+\tau}, \sigma(\hat{P}_i^{\text{A}}) \cdot P_{\text{VM}}), \tag{10}$$

where $\hat{P}_i^{\text{B}}$ is the prediction of the future sequences of the bounding boxes of the ith pedestrian, O_{enc} is the output of the decoder, P_{VM} is the future motion information of the ego vehicle, $\sigma(\cdot)$ is the mask function, M^{Bi} is a 128-hidden-unit Bi-LSTM block function, and "$\cdot$" represents multiplication at each time step.

To smooth the multiplication and avoid the case in which no information from the ego vehicle is used when the action prediction is 0, we chose the sigmoid function as the mask function.

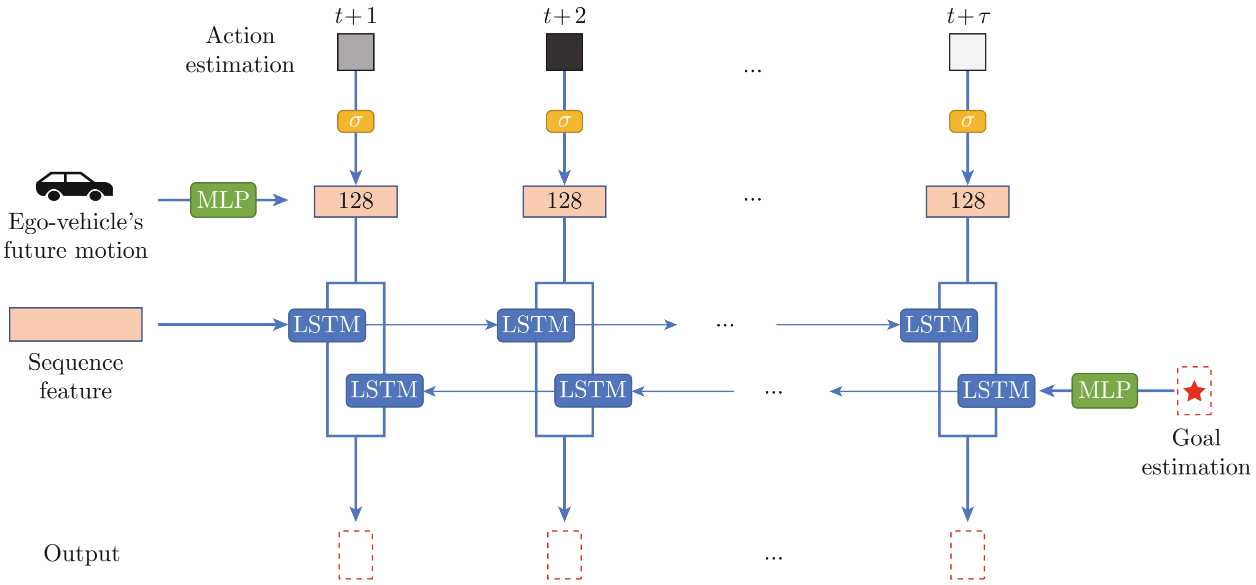


Fig. 3 Network decoder

2.4 Loss Function

In the training stage, three loss functions were adopted in the network: goal estimation, action prediction, and trajectory loss. The goal estimation loss is as follows:

$$L_{\text{goal}} = \|b_{i,t+\tau} - \hat{b}_{i,t+\tau}\|_2. \tag{11}$$

The action prediction loss is as follows:

$$L_{\text{act}} = \sum_{j=t+1}^{t+\tau} [-a_{i,j}\log(\hat{a}_{i,j}) - (1-a_{i,j})\log(1-\hat{a}_{i,j})], \tag{12}$$

where $\hat{a}_{i,j}$ is the action prediction for the ith pedestrian at the time step j.

The trajectory loss is as follows:

$$L_{\text{tra}} = \sum_{j=t+1}^{t+\tau} \|b_{i,j} - \hat{b}_{i,j}\|_2. \tag{13}$$

The entire function training loss is expressed with hyperparameters as follows:

$$L_{\text{total}} = \alpha L_{\text{goal}} + \beta L_{\text{act}} + \gamma L_{\text{tra}}, \tag{14}$$

where α, β, and γ are weights.

3 Experiments

In this section, we describe the evaluation of the proposed method. First, the dataset and its implementation details are presented. Subsequently, we discuss the evaluation experiments and ablation studies conducted.

3.1 Dataset

Recently, numerous public datasets have been used for trajectory prediction. Among them, we chose the PIE dataset to test our model because of its abundant labels on ego-vehicle motion information and pedestrian actions.

The PIE dataset is a first-person view driving dataset containing 909 480 frames, in which 293 437 frames are annotated with 1 842 pedestrians. The dataset includes more than 6 h of video footage of pedestrian, which is divided into training/testing/validation subsets by shot length at a ratio of 50 : 40 : 10. All videos were recorded at 1 920 pixel × 1 080 pixel at 30 frames per second. This is the only dataset that contains both pedestrian action labels and detailed ego-vehicle motion information, which is suitable for our network.

3.2 Implementation Details

The framework was implemented using Pytorch, and the model was trained on four GeForce GTX 1080Ti GPUs. In the training stage, we set the batch size to 64 and total epoch to 50. The model used an exponential learning rate scheduler with a learning rate of 0.001. The observation time was set to 0.5 s, which contains 15 frames, and the periods of the prediction time denoted by t_{p} were set to 0.5 s, 1.0 s, and 1.5 s, which contain 15, 30, and 45 frames, respectively. The hidden unit was set to 128 for the LSTM blocks in the encoder and decoder. To reduce the impact of uncertainties, we conducted each experiment five times and calculated the mean results as the final performance, as shown in Table 1. The hyperparameters set for the loss function are $\alpha = 1, \beta = 5$, and $\gamma = 0.2$.

3.3 Evaluation

The proposed method was evaluated based on the standard metrics: ① the mean-square errors (MSEs) of the bounding box corners in 0.5 s, 1.0 s, and 1.5 s, denoted by $X_{\text{MS},0.5}$, $X_{\text{MS},1.0}$, and $X_{\text{MS},1.5}$, respectively; ② MSE of the bounding box center over the sequence from 0 s to 1.5 s, denoted by $X_{\text{MS,C}}$; ③ MSE of the bounding box center in the last prediction frame, denoted by $X_{\text{MS,CF}}$.

In general, our method outperforms the traditional methods, as listed in Table 1. The action prediction accuracy was 92.5%. The results show that our model presents a lower error than the other models for a longer forecasting time. This may be because the relationship between pedestrian action and ego-vehicle motion may be less close when the prediction time is short. Our action-aware model can better capture the shift of the bounding box over a long time, which indicates that the relationship between the action and ego-vehicle motion becomes closer over time.

Table 1 Model performance on PIE dataset

Method	$X_{\text{MS},0.5}$/pixel	$X_{\text{MS},1.0}$/pixel	$X_{\text{MS},1.5}$/pixel	$X_{\text{MS,C}}$/pixel	$X_{\text{MS,CF}}$/pixel
Linear	233	857	2 303	1 565	6 111
LSTM	289	569	1 558	1 473	5 766
Bi-LSTM	159	539	1 535	1 447	5 615
PIE_{traj}[7]	110	399	1 280	1 183	4 780
BiTraP-D[15]	41	161	511	481	1 949
Ours	43	160	457	436	1 683

3.4 Ablation Studies

We conducted ablation experiments on the PIE dataset concerning two branches: input combination and mask function.

To prove the effect of each branch, we conducted experiments with and without each branch, and the results are presented in Table 2. First, by comparing the first and third rows, the results indicate the effectiveness of the action branch. Second, by comparing the first and second rows, the results indicate the validity of the goal estimation. Ultimately, based on the final row, the combination of these two branches shows a synergistic effect.

Table 2 Ablation experiments on different decoder inputs

Method	$X_{\mathrm{MS},0.5}$/pixel	$X_{\mathrm{MS},1.0}$/pixel	$X_{\mathrm{MS},1.5}$/pixel	$X_{\mathrm{MS,C}}$/pixel	$X_{\mathrm{MS,CF}}$/pixel
No estimation	89	523	1 329	1 028	4 764
Only goal estimation	76	320	747	546	3 064
Only action prediction	69	295	658	532	2 374
Goal + action branches	43	160	457	436	1 683

We also experimented with different input combinations. The results in Table 3 show the effectiveness of pose estimation for action prediction. The accuracy of the action prediction can be improved from 78.6% to 92.5% when a 2D pose is added to the network. In addition, we show that the combination of the speed, roll, yaw, pitch, and heading angle yields the best illustration of an ego vehicle. Moreover, the combination of the bounding box, ego-vehicle motion, and pose results in the best performance.

Finally, we used different mask functions for action prediction, and the results are shown in Table 4. We compared the uses of no functions, a linear function that maps (0, 1) to (0.5, 1) linearly, and a sigmoid function. The results indicate that the smoother the mask function, the better the model performance. The sigmoid function was the best choice among the three choices. The results satisfy our expectation because we aim to avoid the situation in which the weight for the ego-vehicle motion is 0, which indicates that no ego-vehicle motion is used in the decoder.

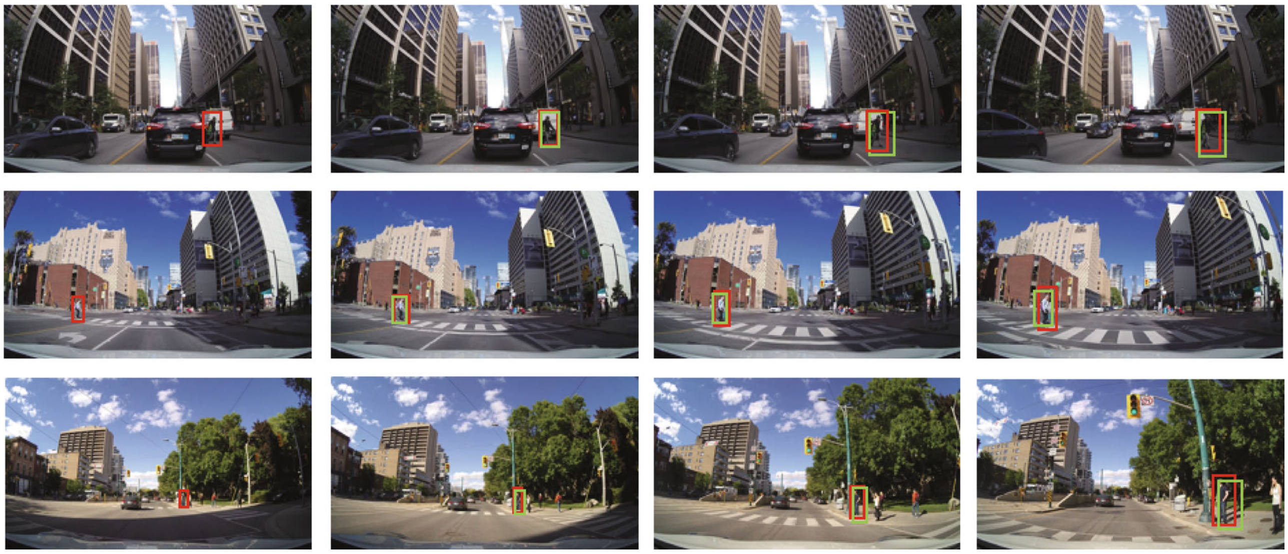
Figure 4 shows the performance on the PIE dataset,

Table 3 Ablation experiments on different inputs

Input	Is the input applied?			
Bounding box	√	√	√	√
Speed		√	√	√
Roll, yaw, pitch			√	√
Heading angle			√	√
Pose	√			√
Accuracy for action prediction/%	87.3	76.5	78.6	92.5
$X_{\mathrm{MS},1.5}$/pixel	625	571	485	457

Table 4 Ablation experiments on mask function

Function	$X_{\mathrm{MS},0.5}$/pixel	$X_{\mathrm{MS},1.0}$/pixel	$X_{\mathrm{MS},1.5}$/pixel	$X_{\mathrm{MS,C}}$/pixel	$X_{\mathrm{MS,CF}}$/pixel
No functions	78	290	660	568	2 303
Linear	46	172	460	458	1 758
Sigmoid	43	160	457	436	1 683



$t_{\mathrm{p}}=0\,\mathrm{s}$ $t_{\mathrm{p}}=0.5\,\mathrm{s}$ $t_{\mathrm{p}}=1.0\,\mathrm{s}$ $t_{\mathrm{p}}=1.5\,\mathrm{s}$

Fig. 4 Visualization examples of pedestrain trajectory predictions (red for ground truth and green for prediction)

where the pedestrians in the first and second rows are estimated as "walking" over the entire prediction time, and the pedestrian in the third row is estimated as "standing."

4 Conclusion

In this study, we propose an action-aware network for pedestrian trajectory prediction in the first-person view based on an encoder-decoder architecture with LSTM blocks. Unlike bird's-eye view datasets, the ego-vehicle motion should be considered for the movement of bounding boxes in first-person view datasets, particularly for standing pedestrians. We designed an action prediction branch and goal estimation at the end of the encoder, and designed a corresponding loss function. The main part of the decoder is a Bi-LSTM with ego-vehicle motion information as the input, where the action prediction is applied as a mask for the ego-vehicle motion.

The method is evaluated on a first-person view dataset and exhibits a competitive performance, and the ablation study demonstrates the significance of the action prediction.

References

[1] MALLA S, DARIUSH B, CHOI C. TITAN: future forecast using action priors [C]//*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA: IEEE, 2020: 11183-11193.

[2] ZHANG T L, TU H Z, QIU W. Developing high-precision maps for automated driving in China: Legal obstacles and the way to overcome them [J]. *Journal of Shanghai Jiao Tong University* (*Science*), 2021, **26**(5): 658-669.

[3] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The KITTI dataset [J]. *The International Journal of Robotics Research*, 2013, **32**(11): 1231-1237.

[4] SONG X B, WANG P, ZHOU D F, et al. Apollo-Car3D: A large 3D car instance understanding benchmark for autonomous driving [C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA: IEEE, 2019: 5447-5457.

[5] HU Y K, WANG C X, YANG M. Decision-making method of intelligent vehicles: A survey [J]. *Journal of Shanghai Jiao Tong University*, 2021, **55**(8): 1035-1048 (in Chinese).

[6] SHI Q, ZHANG J L, YANG M. Curvature adaptive control based path following for automatic driving vehicles in private area [J]. *Journal of Shanghai Jiao Tong University* (*Science*), 2021, **26**(5): 690-698.

[7] RASOULI A, KOTSERUBA I, KUNIC T, et al. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction [C]//*2019 IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE, 2019: 6261-6270.

[8] RASOULI A, KOTSERUBA I, TSOTSOS J K. Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior [C]//*2017 IEEE International Conference on Computer Vision Workshops*. Venice: IEEE, 2017: 206-213.

[9] PELLEGRINI S, ESS A, SCHINDLER K, et al. You'll never walk alone: Modeling social behavior for multi-target tracking [C]//*2009 IEEE 12th International Conference on Computer Vision*. Kyoto: IEEE, 2009: 261-268.

[10] LEAL-TAIXÉ L, FENZI M, KUZNETSOVA A, et al. Learning an image-based motion context for multiple people tracking [C]//*2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH: IEEE, 2014: 3542-3549.

[11] ALAHI A, GOEL K, RAMANATHAN V, et al. Social LSTM: Human trajectory prediction in crowded spaces [C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV: IEEE, 2016: 961-971.

[12] LIANG J W, JIANG L, NIEBLES J C, et al. Peeking into the future: Predicting future person activities and locations in videos [C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA: IEEE, 2019: 5718-5727.

[13] SIVARAMAN S, TRIVEDI M M. Dynamic probabilistic drivability maps for lane change and merge driver assistance [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2014, **15**(5): 2063-2073.

[14] LI N, YAO Y, KOLMANOVSKY I, et al. Game-theoretic modeling of multi-vehicle interactions at uncontrolled intersections [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, **23**(2): 1428-1442.

[15] YAO Y, ATKINS E, JOHNSON-ROBERSON M, et al. BiTraP: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation [J]. *IEEE Robotics and Automation Letters*, 2021, **6**(2): 1463-1470.

[16] WANG C H, WANG Y C, XU M Z, et al. Stepwise goal-driven networks for trajectory prediction [J]. *IEEE Robotics and Automation Letters*, 2022, **7**(2): 2716-2723.

[17] MANGALAM K, GIRASE H, AGARWAL S, et al. It is not the journey but the destination: Endpoint conditioned trajectory prediction [M]//Computer Vision – ECCV 2020. Cham: Springer, 2020: 759-776.

[18] REHDER E, KLOEDEN H. Goal-directed pedestrian prediction [C]//*2015 IEEE International Conference on Computer Vision Workshop*. Santiago: IEEE, 2015: 139-147.

[19] RHINEHART N, MCALLISTER R, KITANI K, et al. PRECOG: Prediction conditioned on goals in visual multi-agent settings [C]//*2019 IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE, 2019: 2821-2830.

[20] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. *Neural Computation*, 1997, **9**(8): 1735-1780.

[21] GUPTA A, JOHNSON J, LI F F, et al. Social GAN: Socially acceptable trajectories with generative adversarial networks [C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, 2018: 2255-2264.

[22] KOSARAJU V, SADEGHIAN A, MARTÍN-MARTÍN R, et al. Social-BiGAT: Multimodal trajectory forecasting using bicycle-GAN and graph attention networks [C]//*Advances in Neural Information Processing Systems*. Vancouver, BC: Neural Information Processing Systems Foundation, 2019: 137-146.

[23] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]//*Advances in Neural Information Processing Systems*. Montreal: Neural Information Processing Systems Foundation, 2014: 2672-2680.

[24] SHAFIEE N, PADIR T, ELHAMIFAR E. Introvert: Human trajectory prediction via conditional 3D attention [C]//*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, TN: IEEE, 2021: 16810-16820.

[25] DU L, DING X, LIU T, et al. Modeling event background for if-then commonsense reasoning using context-aware variational autoencoder [C]//*2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Hong Kong: Association for Computational Linguistics, 2019: 2682-2691.

[26] ZHAO T C, ZHAO R, ESKENAZI M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders [C]//*55th Annual Meeting of the Association for Computational Linguistics*. Vancouver: Association for Computational Linguistics, 2017: 654-664.

[27] SOHN K, LEE H, YAN X. Learning structured output representation using deep conditional generative models [C]//*Advances in Neural Information Processing Systems*. Montréal: Neural Information Processing Systems Foundation, 2015: 3483-3491.

[28] REYNOLDS D. Gaussian mixture models [M]//Encyclopedia of biometrics. Boston, MA: Springer, 2009: 659-663.

[29] QUAN R J, ZHU L C, WU Y, et al. Holistic LSTM for pedestrian trajectory prediction [J]. *IEEE Transactions on Image Processing*, 2021, **30**: 3229-3239.

[30] NEUMANN L, VEDALDI A. Pedestrian and ego-vehicle trajectory prediction from monocular camera [C]//*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, TN: IEEE, 2021: 10199-10207.

[31] RHINEHART N, KITANI K M, VERNAZA P. R2P2: A reparameterized pushforward policy for diverse, precise generative path forecasting [M]//Computer vision – ECCV 2018. Cham: Springer, 2018: 794-811.

[32] LI J C, MA H B, TOMIZUKA M. Conditional generative neural system for probabilistic trajectory prediction [C]//*2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Macao: IEEE, 2019: 6150-6156.

[33] CHOI C, MALLA S, PATIL A, et al. DROGON: A causal reasoning framework for future trajectory forecast [EB/OL]. (2020-11-06) [2022-04-19]. https://arxiv.org/abs/1908.00024.

[34] DEO N, TRIVEDI M M. Trajectory forecasts in unknown environments conditioned on grid-based plans [EB/OL]. (2021-04-29) [2022-04-19]. https://arxiv.org/abs/2001.00735.

[35] FANG Z J, LÓPEZ A M. Is the pedestrian going to cross? Answering by 2D pose estimation [C]//*2018 IEEE Intelligent Vehicles Symposium*. Changshu: IEEE, 2018: 1271-1276.

[36] CAO Z, SIMON T, WEI S H, et al. Realtime multi-person 2D pose estimation using part affinity fields [C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI: IEEE, 2017: 1302-1310.