Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 4, No. 3, 2023

Design of Real-time Pedestrian Trajectory Prediction System based on Jetson Xavier

Quankai Liu *, Haifeng Sang

School of Information Science and Engineering, Shenyang University of Technology, Shenyang, Liaoning 110870, China
* **Corresponding author**: Quankai Liu (Email: liuqk_sut@163.com)

Abstract: This paper presents a vehicle-mounted real-time pedestrian trajectory prediction system based on the embedded device Jetson Xavier. It achieves low-cost real-time pedestrian trajectory prediction using only the front camera of the vehicle. Firstly, pedestrian detection and tracking are implemented based on YoLoV7, while also estimating pedestrian poses and optical flow to provide multiple information sequences for the trajectory prediction network. Secondly, the pedestrian trajectory algorithm from a driver's perspective is studied, and a trajectory prediction algorithm that considers pedestrian pose, optical flow, and trajectory information is proposed. A novel multi-information fusion network is designed to better integrate multiple features. The algorithm is tested on the JAAD and PIE datasets, and the displacement errors are reduced by 6.35% and 3.28%, respectively, compared to BiTraP. Finally, the algorithm is ported to the embedded device Xavier and installed on a simulated vehicle for testing. By predicting pedestrian future trajectories based on pedestrian detection, collisions can be avoided in advance, improving the safety of autonomous driving. The proposed system has significant practical value.

Keywords: Pedestrian Trajectory Prediction; Attention Mechanism; Embedded Deployment; Automatic Driving.

1. Introduction

In recent years, with the development of the new energy vehicle industry, an increasing number of intelligent cars are equipped with autonomous driving technology. However, the safety of existing autonomous driving algorithms still lags behind fully autonomous driving. The currently mass-produced autonomous vehicles are mainly used in traffic scenarios with simple road elements and few pedestrians. There is still a need for extensive real-world testing in crowded pedestrian and complex traffic scenarios. Pedestrian trajectory prediction is crucial for collision avoidance and improving the safety of autonomous driving. Existing autonomous driving systems mostly detect pedestrians, but autonomous vehicles equipped with pedestrian trajectory prediction systems can not only detect pedestrians but also predict their future positions based on pedestrian detection, thereby avoiding collisions in advance and planning safer and more robust driving paths [1].

Since Helbing [2] proposed the use of manually crafted energy functions to guide pedestrians towards their respective destinations using attractive forces and avoid collisions using repulsive forces, this approach has been widely developed and used [3, 4, 5]. Compared to previous methods for capturing pedestrian interactions using manually crafted functions, deep learning-based methods typically employ pooling mechanisms [6, 7], attention mechanisms [8, 9], and graph neural networks [10, 11] to capture complex pedestrian interaction features. The aforementioned methods for studying pedestrian interactions are based on a top-down bird's-eye view perspective. However, from the driver's perspective, onboard cameras cannot provide a top-down scene view, making it difficult to infer social interactions between pedestrians [12]. From a human eye viewpoint, drivers can easily discern pedestrian intentions, but this judgment is not easily achievable using sensors such as cameras. By predicting pedestrian intentions, we can further estimate pedestrian trajectories and understand their next actions, thereby significantly reducing the risk of accidents. The size of pedestrians from a driver's perspective cannot be ignored, and simplifying them as a point in space and considering only their historical trajectory information can lead to significant prediction errors. Therefore, we introduce pedestrian pose information to represent pedestrian intentions. In the first-person view, the camera moves with the vehicle, so the vehicle's own motion is crucial for improving the accuracy of pedestrian trajectory prediction. Optical flow refers to the movement of the target pixel in the image caused by the movement of the object in the image or the movement of the camera in two consecutive frames of image. To this end, we introduce optical flow estimation to model the vehicle's motion. Existing pedestrian trajectory prediction algorithms are mostly trained and tested on datasets, requiring manual annotation of the datasets. There is currently no complete trajectory prediction system that can perform real-time pedestrian trajectory prediction in any traffic scenario, rather than just testing on datasets.

Therefore, the main contributions of this paper are as follows:

(1) We have designed an onboard pedestrian trajectory prediction system that can predict pedestrian future trajectories in real-time end-to-end. All algorithms are deployed on Xavier for independent operation and mounted on a simulated vehicle. The downstream motors, servos, and traffic lights are controlled through Xavier's GPIO to simulate control over the vehicle.

(2) We introduce pose information to model pedestrian intentions based on single trajectory information and introduce optical flow information to model the vehicle's own motion. A multi-information feature extraction module is used to extract pedestrian trajectory, pose, and optical flow information from the monocular camera's video feed.

(3) We have also designed a novel multi-information fusion network that efficiently combines multiple information features. The pedestrian trajectory prediction module outputs the pedestrian's future trajectory. The proposed model shows

significant improvements in prediction accuracy compared to the BiTraP model on the JAAD and PIE datasets.

2. Design of the Pedestrian Trajectory Prediction System

2.1. The Overall Framework of Pedestrian Trajectory Prediction System

The vehicle-mounted pedestrian trajectory prediction system designed in this paper runs on Jetson Xavier, utilizing the onboard camera for visual information acquisition. The trajectory prediction module extracts multiple information features from the visual information and predicts pedestrian trajectories. The onboard hardware resources of Xavier are utilized to control hardware components such as motors, steering gears, and indicators. The entire system is based on the Pytorch framework under the Linux system and utilizes opensource algorithm libraries such as OpenCV. The software architecture follows a layered, multi-threaded, and modular structure, and a UI interface is designed using PyQt.

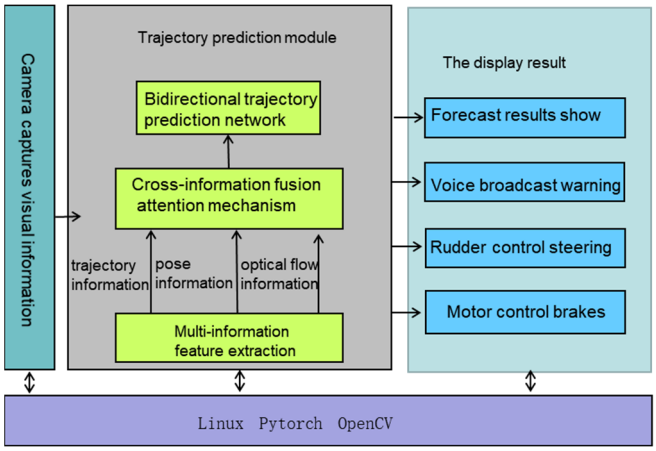

Figure 1. The overall structure of pedestrian trajectory prediction system

2.2. Multi-information Feature Extraction Network

The multi-information feature extraction network mainly extracts pedestrian observation trajectory information, pedestrian skeleton pose information and pedestrian optical flow information from the images collected by the camera. The whole network is divided into two parts: multi-information collection and multi-information coding features.

2.2.1. Pedestrian Multi-information Collection

In this paper, the prediction of pedestrian trajectory is to predict the trajectory of up to 45 (1.5s) frames in the future by observing the historical information of 15 (0.5s) frames of pedestrians. Therefore, this paper needs to collect 15 frames of historical trajectory information, attitude information, and optical flow information containing pedestrian size. Therefore, based on the target detection algorithm YoloV7 [13], combined with the pose estimation algorithm OpenPose [14] and the optical flow estimation algorithm GMFlow [15], this paper designs the Y-Pose algorithm and Y-Flow algorithm based on the improved YoloV7, so as to realize the real-time acquisition of pedestrian pose information, optical flow information and trajectory information by using the camera.

2.2.2. Pose Encoder

For pedestrians i, let's assume that their skeleton vector at a certain moment is P_t^i. The observed skeleton information can be represented as $P_{obs}^i = [P_{t-t_{obs}+1}, \ldots, P_t^i]$, where t_{obs} represents the observation time. The pedestrian's pose can be represented by the skeleton information. To encode the pedestrian's pose information, a Gated Recurrent Unit (GRU) is used, resulting in encoded features P_e.

$$P_e^i = f_{GRU}(P_{obs}^i, W_p) \tag{1}$$

Where P_{obs}^i and P_e^i are the observed pose sequence and pose encoding features of the pedestrian, respectively. f_{GRU} is the GRU function, W_p is learnable parameter matrix.

2.2.3. Optical Flow Encoder

For pedestrians i, let's assume that their optical flow vectors at a certain moment and the previous moment are F_t^i. The observed optical flow sequence can be represented as $F_{obs}^i = [F_{t-t_{obs}}, \ldots, F_t^i]$. The optical flow estimation produces a three-dimensional flow matrix [height, width, 2], where height and width represent the height and width of the image, and the third dimension represents the displacement of the corresponding pixel in the X and Y directions, with positive and negative values indicating the direction. Since the direction of pedestrian movement is uncertain, negative values may appear in the flow matrix. The information encoding network adopts the ReLU activation function, which filters out negative values, leading to distortion of the optical flow information. Therefore, a polar coordinate mapping method is used to map the optical flow information from the Cartesian coordinate system to the polar coordinate system. A Convolutional Neural Network (CNN) is used to extract features from the optical flow sequence, and the obtained features are encoded by a GRU network, resulting in the final local flow encoding F_e.

$$F_e^i = f_{GRU}(f_{CNN}(F_{obs}^i, W_c), W_f) \tag{2}$$

Where F_{obs}^i and F_e^i are observed optical flow sequence and optical flow encoding features of the pedestrian; f_{CNN} is convolution function, W_c is learnable parameter matrix; f_{GRU} is the GRU function, W_f is learnable parameter matrix.

2.2.4. Trajectory Encoder

For pedestrians i, let's assume that their position vector at a certain moment is T_t^i, which includes the coordinates of the pedestrian's center point and the length and width of the bounding box. The observed trajectory sequence can be represented as $T_{obs}^i = [T_{t-t_{obs}+1}^i, \ldots, T_t^i]$. By using a GRU network, the pedestrian's pose information is encoded, resulting in encoded features T_e.

$$T_e^i = f_{GRU}(T_{obs}^i, W_t) \tag{3}$$

Where T_{obs}^i and T_e^i is the observation pose sequence and pose coding feature of pedestrians, f_{GRU} is the GRU function, W_t is learnable parameter matrix.

2.3. Multi-information Fusion Network

To more effectively integrate multiple types of information, this paper adopts a cross-modal fusion attention mechanism based on the scaled dot-product attention mechanism improved by Transformer [16]. It performs hierarchical fusion of pose features, optical flow features, and trajectory

features. The network architecture is shown below:

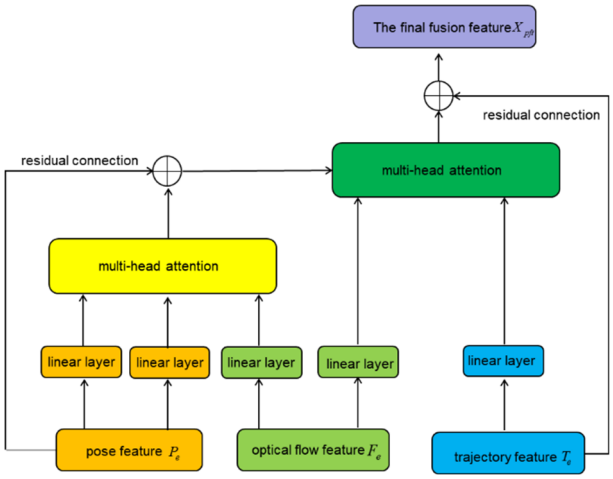

Figure 2. The overall structure of multi-information fusion network based on attention mechanism

When performing multi-information fusion, it is divided into two levels of fusion. The primary fusion involves using attention mechanisms to fuse pose features and optical flow features, while also introducing residual connections. The features obtained from the primary fusion are then finally fused with the trajectory features. The entire fusion process can be represented by the following formula:

$$X_{pf} = \frac{P_e W_{pq}^{\mathrm{T}} P_e W_{pk}^{\mathrm{T}}}{\sqrt{d_k}} F_e W_{fv}^{\mathrm{T}} + P_e \tag{4}$$

$$X_{pft} = \frac{X_{pf} T_e W_{fk}^{\mathrm{T}}}{\sqrt{d_k}} T_e W_{tv}^{\mathrm{T}} + F_e \tag{5}$$

Where P_e, F_e, T_e are pose, optical flow and trajectory coding features, respectively; W_{pq} and W_{pk} is the mapping weight of two linear branches of attitude coding, W_{fq} and W_{fk} is the mapping weight of two linear branches of optical flow coding, W_{tv} is the mapping weight of trajectory coding, X_{pf} and X_{pft} are the features of primary fusion and final fusion respectively.

2.4. Trajectory Decoding Network

The trajectory decoding network in this model is inspired by the decoding part of the BiTraP [17] model, consisting of two parts: the object decoding network and the trajectory decoding network. The object decoding network decodes the fused features to predict future targets and adopts a perceptron model with 2 layers. The bidirectional trajectory decoding network includes forward and backward GRUs. The forward GRU takes the observed fused features as input, while the backward GRU takes the predicted future targets from the object decoder as input. The calculation process is as follows:

$$h_{t+1}^{f} = GRU_f(h_t^f, W_g^i h_t^f + b_f^i) \tag{6}$$

$$h_{t+\delta-1}^{b} = GRU_f(h_{t+\delta}^b, W_b^i \hat{Y}_{t+\delta} + b_b^i) \tag{7}$$

$$\hat{Y}_{t+\delta-1}^{b} = W_f^o h_{t+\delta-1}^f + W_b^o h_{t+\delta-1}^b + b^o \tag{8}$$

Where f, b, i, o indicate "forward," "backward," "input" and "output" respectively, and h_t^f and $h_{t+\delta}^b$ are initialized by passing h_t through two different fully-connected networks.

3. Experiments

3.1. Datasets

Joint Attention in Autonomous Driving (JAAD) is a dataset used for studying joint attention in the context of autonomous driving. It focuses on the behaviors of pedestrians and drivers at intersections and the factors that influence them. The dataset contains 346 video clips, with a frame rate of 30 frames per second, and a duration of 5-10 seconds, totaling 82,032 frames. The videos were captured using three in-car cameras and cover typical scenarios of daily urban driving in various weather conditions in North America and Eastern Europe. The annotations in this dataset include 2,793 pedestrians and 378,643 2D bounding boxes.

Pedestrian Intention Estimation is a dataset specifically designed for studying pedestrian behavior in traffic. It consists of over 6 hours of recorded videos capturing typical traffic scenarios, recorded using in-car cameras, and providing accurate vehicle information from OBD sensors. This dataset contains over 300,000 annotated video frames and 1,842 pedestrian samples, making it the largest publicly available dataset for studying pedestrian behavior in traffic.

3.2. Implementation Details

We use the Adam optimizer with default parameters and initial learning rate 0.001, which is dynamically reduced based on the validation loss. Our models are optimized end-to-end with batch size 128 and the training is terminated after 100 epochs. The algorithm in this paper was initially developed and tested on an Ubuntu 20.4 system with an NVIDIA 3080 graphics card, and then it was ported to the embedded device, Xavier.

Jetson Xavier is equipped with a six-core 64-bit processor and an NVIDIA Volta architecture GPU, along with a rich set of I/O interfaces. The entire algorithm is based on the PyTorch framework and utilizes PyQt5 to design the program's user interface.

To improve the inference speed of the model on the embedded device, the network model is pruned by removing channels that have small weights in the convolutional kernels. Additionally, the model is converted to an ONNX model, and the TensorRT inference engine is utilized to accelerate model inference.

3.3. Evaluation Protocols

Our evaluation metrics include Average Displacement Error (ADE), which measures the accuracy of the entire trajectory, and Final Displacement Error (FDE), which measures the accuracy only at the endpoint of the trajectory. We use Mean Squared Error (MSE) to evaluate our performance on the JAAD and PIE datasets, calculated based on the upper-left and lower-right coordinates of the bounding box. Additionally, we utilize Center Mean Squared Error (CMSE) and Center Final Mean Squared Error (CFMSE) as evaluation metrics on the JAAD and PIE datasets. These two-error metrics are similar to MSE, but they are computed based on the centroids of the bounding boxes.

3.4. Experimental Results and Analysis

3.4.1. Comparison with Related Methods

The baseline models compared in this paper include: early trajectory prediction models such as (1) linear Kalman filter, (2) LSTM, (3) Bayesian LSTM model [18], (4) FOL-X [19]; recent prediction models such as (1) PIE$traj$ [20](2) BtriP [17]

and the current state-of-the-arts (SOTA) models SGNet [21].

The trajectory prediction algorithm proposed in this paper can predict the trajectories of pedestrians in the next 0.5s, 1.0s, and 1.5s, as shown in Table 1. This algorithm improves upon the BiTraP by considering pedestrian pose information, optical flow information, and trajectory information. It effectively integrates these features using a cross-modal attention mechanism. From the table, it can be seen that the proposed algorithm performs better than SGNet-D on the JAAD and PIE datasets. The algorithm is tested on the JAAD and PIE datasets, and the displacement errors are reduced by 6.35% and 3.28%, respectively, compared to BiTraP.

Table 1. Deterministic results on JAAD in terms of MSE/CMSE/CFMSE. ↓ denotes lower is better.

Method	ADE 0.5s↓	ADE 1.0s↓	ADE 1.5s↓	CADE 1.5s↓	CFDE↓
Linear	233	857	2303	1565	6111
LSTM	289	569	1558	1573	5766
PIE	110	399	1280	1183	4780
BiTraP	93	378	1206	1105	4565
SGNet-D	87	350	1121	1065	4355
Ours	85	350	1105	1057	4275

It also shows a slight improvement compared to the latest model SGNet-D, especially in long-term (1.5s) trajectory prediction tasks. The proposed algorithm significantly reduces the final error by approximately 8.9 pixels on the JAAD dataset and 2.4 pixels on the PIE dataset for the 1.5s prediction task.

Table 2. Deterministic results on PIE in terms of MSE/CMSE/CFMSE. ↓ denotes lower is better.

Method	ADE 0.5s↓	ADE 1.0s↓	ADE 1.5s↓	CADE 1.5s↓	CFDE↓
Linear	123	477	1365	950	3983
LSTM	172	330	911	837	3352
PIE	58	200	636	596	2447
BiTraP	41	161	511	481	1949
SGNet-D	37	148	478	450	1891
Ours	36	145	471	442	1885

Table 3 compares the parameter count and inference time of mainstream models on the entire test set. All tests were conducted on the same hardware environment, with batch size set to 1024. From the table, it can be observed that although the proposed algorithm has a slightly increased parameter count compared to BiTraP-D and a slightly decreased inference speed, it still demonstrates significant advantages compared to the latest model SGNet-D.

Table 3. The comparison of parameter quantity and reasoning speed between this algorithm and other mainstream algorithms

Method	Parameters (pieces)	JAAD speed (second)	PIE speed (seconds)
PIE	1240473	---	2.09
BiTraP	1427112	1.10	2.73
Ours	3243421	2.92	6.49
SGNet-D	7622406	33.29	85.81

3.4.2. Qualitative Results

The hardware setup of the simulated car used in this paper is shown in Figure 3. It is equipped with the embedded device Xavier as the computing and control center. A high-definition camera is used to capture videos, and the steering of the front wheels of the car is controlled by the GPIO of Xavier using a servo. To display the prediction results, a high-definition display screen is also equipped.

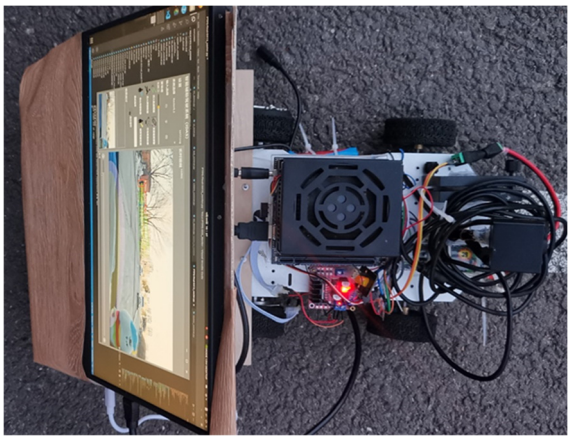

Figure 3. Hardware structure of simulation vehicle equipped with pedestrian trajectory prediction system

Using the trajectory prediction system designed in this paper, the predicted results of the video in a traffic scenario are shown in the figure. From Figure 4, it can be observed that the pedestrian skeleton recognition is complete and the predicted future trajectory matches expectations. The system also displays a cautionary zone, and when the future position of the pedestrian is within this zone, the color of the pedestrian's bounding box changes from green to red. At the same time, a voice warning is issued, and Xavier controls the motor detection to avoid collisions.

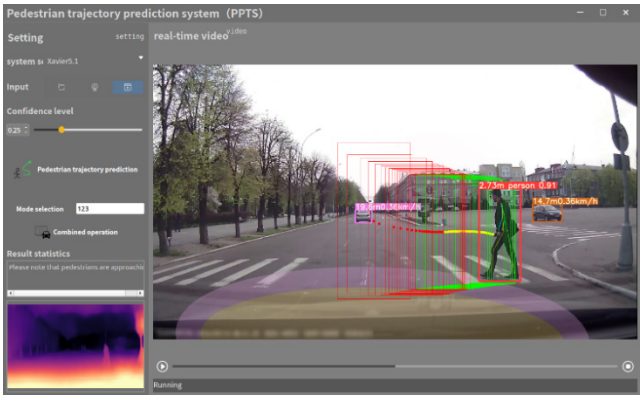

Figure 4. Pedestrian trajectory prediction results

4. Conclusion

In this paper, a vehicle-mounted pedestrian trajectory prediction system based on Jetson Xavier is designed. The pedestrian trajectory prediction from the driving perspective is studied in detail. The attitude information and optical flow information are introduced on the basis of a single trajectory information. A novel multi-center fusion attention mechanism is designed to more effectively fuse multi-information features and improve the accuracy of pedestrian trajectory prediction. The designed algorithm is transplanted into the embedded device Jetson Xavier to realize the whole process and real-time operation from pedestrian detection, pedestrian tracking, pedestrian attitude estimation, pedestrian optical flow estimation, pedestrian trajectory acquisition to pedestrian trajectory prediction, vehicle control, etc. The designed system is mounted on the simulation car, and tested for different scenarios to verify the function of the system. The future trajectory of the pedestrian is displayed to the driver through the display, and the warning area is delineated. When the future trajectory of the pedestrian enters the warning area, the system will issue a voice broadcast to warn the driver to pay attention to the pedestrian. Through experiments, it is found that when the speed is not fast, the real-time prediction error of pedestrian trajectory is small and has good practicability.

References

[1] Rudenko A, Palmieri L, Herman M, et al. Human motion trajectory prediction: A survey[J]. The International Journal of Robotics Research, 2020, 39(8): 895-935.

[2] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics[J]. Physical review E, 1995,51(5):4282-4286.

[3] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model[C]. IEEE Conference on Computer Vision and Pattern Recognition,2009: 935–942.

[4] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking[C]. IEEE International Conference on Computer Vision and Pattern Recognition, 2009:261–268.

[5] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? [C], IEEE International Conference on Computer Vision and Pattern Recognition, 2011:1345–1352.

[6] Gianluca Antonini, Michel Bierlaire, and Mats Weber. Discrete choice models of pedestrian walking behavior[J]. Transportation Research Part B: Methodological, 2006, 40(8): 667–687.

[7] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks[C]. IEEE Conference on Computer Vision and Pattern Recognition,2018:2255–2264.

[8] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds[C]. IEEE International Conference on Robotics and Automation, 2018:1–7.

[9] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Soft+ hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection[J]. Neural networks, 2018,108:466–478.

[10] Mohamed A, Qian K, Elhoseiny M, et al. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1442.

[11] Wang C, Cai S, Tan G. Graphtcn: Spatio-temporal interaction modeling for human trajectory prediction[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 3450-3459.

[12] Yagi T, Mangalam K, Yonetani R, Sato Y. Future Person Localization in First-Person Videos[C]. IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake: IEEE, 2018: 7593-7602.ong Beach: IEEE. 2019: 2960-2963.

[13] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv preprint arXiv:2207.02696, 2022.

[14] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Hawaii: IEEE, 2017:1302-1310.

[15] Xu H, Zhang J, Cai J, Rezatofighi H, Tao D. GMFlow:Learning Optical Flow via Global Matching[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE. 2022: 8111-8120.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[C]. Conference and Workshop on Neural Information Processing Systems. California: MIT Press,2017: 5998-6008.

[17] 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.

[18] K. Ramanathan-V. Robicquet A. Li FF. Savarese S. Alahi, A. Goel. Social lstm: Human trajectory prediction in crowded spaces. In IEEE Conference on Computer Vision and Pattern Recognition, pages 961-971, 2016.

[19] MZ. Wang YC. Crandall DJ. Atkins EM. Yao, Y. Xu. Unsupervised traffic accident detection in first-person videos. In IEEE International Conference on Intelligent Robots and Systems., pages 273–280, 2019.

[20] I. Kunic-T. Tsotsos J. Rasouli, A. Kotseruba. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV)., pages 6261–6270, 2019.

[21] O Bideau. Halawa, M. Hellwich. Action-based contrastive learning for trajectory prediction. In European Conference on Computer Vision, pages 143–159, 2022.