ITC 3/53
Information Technology and Control
Vol. 53 / No. 3/ 2024
pp. 899-915
DOI 10.5755/j01.itc.53.3.34807

FHPE-Net: Pedestrian Intention Prediction Using Fusion with Head Pose Estimation Based on RNN

Received 2023/08/07 | Accepted after revision 2023/12/04

HOW TO CITE: Yang, Z., Guo, Z., Zhang, R., Guo, J., Zhou, Y. (2024). FHPE-Net: Pedestrian Intention Prediction Using Fusion with Head Pose Estimation Based on RNN. *Information Technology and Control*, 53(3), 899-915. https://doi.org/10.5755/j01.itc.53.3.34807

FHPE-Net: Pedestrian Intention Prediction Using Fusion with Head Pose Estimation Based on RNN

Zhiyong Yang
The College of Computer and Information Science, Chongqing Normal University, and the College of Big Data and Internet of Things, Chongqing Vocational Institute of Engineering, Chongqing, 402246, China;
e-mail: zyy@cqvie.edu.cn

Zihang Guo, Ruixiang Zhang, Jieru Guo
Chongqing Normal University School of Computer and Information Science, Chongqing, 401331, China

Yu Zhou
The College of Finance and Tourism, Chongqing Vocational Institute of Engineering, Chongqing, 402246, China

Corresponding author: zyy@cqvie.edu.cn

Accurate real-time prediction of pedestrian crossing intent during the autonomous driving process is crucial for ensuring the safety of both pedestrians and passengers, as well as improving riding comfort. However, existing methods for pedestrian crossing intent detection mostly rely on extracting complete pose information of pedestrians, leading to reduced accuracy when pedestrians are occluded. To address this issue, this paper proposes FHPE-Net: a lightweight, multi-branch prediction model that utilizes only the head pose features of pedestrians. In pedestrian crossing scenarios, pedestrian behavior is highly influenced by surrounding vehicles and the environment. FHPE-Net encodes pedestrian head poses and global context semantic image sequences to comprehensively capture spatiotemporal interaction features between pedestrians, vehicles, and the environment, thereby enhancing the accuracy of pedestrian crossing intent prediction. To improve the robustness of the FHPE-Net method, this study further processes bounding box positions and vehicle velocity features, making it more stable and reliable in complex traffic scenarios. Finally, a novel U-BiGRUs module is introduced for feature fusion, and an optimal fusion strategy is employed to achieve the best predictive performance in terms of F1 score and accuracy (ACC). Extensive ablation experiments are conducted on the PIE dataset, and performance analysis demonstrates that FHPE-Net achieves an accuracy of 90%, outperforming baseline methods such as PCPA and Multi-RNN, while using only pedestrian head pose features. This research holds significant guidance in enhancing traffic safety and optimizing urban traffic management. Furthermore, it provides essential technological support for advancing the commercialization of autonomous driving.

KEYWORDS: pedestrian action, autonomous vehicles, transport safety, fusion strategy.

1. Introduction

The widespread use of artificial intelligence (AI) and deep learning in recent years has led to the rapid development of autonomous driving technology. However, the interplay between vehicular systems and vulnerable road users (VRUs) (e.g., pedestrians) remains a significant barrier to developing fully intelligent driving vehicles for various applications. Many researchers are still exploring solutions for optimal interaction between vehicular systems and pedestrians, especially in detecting pedestrian crossing intentions. In L4-level autonomous driving (vehicles can be fully autonomous under certain conditions), a pedestrian crossing is one of the behaviours that can be solved most urgently.

Initially, researchers employed Convolutional Neural Networks (CNNs) [26] to redefine pedestrian intention prediction tasks as static image classification problems, ultimately relying solely on the final frame of the observed video to predict pedestrian crossing intentions.

This method neglected crucial information regarding the temporal orientation of video frames. Subsequently, with the development of Recurrent Neural Networks (RNNs), researchers began to predict pedestrian crossing intentions by analyzing the motion consistency of pedestrians' visual features over short time frames [17], [19], [27]. This resulted in a variety of methods to merge different features [14], [25], [28-30], [40], such as detected pedestrian boundary frames, human postures, behaviour, appearance, and current information about the vehicle.

Recently, the most recent benchmark for predicting pedestrian intent [20] was released, and the PCPA model outperformed all others on the widely used PIE dataset [31]. However, the integration of different forms of perceptual modal information extracted from additional networks can lead to large model sizes and slow inference. For effective application in real-world autonomous driving scenarios, however, the optimal decision model must operate effectively in real-time. To address this problem, a recent study [15] proposed a solution that uses only one add-on network [7] to extract pedestrian pose information for predicting pedestrian behaviour and understanding pedestrian intentions. Although the identification of human skeletal points can determine the movements of pedestrians before crossing the road and thus predict their crossing intentions, several obstacles, such as complex environmental conditions, occlusion, and varying distances between pedestrians and vehicles, make it difficult to accurately detect human skeletal points, thereby reducing the overall accuracy of motion recognition.

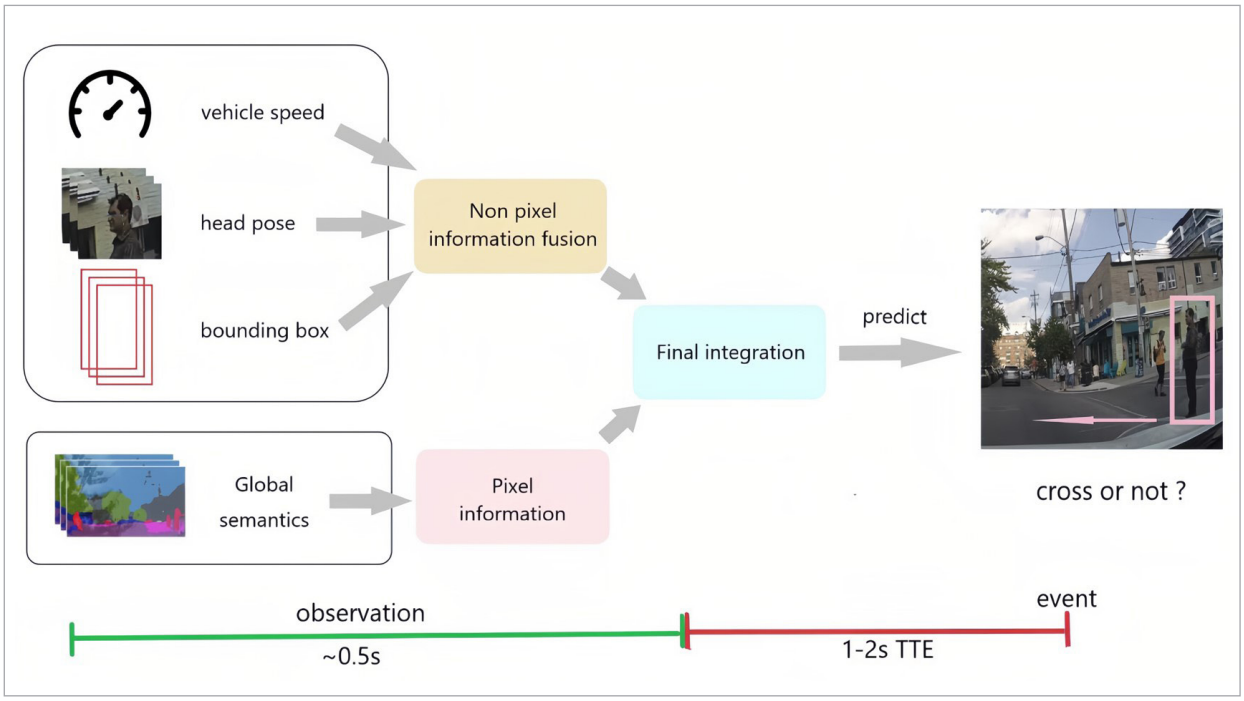
Another recent study [24] used head pose orientation features, including yaw, pitch, and roll, and applied a clustering algorithm to predict pedestrian crossing intentions with good performance. These results indicate that head pose is essential in predicting pedestrian crossing intentions. However, this method relies heavily on accurate head pose estimation, and incorrect estimates can significantly affect the accuracy of prediction. Furthermore, the current fusion strategy may not be optimal, necessitating further optimisation. In addition, identifying pedestrians' intentions only through pedestrian detection, tracking, trajectory prediction, and action recognition without considering contextual semantic information makes it difficult to accurately determine their crossing intentions as they are closely related to the traffic environment.

The existing methods for pedestrian crossing intent detection encounter challenges when pedestrians are occluded, as they heavily rely on extracting complete pose information, leading to reduced accuracy in such scenarios. To address this limitation and enhance the accuracy of pedestrian crossing intent prediction, we propose FHPE-Net, a lightweight, multi-branch prediction model that leverages the head pose features of pedestrians. Figure 1 illustrates how the proposed model integrates several key features, including pedestrian head posture direction features (yaw, pitch, and roll), global context semantic information (semantic segmentation of road, pedestrian, and vehicle), real-time vehicle speed, and pedestrian boundary box information. FHPE-Net employs a fusion mechanism that combines pixel-level and non-pixel-level information to achieve precise pedestrian crossing intention prediction.

The main contributions of the thesis can be summarized as follows:

– Firstly, the paper proposes FHPE-Net, a lightweight and efficient prediction model that focuses solely on utilizing the head pose features of pedestrians for pedestrian crossing intent detection. By avoiding

Figure 1
Predicting pedestrian behaviour, the goal is to predict whether a pedestrian will start crossing the street given an observation t of length m



the need for complete pose information, the model addresses the challenge of reduced accuracy in occluded pedestrian scenarios.

- Secondly, this paper introduces a novel U-BiGRUs module for feature fusion: an asymmetric bidirectional recursive architecture aimed at incorporating different features by leveraging bidirectional spatiotemporal context and long-term spatiotemporal information. Through extensive ablation experiments involving mixed fusion strategies, input configurations (adding or reducing input channels), and encoder options (attention mechanism and RNN), the paper determines the optimal model distribution. The U-BiGRUs module effectively integrates diverse information for enhanced feature fusion, enhancing the overall performance of the proposed pedestrian crossing intent prediction model, FHPE-Net.
- Finally, the proposed model was extensively compared and evaluated on the widely used Pedestrian Intention and Trajectory Estimation (PIE) [27] dataset, demonstrating its effectiveness. FHPE-Net outperformed the compared baseline methods.

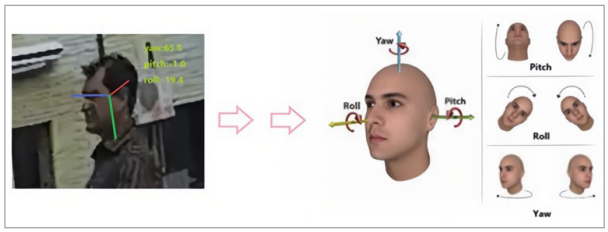
The rest of the paper is organized as follows. Section 2 presents some related work and Section 3 gives the research methodology. Then Section 4 describes the experimental procedure and details. Then Section 5 presents the experimental results and analysis. Finally, Section 6 summarizes the paper.

2. Related Work

2.1. Head Pose Estimation

Head pose estimation is a computer vision and pattern recognition technique that determines the orientation of a person's head in a digital image using yaw, pitch, and roll angles in a spatial coordinate system, as shown in Figure 2. In pedestrian crossing scenarios, people often look around to observe oncoming traffic and determine whether it is safe to cross the street. Similarly, a driver must move their head to examine the surrounding traffic when driving a car. People can express agreement or disagreement by nodding or shaking their heads, highlighting the importance of head posture in human behaviour analysis. As a result, researchers

Figure 2
Example of head posture



have developed several excellent algorithms for head pose estimation, as shown in previous studies [13], [44]. The increasing maturity of these algorithms has led to the widespread use of head pose in human behaviour analysis and understanding, such as attention detection and human-computer interaction.

2.2. Pedestrian Crossing Prediction

Behaviour detection and prediction is a widely studied topic in computer vision [41], [43], [9], [1]. Specifically, in order to predict whether or not a pedestrian will begin crossing the street shortly, the job is formulated as a binary classification issue, as shown in Figure 1. Previous work has transformed the pedestrian prediction task into a static image classification problem using convolutional neural networks. However, this method needed to incorporate the temporal and spatial information essential for accurate prediction. Recent methods have used recurrent neural networks (RNNs), particularly variants such as GRU and LSTM, to explore the temporal consistency of RGB video frames [43]. In addition, methods that combine multiple sources of information have been proposed, using different strategies to integrate different sources of information [40], [42]. For example, one study proposed SF-GRU [28], which uses a hierarchical GRU architecture to merge five sources of features: the appearance of pedestrians, the environment, the pose of the pedestrian's skeleton, the bounding box, and vehicle speed. Finally, a dense layer predicts the pedestrian's crossing intention. However, these approaches use additional networks to extract feature sources, which significantly increases the latency of model recognition and requires significant additional computational resources.

A recent study [15] proposed a novel method for predicting behaviour using only an add-on network. This approach is based on human kinematics and predicts pedestrian intentions by detecting changes in 2D skeletal joints. However, this method relies too heavily on human pose features and neglects the influence of other features. Recently, many datasets, such as the PIE dataset, have provided more annotated information that can be used for feature fusion. A recent study has proposed a common evaluation protocol and pattern input to advance the research on pedestrian behaviour prediction and enable fair comparisons between proposed methods [20]. These efforts have helped to advance the field and provide more comprehensive support for future research.

3. Method

3.1. Problem Formulation

The following describes the tasks in this paper aimed at predicting pedestrian crossing intentions modelling and analysis of vehicle pedestrian interactions in a continuous time series. A model is constructed to derive the target pedestrian behaviour probability $A_i^{n+t} \in \{0,1\}$ using a series of video frames acquired from within the field of view in front of the car and information related to the car's movement, where t represents the specific moment when the last observed frame occurs. n are the frames observed before a crossing or non-crossing (C/NC) event occurs. First, the model extracts salient features such as pedestrian bounding boxes, head pose orientation and global context (semantic segmentation). These channels and vehicle speeds are independently used as inputs to the prediction model. Thus, the model constructed in this paper contains the following input sources:

2D positioning trajectory of the pedestrian in the coordinates of the enclosing box (top left and bottom right points):

$$B_i = \left\{ b_i^{t-m}, b_i^{t-m+1}, ..., b_i^t \right\}. \quad (1)$$

Head pose sequence for pedestrian i:

$$P_i = \left\{ \theta_i^{t-m}, \theta_i^{t-m+1}, ..., \theta_i^t \right\}. \quad (2)$$

Global semantic segmentation sequences:

$$K_g = \left\{ k_g^{t-m}, k_g^{t-m+1}, ..., k_g^t \right\}. \quad (3)$$

Vehicle ego speed sequence:

$$S = \left\{ v^{t-m}, v^{t-m+1}, ..., v^t \right\}. \quad (4)$$

For each source, there exists a sequence of length m+1. Figure 3 shows the input sources.

3.2. Input Acquisition

3.2.1. Bounding Box Coordinates

The 2D localisation trajectory, denoted B_i, represents the positional changes of the pedestrian target in the image plane. It is possible to extract these trajectories using object detection systems such as YOLO [32] or object tracking systems such as SORT [39]. In order to maintain the focus of this paper, the pedestrian detection and tracking task is not explored, and therefore the B_i trajectories in the dataset are used directly. Specifically, the 2D localisation trajectory $B_i=\left\{b_i^{t-m},b_i^{t-m+1},...,b_i^t\right\}$ consists of the bounding box coordinates of the target pedestrian. i.e.

$$B_i^{t-m}=\left\{x_{i\alpha}^{t-m},y_{i\alpha}^{t-m},x_{i\beta}^{t-m},y_{i\beta}^{t-m}\right\}, \quad (5)$$

where $x_{i\alpha}^{t-m},y_{i\alpha}^{t-m}$ indicates the top left corner dot and $x_{i\beta}^{t-m},y_{i\beta}^{t-m}$ indicates the bottom right corner dot.

3.2.2. Head Pose Estimation

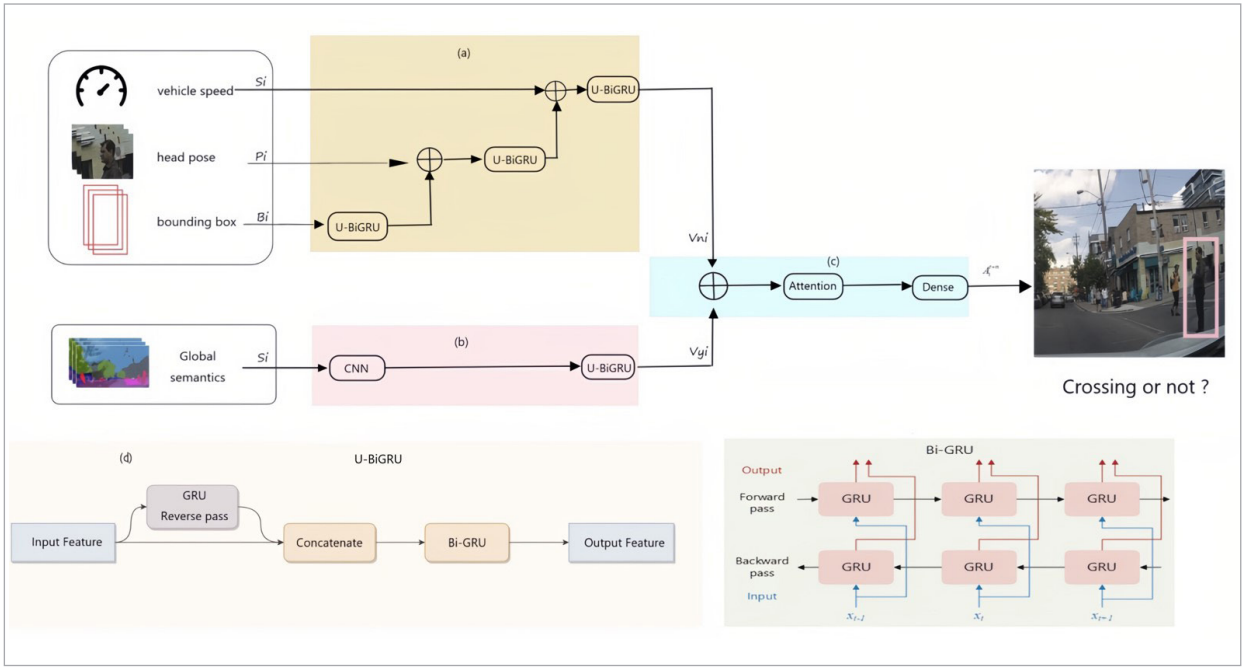
Pedestrian head poses orientation features reflect the variations in head pose as the target pedestrian crosses the road, including changes in yaw, pitch and side tilt angles per frame. These features can be obtained using lightweight head pose estimation algorithms such as WHENet[44]. This work estimates the head pose using the available head pose data within the PIE dataset at $P_i=\left\{\theta_i^{t-m},\theta_i^{t-m+1},...,\theta_i^t\right\}$. Specifically, The vector P is a 3D vector of three yaw angles, including yaw, pitch, and roll. i.e.

$$P_i^{t-m}=\left\{\left|\sigma_{ir}^{t-m}\right|,\left|\sigma_{ip}^{t-m}\right|,\left|\sigma_{iy}^{t-m}\right|\right\}, \quad (6)$$

where σ_{ir}^{t-m}, σ_{ip}^{t-m}, σ_{iy}^{t-m} are the angles of rotation, pitch, and yaw. The range of pitch angle and angle of rotation covers $(-99^{\circ},99)$ and the range of yaw cover $(-180^{\circ},180)$.

Figure 3

The network architecture of FHPE-Net: The model's inputs include global contextual information, bounding boxes, pedestrian head pose, and real-time vehicle speed. The red part (b) shows the pixel-level information. The global contextual information is feature extracted using CNN and then fed into the U-BiGRU module for coding. The yellow part shows (a) the fusion of non-pixel level information, and these features are encoded using U-BiGRU and fused layer by layer. The blue part (c) shows the final fusion, where pixel-level and non-pixel-level features are concatenated and fed to the Attention module and finally to the Dense layer for final prediction. U-BiGRU block: the input features are passed backwards through the first GRU layer, then concatenated with the original input and encoded together in a Bi-GRU layer for the final output



To prevent negative angles from having an effect on the training of the model, we take the absolute values of rotation, pitch and yaw angles as initial inputs to the model. Therefore, the changed rotation and pitch angle range cover $(0^{\circ},99)$ and the yaw range covers $(0^{\circ},180)$.

3.2.3. Global Semantic Information

Global Semantic Segmentation Information, $K_g=\left\{k_g^{t-m},k_g^{t-m+1},...,k_g^{t}\right\}$, provides a visual representation of the complex interactions between road users and the surrounding environment. This study leverages semantic mask information at the pixel level to capture global context. Semantic masks allow different objects in an image to be classified and localized by assigning pixel values to all the pixels associated with a particular object. Since the PIE dataset is devoid of semantic mask annotations, this paper introduces the lightweight DeepLabV3 model [25], pretrained on the Cityscapes dataset [29], in order to extract semantic masking information. The paper retains only critical information, including roads, buildings, pedestrians, and vehicles, to represent the global semantic segmentation information. To facilitate model learning, target pedestrians are masked using a unique label (with the mask region bounding box obtained from B_i). The semantic segmentation information is scaled as [224,224] for all input frames.

3.2.4. Ego-vehicle Speed

The speed S of the vehicle in real-time is an important factor in pedestrian crossing decisions. It can be accessed directly from the existing vehicle network. Since the dataset includes annotations of the ego's velocity of the vehicle, this study directly uses the labels of the true values to represent the velocity $S=\left\{v^{t-m},v^{t-m+1},...,v^{t}\right\}$ of the vehicle.

The following describes the tasks in this paper aimed at predicting pedestrian crossing intentions—modelling and analysis of vehicle-pedestrian interactions in a continuous time series. A model is constructed to derive the target pedestrian behaviour probability $A_i^{n+t}\in\left\{0,1\right\}$ using a series of video frames acquired from within the field of view in front of the car and information.

3.3. Model Architecture

As shown in Figure 3, a comprehensive prediction model framework covering the entire forecasting process is presented. The hybrid fusion method integrates CNN, RNN, and feature branch fusion modules to train a holistic model.

3.3.1. CNN Encodes Global Semantic Information

In this work, an efficient neural network model is employed as a tool for extracting global environmental visual features, aiming to achieve high efficiency. MobileNet [33] employs a distinct architectural design, decomposing standard convolutions into depthwise convolutions and 1x1 pointwise convolutions. Standard convolutions use kernel filters across all input channels and merge them in one step, while depthwise convolutions separate kernel filters for each input channel and combine them using pointwise convolutions. This separation and feature fusion method reduces computational complexity and model size. Additionally, since the global semantic segmentation map already serves as intermediate features for image recognition, a shallower (fewer layers) and narrower (fewer channels) network is sufficient for modeling the spatiotemporal dynamics of global visual features. Based on these principles, we adopt the lightweight convolutional neural network, MobileNet, as the backbone network for visual feature extraction. This model is pre-trained on the widely used ImageNet dataset [9], as using pre-trained models speeds up the training process, enabling accurate predictions with a relatively small dataset. In our experiments, we train this module using a batch of consecutive video clips $X\in R^{l\times c\times h\times w}$ as input. Here, h and w represent the height and width of the images; l stands for the number of frames observed; c represents the number of image channels. Once the video clips are loaded, they are fed into the MobileNet network to generate a final feature vector of dimension $\left[l,1280\right]$ as a visual feature sequence. Using the MobileNet model, our CNN module extracts crucial visual features from the input data, enabling us to effectively predict pedestrian crossing intentions in complex urban environments

3.3.2. RNN Module: U-BiGRUs

In this study, to construct the RNN module, we employed a combination of Gated Recurrent Units (GRU) [12] and Bidirectional Gated Recurrent Units (Bi-GRU) [4], as depicted in Figure 3. We chose GRU over Long Short-Term Memory (LSTM) [34] due to

its superior computational efficiency and architectural simplicity. Recurrent Neural Networks (RNNs) are an extension of feedforward networks. RNNs possess recurrent hidden states that allow them to learn temporal dependencies in sequential data. This inherent temporal depth has proven highly advantageous for tasks such as pedestrian trajectory prediction, where single-layer RNNs are applied to point coordinates in space. Besides temporal depth, spatial depth of RNNs can be increased by stacking multiple layers of RNN units on top of each other. This approach has been shown as an effective method to enhance sequential data modeling for complex tasks [28], particularly in video sequence analysis [17], where networks model dependencies between visual features of consecutive video frames. Given the multimodal nature of pedestrian motion prediction that depends on dynamic and visual scene information, we employed a cascade fusion approach, gradually integrating features at each level based on their complexity. In other words, we input complex visual features of the scene, which can benefit more from the spatial depth and dynamic features of lower-level networks, such as head pose and velocity, at higher levels of the network. Meanwhile, compared to using GRU or Bidirectional GRU modules separately, the output layer of the RNN module constructed with U-BiGRU can better access past and future information. This allows the model to better discern which data will contribute to future predictions, thereby enhancing performance. Similarly, context features are processed in parallel through the same architecture. The GRU employed in this study consists of 256 hidden units, resulting in a feature vector of dimension [1, 256]. The combination of GRU and BiGRU in the RNN module efficiently captures temporal correlations in the data and accurately predicts pedestrian crossing intentions. The formula calculation is as follows: Calculate the update gate vector:

$$r_t=\sigma_g\left(W_r\overleftarrow{x_t}+b_{Wr}+N_r h_{t-1}+b_{Nr}\right), \tag{7}$$

$$z_t=\sigma_g\left(W_z\overleftarrow{x_t}+b_{Wz}+N_z h_{t-1}+b_{Nz}\right), \tag{8}$$

$$\tilde{h}_t=\tanh(W_h\overleftarrow{x_t}+b_{Wh}+N_h(r_t e h_{t-1})+b_{Nh}), \tag{9}$$

$$h_t=(1-z_t)e\tilde{h}_t+z_t e h_{t-1}, \tag{10}$$

where $\overleftarrow{x_t}$ is the reverse vector of the input vector x_t, h_{t-1} indicates the hidden state of the previous moment. W, N and b are trainable weights and bias terms, respectively. σ_g is the sigmoid function. $\odot$ denotes the element-by-element multiplication operation of a vector.

$$\overleftarrow{h_t}=\mathrm{GRU}\left(\overleftarrow{x_t},\overleftarrow{h_{t-1}}\right), \tag{11}$$

$$\tilde{x}_t=\left[\overrightarrow{x_t};\overleftarrow{h_t}\right], \tag{12}$$

$$H_t=\left[\mathrm{GRU}\left(\vec{\tilde{x}}_t,\overrightarrow{H_{t-1}}\right);\mathrm{GRU}\left(\overleftarrow{\tilde{x}}_t,\overleftarrow{H_{t-1}}\right)\right], \tag{13}$$

where $\overleftarrow{x_t}$ is the inverse vector of the input vector x_t, and $\tilde{x}_t$ is the vector in which $\overrightarrow{x_t}$ and $\overleftarrow{h_t}$ have been combined. $\mathrm{GRU}(\ )$ denotes the non-linear transformation of the input vector that encodes $\overrightarrow{x_t}$, $\widetilde{x_t}$ into the corresponding state of the hidden layer of the GRU. $\mathrm{GRU}\left(\vec{\tilde{x}}_t,\overrightarrow{H_{t-1}}\right)$ is the state of the hidden layer of the forward GRU at time t, and $\vec{\tilde{x}}$ is the input data; $\mathrm{GRU}\left(\overleftarrow{\tilde{x}}_t,\overleftarrow{H_{t-1}}\right)$ is the hidden-layer state of the backward GRU at time t, and $\overleftarrow{\tilde{x}}$ is the inverse input data. The final hidden state can be obtained by sewing together the forms of the forward and backward hidden layers, where $[;]$ denotes the operation of splicing.

3.3.3. Hybrid Fusion of All the Features Branches

In real-world scenarios, it is often necessary to partition and combine different knowledge modules at various abstraction levels based on the nature and complexity of the problem to achieve partial or holistic coordination for effective problem-solving. In fact, most problems are composed of subproblems. Thus, the ability to consider problems at different abstraction levels simultaneously is crucial for efficient and robust learning.

Based on the aforementioned principles, this paper proposes a hybrid fusion strategy that effectively integrates feature information from different modes, structured at both pixel-level and non-pixel-level abstractions, as illustrated in Figure 3. This architecture consists of two branches, one for pixel-level feature processing and the other for non-pixel-level feature fusion. Pixel-level features, as shown in Figure 3(b), involve the extraction of spatial feature information for global semantic segmentation, as discussed

in the previous section, using a CNN module. Subsequently, the U-BiGRU module encodes temporal features to obtain pixel-level feature vectors V_{yi}. In contrast, the non-pixel feature branch encompasses three types of feature information: the position of the target pedestrian, head pose orientation, and vehicle velocity. Based on their complexity and abstraction levels, basic features are extracted using the U-BiGRU module and a cascade structure, as depicted in Figure 3(d). Next, a non-pixel-level feature vector V_{ni} is obtained through feature fusion, as shown in Figure 3(a). The non-pixel-level feature V_{ni} and the pixel-level feature V_{yi} are concatenated and fed into a modality attention block [20], as illustrated in Figure 3(c). This process involves merging multiple modality inputs into a representation, maximizing feature informativeness by weighting the information from each modality. We represent the sequence features (e.g., encoder-based outputs from an RNN) as hidden states $h=\{h_1,h_2,...,h_f\}$. The calculation of attention weights is as follows:

$$\alpha_{fs}=\frac{\exp(\text{score}(h_f,\overline{h_s}))}{\sum_{s'}^{S}\exp(\text{score}(h_f,\overline{h_{s'}}))}. \quad (14)$$

$\text{score}\left(h_f,\overline{h_s}\right)=h_f^T W\overline{h_s}$ and W are weight matrix. The attention weights swap the last hidden state, h_f, along with each of the previous hidden states, $\overline{h_s}$, of the source state. The output vector of the attention module is

$$a_f=f(c_f,h_f)=\tanh(W_c\left[c_f;h_f\right]), \quad (15)$$

where W_c is a weight matrix, c_f is the weighted hidden state sum of all attentions and $c_f=\sum_s \alpha_{fs}\overline{h_s}$. In this work, the attention module output vector is of dimension [1,256].

Finally, the output of the modality attention block is passed to a dense layer to predict the ultimate action. Overall, the proposed hybrid fusion strategy effectively integrates pixel-level and non-pixel-level features, utilizing the U-BiGRU block and modality attention to obtain a comprehensive and information-rich feature vector for predicting the final action. The final fusion formula is expressed as follows:

$$A_i^{t+n}=f_{dense}(f_{attention}(V_{ni};V_{yi})). \quad (16)$$

4. Experiment

4.1. Dataset and Benchmark

The effectiveness of our suggested multi-branching strategy and its architecture for analysing traffic pedestrian behaviour was validated by experiments using the Pedestrian Intention Estimation (PIE) [27] dataset in this study. The dataset includes more than 6 hours of continuous video footage recorded by in-vehicle cameras, providing a rich source of typical traffic scenarios covering complex road structures and crowded urban environments. The dataset includes annotations for 1842 pedestrian trajectories, providing comprehensive coverage of pedestrians near curbs and intersections that may have crosswalk intentions. The overall ratio of non-cross-over events to cross-over events in the dataset is 2.5:1. The dataset also contains a large amount of annotated information, including bounding boxes for pedestrians, head posture and vehicle sensor information such as vehicle velocity and yaw angle. Thus, for each pedestrian sample, we identified an event point for those who crossed in front of a vehicle the instant they began to cross the street. The data were randomly divided into training and test sets in a 6:4 ratio, and performance was assessed using established metrics such as ACC, AUC, F1, precision and recall. These metrics are widely used in binary event prediction and reflect the balanced accuracy of the algorithm.

4.2. Implementation Details

This paper adopts a benchmark implementation based on the PCPA model [20], which encompasses most pedestrian intent prediction methods. Specifically, we use a U-BiGRU model with 256 hidden units and a cascade structure to encode all features except global semantic information. To reduce overfitting, we set the dropout rate of the RNN module to 0.2 and added an L2 regularization of 0.001 to the final dense layer. The number of observation frames is set to 16. We use a binary cross-loss function and the Adam optimization algorithm [17], with a learning rate of 5×10^{-5} and a batch size of 32. The training process is performed over 60 epochs.

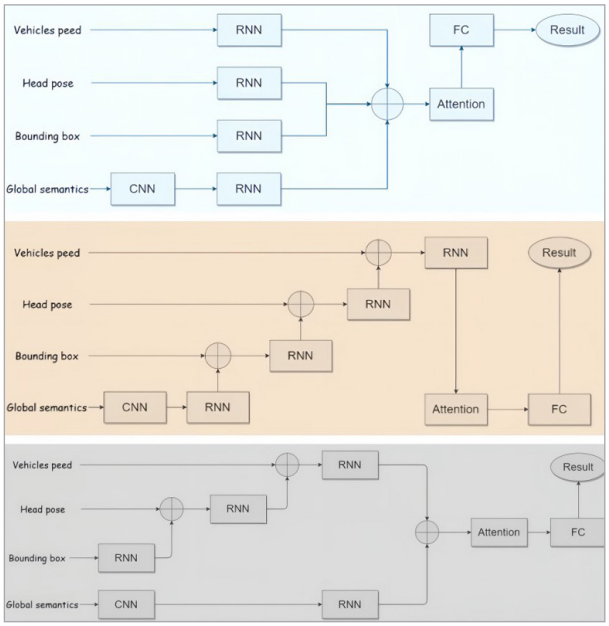
4.3. Baseline and Ablation Experiments

This study conducted extensive experiments to assess the effectiveness of FHPE-Net under different

feature fusion strategies. In the baseline model comparison experiments, we initially compared FHPE-Net to the standard baseline criterion for pedestrian crossing prediction, as detailed in Table 1. Next, to validate the U-BiGRU approach proposed in this paper, we compared it to models in the baseline that utilized recurrent neural networks for encoding (including Multi-RNN, Stacked-RNN, Hierarchical-RNN, and PCPA). The corresponding results are presented in Figure 4. Finally, to further verify the effectiveness of FHPE-Net in detecting occluded pedestrians, we conducted comparison experiments with partially occluded pedestrians in the PIE dataset against the baseline models, as shown in Figure 6.

Figure 4

Schematic diagram of the fusion method. The pictures from top to bottom show the direct connection structure, the Grade connection structure and the hybrid connection structure



In addition to the baseline model comparisons, this paper also conducted ablation studies. Firstly, to validate the effectiveness of the multi-feature fusion strategy proposed in this paper, we performed ablation experiments on the network fusion architecture and RNN encoders, with relevant results provided in Table 2. In these experiments, we explored different types of RNN encoders, including GRU, BiGRU, and U-BiGRU, as well as different fusion architectures as illustrated in Figure 4, including pixel-level and non-pixel-level fusion, cascaded and direct fusion structures, to investigate the optimal model configurations. Secondly, we examined the influence of different types of data sources on FHPE-Net's performance. Finally, through visualizing the results, we demonstrated the actual performance of the model, further illustrating its strengths and weaknesses.

5. Results

5.1. Baseline Comparison of Intent Prediction

Table 1 presents the quantitative results on the PIE dataset, offering a comprehensive comparison between our proposed model and baseline models. We observe four distinct types of models, including 2D convolution models, recurrent models, 3D convolution models, and multi-modal fusion models. Firstly, the initial set of models, such as ATGC, exclusively employ static images as input for pedestrian crossing prediction. Other models utilize recurrent neural networks (RNNs), stacking multiple image frames as historical data to enhance accuracy. In contrast, models like ID3, which are 3D convolution models, exhibit remarkable performance, but they come with relatively higher computational costs. Additionally, the PCPA model demonstrates exceptional multi-modal capabilities, amalgamating the advantages of 3D convolution models. FHPE-Net diverges from the baseline models by not only incorporating pixel-level and non-pixel-level fusion methods in feature fusion strategies but also introducing head pose information in the input sources.

Experimental results indicate that, in comparison to the PCPA model, FHPE-Net achieves a 3% improvement in accuracy, a 1% increase in AUC, and a 5% boost in F1 score. It is worth emphasizing that the F1 score, as a comprehensive metric that considers both recall and precision, stands as a pivotal indicator for evaluating a model's performance in binary classification tasks. In this respect, our proposed model attains the highest score among all the compared models.

Table 1
Quantitative results of the baseline model and the latest model and its variants based on the PIE dataset Solid lines separate the different types of architecture. Models correspond to the type of network used in the given method

Model Name	Model Variants	Head	ACC	AUC	F1	P	R
Static	VGG16 [35]	No	0.71	0.60	0.41	0.49	0.36
	Resnet50 [15]	No	0.70	0.59	0.38	0.47	0.32
ATGC [38]	AlexNet	No	0.59	0.55	0.39	0.33	0.47
Con-LSTM [34]	VGG16	No	0.58	0.55	0.39	0.32	0.49
	Resnet50	No	0.54	0.46	0.26	0.23	0.29
Single-RNN [18]	GRU	No	0.81	0.75	0.64	0.67	0.61
	LSTM	No	0.83	0.77	0.67	0.70	0.64
Multi-RNN [28]	GRU	No	0.83	0.80	0.71	0.69	0.73
Stacked-RNN [27]	GRU	No	0.82	0.78	0.67	0.67	0.68
Hierarchical-RNN [41]	GRU	No	0.82	0.77	0.67	0.68	0.66
C3D[37]	RGB	No	0.77	0.67	0.52	0.63	0.44
I3D[5]	RGB	No	0.80	0.73	0.62	0.67	0.58
	Optical flow	No	0.81	0.83	0.72	0.60	**0.90**
Two-Stream [36]	VGG16	No	0.64	0.54	0.32	0.33	0.31
PCPA [19]	Temp.+ mod. Attention	No	0.87	0.86	0.77		
Ours (FHPE-Net)	MobileNetV2 + U-BiGRU	Yes	**0.90**	**0.87**	**0.82**	**0.79**	0.86

Table 2
Quantitative results of ablation experiments based on the PIE dataset

Model Name	Model Variants			PIE				
	Encoder	Head	Fusion Approach	ACC	AUC	F1	P	R
Ours1	CNN + GRU	Yes	Grade connection	0.86	0.85	0.76	0.72	0.81
Ours2	CNN + GRU	Yes	Direct connection	0.85	0.83	0.75	0.70	0.80
Ours3	CNN + GRU	Yes	Hybrid connection	0.86	0.83	0.76	0.74	0.77
Ours4	CNN + BiGRU	Yes	Grade connection	0.86	0.85	0.77	0.73	0.83
Ours5	CNN + BiGRU	Yes	Direct connection	0.87	0.83	0.75	0.76	0.75
Ours6	CNN + BiGRU	Yes	Hybrid connection	0.85	0.86	0.77	0.68	**0.90**
Ours7	CNN + U-BiGRU	Yes	Grade connection	0.87	0.84	0.76	0.77	0.76
Ours8	CNN + U-BiGRU	Yes	Direct connection	0.88	0.83	0.77	**0.80**	0.74
Ours9	CNN + U-BiGRU	Yes	hybrid connection	**0.90**	**0.87**	**0.82**	0.79	0.86

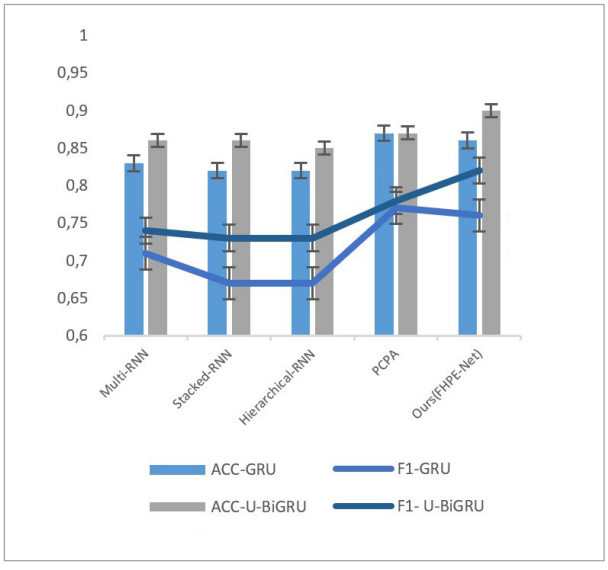
5.2. Comparison of Results of U-BiGRU in Baseline Models

To validate the effectiveness of the U-BiGRU method proposed in this study for pedestrian crossing intention prediction, a series of experiments were conducted with a specific focus on the importance of bidirectional temporal modeling and long-term contextual information. Several commonly used baseline models were selected for comparison, all of which employed recurrent neural networks (RNNs) in the encoding process. These baseline models included Multi-RNN, Stacked-RNN, Hierarchical-RNN, and PCPA. The performance of these models was evaluated by replacing the GRU method in their architectures with

the U-BiGRU method proposed in this paper while keeping all other experimental settings constant. As depicted in Figure 5, the experimental results reveal notable improvements in the performance metrics (ACC and F1) of the Multi-RNN model, with increases of 3%. Similarly, the Stacked-RNN model and the Hierarchical-RNN model exhibited performance improvements of 4% and 6%, and 3% and 5%, respectively. However, among the four baseline methods mentioned, the PCPA model's performance improvement was relatively less pronounced, with no improvement in ACC and a 1% increase in F1. This is attributed to the fact that the PCPA model's design did not adequately consider the global contextual background, including information regarding roads and other road users, which is an essential factor in the task of predicting pedestrian crossing intentions. Nevertheless, these experimental results unmistakably underscore the vital value of the U-BiGRU method in pedestrian intention prediction. The approach presented in this paper enhances the performance of the baseline models, especially when considering bidirectional temporal modeling and long-term contextual information. These results strengthen our confidence in the U-BiGRU method and emphasize its effectiveness in predicting pedestrian crossing intentions.

Figure 5

Comparative experimental results of GRU and U-BiGRU in the baseline model including Multi-RNN, Stacked-RNN, Hierarchical-RNN and PCPA

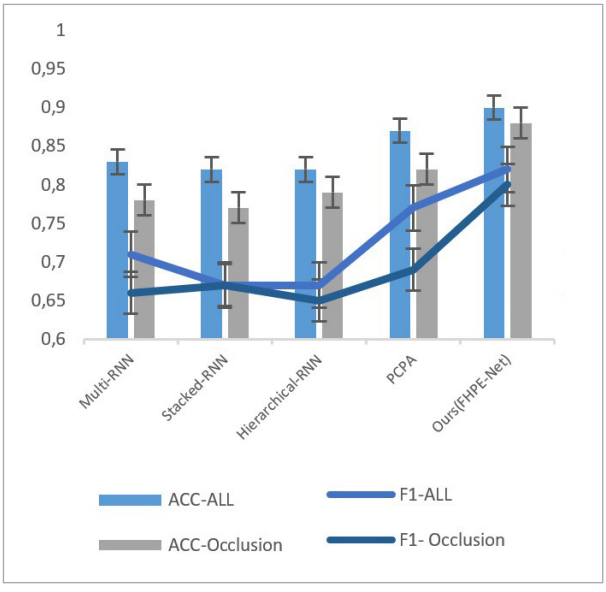


5.3. Comparison of Results of Baseline Models Under Pedestrian Occlusion

The FHPE-Net proposed in this study is designed to efficiently predict pedestrian crossing intentions in the presence of occlusions. To assess the effectiveness of FHPE-Net in detecting pedestrian crossing intentions when occlusions are present, we conducted a series of systematic experiments and compared its performance against several high-performing baseline models. These baseline models include Multi-RNN, Stacked-RNN, Hierarchical-RNN, and the PCPA model, with the first three being commonly used sequential modeling approaches, and the PCPA being an advanced model for pedestrian crossing intention prediction. In this experiment, we selected 427 pedestrian crossing segments with partial occlusions (occlusion rates ranging from 0.25 to 0.75) from the PIE dataset. As shown in Figure 6, the experimental results indicate that FHPE-Net exhibits significant performance advantages, achieving higher values in terms of accuracy and F1 score compared to the baseline models. Of particular note, FHPE-Net demonstrates exceptional robustness when predicting pedestrian crossing intentions in the presence of

Figure 6

Comparative experimental results of occluded pedestrian data set in baseline model including Multi-RNN, Stacked-RNN, Hierarchical-RNN and PCPA. **ALL:** All pedestrians, **Occlusion:** Occlusion of pedestrians



occlusions, experiencing less interference from occlusion factors compared to other models.

5.4. Ablation Experiment Results

5.4.1. Network Fusion Architecture and Encoder Ablation Results

The task of predicting pedestrian street-crossing behavior has historically been dedicated to the integration of multiple information sources, employing various strategies to consolidate different data streams. Presently, the majority of highly accurate pedestrian street-crossing prediction methods have achieved preeminence within their respective datasets, primarily attributable to their ingenious fusion strategies. This unequivocally underscores the critical importance of fusion strategies in the context of multi-feature integration for pedestrian street-crossing prediction tasks.

Concurrently, in order to evaluate the U-BiGRU model proposed in this study for pedestrian street-crossing intention prediction, leveraging the significance of bidirectional temporal modeling and long-term contextual comprehension, we conducted network fusion architecture and encoder ablation experiments. As illustrated in Table 2, we initially investigated the effects of substituting U-BiGRU in the encoder component with GRU and bidirectional GRU. Experimental results indicate that both of these enhancement methods led to a performance decrease, reaffirming our belief that an effective model for pedestrian behavior prediction should concurrently address long-term dependencies and multi-scale temporal characteristics.

Furthermore, we explored various fusion strategies, including pixel-level and non-pixel-level fusion, cascading, and direct connection structures, among others. According to the experimental data presented in Table 2, the hybrid fusion strategy of CNN+U-BiGRU in 'Ours9' exhibited the most exceptional performance. This highlights the model's capacity to simultaneously consider the problem from different levels of abstraction as essential for achieving efficient and robust learning in the task of pedestrian intention prediction.

5.4.2. Input Source Ablation Results

Due to the complexity of real-world traffic scenarios, accurate identification for the task of predicting pedestrian crossings typically requires the integration of multiple features. However, the integration of different forms of perceptual modal information from additional networks may lead to substantial increases in model size and slow inference speeds. Therefore, achieving real-time effectiveness of the optimal decision model is crucial in the context of autonomous driving scenarios. In this section, we conducted a series of ablation studies to thoroughly analyze the impact of different data types on FHPE-Net.

Table 3
Quantitative results of ablation experiments based on the PIE dataset

Model Variants	ACC	AUC	F1
FHPE-Net without head pose	0.82	0.83	0.72
FHPE-Net without speed	0.86	0.85	0.76
FHPE-Net without bounding box	0.88	0.84	0.77
FHPE-Net without Global Semantic	0.77	0.80	0.68
FHPE-Net replace contexts semantics with contexts image	0.83	0.80	0.71
FHPE-Net with 2D skeleton pose	0.87	0.85	0.77
FHPE-Net with local context	0.87	0.83	0.76
FHPE-Net	**0.90**	**0.87**	**0.82**

As shown in Table 3, when the model excludes head pose information, the performance metrics of the model (including ACC, AUC, and F1) decrease by 8%, 4%, and 10%, respectively. This once again emphasizes that head pose information is a critical factor affecting the interaction features between target pedestrians and moving vehicles in the task of predicting pedestrian intentions. Similarly, removing vehicle speed, pedestrian position, and global semantic information results in a decrease in model performance. Particularly, when replacing global semantic information with contextual images, ACC, AUC, and F1 decrease by 13%, 7%, 14%, and 7%, 7%, 11%, respectively. Thus, we can emphasize that environmental semantic information is a key factor influencing the interaction features between target pedestrians and the traffic environment.

Secondly, we attempted to extend the feature extraction network by introducing human 2D skeletal pose as non-pixel-level input to the model. However, performance metrics decreased by 3%, 2%, and 5%,

confirming the theoretical standpoint of this study that, influenced by factors such as complex environmental conditions, occlusion, and variations in distance between pedestrians and vehicles, accurate detection of human skeletal points is challenging, thus reducing the overall accuracy of motion recognition.

Finally, we attempted to improve the proposed method by introducing local contextual features of the target pedestrian. Although the results did not meet expectations, with performance metrics decreasing by 3%, 4%, and 6%, this may be due to reasons similar to previous attempts. The experiments described in Table 3 represent some intriguing explorations, and despite not achieving the initial expectations, they once again underscore that the proposed model can achieve better performance in situations with fewer features.

5.5. Visualization Results

5.5.1. Challenging Visualization

Figure 7 presents qualitative results of the proposed pedestrian crossing intention prediction model in this paper, comparing it with the PCPA model. As illustrated in the examples, our method accurately predicts pedestrian crossing intentions compared to the predictions made by PCPA. After careful consideration of these cases, the following arguments emerge: pedestrian head pose and overall environmental context can effectively contribute to solving the prediction problem of pedestrian crossing intentions:

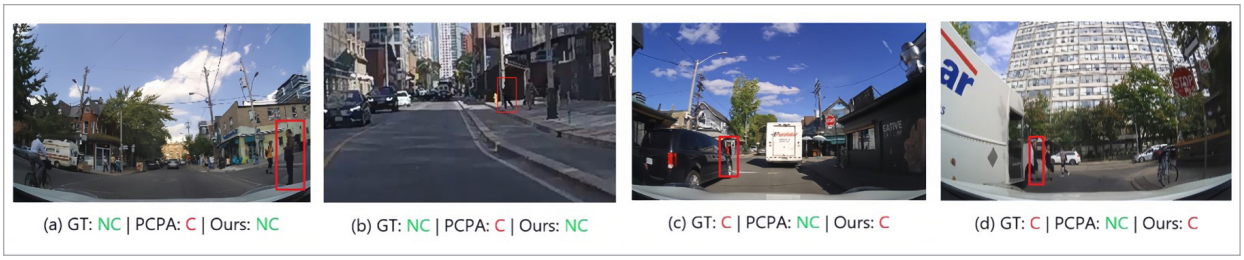
In Figure 7(a), pedestrians standing in an unknown direction are depicted. The pedestrians are positioned at the edge of the street, indicating an intention to cross. However, as their head poses do not exhibit an interactive posture with the moving vehicles, the PCPA model fails to consider the influence of pedestrian head poses, leading to an incorrect judgment. In contrast, our model fully utilizes the information provided by pedestrian head features regarding pedestrian-vehicle interactions, accurately determining that the pedestrian in question does not have an intention to cross.

In Figure 7(b), pedestrians with shadows forming visual disparities are shown. Due to the target pedestrian being in a shadowed position, the shadow may create contours similar to the pedestrian, making it challenging for the PCPA model to distinguish between the pedestrian and the background, resulting in misjudgment. Conversely, FHPE-Net, by leveraging semantic segmentation information from environmental context, is able to accurately differentiate these elements from the semantic region of the pedestrian, even when shadows or visual disparities create some resemblance. Additionally, the model integrates semantic segmentation to provide more contextual information, determining that the pedestrian does not have an intention to cross.

As shown in Figures 7(c)-(d), pedestrians with large body regions obscured below the head. Since the PCPA model relies on local body features and pose information to determine pedestrian intentions, occlusion of the pedestrian's body can lead to unreliable judgments. However, our approach utilizes both pedestrian head features and overall global context semantic information, enabling accurate prediction even in situations of occlusion as illustrated in Figure 7(c)-(d). This is because the head pose provides information about the interaction between the target pedestrian and moving vehicles, while global context semantic information provides information about the interaction between the target pedestrian and the traffic scene. Thus, combining head pose with global

Figure 7

Example diagram comparing pedestrian crossing intentions, with GT indicating true label values. NC indicates that the target pedestrian has not crossed the street. C indicates that the target pedestrian crossed the street



context, FHPE-Net significantly improves the accuracy of pedestrian crossing intention prediction, even in challenging scenarios such as body part occlusion, visual disparities, or ambiguous directions. This result highlights the robustness of our proposed method in various real-world scenarios.

5.5.2. Limitation Visualization

As shown in Figure 8, this study conducted a detailed analysis of failure cases of the proposed model. Figures 8(a)-(b) depict pedestrian crossing scenarios under overcast conditions. FHPE-Net inaccurately predicted pedestrian crossing intentions in these cases. Our analysis indicates that the primary reason lies in insufficient lighting conditions on overcast days, combined with the relatively long distance between the target pedestrians and vehicles. This resulted in inaccuracies in the head pose estimation algorithm when extracting pedestrian head pose information. Due to the lower quality of extracted head pose, FHPE-Net failed to capture crucial information affecting pedestrian intention decisions, namely, head pose. However, as these pedestrians were far from vehicles and did not interact with them, they did not pose any impact on driving vehicles.

Figure 8(c) illustrates a scenario, where pedestrians, standing at the edge of a crossroads, exhibit head poses suggesting interaction with moving vehicles. Nevertheless, FHPE-Net incorrectly predicted that these pedestrians had an intention to cross. We speculate that the pedestrian might have initially intended to cross but eventually abandoned the plan due to the high volume of oncoming traffic. Hence, the traffic situation should be a significant factor to be considered in future tasks involving predicting pedestrian intentions, especially at intersections without traffic signals.

Figure 8(d) presents a scenario of pedestrians not crossing in nighttime conditions, but FHPE-Net erroneously predicted that they had an intention to cross. This misjudgment is attributed to severe low light conditions under adverse weather (nighttime, rainy or snowy weather), affecting the feature extraction process of the model. Both PCPA and our model made incorrect judgments under these conditions. Based on these failure cases, this study attributes the main challenges to the limitations of vision-based depth estimation and pedestrian detection technologies in low light and adverse weather conditions. To enhance the applicability of these technologies, we can consider the integration of image restoration methods (such as de-raining, de-fogging, low-light enhancement, and image super-resolution), or in future research, contemplate the amalgamation of depth and radar technologies [5], [21] as alternatives to traditional image recognition methods. Radar technology employs radio waves to perceive the surrounding environment and offers stability under various weather conditions [3]. Radar sensors accurately detect the position, speed, and distance of objects, irrespective of visual constraints. This inherent robustness makes radar an intriguing alternative, particularly for nighttime driving, fog, heavy rain, or other low-visibility conditions where visual recognition may be severely hampered. The fusion of laser radar and visual perception systems combines the advantages of rapid dynamic object measurement with radar and the capability of visual recognition of obstacles. This collaborative approach enhances the system's comprehensive understanding of the road environment, thereby improving the accuracy and reliability of pedestrian intent prediction.

Figure 8

Example diagram comparing pedestrian crossing intentions, with GT indicating true label values. NC indicates that the target pedestrian has not crossed the street. C indicates that the target pedestrian crossed the street

5.6. Ethical Issues in Intention Prediction

When it comes to autonomous driving technology, careful consideration of the ethical issues involved is paramount. This research aims to delve deeply into the profound societal and ethical implications of autonomous driving technology, underscoring the importance of this examination. The ethical concerns arising from autonomous vehicles are wide-ranging, encompassing various facets including safety, privacy, and justice, among others, which demand meticulous attention and resolution.

Firstly, concerning safety issues, the pedestrian crossing intention prediction algorithm developed in this study is designed to provide effective alerts to guide drivers in taking necessary measures to avoid potential traffic accidents. This proactive approach seeks to reduce risks beforehand and enhance pedestrian safety.

Secondly, regarding privacy concerns, it is worth noting that the data used in this research originates from the publicly available Pedestrian Intention Estimation (PIE) dataset. During the data usage process, we strictly adhere to data privacy protection principles, utilizing only the necessary data that fulfils the requirements of this research. Furthermore, we explicitly confine the use of this data within the scope of academic research to ensure the privacy of the data is rigorously safeguarded.

Lastly, with regards to issues of social justice, the focal point of this research is the FHPE-Net algorithm, dedicated to improving overall urban traffic safety while optimizing urban traffic management. Moreover, this model is adaptable to different age groups and special pedestrian categories, such as the elderly, children, and disabled individuals, allowing for personalized fine-tuning to better serve these specific demographics.

The purpose of this discourse is to emphasize the significance of ethical issues within the field of autonomous driving technology and to showcase this research's rigorous methods and commitment to addressing these concerns. This, in turn, contributes to ensuring that our research is fully cognizant of ethical and social justice considerations and contributes to the overall well-being of society.

6. Conclusion

This paper presents FHPE-Net, designed to efficiently and accurately predict obscured pedestrian intent. The model employs pixel-level and non-pixel-level fusion strategies and integrates the U-BiGRU module to effectively fuse pedestrian head poses, bounding box positions, vehicle speed, and environmental information. Extensive comparative experiments validate the outstanding performance of FHPE-Net in pedestrian intent prediction and the effectiveness of U-BiGRU. Furthermore, ablation results suggest that combining head poses with global context and considering the problem from different abstract levels can enhance the accuracy of predicting pedestrians' intent to cross the street. Future research will focus on the integration of deep learning with radar technology to enhance the model's stability in adverse environments. While this paper's architecture is primarily applied to pedestrian intent prediction, similar methods may potentially bring benefits to other activities requiring head pose recognition.

Acknowledgment

This work was supported by the Fundamental Research Funds for the Program for Innovation Research Groups at Institutions of Higher Education in Chongqing (CXQT21032), the Fundamental Research Funds for the Natural Science Foundation of Chongqing, China (cstc2021ycjh-bgzxm0088) and the Fundamental Research Funds for the Science and Technology Research Project of Chongqing Municipal Education Commission KJZD-M202303401 .

Declaration of Interest Statement

The authors report there are no competing interests to declare.

References

1. Abughalieh, K. M., Alawneh, S. G. Predicting Pedestrian Intention to Cross the Road. IEEE Access, 2020, 8, 72558-72569. https://doi.org/10.1109/ACCESS.2020.2987777
2. Bhattacharyya, A., Fritz, M., Schiele, B. Long Term on Board Prediction of People in Traffic Scenes under Uncertainty. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City,

UT, USA, 2018, 4194-4202. https://doi.org/10.1109/CVPR.2018.00441

3. Bing, L., X. Zhang, J. P. Munoz, J. Xiao, X. Rong, Tian, Y. Assisting Blind People to Avoid Obstacles: An Wearable Obstacle Stereo Feedback System Based on 3D Detection. 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), Zhuhai, China, 2015, 2307-2311. https://doi.org/10.1109/ROBIO.2015.7419118
4. Brakel, P., Stroobandt, D., Schrauwen, B. (2013) Bidirectional Truncated Recurrent Neural Networks for Efficient Speech Denoising. Proceedings of Interspeech, 2013, 2973-2977. https://doi.org/10.21437/Interspeech.2013-272
5. Buchman, D., Drozdov, M., Krilavičius, T., Maskeliūnas, R., Damaševičius, R. Pedestrian and Animal Recognition Using Doppler Radar Signature and Deep Learning. Sensors, 2022, 22(9), 3456. https://doi.org/10.3390/s22093456
6. Carreira, J., Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. https://doi.org/10.1109/CVPR.2017.502
7. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1), 172-186. https://doi.org/10.1109/TPAMI.2019.2929257
8. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv.org, 2017, arXiv:1706.05587.
9. Chen, T., Tian, R., Ding, Z. Visual Reasoning using Graph Convolutional Networks for Predicting Pedestrian Crossing Intention. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 2021, 3096-3102. https://doi.org/10.1109/ICCVW54120.2021.00345
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, 3213-3223. https://doi.org/10.1109/CVPR.2016.350
11. Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, 248-255. https://doi.org/10.1109/CVPR.2009.5206848
12. Dey, R., Salem, F. M. Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks. 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, USA, 2017, 1597-1600. https://doi.org/10.1109/MWSCAS.2017.8053243
13. Dhingra, N. LwPosr: Lightweight Efficient Fine Grained Head Pose Estimation. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, 1204-1214. https://doi.org/10.1109/WACV51458.2022.00127
14. Fang, Z., Lopez, A. M. Is the Pedestrian going to Cross? Answering by 2D Pose Estimation. 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 2018, 1271-1276. https://doi.org/10.1109/IVS.2018.8500413
15. Gesnouin, J., Pechberti, S., Stanciulcscu, B., Moutarde, F. Trouspi-net: Spatio-Temporal Attention on Parallel Atrous Convolutions and U-GRUs for Skeletal Pedestrian Crossing Prediction. 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 2021, 01-07. https://doi.org/10.1109/FG52635.2021.9666989
16. He, K., Zhang, X., Ren, S., Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. https://doi.org/10.1109/CVPR.2016.90
17. Joe, Y.-H. N., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G. Beyond Short Snippets: Deep Networks for Video Classification. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, 4694-4702. https://doi.org/10.1109/CVPR.2015.7299101
18. Kingma, D. P., Ba, J. Adam: A Method for Stochastic Optimization. arXiv.org, 2014, arXiv:1412.6980.
19. Kotseruba, I., Rasouli, A., Tsotsos, J. K. Do They want to Cross? Understanding Pedestrian Intention for Behavior Prediction. 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 2020, 1688-1693 2020. https://doi.org/10.1109/IV47402.2020.9304591
20. Kotseruba, I., Rasouli, A., Tsotsos, J. K. Benchmark for Evaluating Pedestrian Action Prediction. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021, 1257-1267. https://doi.org/10.1109/WACV48630.2021.00130
21. Kulikajevas, A., Maskeliūnas, R., Damaševičius, R., Ho, E. S. L. 3D Object Reconstruction from Imperfect Depth Data Using Extended YOLOv3 Network. Sensors, 2020, 20(7), 2025. https://doi.org/10.3390/s20072025
22. Lorenzo, J., Parra, I., Wirth, F., Stiller, C., Llorca, D. F., Sotelo, M. A. RNN-Based Pedestrian Crossing Prediction Using Activity and Pose-Related Features. 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 2020, 1801-1806. https://doi.org/10.1109/IV47402.2020.9304652
23. Luong, T., Pham, H., Manning, C. D. Effective Approaches to Attention-Based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015, 1412-1421. https://doi.org/10.18653/v1/D15-1166

24. Perdana, M. I., Anggraeni, W., Sidharta, H. A., Yuniarno, E. M., Purnomo, M. H. Early Warning Pedestrian Crossing Intention from its Head Gesture Using Head Pose Estimation. 2021 International Seminar on Intelligent Technology and Its Applications (ISITIA), Surabaya, Indonesia, 2021, 402-407. https://doi.org/10.1109/ISITIA52817.2021.9502231

25. Piccoli, F., Balakrishnan, R., Perez, M. J., Sachdeo, M., Nunez, C., Tang, M., Andreasson, K., Bjurek, K., Dass Raj, R., Davidsson, E., Eriksson, C., Hagman, V., Sjoberg, J., Li, Y., Srikar Muppirisetty, L., Roychowdhury, S. Fussi-Net: Fusion of Spatio-Temporal Skeletons for Intention Prediction Network. 2020 54th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2020, 68-72. https://doi.org/10.1109/IEEECONF51394.2020.9443552

26. Rasouli, A., Kotseruba, I., Tsotsos, J. K. Are They going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 2017, 206-213. https://doi.org/10.1109/ICCVW.2017.33

27. Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J. Pie: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, 6261-6270. https://doi.org/10.1109/ICCV.2019.00636

28. Rasouli, A., Kotseruba, I., Tsotsos, J. K. Pedestrian Action Anticipation Using Contextual Feature Fusion in Stacked RNNS. arXiv.org, 2020. https://arxiv.org/abs/2005.06582

29. Rasouli, A., Yau, T., Rohani, M., Luo, J. Multi-modal Hybrid Architecture for Pedestrian Action Prediction. 2022 IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 2022, 91-97. https://doi.org/10.1109/IV51971.2022.9827055

30. Rasouli, A., Rohani, M., Luo, J. Pedestrian Behavior Prediction Via Multitask Learning and Categorical Interaction Modeling. arXiv.org, 2020, arXiv:2012.03298.

31. Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J. Pie: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) Seoul, Korea (South), 2019, 6261-6270. https://doi.org/10.1109/ICCV.2019.00636

32. Redmon, J., Farhadi, A. Yolov3: An Incremental Improvement. arXiv.org, 2018, arXiv:1804.02767.

33. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

34. Sak, H., Senior, A., Beaufays, F. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. Interspeech, 2014, 338-342. https://doi.org/10.21437/Interspeech.2014-80

35. Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., Woo, W. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. arXiv.org, 2015, arXiv: 1506.04214

36. Simonyan, K., Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv.org, 2014, arXiv: 1409.1556.

37. Simonyan, K., Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. arXiv.org, 2014, arXiv: 1406.2199.

38. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 IEEE International Conference on Computer Vision (ICCV), 2015. https://doi.org/10.1109/ICCV.2015.510

39. Wojke, N., Bewley, A., Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017, 3645-3649. https://doi.org/10.1109/ICIP.2017.8296962

40. Yang, D., Zhang, H., Yurtsever, E., Redmill, K. A., Ozguner, U. Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention. IEEE Transactions on Intelligent Vehicles, 2022, 7(2), 221-230. https://doi.org/10.1109/TIV.2022.3162719

41. Yao, Y., Atkins, E., Johnson-Roberson, M., Vasudevan, R., Du, X. Coupling Intent and Action for Pedestrian Crossing Behavior Prediction. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021, 1238-1244. https://doi.org/10.24963/ijcai.2021/171

42. Yong, D., Wang, W., Wang, L. Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, 1110-1118. https://doi.org/10.1109/CVPR.2015.7298714

43. Zhang, S., Abdel-Aty, M., Wu, Y., Zheng, O. Pedestrian Crossing Intention Prediction at Red Light Using Pose Estimation. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(3), 2331-2339. https://doi.org/10.1109/TITS.2021.3074829

44. Zhou, Y., Gregson, J. WHENet: Real-Time Fine-Grained Estimation for Wide Range Head Pose. arXiv.org, 2020, arXiv: 2005.10353.