Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 9940126, 18 pages
https://doi.org/10.1155/2021/9940126

Research Article

At the Traffic Intersection, Stopping, or Walking? Pedestrian Path Prediction Based on KPOF-GPDM for Driving Assistance

Xudong Long,[1] **Weiwei Zhang**,[2] **Bo Zhao**,[1] **and Shaoxing Mo**[1]

[1]*Shanghai Engineering and Technology University, Shanghai 200000, China*
[2]*Tsinghua University, Beijing 100000, China*

Correspondence should be addressed to Weiwei Zhang; zwwsues@163.com

Received 30 March 2021; Revised 2 July 2021; Accepted 14 July 2021; Published 30 July 2021

Academic Editor: Peng Hang

Pedestrian detection has always been a research hotspot in the Advanced Driving Assistance System (ADAS) with great progress in recent years. However, for the ADAS, we not only need to detect the behavior of pedestrians in front of the vehicle but also predict future action and the motion trajectory. Therefore, in this paper, we propose a human key point combined optical flow network (KPOF-Net) in the vehicle ADAS for the occlusion situation in the actual scene. When the vehicle encounters a blocked pedestrian at a traffic intersection, we used self-flow to estimate the global optical flow in the image sequence and then proposed a White Edge Cutting (WEC) algorithm to remove obstructions and simply modified the generative adversarial network to initialize pedestrians behind the obstructions. Next, we extracted pedestrian optical flow information and human joint point information in parallel, among which we trained four human key point models suitable for traffic intersections. At last, KPOF-GPDM fusion was proposed to predict the future status and walking trajectories of pedestrians, which combined optical flow information with human key point information. In the experiment, we did not merely compare our method with other four representative approaches in the same scene sequences. We also verified the accuracy of the pedestrian motion state and motion trajectory prediction of the system after fusion of human joint points and optical flow information. Taking into account the real-time performance of the system, in the low-speed and barrier-free environment, the comparative analysis only uses optical flow information, human joint point information, and KPOF-Net three prediction models. The results show that (1) in the same traffic environment, our proposed KPOF-Net can predict the change of pedestrian motion state about 5 frames (about 0.26 s) ahead of other excellent systems; (2) at the same time, our system predicts the trajectory of the pedestrian more accurately than the other four systems, which can achieve more stable minimum error ±0.04 m; (3) in a low-speed, barrier-free experimental environment, our proposed trajectory prediction model that integrates human joint points and optical flow information has higher prediction accuracy and smaller fluctuations than a single-information prediction model, and it can be well applied to automobiles' ADAS.

1. Introduction

In the automatic driving scene, efficient detection of vehicles and pedestrians around the vehicle has become the basic ability of autonomous vehicles [1]. Recently, some researchers focused their attention on the understanding of pedestrians' behaviors and intentions in front of the car and conducted simulation experiments in the Atlanta world assumption [2]. If the collision between pedestrians and vehicles can be predicted in advance, many unnecessary traffic accidents can be averted. For example, in the ADAS of Mercedes-Benz E-class and S-class car models [3], a pedestrian prediction algorithm based on stereo vision is introduced, which is applied in emergency braking in dangerous scenarios. In complex scenarios, such as intersections and crosswalks, it is necessary to accurately estimate the current and future positions of pedestrians relative to the moving vehicle. In the process of pedestrian trajectory prediction, we need to consider several influencing factors. First of all, for pedestrians showing more random during walking on the road, such as the interaction between people, a particular pedestrian trajectory is affected by the position of other pedestrians. In addition, people with social attributes will also have an impact on the final trajectory, and the

quantification of these indicators is a cumbersome process. Secondly, the pedestrian movement in the eyes of the ADAS is regarded as the common result of pedestrian movement and vehicle movement. Therefore, the prediction range of the active pedestrian prediction system is very short with even a small improvement eliciting a significantly improved performance. This article focuses on the prediction of pedestrian positions at intersections and crosswalks. With the method of fusion of optical flow information and joint point models, the paper first predicts whether the pedestrian's state in the future is standing or stopping, and then forecasts the pedestrian's trajectory and position in the future.

With auxiliary information provided for the ADAS, it discusses the status and location information of pedestrians in the future, adjusts the speed in advance to avoid traffic accidents due to proximity between the car and the pedestrian.

In summary, we highlight our main contributions as follows:

(i) We propose KPOF-Net, a novel framework of pedestrian trajectory prediction algorithm, which combines optical flow information and pedestrian joint models to collaboratively predict the state and trajectory of pedestrians in the future.

(ii) By evaluating the complexity of the application scene, we propose using an optical flow estimation module to estimate the optical flow of pedestrians with occlusion through the self-flow network, and then design a WEC algorithm based on Canny to remove occluded objects, and finally modify UCTGAN slightly to generate complete pedestrian optical flow information.

(iii) Through a large number of observations and researches on pedestrian motion status at intersections and crosswalks, we used the posture change information when pedestrian motion status changes to train four human joint point models of standing, stopping, standing tendency, and stopping tendency. The human joint point model trained at this time can more accurately predict the motion state and motion trajectory of pedestrians in the future.

(iv) Considering that a single optical flow cannot obtain detailed information of the pedestrian's posture at a traffic intersection, it can only roughly predict the pedestrian's motion state and trajectory in the future. The KPOF-GPDM prediction algorithm is proposed, which integrates optical flow and human joint point information, and it combines the movements of the upper and lower limbs of pedestrians in different traffic situations to predict the movement state and trajectory of pedestrians in the future, providing more efficient and active safety data for the ADAS.

(v) We propose ADE and FDE evaluation methods based on Euclidean distance, and compare and analyze the effect of KPOF-Net fusion of optical flow information and joint point model in improving the accuracy of pedestrian trajectory prediction in an unobstructed experimental environment.

2. Related Work

In this section, we provide a review of the optical flow and key point prediction approaches for pedestrian trajectory prediction under occlusion. We focus on learning related research to solve the problem of pedestrian trajectory prediction at intersections and sidewalks.

2.1. Optical Flow. Optical flow estimation is mainly divided into three categories: Supervised Learning of Optical Flow, Unsupervised Learning of Optical Flow, and Self-Supervised Learning. FlowNet [4] is the first end-to-end optical flow learning framework, which takes continuous images as input and dense optical flow graphs as output. SpyNet [5] uses a pyramid network with a compact space structure to scale the image to deal with the large-scale displacement of the object. LiteFlowNet [6] achieves lightweight by distorting the feature objects extracted by CNNs [4, 6, 7]. However, this type of method needs to use the rules [4, 8] to pretrain multiple synthetic datasets, which consumes a lot of time, and it involves low-speed, offline operation, and not real time. Moreover, the result is too dependent on the pretraining results of the synthetic dataset, and its optical flow accuracy does not meet our scenario requirements. The unsupervised learning method mainly uses the principle of constant brightness [9] and spatial smoothness [10], by measuring the pixel difference between the initial image and the test image, which can handle optical flow estimation with obstructions. Janai et al. [11] use multiple frames of images to jointly derive optical flow images. However, the detection accuracy of this scheme needs to be improved. DDFlow [12] proposes an optical flow data distillation method to learn the optical flow of occluded objects, but this type of method has limitations, which means it can only handle occluded objects under specific circumstances and cannot be applied to all scenes. Self-supervised learning adopts the data itself as a supervised signal, which is widely used to learn features from unlabeled data [13], and is often used to deal with image restoration [14], image coloring [15], and stitching problems [16]. Doersch and Zisserman [17] combine feature learning based on low-level motion cues. The study in [18] proposes S4L-Rotation and S4L-Exemplar algorithms to deal with the classification loss problem. However, the linear classifier obtained by this method is very dependent on the adjustment strategy of the learning rate and has uncertainty. Self-flow [19] takes reliable predictions of nonoccluded pixels as the self-supervision signal to guide our optical flow learning of occluded pixels. The network not only has a simple structure but also changes the image pyramid to a feature pyramid and uses a multiframe input method, which increases the information input of the network, and has good effects in terms of progress and real-time performance. In the scene of intersection and sidewalk, considering the comprehensive robustness of the automotive ADAS, the self-flow network is used to extract the optical flow information of pedestrians, and at the same time, when pedestrians are blocked by obstructions, a WEC algorithm based on Canny [20] was proposed on the basis of self-flow by

comprehensively considering the complexity and timeliness of the optical flow system as well as the accuracy of trajectory tracking in the presence of obstacles. To obtain smooth pedestrian optical flow, an UCTGAN network [21] was also used to jointly recover pedestrian optical flow in blocked scenes.

2.2. Key Point Prediction. In the past few decades, human pose estimation [22] and pedestrian trajectory prediction [23] have made rapid development. In the KF [24], the current state of a dynamic system can be propagated to the future by means of the underlying linear dynamical model without the incorporation of new measurements. IMM KF [24] introduces a similar method to predict pedestrian trajectories in a multithreaded dynamic model. Choi and Savarese [25] propose a framework that can track multiple objects, recognize the atomic activities performed by individuals, such as walking or standing, identify interactions between pairs of individuals (i.e., interactive activities), and understand the activities of groups of individuals. However, this method could appear to be biased in assigning interaction labels. In the research process, Hu et al., respectively, proposed an improved Bernoulli heatmap [26] and a new convolutional recurrent network model [27] to estimate the joint point information of various parts of the human body. Although these methods can quickly and accurately construct a human head joint point model, the performance needs to be improved when processing large-angle samples. Karasev et al. [28] proposed to use the Jump-Markov process to model the pedestrian's movement and infer the state of the pedestrian through the Rao-Blackwellized filter. However, the predictable change event types of this scheme are limited, and it cannot be widely used in various traffic scenarios. Anca Marginean et al. [29] proposed a set of pose-based and recursive framework-based algorithms to deal with imbalances in pedestrian estimation. When our scene is set at intersections and sidewalks, Keller and Gavrila [30] used the Gaussian dynamics model and probabilistic hierarchical trajectories based on dense optical flow to obtain pedestrian characteristics. Goldhammer et al. [31] proposed the use of polynomial least squares approximation and multilayer perceptron (MLP) to predict the trajectory of pedestrians in the next 2.5 s. However, the accuracy of pedestrian trajectories predicted by this method needs to be improved. The study in [32] also used a similar method to predict the trajectory of pedestrians riding bicycles. Alahi et al. [33] proposed an algorithm based on LSTM to predict the trajectory of pedestrians by considering the interdependence of pedestrians. Urtasun et al. [34] used a GPDM to track a small number of 3-D body points that have been derived using an image-based tracker and the system is trained with one gait cycle from six subjects and is able to handle several frames of occlusions. However, the exported 3D body points are limited, and the occlusion processing effect is not good. Minguez et al. [35] used balanced GPDMs for intention detection and trajectory forecasting of pedestrians based on 3D poses, through training four types of postures, namely, starting, stopping, standing, and walking, to predict pedestrian trajectory movements. However, this method only considers the situation that the pedestrian is always in the same motion state, and ignores the detailed information of the body when the motion state of the pedestrian changes. Kress et al. [36] proposed the use of 3D human poses for trajectory forecasting of vulnerable road users (VRUs), such as pedestrians and cyclists, in road traffic. The 3D poses represent that the entire body posture of the VRUs can provide important posture information for pedestrian trajectory prediction. The above methods can predict the trajectory and movement classification of pedestrians at intersections and crosswalks, but there is still room for improvement in accuracy. Based on B-GPDMs [35], this paper trains a human joint point model with walking and stopping trends, and then couples optical flow information and joint point information to predict the pose information and movement classification of future pedestrians. Although all the above schemes can predict the motion behavior and motion trajectory of pedestrians, the accuracy of the prediction of the motion state and motion trajectory of pedestrians at traffic intersections is limited.

In summary, the optical flow information prediction model can also predict the movement state and trajectory of pedestrians at traffic intersections. But, the optical flow information lacks detailed information about pedestrians in the process of movement, which makes it impossible to accurately predict the spatial position of pedestrians in the future. Therefore, under the premise of considering obstructions, this paper combines optical flow information and human body joint point information to propose a KPOF-Net prediction model to collaboratively predict the motion state and trajectory of pedestrians in the future.

3. Overview of KPOF-Net

3.1. Main Network. Figure 1 summarizes the main framework of KPOF-Net, composed of three main parts of occlusion object removal, pedestrians' state estimation, and trajectory prediction. In occlusion object removal, we use self-supervised learning method of self-flow [19] to detect the optical flow information of pedestrians at intersections and crosswalks, and then propose a Canny-based White Edge Cutting (WEC) algorithm to remove obstructions, and restore the pedestrian posture behind the obstructions by modifying the UCTGAN network. In pedestrians' state estimation, we use Carnegie Mellon University (CMU) [37] to train pedestrians' key point station, e.g., stopping, walking, stopping tendency, and walking tendency. In trajectory prediction, we propose the KPOF-GPDM method, which combines pedestrians' key point and optical flow information to predict pedestrians' future trajectory.

3.2. Pedestrian in Painting behind Occlusion. In actual scenes, pedestrians may be obscured by luggage, handbags, trash cans, stone pillars, animals, and other objects at intersections and crosswalks. In the process of extracting optical flow information, the pedestrian mask cannot be completely obtained. It has great influence on the pedestrian

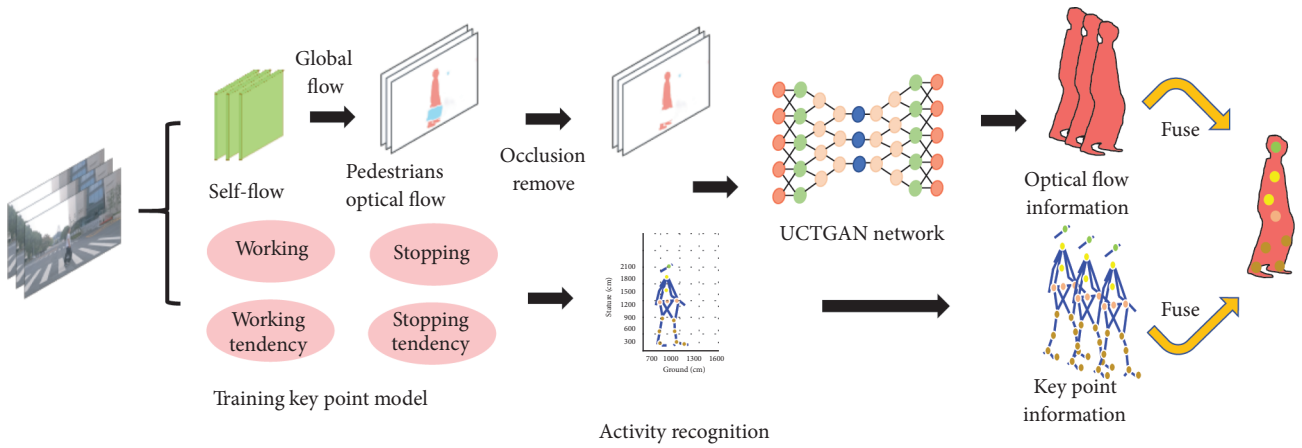

Figure 1: KPOF-Net is composed of an optical flow module and a joint point module. On the one hand, the optical flow extraction of pedestrians and obstructions is completed through self-flow, WEC is proposed to remove obstructions, and then the UCTGAN network is used to generate a complete pedestrian optical flow diagram. On the other hand, it is based on B-GPDM training four types of joint models of walking, stopping, walking tendency, and stopping tendency. Finally, the pedestrian optical flow information is combined with the joint point model to predict the pedestrian's posture and trajectory.

trajectory prediction at the back end, and an accurate trajectory route cannot be obtained. Therefore, we introduce the self-flow network to extract the optical flow information of the pedestrians at the intersection, integrate the WEC to remove the contour information of the obstruction, and then initialize the pedestrian optical flow information after the obstruction through the modified confrontation generation network.

3.2.1. Self-Supervised Learning Method, Self-Flow. Self-flow net is an excellent method to get the objects' optical flow information behind the occlusion. It builds on PWC-Net [38] and extends it to multiframe optical flow estimation. PFC-Net uses pyramid processing to improve the resolution from coarse to thin, and uses feature distortion, cost volume constructs to estimate the optical flow of each layer. Based on these principles, it has achieved state-of-the-art performance with a compact model size.

As shown in Figure 2, the reason why we chose the self-flow network can be seen. First, it uses three images as input to generate multi-frame optical flow estimation of three feature representations of F_{t-1}, F_t, and F_{t+1}. Then, self-flow uses the initial backward flow and backward cost volume information for the previous frame. I_{t-1} can provide effective information about the occlusion, especially the area that is occluded in I_{t+1} but not occluded in I_{t-1}, and self-flow combines this information to get a more accurate optical flow estimation.

At the same time, the self-flow network uses five frames of images as input to perform consistency checks when estimating the optical flow between two frames, thereby inferring the occlusion map between two consecutive images. For the forward-backward consistency check, when the mismatch between the forward flow and the reverse forward flow is too large, the self-flow network considers a pixel to be occluded. A pixel is considered occluded whenever it violates the following constraint:

$$\left|w_{t\longrightarrow t+1}+\hat{w}_{t\longrightarrow t+1}\right|^2<\alpha_1\left(\left|w_{t\longrightarrow t+1}\right|^2+\left|\hat{w}_{t\longrightarrow t+1}\right|^2\right)+\alpha_2, \quad (1)$$

when $\alpha_1 = 0.01$ and $\alpha_2 = 0.05$, we get accurate optical flow information of pedestrians and obstructions.

3.2.2. Removal of Occlusion Region. In the video sequence at intersections and crosswalks, obstruction objects can be divided into two categories as static occlusion (such as stone pillars, trash cans, railings) and dynamic occlusion (such as suitcases, luggage bags, animals). In self-flow network, static occlusion will not generate optical flow for no motion between frames. The dynamic occlusion will produce striking interference optical flow, which is difficult to eliminate. Then, in this part, we only consider the situation when pedestrians are blocked by dynamic objects. A White Edge Cutting (WEC) algorithm is based on Canny, which removes the optical flow information in the occluded area:

$$H_{ij}=\frac{1}{2\pi\sigma^2}e^{\left(-\left(\left((i-(k+1))^2+(j-(k+1))^2\right)/2\sigma^2\right)\right)}, \quad 1\le i,j\le(2k+1), \quad (2)$$

$$\begin{aligned} G&=\sqrt{G_x^2+G_y^2},\\ \theta&=\arctan\left(\frac{G_y}{G_x}\right). \end{aligned} \quad (3)$$

In formulas (2) and (3), H_{ij} represents the Gaussian convolution kernel, (i, j) represents the pixel coordinates, k is the dimension of the convolution kernel, G represents the gradient descent value, G_x and G_y represent the bias value in the (x, y) direction, and θ represents the gradient descent direction.

The pixels in the optical flow image are filtered by Gaussian filter (2) to calculate the wave recorder core to obtain the pixel threshold with weight, then (3) is used to calculate the gradient value and gradient direction, and finally the WEC algorithm is considered to remove the optical

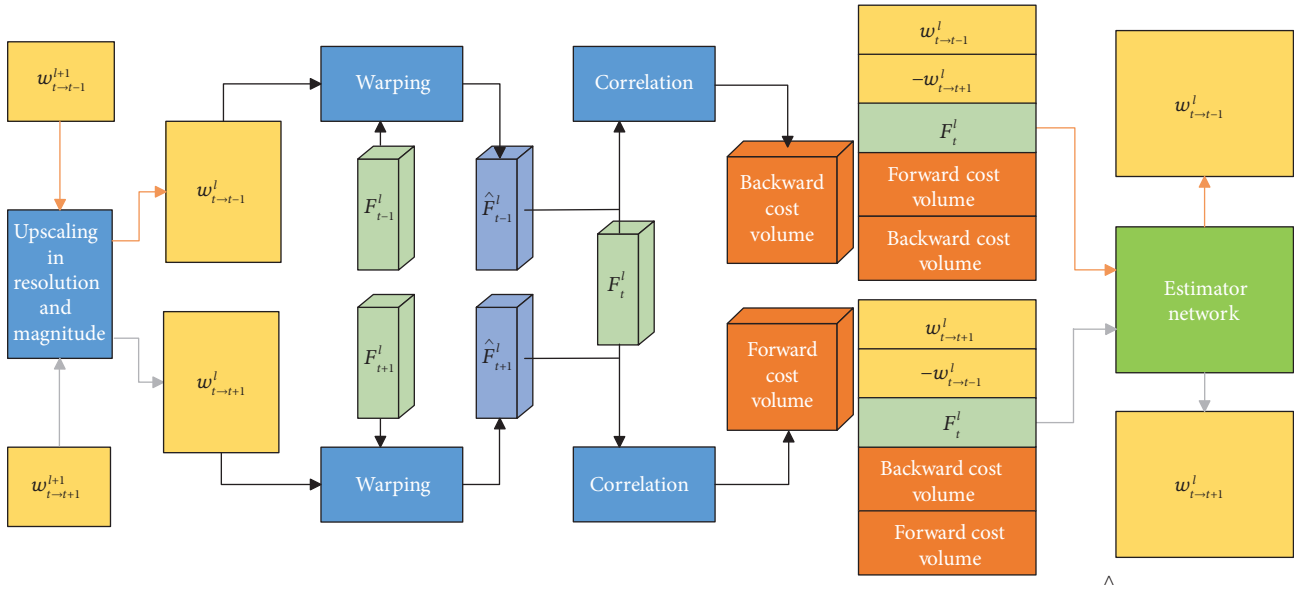

FIGURE 2: Self-flow network architecture at each level (similar to PWC-Net). w^l denotes the initial coarse flow of l and $\hat{F}^l$ denotes the warped feature representation.

flow in the obstructed area information, as the result shown in Figure 3.

In the process of detecting WEC, we take a small portion of the optical flow into the pixel level. In the field of 3×3, we can find that at the junction of two objects, the optical flow information will be close to white or white, so we use the principle to connect the pixel values that tend to be white at the boundary to obtain the exact boundary between the pedestrian and the dynamic object, and remove this part of the area:

$$\text{difference} = \frac{\left|S_{\text{WEC}} - S_{\text{origin}}\right|}{S_{\text{origin}}} * 100\%. \tag{4}$$

We use formula (4) to calculate the percentage of the area difference before and after the occlusion segmentation to show the performance of the WEC segmentation occlusion. In the formula, S_{WEC} represents the area of the occlusion after division, and S_{origin} represents the area of the occlusion before division. From Table 1, it is found that the WEC method can control the area difference within 5%. Although the area difference increases when facing small obstructions, it can still separate different types of obstructions from pedestrians.

3.2.3. Pedestrians Inpainting according to the UCTGAN Network. When pedestrians at the intersections and crosswalks are blocked by static objects, they get incomplete optical flow information. When they are occluded by a dynamic target, the WEC algorithm is used to remove the occluded area to get the pedestrian optical flow map of the incomplete area. In view of the above situation, this paper integrates the GAN network to generate a complete pedestrian optical flow image, as shown in Figure 4.

We simply modified the network based on UCTGAN, and deleted the multi-scale scheme of the original network according to the actual needs of the experiment. The image area of size 256×256 is directly extracted from the external square center of the occlusion area and input to the network. In this way, multiple calculations on different scales of the network can be avoided and the operation efficiency of the network is greatly improved.

The UCTGAN network is trained in an end-to-end fashion, which consists of two branches. The UCTCAN framework mainly includes three network modules: manifold projection module E_1, conditional encoder module E_2, and generation module G. The primary branch consists of a manifold projection module E_1 and a generation module G, which is responsible for learning one-to-one image mapping between two spaces in an unsupervised way by projecting instance image space S_i and conditional completion image space S_i into one common latent manifold space S_m. The second branch consists of a conditional encoder module E_2, which acts as conditional constraint similar to the conditional label. The UCTGAN framework could maximize the conditional log-likelihood of the training instances, which involves a variational lower bound:

$$\begin{aligned}\log p(I_c|I_m) \geq & -KL\left(f_\varphi(Z_c|I_i, I_m) \parallel f_\phi(Z_c|I_m)\right) \\ & + E_{Z_c \sim f_\varphi(Z_c|I_i, I_m)}\left[\log g_\theta(I_c|Z_c, I_m)\right],\end{aligned} \tag{5}$$

where I_c, I_m, and I_i are instance image, masked image, and the repaired image, respectively. Z_c is the latent vector of I_i in space S_m. f_φ, f_ϕ, and g_θ are the posterior sampling function, conditional prior, and likelihood, respectively, where φ, ϕ , and ϑ are the corresponding deep network parameters.

One of the reasons why we choose the UCTGAN network is its series of training loss, including L_{ccl} Condition Constraint Loss, L_{KL} KL Divergence Loss, L_{rec}

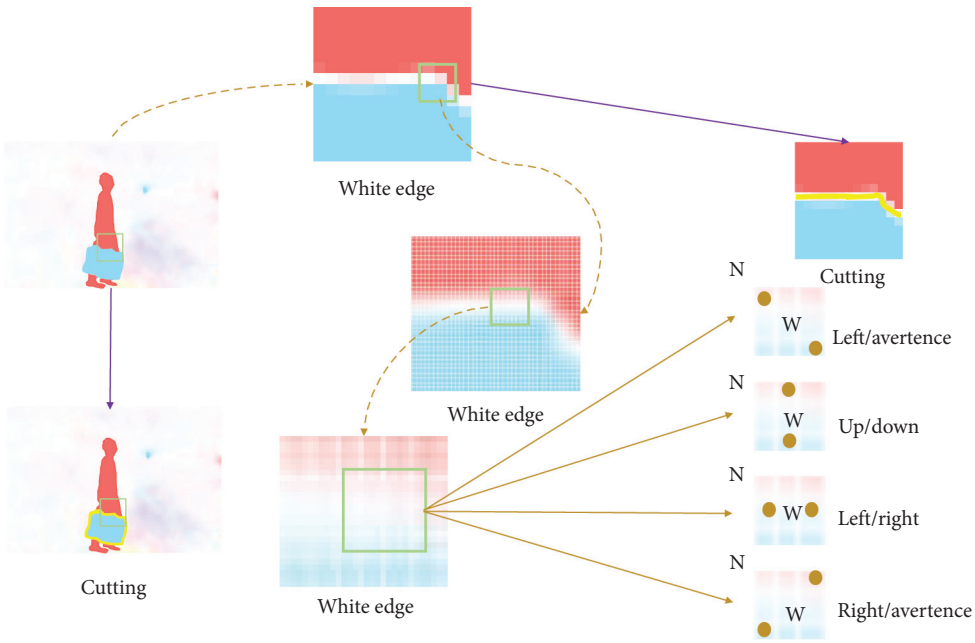

Figure 3: In the WEC detection process, a small part of the optical flow is taken and enlarged to the pixel level. In the $3 * 3$ pixel matrix, there is a transition zone that is close to white at the intersection of the obstruction and the pedestrian in the light flow. Based on this, WEC determines four types of symbols: left/avertence, up/down, left/right, and right/avertence. The N in the figure indicates that the value signs of the corresponding two points are different. When the four pairs of values are opposite in sign and the absolute value of the difference is less than a certain threshold, W is recorded as the WEC boundary. All W points are mapped to the original image to obtain the boundary of the occlusion in the input image.

Table 1: The area value ratio before and after the occluder is moved by WEC.

Object	Trunk (%)	Animal (%)	Handbag (%)	Motorcycle (%)	Bicycle (%)
Difference	2.37	4.32	3.21	3.56	4.77

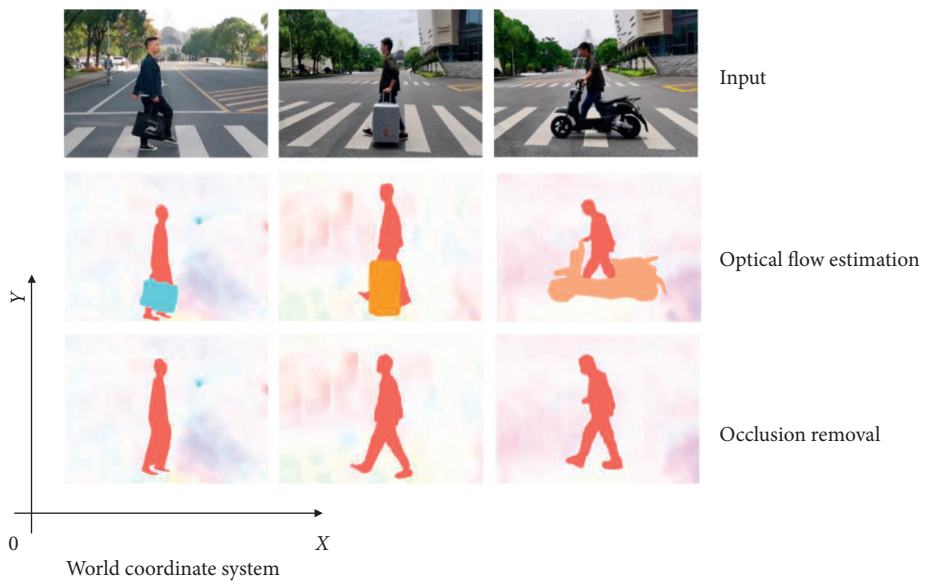

Figure 4: Pedestrian movement spectrogram. In the spatiotemporal sequence, pedestrians with obstructions obtain the initialized pedestrian optical flow diagram through three steps, which include picture input, optical flow estimation, and occlusion removal.

Reconstruction Loss, and L_{adv} Adversarial Loss. The total loss function L_{total} of UCTGAN consists of four groups of component losses, as shown in formula (6). Among them λ_{rec}, λ_{ccl}, λ_{adv}, and λ_{KL} are the hyperparameters corresponding to the constraints, which play a relatively important role in each constraint:

$$L_{\text{total}} = \lambda_{\text{rec}}\left(L^{g}_{\text{rec}} + L^{l}_{\text{rec}}\right) + \lambda_{\text{ccl}}\left(L^{a}_{\text{ccl}} + L^{f}_{\text{ccl}}\right) + \lambda_{\text{adv}}L_{\text{adv}} + \lambda_{KL}\left(L^{i}_{KL} + L^{m}_{KL}\right). \tag{6}$$

The condition constraint losses L^{a}_{ccl} and L^{f}_{ccl} encourage consistency and integrity between completion contents and known contents, reconstruction losses L^{g}_{rec}, L^{l}_{rec}, and L^{l}_{rec} encourage one-to-one mapping between the instance image and the repaired image, and avoid falling into mode collapse, and adversarial loss L_{adv} makes repaired images fit in with the distribution of the training dataset. The loss of all these prompted us to get excellent pedestrian optical information. If you are interested in the specific details of how to inpainting pedestrian optical flow information after proposing dynamic objects in this article, you can do intensive reading UCTGAN network [21].

3.3. Pedestrians' Key Point Model Establishment. In actual scenes, at intersections and crosswalks, pedestrians will produce corresponding stance based on the current status of traffic lights, traffic flow, and their own consciousness. This posture information can help us predict the state of pedestrians in the future. Therefore, in this paper, we extract human joint point information based on the hidden Markov joint point recognition model in Minguez et al. [35], and then train a joint point model that adapts to the human posture in the intersection and crosswalk scenes, including stopping, walking, stopping tendency, and walking tendency.

3.3.1. Data Set Description. In this section, our main goal was to train accurate models with different pedestrian dynamics. For this, we used high-frequency, low-noise datasets released by Carnegie Mellon University (CMU) [39]. On the one side, the high frequency of the dataset helps the algorithms to properly learn the dynamics of different activities and increases the probability of finding a similar test observation in the trained data without missing intermediate observations. On the other side, low-noise models improve the prediction when working with noisy test samples. The pedestrian motion simulation dataset contains a typical pedestrian motion sequence package. Among this, we collect the three-dimensional coordinates of 41 joints along the body with a frequency of 120 Hz. However, according to the actual situation of intersections and crosswalks, we focused on using part of the joint point information of the legs and the body. At the same time, according to our actual needs, we selected four categories from the CMU dataset that meets stopping, walking, stopping tendency, and walking tendency sequence; a total of 200 sequences were extracted, which consisted of 143,207 pedestrian poses from 25 different subjects. See Table 2 for details.

Pedestrian skeleton estimation algorithm, based on point clouds extracted from a stereo pair and geometrical constraints, was implemented to test the proposed method with noisy observations. The algorithm is based on references [40, 41], and the specific details can be learned from the literature [42].

3.3.2. Learning Pedestrians' Key Point Model. After extracting the human body joint point information, we need to identify the human body joint point model of the corresponding joint point state in the intersection and crosswalk scenes. In this part, we use the B-GPDMs method in the literature [35] to identify the joint points of pedestrians.

Mínguez et al. just trained four models suitable for their experimental needs, including walking, starting, stopping, and standing. However, these four models are only limited to the joint point model in which the pedestrian has been in an upcoming motion state, and did not consider the detailed information of the joint point when the pedestrian's motion state changes, resulting in the system being unable to predict accurate pedestrian trajectory information. So, we use the B-GPDMs algorithm to train four types of joint point models: standing, walking, stopping tendency, and walking tendency. This is because when pedestrians pass crosswalks and intersections, their movement is not restricted to only two states of walking and stopping alone, as they will constantly judge the current traffic situation to change their own motion state. When there is a vehicle in front, pedestrians in the walking state will collect forward environment information in real time by leaning their upper limbs forward, and the distance between the lower limbs will continue to shrink. When the front is passable, the upper limbs also lean forward to collect the front environmental information in real time, while the distance between the lower limbs is increasing.

In the process of learning all the sequences contained in the CMU dataset, since the coordinate system of these sequences is affected by the sensor, we deleted the 3D data of each observation and obtained the coordinate system with the pedestrian as the origin, which allows us to deal with pedestrians in any location. Then, by subtracting the mean and dividing each mean by the standard deviation to scale, it is more convenient to obtain zero mean and unit variance data. Since B-GPDM needs to use the smallest posterior function to iterate, we give appropriate initialization potential positions, hyperparameters, and constants according to the literature [30]. We initialize the potential coordinates through PCA [43], the kernel parameters and the corresponding values in the constants, and finally used the dataset in Table 2 to learn the four types of human joint point models suitable for intersections and crosswalks in Figure 5.

3.4. Pedestrian Path Prediction by KPOF-GPDM. Pedestrians are prone to wandering at intersections. At crosswalks, they may also stand or walk due to the status of traffic lights and vehicle driving on the road. However, the use of optical flow information alone can lead to the loss of some posture information of the walking or stopping state. Therefore, we introduce the human joint point information into the optical flow information prediction algorithm to form the KPOF-GPDM algorithm in this article. The algorithm supplements the detailed information of the pedestrian's posture during the movement and can more accurately predict the pedestrian's movement state and movement trajectory in the future.

Table 2: Pedestrian data sequences.

	Sequence				
Orientation	Stopping	Walking	Stopping tendency	Walking tendency	Total
Left to right	18	15	32	36	101
Right to left	15	17	37	30	99
Total	33	32	69	66	200

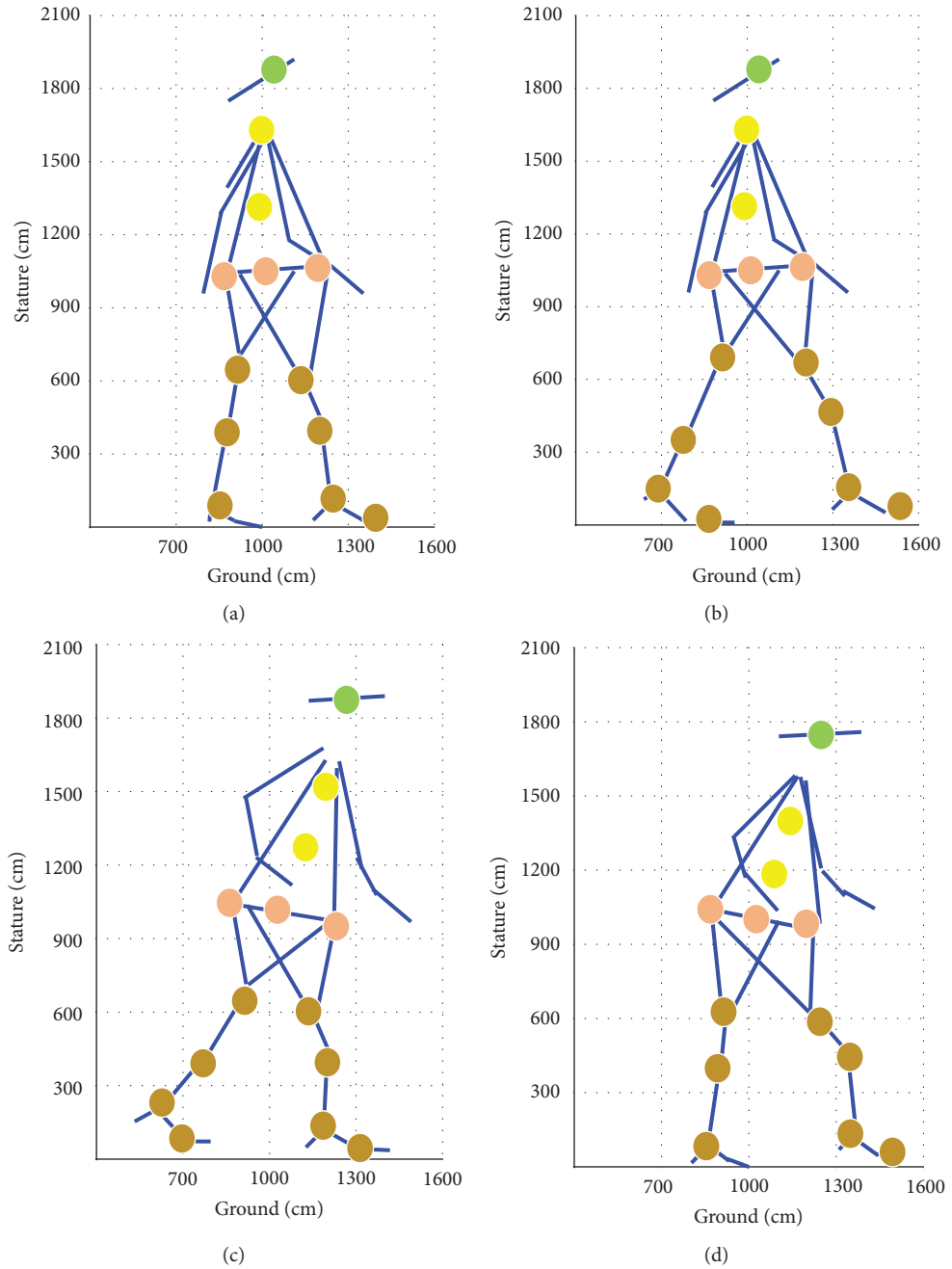

Figure 5: Four types of human joint point models are shown. When the pedestrian is in the stopping state, the whole body is in a vertical posture, and the direct distance between the legs is 0–30 cm. When the pedestrian is in the walking state, the whole body is basically in a vertical posture, and the distance between the feet is 50–70 cm. When a pedestrian has a change in the state of motion, the upper limbs will lean forward, but the distance between the pedestrian's legs in (c) is 40–60 cm, and the distance between the legs in (d) is about 0–30 cm. (a) Stopping. (b) Walking. (c) Stopping tendency. (d) Walking tendency.

KPOF-GPDM integrates pedestrian optical flow characteristics and joint point information to predict the pedestrian's lateral motion state at crosswalks and intersections and the motion trajectory in world coordinates. Firstly, we construct the pedestrian's lateral attitude information in the world coordinate system. Then in the same world coordinate system, the pedestrian's movement state at the intersection and crosswalk, the detailed information of the upper and lower limbs are extracted, and the pedestrian optical flow information and the joint point information are merged. Secondly, the underlying spatial dynamic model GPDM is used to reduce the dimension of the feature information. Finally, trajectory prediction and motion feature reconstruction are performed in low dimensional space.

3.4.1. The Lateral Position of Pedestrians in World Coordinates. We need pedestrian distance information when we establish the mapping between pedestrian dynamic optical flow features and real pedestrian speed. In monocular ranging model, we assume that the road surface is flat and pedestrians walk upright. In this process, we need to calibrate the camera's internal and external parameters. The camera's internal parameters are fixed. The camera's height and pitch angle will remain unchanged once the camera is fixed on the vehicle. Based on the above premises, the world coordinate system can be established. The projection of the camera's optical axis on the ground is the center, the direction of the vehicle is X axis, and Y axis is perpendicular to the ground.

We regard the center of the pedestrian's projection on the ground as the ranging point. Correspondingly, we use the projection of the pedestrian's abdominal transverse center on the bottom of the pedestrian mask to calculate the ranging points $p(u, v)$. Among them, u is the maximum value in the Y direction of the mask area, and v is the average value of the x-axis in the mask area:

$$\begin{aligned} u &= \max\left(y_{\text{mask}}\right), \\ v &= \text{mean}\left(x_{\text{mask}}\right). \end{aligned} \tag{7}$$

According to reference [44], the focal length of the camera is f and the ranging point is $p(u, v)$. Then, the world coordinates of pedestrians can be obtained as follows:

$$\begin{aligned} Y &= \frac{H}{\tan\left(\alpha - \arctan\left(\left(v - v_0\right)/f_y\right)\right)}, \\ X &= \frac{\left(u - u_0\right) \times H}{\sqrt{f_x^2 + \left(v - v_0\right)^2} \times \sin\left(\alpha - \arctan\left(\left(v - v_0\right)/f_y\right)\right)}. \end{aligned} \tag{8}$$

With the angle between the optical axis and the horizontal road surface α, the height between the camera and the ground is H. $f_y = f/d_y$; $f_x = f/d_x$; and d_x and d_y represent the pixels distance in image coordinates u and v. Then, the distance between pedestrians and vehicle is computed using $D(P) = \sqrt{X^2 + Y^2}$.

3.4.2. Key Point Feature Fuse. In the fusion module of optical flow information and connection point model, firstly, pedestrian horizontal optical flow is the transverse component of dynamic optical flow. The pedestrian mask area obtained by self-flow can locate the pedestrian position and obtain more accurate optical flow characteristics. At the same time, complete pedestrian mask can be repaired through the UCTGAN network. Simultaneously, the pedestrian connection point model we trained in Section 3.3 is used to identify the pedestrian connection points in the image sequence. Finally, in the same world coordinate system, pedestrian optical flow information and human body node information are fused; the specific process is shown in Figure 6.

When only using optical flow information to predict pedestrian trajectories, the predicted pedestrian optical flow velocity V_{of} can help the ADAS to predict the motion trajectory of the behavior in the future to a certain extent. But, at this time, V_{of} reflects the overall speed of a pedestrian, while the detailed information of the upper and lower limbs of the pedestrian in the process of walking and parking is lost, and the pedestrian track with higher accuracy cannot be obtained in the future. Therefore, this article introduces the human body joint point information to form a trajectory prediction model that combines optical flow information with human optical nodes. When pedestrians pass through an intersection or crosswalk, they will not only collect current traffic information in real time by leaning forward but also adjust the movements of their upper and lower limbs to reflect changes in their own movement status. For example, when there are vehicles ahead, the upper limbs of the pedestrian in the walking state will lean forward to judge the forward traffic situation, the lateral velocity V_{Tkp_i} of the joint points of the upper limb of the human body will gradually slow down. The lateral spacing between the joint points of the lower limbs gradually decreases, the lateral velocity $V_{Dk\ p_i}$ gradually decreases. At this time, the pedestrian movement status changes from walking to stopping. When the front traffic condition is good, the upper limbs of pedestrians will also lean forward to judge the current traffic situation, and the lateral velocity V_{Tkp_i} of the joint points of the upper limbs will increase positively. The distance between the joint points of the lower limbs and the lateral velocity are also gradually increasing. At this time, the pedestrian motion state shows a trend of gradually changing from the stopping state to the walking state.

In the traditional optical flow trajectory prediction, the car's ADAS only uses V_{of} in equation (9) to calculate the pedestrian's speed on road, and then roughly predicts the pedestrian's motion state and trajectory in a Gaussian low-dimensional space, unable to provide effective anti-collision data. After introducing the information of human joints, the ADAS uses the detailed information of the joints of the upper and lower limbs of the human body to estimate the trend of pedestrian movement in the future. It also uses formula (10) to calculate the speed of the upper and lower limbs of the pedestrian, which reflects the detailed information of the pedestrian when facing different traffic con-

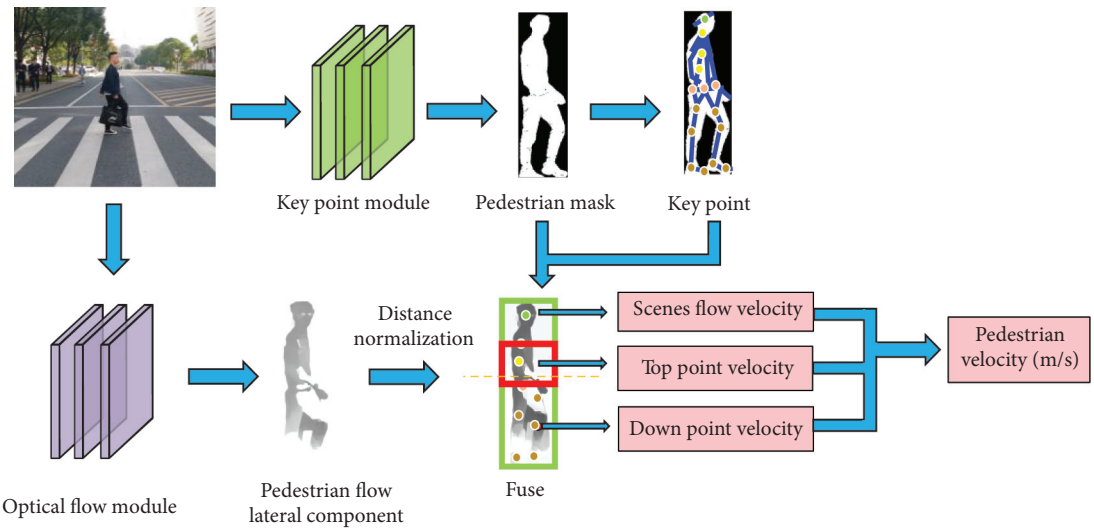

Figure 6: The fusion process of optical flow information and node model can improve the accuracy of pedestrian trajectory prediction through the fusion of pedestrian nodes with different motion states and different postures.

ditions, and provides more accurate data input for the back-end motion state and trajectory prediction. At this time, the three calculation data of V_{Tkp_i}, $V_{Dk\ p_i}$, and $V_X(p)$ are used to predict the pedestrian's motion state, and the trajectory prediction can better reflect the real pedestrian state and provide more effective active safety data for the car's ADAS.

With the pedestrian's lateral optical flow V_{of}, lateral velocity of key point V_{Tkp_i} and $V_{Dk\ p_i}$, $(V_{Tkp_i}, V_{Dk\ p_i} \in V_{kp_i})$, and the camera interval of each frame Δt, the pedestrians' walking speed $V_X(p)$ and X_t can be calculated by the following formula:

$$V_X(p) = \frac{\tau \cdot V_{of}(p)}{D \cdot \Delta t}, \tag{9}$$

$$V_{T/Dk\ p_i} = \frac{1}{n} \cdot \frac{\tau \cdot \sum_{i=0}^{n}\left(V_{kp_i}\right)}{D \cdot \Delta t}, \tag{10}$$

$$V_X(p) = \frac{1}{(n+1)} \frac{\tau \cdot \left(V_{of}(p) + \sum_{i=0}^{n}\left(V_{kp_i}\right)\right)}{D \cdot \Delta t}, \tag{11}$$

where $V_X(p)$ is the velocity of pedestrians, the function D represents pedestrian distance, and τ is a constant. The average value of velocity optical flow in pedestrian upper body can be regarded as the average speed of the pedestrian $\overline{v}$. We resize $V_X(p)$ to 32×16 pixel. We construct a feature vector $y_t \in \mathbb{R}^D$, $(D = 515)$ that includes position, average speed, and velocity optical flow.

3.4.3. Gaussian Model. GPDM is a latent variable model. It established the mapping relation from a latent space x_t to observation space y_t and a latent dynamical model which account for the temporal dependence on pedestrian motion features [39].

The observation space $Y = [y_1, y_2, \ldots, y_N]^T$ is N frames motion feature vector. $X = [x_1, x_2, \ldots, x_N]^T$ is the dynamic mapping on latent positions. The mapping relation can be described:

$$p(Y|X, \overline{\beta}, W) = \frac{|W|^N}{\sqrt{(2\pi)^{N\times 16}|K_Y|^{16}}} \exp\left(-\frac{1}{2} T_r\left(K_Y^{-1} Y W^2 Y^T\right)\right), \tag{12}$$

where K_Y is a kernel matrix of size $N \times N$ constructed by the kernel function κ_Y. The parameter of the kernel matrix is $\overline{\beta} = \{\beta_1, \beta_2, \beta_3\}$. For our data, we use the RBF (radial basis function) kernel $\kappa_Y(x, x') = \beta_1 \exp(-(\beta_2/2)\|x - x'\|^2) + \beta_3^{-1}\delta_{x,x'}$; W is a $D \times D$ diagonal matrix that represents the weight of different dimensions of y_t. Assuming that the dynamics of the data in the latent space x_t satisfies the first-order Markov model, the dynamics of the time series data is incorporated using

$$p(X|\overline{\alpha}) = \frac{p(x_1)}{\sqrt{(2\pi)^{(N-1)\times d}|K_X|^d}} \exp\left(-\frac{1}{2} Tr\left(K_X^{-1} X_{2:N} X_{2:N}^T\right)\right), \tag{13}$$

where $X_{2:N} = [x_2, \ldots, x_N]^T$, the kernel matrix K_X is $(N-1) \times (N-1)$ that is constructed from $X_{1:N-1} = [x_1, \ldots, x_{N-1}]^T$ and defined by a kernel function $\kappa_X(x, x')$. We use RBF and a linear kernel in the kernel function with kernel hyperparameters $\overline{\alpha} = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$:

$$\kappa_X(x, x') = \alpha_1 \exp\left(-\frac{\alpha_2}{2}\|x - x'\|^2\right) + \alpha_3 x^T x' \alpha_4^{-1} \delta_{x,x'}. \tag{14}$$

Latent mapping and latent dynamics model combined with time series observations:

$$p(X, Y, \overline{\alpha}, \overline{\beta}, W) = p(Y|X, \overline{\beta}, W) p(X|\overline{\alpha}) p(\overline{\alpha}) p(\overline{\beta}) p(W). \tag{15}$$

The process of GPDM inference is finding hidden space variables X and kernel parameters $\{\overline{\alpha}, \overline{\beta}\}$ by minimizing the negative logarithm joint posterior $-\ln p(X, \overline{\alpha}, \overline{\beta}|Y)$. It can be optimized by the scaled conjugate gradient (SCD) algorithm. The dimension of latent space $d = 3$. Figure 7 illustrates this

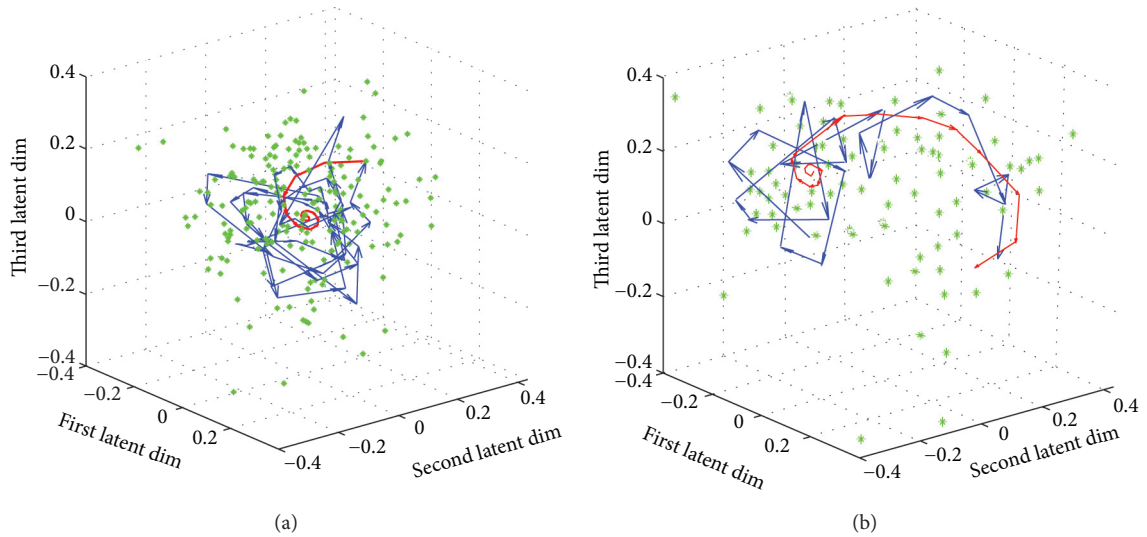

FIGURE 7: (a) The hidden space 3D trajectory obtained by walking data training and (b) hidden space 3D trajectory obtained by stopping data training. The green lines represent the trajectories of pedestrian walking and stopping features under hidden space projection, respectively, and red represents the average predicted trajectory learned by KPOF-GPDM.

mean prediction of a point for several frames on the low-dimensional space.

The motion state of the pedestrian at time t is described by $\phi_t = [x_t, X_t]$, where $x_t \in \Re^d$ is a point in the low-dimensional space and X_t is the horizontal position of the pedestrian in practice. Given an observed motion feature y_t and observed lateral position Y_t, the probability of a pedestrian state ϕ_t is computed by

$$p(\phi_t|y_t, Y_t) = \eta p(y_t, Y_t|\phi_t) \int p(\phi_t|\phi_{t-1}) p(\phi_{t-1}|y_{t-1}, Y_{t-1}) d_{\phi_{t-1}}, \tag{16}$$

with normalization constant η. The probability $p(\phi_t|\phi_{t-1})$ of observing a future state is computed from the GPDM latent space mean prediction.

3.4.4. Motion Feature Reconstruction. KPOF-GPDM can be obtained by Bayesian law which can generate new observation sequences. With the trained model $\Gamma = \{Y, X, \overline{\alpha}, \overline{\beta}, W\}$, a new observation sequence and the joint conditional distribution of the scene stream feature corresponding to the hidden space feature sequence is expressed as

$$p(Y^*, X^*|\Gamma) = p(Y^*|X^*, \Gamma) p(X^*|\Gamma). \tag{17}$$

The new latent variable sequence x^* can be predicted by maximizing the formula (16). The process of predicting a new sequence by the first latent variable X_1 requires two steps:

(a) A new latent space variable is predicted based on the data at the previous time.

$$\mu_X(x^*) = k_X(x_{t-1})^T K_X^{-1} X_{2:N}, \tag{18}$$

where the vector $k_X(x)$ is containing $k_X(x, x_i)$ in the ith entry, and x_i is the ith training vector.

(b) The new data in the observation space is constructed using

$$\mu_Y(x^*) = k_Y(x^*)^{\mathrm{T}} K_Y^{-1} Y. \tag{19}$$

Figure 8 shows the reconstructed scene flow when pedestrians cross the road, and Figure 9 shows the pedestrian velocity in the future.

4. Experiment

We deploy our system on PX2 mobile devices, using its TensorRT neural network inference engine and cuDNN deep neural network library to improve its real-time performance. During the experiment, on the one hand, our system and four excellent pedestrian trajectory prediction systems (KF, IMM-KF, HoM/Traj [3], and SFlowX/GPDM [30]) are placed in the same video sequence to compare their pedestrian trajectory position prediction accuracy and pedestrian action classification probability accuracy. On the other hand, our fusion model is compared with the prediction model using only one piece of information to verify the improvement in our system performance. In this section, we use a monocular camera (baseline 33 cm, 30 fps) mounted on the inside of the windshield and behind the rearview mirror to collect video data at a busy intersection in the campus. Video data are divided into two scenarios, both of which are pedestrians crossing the road on the crosswalk without being covered. In the first scene, Figure 10(a), when pedestrians stop at the side of the road, they observe the traffic flow on the road and decide whether to stay in place or walk through the crosswalk for the next stage of action. In the

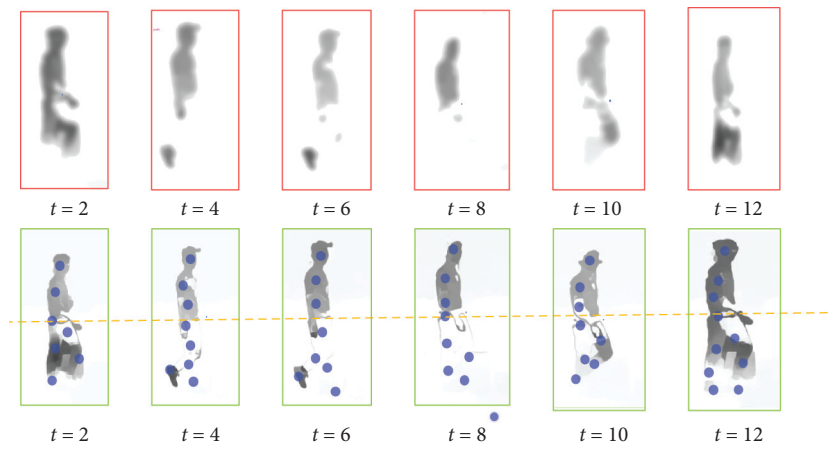

Figure 8: (Top row) Reconstructed optical flow based on current state ($t = 0$) and state predictions ($t = 2, \ldots, 12$) in low-dimensional latent space. (Bottom row) Optical flow that is (will be) actually measured at the corresponding time steps.

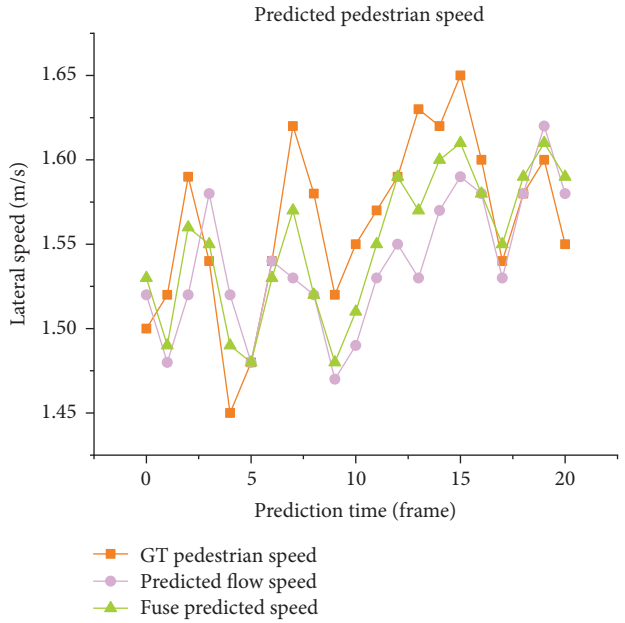

Figure 9: The comparison between using only optical flow and using the fusion model to predict pedestrian speed and the true value shows that KPOF-Net can predict pedestrian speeds that are closer to the true value.

(a) (b)

Figure 10: Campus experiment scene. (a) Pedestrians stand by the roadside. (b) Pedestrians pass the crosswalk.

second scene, Figure 10(b), when pedestrians pass the crosswalk, they judge the traffic situation in real time and change their own motion state.

In order to evaluate the performance of the pedestrian trajectory prediction and action classification in this paper, we first evaluate the performance of the pedestrian motion

classification model through experiments. Through manual annotation of pedestrians in the images of the shape, we obtained the coordinates of the ground truth (GT) of pedestrians in the world coordinate system. Since pedestrian basically walk upright, set the pedestrian's standing point as the center origin instead of the pedestrian's center of gravity, so that we can obtain the horizontal and vertical coordinates of the pedestrian on the ground in order to obtain more reliable basic data. Due to the displacement, it is not reliable to show an action classification based on starting and stopping (with displacement of starting and stopping), and we use the key information to get body posture. At the beginning, the legs are in a separate position, and at the stop, the legs shift from closed to walking, and the arms and legs swing alternately. By combining posture and displacement, we can better classify movement types. In the experiment, the tester must choose to stop or cross the road, and the probability of pedestrian action classification is expressed by the values of [0,1], and the probability value is calculated by formula (16). In terms of alignment along the time axis, for each trajectory in which the pedestrian is stopping, the moment of the last placement of the foot is labeled as the stopping moment. The time-to-stop (TTS) value is used to count the number of frames before the event, meaning that the TTS value for the frames before the stop event is positive and the TTS value for the frames after the stop event is negative. In sequences in which the pedestrian continues walking, the closest point to the curbstone (with closed legs) is labeled. Analogous to the TTS definition, the latter is called the time-to-curb (TTC) value. We assume that TTC = 0 represents the time when pedestrians change from walking to stopping; when the TTC is bigger than 0, it represents the previous frames of stopping, and when the TTC is less than 0, it represents the frames after stopping.

4.1. Pedestrian Action Classification. We compare the predicted probabilities of the system with three excellent systems in the same video sequence set for its action classification, and judge whether pedestrians will stop or walk in a short time in the future. When pedestrians are at intersections or crosswalks, they will choose to stand or walk because of the current status of traffic lights, road traffic, and their own decision-making awareness. Figure 11 shows the predicted probabilities of whether the five systems are walking or stopping in the future in the same sequence set.

In order to fully test the performance of the five systems for pedestrian motion classification, we tested multiple video datasets on campus. Pedestrians wandered at intersections and choose to stop at the crosswalk due to road traffic conditions. When the current safe traffic environment is determined, they resumed walking and other movements. For each test sequence, we used a slider between 0 and 1 to provide probability (confidence), which is displayed in the most intuitive way.

In Figure 11, we can see that when the car is in motion, it has an impact on the predicted probability of our state change. When the pedestrian state is about to change, the prediction probability of the pedestrian state change in the dynamic scene of the vehicle is lower than that in the static scene, and the prediction ability is reduced. However, our system has the least decline range and the least impact. At the same time, regardless of whether the car is in a stopping or moving state, when the pedestrian state changes (from walking state to stopping state or from stopping state to walking), each system starts with a low probability, and predicts the probability gradually as the pedestrian state changes increase. However, whether the car is in motion or stopped, it can be clearly seen that our system is more sensitive than other systems, reacts more quickly, and can keenly grasp the characteristics of the human body when the pedestrian's state changes, so as to predict the change of the pedestrian's state more quickly.

In the classification and discrimination of pedestrian movement, since our system has learned four types of human joint point models in Section 3.3, it not only includes two basic models of standing and walking but also two pedestrian joint point models of standing tendency and walking tendency. When pedestrians are about to stop or walk at intersections or crosswalks, they will judge the traffic situation ahead by leaning forward to prepare for changes in their own state. As a result, we can capture more human body posture information, prompting our system to achieve better prediction results than other systems in experiments.

At each moment of the input trajectory, we determine the category membership degree by estimating the stopping probability through the threshold, adjusted the parameters through experiments, and set the appropriate threshold. When the probability of our state change is greater than the threshold, we determine that our pedestrian state is walking or stopping. In our experimental results, when the car is in a stopped state, there is a probability of 0.402 to predict the future state of the pedestrian 7 frames before the state change.

4.2. Pedestrian Trajectory Prediction. We also attach importance to the system's ability to predict the accuracy of pedestrian location. Accurate location information can establish an excellent pedestrian prediction model and provide auxiliary information for ADAS functions. During the test, we considered that the state of the car's movement also affects the ADAS's prediction of the change of the pedestrian's movement state. Therefore, in Table 3, we collected video sequences of the car in different states of movement. We compared the positioning accuracy between systems by the average value and standard deviation of the RMSE of each video sequence. The range of pedestrian frames is [−20, 15], when frame 0 means the manually labeled TTS/TTC moment. The position between [0,15] is predicted by the system, which represents the comparison between our positioning accuracy. It can be seen from Table 4 that when the pedestrian is in motion, all systems can capture enough pedestrian posture information, and obtain a smaller posture prediction error compared to the pedestrian stopped state. At the same time, no matter whether the pedestrian is walking or stopping, our system can extract more pedestrian pose information by fusing optical flow information and

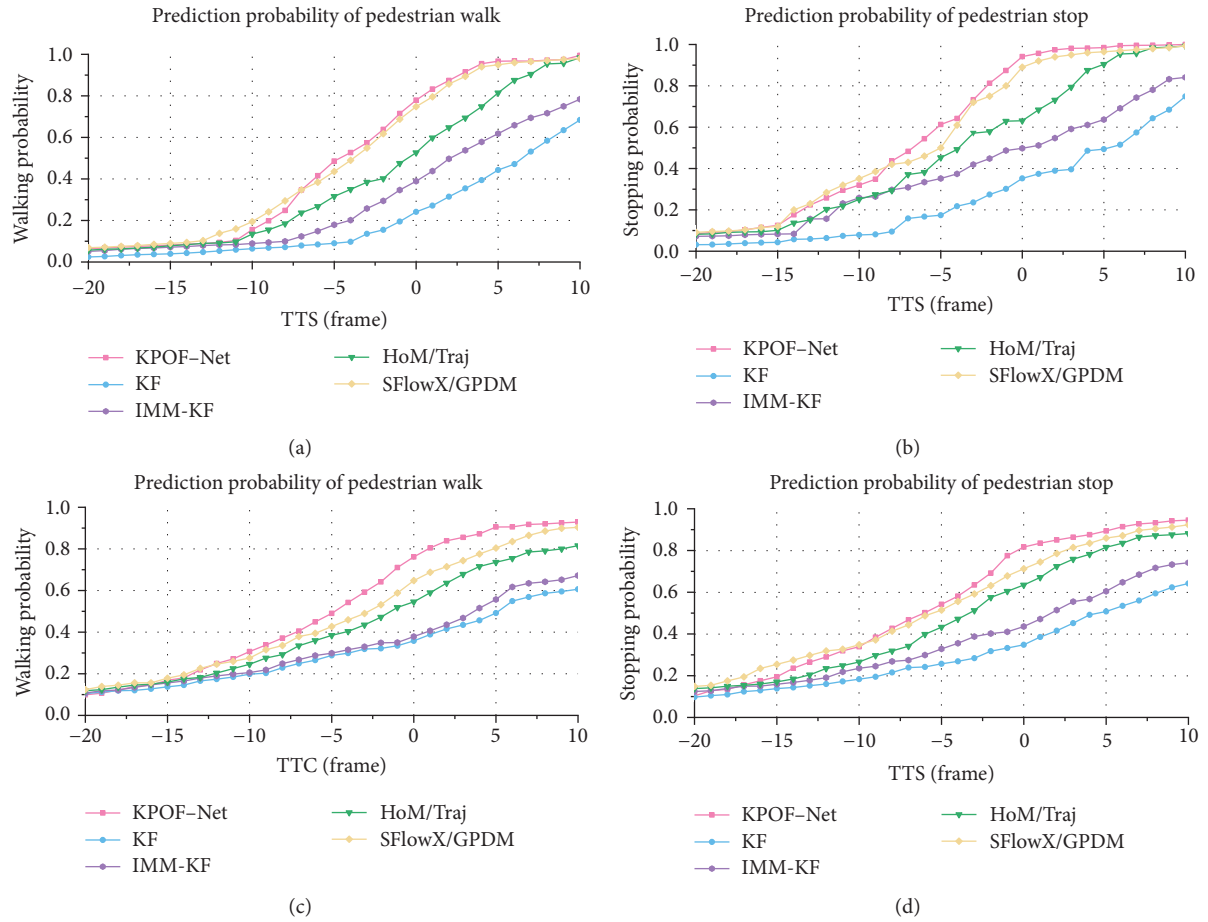

Figure 11: (a) Prediction probability of pedestrian walk. (b) Prediction probability of pedestrian stop. (c) Prediction probability of pedestrian walk. (d) Prediction probability of pedestrian stop. (a, b) The car is in a stopped state and is stopped in front of the sidewalk, and the probability of the state change at the pedestrian intersection and sidewalk is measured. (c, d) The car is in a state of motion, and when it is gradually approaching the sidewalk, we predict the probability of the pedestrian's state change.

Table 3: Data source method.

Sequence	Vehicle standing	Vehicle moving	Vehicle standing + moving
Ped. Stopping	11	5	16
Ped. Walking	9	4	13

Table 4: Pedestrian lateral trajectory prediction error.

Systems		State			
		Walking		Stopping	
		0	15	0	15
KF	Mean	0.28	0.62	0.43	1.27
	±Std	0.05	0.25	0.09	0.24
IMM-KF	Mean	0.34	0.58	0.62	1.15
	±Std	0.06	0.34	0.15	0.31
HoM/Traj	Mean	0.22	0.43	0.31	0.82
	±Std	0.03	0.13	0.09	0.24
SFlow/GPDM	Mean	0.17	0.51	0.37	0.54
	±Std	0.06	0.27	0.08	0.18
KPOF-Net	Mean	0.15	0.38	0.27	0.42
	0 ± Std	0.04	0.13	0.05	0.14

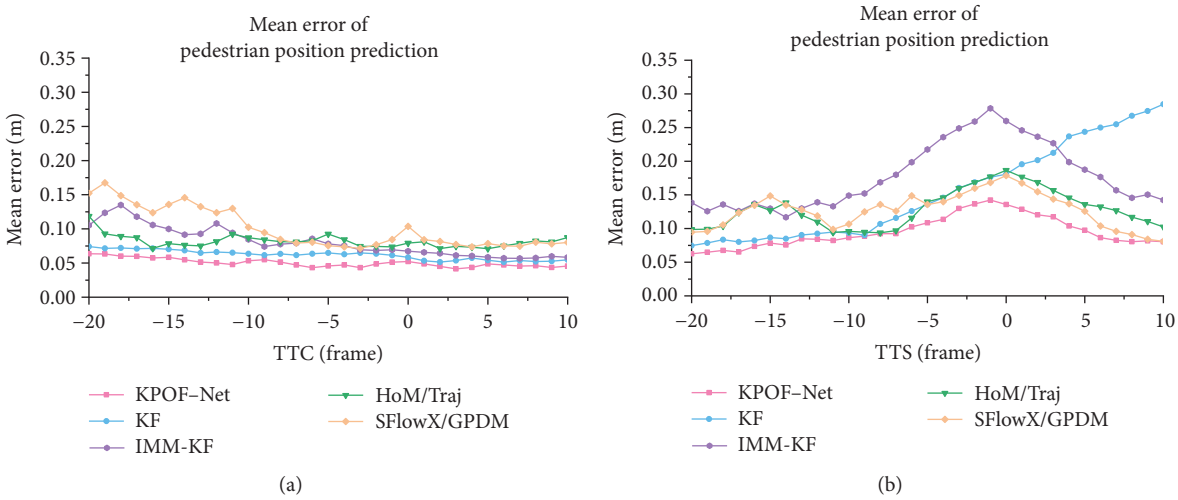

Figure 12: (a) Graph shows the performance of KPOF-Net's pedestrian trajectory prediction accuracy. (b) Curve shows that KPOF-Net is better than others when the state changes. (a, b) The cars are slowly approaching the crosswalk. An excellent system has a faster response speed and can handle the trajectory prediction error caused by the state change in time.

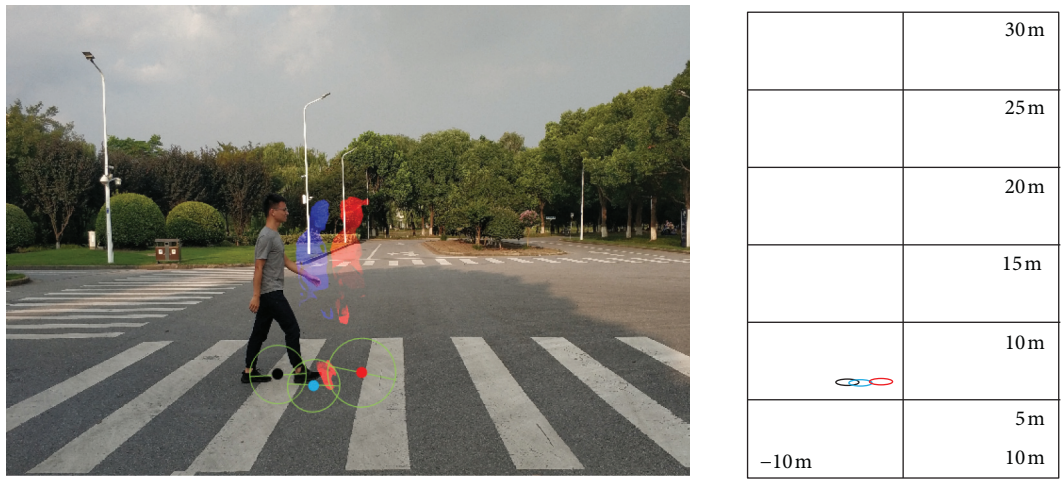

Figure 13: The trajectory prediction model at 0.25 s and 0.5 s when traveling without occlusion.

Table 5: Euclidean distance error analysis.

Metric	Datasets	Point	Flow	Point + flow
Average displacement error (m)	Video 1	0.1832	0.2418	0.1047
	Video 2	0.2463	0.1716	0.0781
	Video 3	0.2314	0.2159	0.0948
Final displacement error (m)	Video 1	0.1657	0.2314	0.1164
	Video 2	0.2591	0.1843	0.0841
	Video 3	0.2534	0.2317	0.0973

joint point information, which enables us to predict the position of pedestrians more accurately. Also, it can be found from Figure 12 that when pedestrians are always in the same state of motion, KPOF-Net shows better prediction performance than other excellent systems, and can always maintain a low pedestrian position prediction error. When the pedestrian motion status changes, with a strong anti-interference ability, our KPOF-Net can suppress the increase in pedestrian position prediction error caused by the motion status change, which is faster than other systems.

4.3. Pedestrian Prediction Model. After removing dynamic object occlusions and fusing the repaired optical flow information with the human joint point model, our system has improved pedestrian action classification and pedestrian trajectory prediction compared with several other excellent systems with excellent experimental results. At the same time, in order to verify the performance improvement of the prediction model that integrates optical flow information and joint point information, we run the KPOF-Net model, single optical flow information prediction model, and joint point prediction model in pedestrian unobstructed video sequences. In this experiment, consider the system's real-time performance and trajectory prediction accuracy on

common hardware devices. On the one hand, the judgment logic we set in the algorithm will skip the WEC link directly when the pedestrian is in an unobstructed state. On the other hand, the experiment in this section is carried out at a low vehicle speed of 15 km/h. And, we propose Average Displacement Error (ADE) and Final Displacement Error (FDE) evaluation rules to verify the performance improvement of our system.

The results of prediction experiments are shown in Figure 13, in which the blue optical flow is the predicted pedestrian position after 0.25 s and the red optical flow is the predicted pedestrian position after 0.5 s.

Next, we make path prediction and select the most accurate activity model to estimate the future state of pedestrians, and two error indexes are used to evaluate the overall position prediction. The comparison results are shown in Table 5:

(1) Average displacement error (ADE): the average Euclidean distance between the predicted location and the actual location over time.

(2) Final displacement error (FDE): the Euclidean distance between the Final predicted location and the real point on the ground.

By comparing the ADE and FDE of the three model systems in multiple video sequences, we can find that when only the optical flow information and joint point information are used to predict the pedestrian trajectory, the ADE and FDE values fluctuate around 0.2 m. There is still a large deviation between the position information predicted by using only the optical flow information or optical node information and the actual position of the pedestrian. The prediction accuracy of KPOF-Net can be set to reach 0.1 m, which can effectively predict the trajectory of pedestrians in the future.

5. Conclusion

At intersections and crosswalks, in the optical flow module, we use the self-flow network to obtain pedestrian optical flow information containing obstructed objects, and then propose the WEC algorithm to segment the obstructed objects from pedestrians, and finally use the UCTGAN Network to restore the pedestrian optical flow image. In the human body joint point module, four human body joint point models are trained: standing, stopping, walking tendency, and stopping tendency. After completing the optical flow module and the human optical node module, we merge the two modules to form a KPOF-Net network for pedestrian trajectory prediction. The network supplements the detailed information of the body movement of pedestrians when passing intersections and crosswalks in the optical flow information of the human body. At the same time, we compare the KPOF-Net system with KF, IMM-KF, HoM/Traj, and SFlowX/GPDM, four excellent pedestrian trajectory prediction systems. Experiments show that after our KPOF-Net system integrates optical flow information and human body joint point information, both the probability of pedestrian state prediction and the accuracy of pedestrian trajectory are improved. Even in the time when the pedestrian state changes, the prediction accuracy of pedestrian estimation fluctuates greatly, and it can respond quickly, restrain the increase of the error, and restore the accuracy to a normal value. It can be found from the full text that the KPOF-Net prediction model after fusion of human joint point information could provide accurate auxiliary information for our advanced driving assistance system.

Data Availability

Some relevant data are available on the website https://github.com/604627144/KPOF-GPDM, wherein some codes, pedestrian data, and experimental results will be presented.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] G. Chen, F. Wang, S. Qu et al., "Pseudo-image and sparse points: vehicle detection with 2D LiDAR revisited by deep learning-based methods," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13. In press, 2020.

[2] Y. Liu, G. Chen, and A. Knoll, "Globally optimal vertical direction estimation in Atlanta World," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, In press.

[3] Mercedes-Benz E-Class, *Passenger car yearbook*, pp. 117-118, SAE, Warrendale, PA, USA, 2014.

[4] A. Dosovitskiy, P. Fischer, I. Eddy et al., "Flownet: learning optical flow with convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, December 2015.

[5] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[6] T.-W. Hui, X. Tang, and C. Change Loy, "LiteFlowNet: a lightweight convolutional neural network for optical flow estimation," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*, Salt Lake City, UT, USA, June 2018.

[7] G. Chen, H. Cao, J. Conradt, H. Tang, F. Rohrbein, and A. Knoll, "Event-based neuromorphic vision for autonomous driving: a paradigm shift for bio-inspired visual sensing and perception," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 34–49, 2020.

[8] N. Mayer, I. Eddy, P. Hausser et al., "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.

[9] J. Y. Jason, A. W. Harley, and K. G. Derpanis, "Back to basics: unsupervised learning of optical flow via brightness constancy and motion smoothness," in *Lecture Notes in Computer Science, Computer Vision – ECCV 2016 Workshops*, Springer, Berlin, Germany, 2016.

[10] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, and H. Zha, "Unsupervised deep learning for optical flow estimation," in *Proceedings of the Thirty-First AAAI Conference on Artificial*

Intelligence (AAAI-17), San Francisco, CA, USA, February 2017.

[11] J. Janai, F. G¨uney, A. Ranjan, J. B. Michael, and A. Geiger, "Unsupervised learning of multi-frame optical flow with occlusions," in *Computer Vision—ECCV 2018, Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2018.

[12] P. Liu, I. King, M. R. Lyu, and X. Jia, "Ddflow: learning optical flow with unlabeled data distillation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8770–8777, 2019.

[13] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: a survey," 2019, https://arxiv.org/abs/1902.06162.

[14] D. Pathak, P. Krahenbuhl, J. Donahue, Trevor Darrell, and A. A. Efros, "Context encoders: feature learning by inpainting," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.

[15] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[16] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Computer Vision—ECCV 2016, Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2016.

[17] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[18] X. Zhai, A. Oliver, A. Kolesnikov et al., "S4L: self-supervised semi-supervised learning," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, November 2019.

[19] P. Liu, M. Lyu, I. King et al., "SelFlow: self-supervised learning of optical flow," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019.

[20] J. Canny, "A computational approach to Edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[21] L. Zhao, Q. Mo, S. Lin et al., "UCTGAN: diverse image inpainting based on unsupervised cross-space translation," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5740–5749, Seattle, WA, USA, June 2020.

[22] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: a survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, Article ID 102897, 2020.

[23] Z. Sun, "A survey of multiple pedestrian tracking based on tracking-by-detection framework," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, 2021.

[24] Y. Bar-Shalom, X. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*, Wiley, Hoboken, NJ, USA, 2001.

[25] W. Choi and S. Savarese, "Understanding collective activitiesof people from videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1242–1257, 2014.

[26] Z. Hu, "Deep convolutional neural network-based Bernoulli heatmap for head pose estimation," *Neurocomputing*, vol. 436, pp. 198–209, 2021.

[27] Z. Hu, "A CRNN module for hand pose estimation," *Neurocomputing*, vol. 333p. 156, 168.

[28] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto, "Intent-aware long-term prediction of pedestrian motion," in *Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2543–2549, Stockholm, Sweden, May 2016.

[29] A. Marginean, R. Brehar, and M. Negru, "Understanding pedestrian behaviour with pose estimation and recurrent networks," in *Proceedings of the 2019 6th International Symposium on Electrical and Electronics Engineering (ISEEE)*, pp. 1–6, Galati, Romania, October 2019.

[30] C. G. Keller and D. M. Gavrila, "Will the pedestrian cross? a study on pedestrian path prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 494–506, 2014.

[31] M. Goldhammer, S. K¨ohler, K. Doll, and B. Sick, "Camera based pedestrian path prediction by means of polynomial least-squares approximation and multilayer perceptron neural networks," in *Proceedings of the SAI Intelligent Systems Conference (IntelliSys), 2015*, pp. 390–399, London, UK, November 2015.

[32] S. Zernetsch, S. Kohnen, M. Goldhammer, K. Doll, and B. Sick, "Trajectory prediction of cyclists using a physical model and an artificial neural network," in *Proceedings of the 2016 Intelligent Vehicles Symposium (IV)*, pp. 833–838, Gothenburg, Sweden, June 2016.

[33] K. G. Alahi, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–971, Las Vegas, NV, USA, June 2016.

[34] R. Urtasun, D. J. Fleet, and P. Fua, "3D people tracking with Gaussian process dynamical models," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 238–245, New York, NY, USA, June 2006.

[35] R. Quintero Minguez, I. Parra Alonso, D. Fernandez-Llorca, and M. A. Sotelo, "Pedestrian path, pose, and intention prediction through Gaussian process dynamical models and pedestrian activity recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1803–1814, 2019.

[36] V. Kress, S. Zernetsch, K. Doll, and B. Sick, "Pose based trajectory forecast of vulnerable road users," in *Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1200–1207, Xiamen, China, December 2019.

[37] CMU graphics lab motion capture database, http://mocap.cs.cmu.edu/.

[38] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*, Salt Lake City, UT, USA, June 2018.

[39] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 283–298, 2008.

[40] R. Quintero, J. Almeida, D. F. Llorca, and M. A. Sotelo, "Pedestrian path prediction using body language traits," in *Proceedings of the 2014 IEEE Intelligent Vehicles Symposium*, pp. 317–323, Dearborn, MI, USA, June 2014.

[41] S. Worrall, "Multi-sensor detection of pedestrian position and behaviour," in *Proceedings of the 23rd World Congress on Intelligent Transport Systems*, pp. 1–12, Tokyo, Japan, 2016.

[42] R. Quintero, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, "Pedestrian intention recognition by means of a Hidden Markov Model and body language," in *Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–7, Yokohama, Japan, October 2017.

[43] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[44] J.-T. Jun-Tao Xue, S.-P. Shi-Ming Wang, and S.-M. Wang, "Research of vehicle monocular measurement system based on computer vision," in *Proceedings of the 2013 International Conference on Machine Learning and Cybernetics*, pp. 957–961, Tianjin, China, July 2013.