IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 February 2024, accepted 23 March 2024, date of publication 27 March 2024, date of current version 3 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3381960

RESEARCH ARTICLE

Pedestrian Trajectory Prediction and Pose Estimation Considering Behavioral Characteristic Relationships and Noise Inhibition

JINCAO ZHOU[1], WEIPING FU[1,2], BENYU NING[1], AND SIYUAN HE[1]
[1]College of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, Xi'an 710048, China
[2]College of Engineering, Xi'an International University, Xi'an 710077, China

Corresponding author: Jincao Zhou (jzhou324@xaut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 52202501, and in part by the Natural Science Foundation of Shaanxi Province under Grant 2022JQ-546.

ABSTRACT In the domain of pedestrian trajectory prediction(PTP) from a roadbed perspective, the visibility of pedestrian feature points is inevitably compromised by external noise interference, impacting both pedestrian pose estimation(PPE) and PTP. This paper presents an innovative model for PTP. The model not only tackles noise interference but also takes into account the inherent correlation between pedestrian pose features and trajectory coordinates. To tackle the challenge of noise interference during pedestrian crossing, we reframe it as an anomalous feature detection problem using the Graph Deviation Network (GDN). Subsequently, we enhance the Long Short-Term Memory (LSTM) module by incorporating a time-domain anomaly suppression module, resulting in the development of an Anomaly Inhibition-LSTM (AI-LSTM) with robust noise suppression capabilities. Finally, by integrating the predicted values of behavioral pose and trajectory position, considering the behavioral characteristic relationship resolved by the GDN algorithm, we achieve accurate prediction and pose estimation of pedestrian crossing trajectories amidst noise interference. Experimental results demonstrate superior performance of our algorithm in the PPE task when compared to GDN and LSTM algorithms. In the PTP task, our algorithm exhibits performance comparable to the Transformer-based method, with the added advantage of improved interpretability.

INDEX TERMS Trajectory prediction, pose estimation, autonomous vehicles.

I. INTRODUCTION

Accurately predicting pedestrian trajectories is essential for the safe operation of autonomous driving systems, aiding in collision avoidance and playing a critical role in intelligent path planning, human-computer interaction, and urban planning [1]. In recent years, researchers have discovered the utility of PPE in understanding pedestrians' action intentions, identifying movement patterns, modeling spatial relationships, and enhancing environmental perception. Leveraging PPE significantly contributes to the accurate prediction of pedestrians' future trajectories, thereby enhancing the overall accuracy and reliability of PTP [2]. Among various methods for PPE, vision-based PPE has gained prominence in intelligent transportation due to its non-contact, real-time capabilities, multi-person pose estimation, complex pose modeling, and scalability. However, vision-based PPE encounters challenges such as occlusion, pose complexity, and data noise. Figure 1 illustrates how the loss or error of pedestrian joint points under external interference results in inaccuracies in subsequent PPE. The issue of addressing PPE features containing noise or obscuration in predicting pedestrian trajectories has not received sufficient attention.

The associate editor coordinating the review of this manuscript and approving it for publication was Shun-Feng Su.

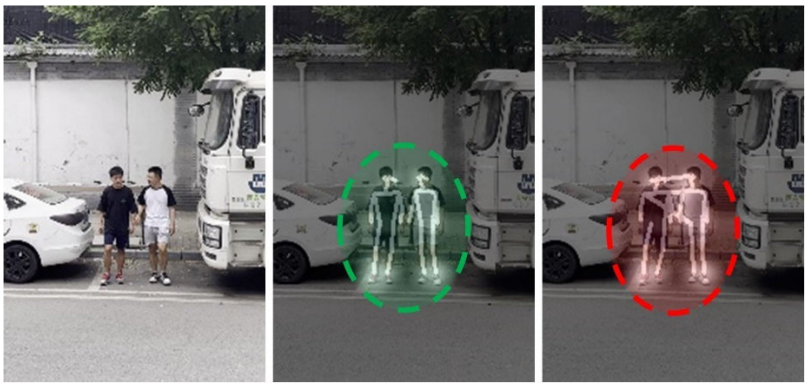


FIGURE 1. Pedestrian pose estimation results under external noise interference. Where (a) denotes the original figure, (b) denotes the pose estimation result without external interference, and (c) denotes the pose estimation result with external interference. These figures were modified from [14].

On one hand, theoretically, it appears viable to employ distinct anti-interference methods for various sources of interference. However, the sources of interference impacting PTP or PPE are diverse. If we establish individual methods for each source, the potential for these methods to transition from laboratory settings to large-scale engineering applications diminishes. On the other hand, we were surprised to discover that, in comparison to other engineering challenges, there are more evident internal connections between pedestrians' pose and positional characteristics during street crossing. For instance, pedestrians commonly rotate their heads to observe vehicles before crossing the street. As they traverse different locations during crossing, their heads turn within specific ranges. Noise interference can be mitigated by leveraging the relationship between pedestrian trajectory and pose. In this context, this method proves highly interpretable and aligns more closely with human cognitive styles.

As a deep learning algorithm for processing graph-structured data, the Graph Neural Network (GNN) can represent the features of nodes by learning their embedding vectors. This embedding vector can then be used to represent nodes' characteristics and their relationship with neighboring nodes, which is crucial to the learning task [3]. Building on the capability of GNN to unveil and characterize relationships between different features, Deng et al. [4] introduced a GNN-based anomaly data detection method. This method involves embedding various channel features and utilizing GNN to establish associations between each channel parameter. Subsequently, it predicts the future data of each channel based on existing data, identifying anomalies when the predicted value deviates significantly from the actual value. While this approach offers advantages in exploring intrinsic connections between nodes and detecting anomalous data, it falls short in reducing the proportion of anomalous data across the entire time series. As a result, direct application to PTP and PPE is limited. However, if the proportion of anomalous data in the entire time series could be reduced by fully leveraging the internal relationships of features, a potential solution to the challenge of PPE and PTP in the presence of noise interference may be achieved.

Additionally, we recognize that LSTM is not only the most widely utilized method for anomaly detection in multi-source time series [5] but also the predominant deep learning-based architecture for trajectory prediction in Intelligent Transport Systems (ITS) [6]. Nevertheless, traditional LSTM units lack the capability to actively suppress noise. If directly applied to trajectory prediction with noise interference, it implies that noise at a certain time node will interfere with subsequent data for an extended period. Hence, the LSTM network structure must be modified when serving as the foundation for our study.

To address the aforementioned challenges, we introduced an algorithm for PPE and PTP by integrating GDN and an enhanced LSTM. Initially, drawing inspiration from the GDN network proposed by Deng et al., we reframed the PPE problem, which includes noise and interference during pedestrian crossing, into an anomaly detection problem. To mitigate the impact of anomalies on the prediction of subsequent features in the time domain, we devised the AI-LSTM by incorporating a time-domain inhibition module into the traditional LSTM network. Ultimately, the prediction results from GDN and pedestrian crossing features were collaboratively employed as inputs for the AI-LSTM.

In this paper, we mainly make the following contributions:

1. In response to the insufficient attention given to noise interference in traditional PTP, we introduced a method that reframes the noise interference problem in the pedestrian crossing process as an anomalous detection task. Consequently, the algorithm demonstrated enhanced interpretability compared to the purely recurrent neural network-based PPE and PTP algorithms;

2. Our proposed method integrates GDN with an AI-LSTM algorithm, taking noise inhibition into account. Our algorithm not only maintains the internal relationship between pose

features and their crossing positions during the pedestrian crossing process but also effectively suppresses the influence of noise information on the features in the time domain.

The remainder of this paper is organized as follows. Section II reviews the related work. Section III presents our proposed method to predict pedestrians' pose and trajec-tory containing noise. Section IV details the experiments and results. Section V summarizes the current work and possible work.

II. RELATED WORK

A. POSE ESTIMATION

Human Pose Estimation (HPE) is a critical task in computer vision, focused on recognizing and estimating pedestrians' poses and movements from images or videos. With the advancement of deep learning techniques, this field has been extensively explored from various angles. For example, Li et al. [7] presented an online processing method Online Knowledge Distillation Human Pose (OKDHP) for pedestrian pose estimation, which obtains high-quality target heat maps by combining multiple receptive fields and performing connectivity changes. Wang et al. [8] constructed a self-supervised calibration mechanism by constructing a 2D to 3D bitmap transformation and a 3D to 2D projection transformation to form a dual learning task. This ensures that the algorithm can adaptively learn from 3D human pose data and external large-scale 2D human pose data. To mitigate constraints to estimating the human poses in general scenarios, Liu et al. [9] prosposed an efficient human pose estimation model (EHPE) with joint direction cues and Gaussian coordinate encoding. Particularly in recent years, the effective resolution of HPE tasks has been further advanced with the application of Transformer theory. For example, Li et al. [10] proposed a new architecture for human pose estimation called channel spatial integrated transformer (CSIT), which focused on spatial features in visual information and innovatively uses a parallel network to combine spatial features with texture features to fully extract information from images through transformer. Zhang et al. [11] leveraged the two intrinsic inductive bias and proposed the ViTAE transformer, which utilizes a reduction cell for multi-scale feature and a normal cell for locality.

In the computer vision field, researchers have successfully estimated human pose in complex environments using various constraints. However, in ITS, pedestrian pose involves not only anatomical constraints and other common constraints in computer vision but also internal correlations with the pedestrian's trajectory. Addressing the challenge of mining the relationship between a pedestrian's trajectory and pose to provide an ITS-oriented and interference-resistant pose estimation approach has not received sufficient attention.

B. TRAJECTORY PREDICTION

In terms of the field of intelligent transportation systems, pose estimation is also widely used in pedestrian trajectory prediction. In traditional research, scholars main-ly used historical trajectories of pedestrians as the main inputs for predicting pedestrian trajectories. For example, Wu et al. [12] proposed a Hierarchical Spatio-Temporal Attention architecture (HSTA) and studied pedestrian trajectory prediction using pedestrian historical trajectories as input. Although these methods using historical pedestrian trajectories as inputs have made some improvements, the accuracy of the algorithms still needs to be improved.

In recent years, more and more studies have incorporated the influence factors such as pedestrian pose estimation, pedestrian intention recognition, and pedestrian action recognition results into pedestrian crossing trajectory prediction, leading to a more significant improvement in trajectory prediction accuracy. For example, Zhang et al. [13] employed 2D pedestrian pose estimation combined with air temperature and traffic light duration as inputs. They realized pedestrian crossing trajectory prediction and intention recognition through SVM and other methods. Zhou et al. [14] used 2D pedestrian pose estimation combined with pedestrian history trajectory and pedestrian crossing intention as inputs to achieve pedestrian crossing trajectory prediction and intention recognition by the IPVO-LSTM algorithm. Kothari et al. [15] combined pedestrian history trajectory with body pose estimation and predicted pedestrian trajectory via the LSTM network. Ahmed et al. [16] present a intent prediction approach for multi-scale pedestrians using 2D pose estimation and a Long Short-term memory (LSTM) architecture. Zhang et al. [17] proposed a method for predicting pedestrian crossing intentions based on spatio-temporal graph convolutional networks using skeleton data (ST CrossingPose). This approach offers a more comprehensive characterization of the spatial and temporal aspects of pedestrian skeleton data compared to manually designed features. Czech et al. [18] introduced a novel method for predicting pedestrian trajectories using in-vehicle camera systems. This approach involves processing pedestrian trajectories derived from multiple input modalities, including pedestrian bounding boxes, body and head orientations, and poses, through independent coding streams.

Unfortunately, despite the widespread use of pedestrian pose parameters for trajectory prediction or intention recognition, the specific issue of interference caused by noise receives limited attention in current research on human trajectory prediction. Furthermore, there is a lack of noise suppression schemes leveraging the intrinsic relationship between the pedestrian's trajectory and pose.

C. ANOMALY DETECTION

Data anomaly detection entails identifying data that deviates significantly from the majority through data mining. This approach has found applications in various contexts and has become a significant area of study for analyzing time series data to extract valuable information.

For example, in the field of medicine. To present a detailed examination on EEG signals with improved

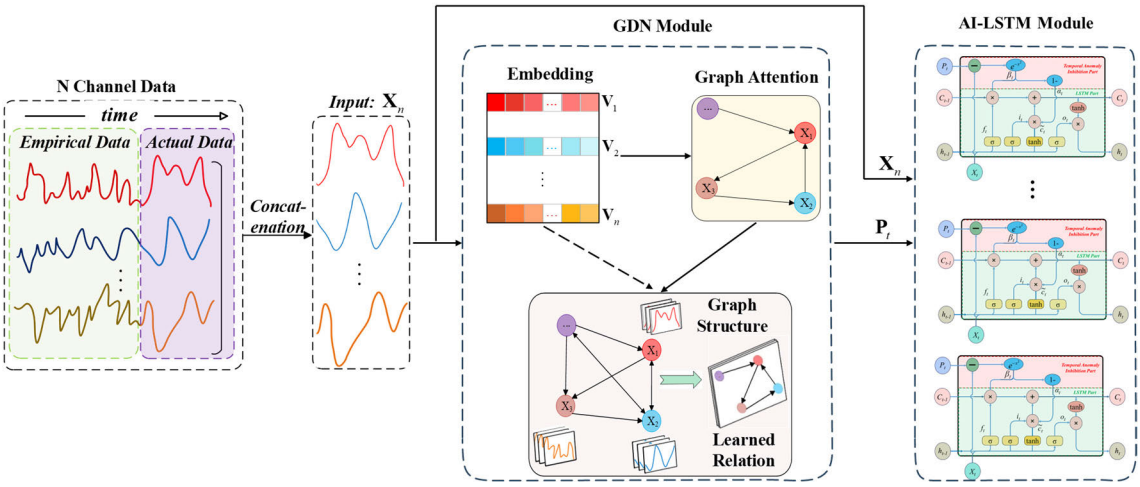


FIGURE 2. Pedestrian pose and trajectory prediction model considering behavioral feature relationships and noise inhibition. Where X_t denotes the actual input value obtained by concatenating the empirical data with the real-time data, and P_t represents the predicted value obtained by X_t after the GDN module.

detection accuracy, Rajinikanth et al. [19] proposed the synchro-extracting-transform (SET), which converts the complex 1D EEG into 2D images using time–frequency transformation. You et al. [20] aimed to present a personalized deep learning-based anomaly detection algorithm for seizure monitoring with behind-the-ear electroencephalogram (EEG) signals. In the financial field, Cheong et al. [21] proposed a spatiotemporal convolutional neural network-based relational network (STCNN-RN) model that can learn the complex correlations between multiple financial time-series data sets, and they used this model to identify abnormal situations. Madurawe et al. [22] proposed a model to detect collusion in stock markets through the application of graph mining and anomaly detection. In the area of network communications security, a DDoS attack may render the server useless for a long period of time causing the services to crash due to extensive load. Sharma et al. [23] proposed an anomaly detection architecture for IoT networks where the detection actually happens on the fog layer.

Despite the widespread use of anomaly detection in various fields, interpreting anomalies remains a significant challenge [5]. This technology is less applied in the field of autonomous driving, and there is a notable absence of research addressing the detection of temporal anomalies due to the loss of pedestrian feature points during pedestrian crossings.

To address these challenges, we propose introducing anomaly detection into the Intelligent Transportation Systems (ITS) domain to enhance pedestrian trajectory prediction and pose estimation under noise interference. Our approach involves constructing a model that considers noise interference, combining GDN and AI-LSTM networks. This model aims to improve trajectory prediction and pose estimation accuracy in noisy environments by identifying and mitigating the impact of noise on pose estimation and trajectory prediction in the time domain, particularly by addressing anomalous time nodes.

III. METHODOLOGY

A. MODEL OVERVIEW

As discussed in the introduction, neither the GDN nor the LSTM algorithms proved capable of addressing the issue presented in this manuscript. Therefore, in this section, we constructed a model that takes into account the relationship between pedestrians' behavioral features and noise inhibition, as shown in Fig. 2.

Initially, the pre-collected empirical data were aligned with actual pedestrian data, and the actual input values were obtained through a sliding window. Subsequently, leveraging the Graph Deviation Network (GDN), we performed correlation learning and parameter prediction for each channel parameter in the pedestrian crossing process. The purpose of this AI-LSTM design is to suppress noise in the temporal domain while predicting pedestrian trajectories and estimating head poses. Finally, the predicted parameters of pedestrian trajectories and head poses $\mathbf{P}_t$ are utilized as the input portion of the anomaly suppression LSTM module. Simultaneously, we enhanced the traditional LSTM by incorporating a time-domain anomaly inhibition module, resulting in the creation of an Anomaly Inhibition-LSTM (AI-LSTM). This AI-LSTM was designed to suppress noise when predicting the pedestrian trajectory in the time domain. Finally, the pedestrian crossing parameter $\mathbf{X}_t$, and the prediction results of the parameters of each channel $\mathbf{P}_t$ were used as the input part of the Anomaly Inhibition-LSTM module.

B. PEDESTRIAN STATE PREDICTION CONSIDERING THE RELATIONSHIP BETWEEN BEHAVIORAL FEATURES AND TRAJECTORIES

Enhancing the accuracy of PTP necessitates addressing target detection failures and image noise interference during image

acquisition. However, this challenge remains unsolved to date. Simultaneously, there is an increasing focus on issues related to time series anomaly detection. Notably, the GDN network proposed by Deng et al. [4] not only considered the correlation of channel parameters at different time nodes in anomaly detection but also provided an interpretable method for anomaly detection. This insight inspired us, considering the dynamic nature of pedestrians during reciprocal crossings. Although the subjects are constantly changing, treating them as the same subject allows us to frame subsequent noise-containing pedestrian poses and trajectories as a form of anomaly detection within a continuous time series. In other words, even though pedestrians may differ each time they cross the street, pedestrians crossing the road at different times can be considered the same pedestrians continuously traversing the road in the context of a complete time series. Specifically, in our model, firstly, the experience parameters $\mathbf{X}_e \in \mathbb{R}^{C*t_e}$ were used as an empirical database, as shown in Equation (1), and through the concatenation operation, the real-time parameters $\mathbf{X} \in \mathbb{R}^{C*t_e}$ were fused with $\mathbf{X}_e$ to obtain a time series variable $\mathbf{X}_t$. It was noted that the $\mathbf{X}_t$ told the head positions and trajectories(both empirical and real-time). Finally, $\mathbf{X}_t$ was used to get prediction values for every channel.

$$\mathbf{X}_t = \mathbf{X} \oplus \mathbf{X}_e, \mathbf{X}_t \in \mathbb{R}^{C*(t_e+t_a)} \tag{1}$$

where $\mathbf{X}_e$ and $\mathbf{X}$ were the empirical and real-time data of each channel respectively, C denoted the number of channels (in our study, the data of head transverse swing angle, head pitch angle, head lateral inclination angle, and horizontal and vertical coordinates of the trajectory position, totaling five channels, respectively),and t_e and t_a denoted the time step between the empirical database and the actual data, respectively. $\mathbf{X}_t$ was used as the input to the GDN network, and after data embedding, graph structure relation learning, and graph attention-based data prediction, the final prediction result $\mathbf{P'}_t$ for each channel parameter was obtained.

$$\mathbf{P'}_t = [\hat{\mathbf{s}}^{(1)}, \hat{\mathbf{s}}^{(2)}, \cdots, \hat{\mathbf{s}}^{(t)}] \tag{2}$$

where $\hat{\mathbf{s}}^{(t)}$ denoted the predicted value at moment t and it was calculated as follows [4]:

$$\hat{\mathbf{s}}^{(\mathbf{t})} = f_\theta \left(\left[\mathbf{v}_1 \odot \mathbf{z}_1^{(t)}, \cdots, \mathbf{v}_N \odot \mathbf{z}_N^{(t)} \right] \right) \tag{3}$$

$$\mathbf{z}_i^{(t)} = \mathrm{ReLU} \left(\alpha_{i,i} \mathbf{W} \mathbf{X}_i^{(t)} + \sum_{j \in \mathcal{N}^{(i)}} \alpha_{i,j} \mathbf{W} \mathbf{X}_j^{(t)} \right) \tag{4}$$

where $\mathbf{z}_i^{(t)} (i = 1, \cdots, N)$ denoted the representations for all N nodes, $\odot$ denoted the Hadamard product (element-wise multiplication of vectors), $\mathbf{X}_i^{(t)}$ and $\mathbf{X}_j^{(t)}$ were input features of node i and j, $\mathbf{W}$ was a trainable weight matrix that applied a linear transformation to each node, f_θ denoted the fully connected operation and $\mathbf{v}_N$ denoted the embedding matrix obtained after the data embedding operation $\mathcal{N}(i) = \{j | \mathbf{A}_{ji} > 0\}$,was the set of neighbors of node i obtained from the learned adjacency matrix $\mathbf{A}_{ji}$, $\mathbf{A}_{ji}$ denoted the learned adjacency matrix, obtained after learning through graph structure relations, and the attention coefficients $\alpha_{i,j}$ were computed as:

$$\alpha_{i,j} = \frac{\exp(\pi(i,j))}{\sum_{k \in \mathcal{N}(i) \cup \{i\}} \exp(\pi(i,k))} \tag{5}$$

where the attention coefficient $\pi(i,j) = \mathrm{LeakyReLU}\left(\mathbf{a}^\top \left(\mathbf{g}_i^{(t)} \oplus \mathbf{g}_j^{(t)}\right)\right)$, $\oplus$ was concatenation operation, $\mathbf{g}_i^{(t)}$ was mainly used to concatenate the embedded data $\mathbf{v}_i$ with the corresponding transformed features $\mathbf{W}\mathbf{X}_i^{(t)}$:

$$\mathbf{g}_i^{(t)} = \mathbf{v}_i \oplus \mathbf{W}\mathbf{X}_i^{(t)} \tag{6}$$

Since $\mathbf{P'}_t$ was a prediction of the overall results of $\mathbf{X}$ and $\mathbf{X}_e$, the predicted value $\mathbf{P}_t$ corresponding to the actual collected value $\mathbf{X}$ also needs to be filtered from $\mathbf{P'}_t$ at this point, as shown in Equation (7).

$$\mathbf{P}_t = \mathbf{X} \sim \mathbf{P'}_t \in \mathbb{R}^{C*t} \tag{7}$$

C. PEDESTRIAN POSE AND TRAJECTORY PREDICTION BASED ON ANOMALY INHIBITION-LSTM

As shown in Fig. 3(a), for the standard LSTMs, f_t and i_t are used in the forgetting gate and input gate, respectively, to control the previous state's forgetting status and the current state's retaining status, respectively, to calculate the state's output at moment t of the cells, as shown in Eq. (8).

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c_t} \tag{8}$$

where $f_t \in [0, 1]$, $i_t \in [0, 1]$.

Inspired by the principle of this algorithm, we added the temporal anomaly inhibition module to the standard LSTM, as shown in the red module in Fig. 3(b), and used the difference $\mathbf{E}_t$ between $\mathbf{P}_t$ and $\mathbf{X}$ as the output parameter for controlling the cells' state at the moment t, as shown in Eq. (9).

$$\mathbf{E}_t = \mathbf{X} - \mathbf{P}_t \tag{9}$$

The fundamental concept is that a higher $\mathbf{E}_t$ value signifies a significant difference between the predicted value and the actual value. In this context, the state at the current moment has a higher probability of being interference noise, and thus, it should be inhibited. Conversely, a lower $\mathbf{E}_t$ value suggests that the state at the current moment does not contain noise. At the same time, to ensure that the final training process can have a faster convergence speed, $\mathbf{E}_t$ is distributed in the interval [0,1]. As shown in Eq. (10), e^{-x^2} is used to process $\mathbf{E}_t$ to obtain the coefficient $\boldsymbol{\alpha}_t$ used to control the degree of forgetting in the previous moment and make $\boldsymbol{\beta}_t = \mathbf{1} - e^{-\mathbf{E}_t^2}$.

$$\boldsymbol{\alpha}_t = e^{-\mathbf{E}_t^2} \tag{10}$$

At the same time, considering $f_t, i_t \in \mathbb{R}^{hidden \times 1}$, it is necessary to perform affine transformation on $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$, thus ensuring that $\boldsymbol{\alpha'}_t, \boldsymbol{\beta'}_t \in \mathbb{R}^{hidden \times 1}$, as shown in Eq. (11).

$$\boldsymbol{\alpha'}_t = \mathbf{W}_\alpha \boldsymbol{\alpha}_t, \boldsymbol{\beta'}_t = \mathbf{W}_\beta \boldsymbol{\beta}_t \tag{11}$$

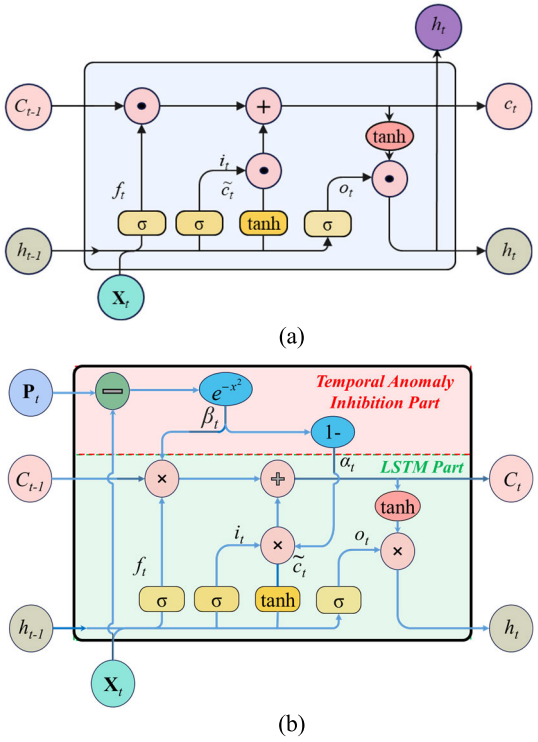


FIGURE 3. Schematic diagrams of LSTM and AI-LSTM structures. (a) denotes a standard LSTM unit and (b) denotes an AI-LSTM unit.

where $\mathbf{W}_\alpha, \mathbf{W}_\beta \in \mathbb{R}^{hidden \times C}$ are learnable affine matrices. Then the output of the cellular state at moment t is

$$c_t = \boldsymbol{\alpha}'_t \odot f_t \odot c_{t-1} + \boldsymbol{\beta}'_t \odot i_t \odot \tilde{c}_t \tag{12}$$

After obtaining the intermediate hidden layer vectors $\tilde{\mathbf{H}} = \left[\hat{\mathbf{H}}_1, \hat{\mathbf{H}}_2, \cdots, \hat{\mathbf{H}}_t\right]$ output from the AI-LSTM unit, we processed them through fully connected and regression operations. Finally, we obtained the predicted values of each channel with the noise interference removed $\mathbf{Y} \in \mathbb{R}^{C*t}$.

$$\mathbf{Y} = \mathbf{W}_y \tilde{\mathbf{H}} + \mathbf{b}_y \tag{13}$$

As shown in Eq. (14), the final loss value *loss* is obtained by target value t_{ij} and predicted value y_{ij}. Where t_{ij} and y_{ij} represented the target value and predicted value of channel j at moment i, respectively.

$$loss = \frac{1}{2} \sum_{i=1}^{t} \sum_{j=1}^{C} (t_{ij} - y_{ij}) \tag{14}$$

IV. EXPERIMENTS

A. EXPERIMENTAL DESIGN AND TRAINING SAMPLE CONSTRUCTION

The algorithm operates under the following constraints: 1. Pedestrian information must be captured from the angle of the road-based equipment; 2. The information must be captured by the same road-based sensors (this restriction is primarily intended to mitigate the influence of information such as the road structure, traffic flow, and traffic laws and regulations on pedestrian poses, ensuring that empirical information aligns with the characteristics of road conditions); and 3. The captured pedestrian information must be in three-dimensional form.

While numerous pedestrian crossing datasets are available, our previous research [14] revealed that existing datasets often do not align with our specific requirements. Consequently, we opted to construct a virtual environment for the dataset used in this paper. As depicted in Fig. 4, prioritizing the safety of the experimental process led us to adopt the offline simulation method. Initially, we created a virtual city road scene using Unreal Engine 4 (UE4). Recognizing that various scenarios may elicit distinct responses from pedestrians, we opted for a typical urban road setting comprising solely of a crosswalk for our experimental investigations. Subsequently, we connected the Oculus Quest2 device to the test host via a Bluetooth module to collect various characteristic parameters of pedestrians. The experimenter simulated street crossing by wearing the device in a safe indoor environment, simultaneously collecting three-channel head and positional coordinates data. As mentioned in the "Introduction" section, the noise we encounter stems from interference, such as the occlusion of jointed body parts during target detection. Moreover, we observed that this interference exhibits high randomness and continuity, consistent with the characteristics of Gaussian noise. Hence, after acquiring each dataset from the experimenter, we introduced Gaussian noise with different σ values to simulate noise interference in a real environment. Notably, the Oculus Quest2 is a VR device enabling the subject to simulate street crossing in a virtual environment, recording information such as the subject's head pose angle, absolute position, and relative position in the virtual environment.

The hardware information we applied in the network training process was as follows: GPU was NVIDIA RTX4070TI, RAM was 12GB, the software platform was pytorch1.5.1, and the other environment information was cuda10.2+cudnn-10.2.The parameter settings during the network training session were as follows: 1. the TOPK value of GDN was 15, the embedding dimension was equal to 96; 2. the settings of LSTM, LSTM, +Attention, Bi-LSTM, and GRU were as follows: the hidden states were equal to 100, the optimizer was Adam, the maximum number of epochs to use for training was set to 30, and we padded sequences on the right; 3. Our method's TOPK value and embedding dimension were 15 and 96, respectively. Moreover, we set The hidden states, the optimizer, and the maximum number of epochs as the same as LSTM.

B. POSE ESTIMATION COMPARISON EXPERIMENT

In the experiments comparing pose estimation, we introduced various levels of Gaussian noise to the pedestrian head pose data. We then combined this noise, characterized by different variances, with the actual values of head poses for each angle. Subsequently, we applied our method and each of the other

(a) (b) (c)

FIGURE 4. Schematic diagram of data acquisition equipment and process. Where (a) represents the virtual scene we built based on UE4, (b) represents the virtual road scene and the experimental equipment, (c) represents the experiment process.

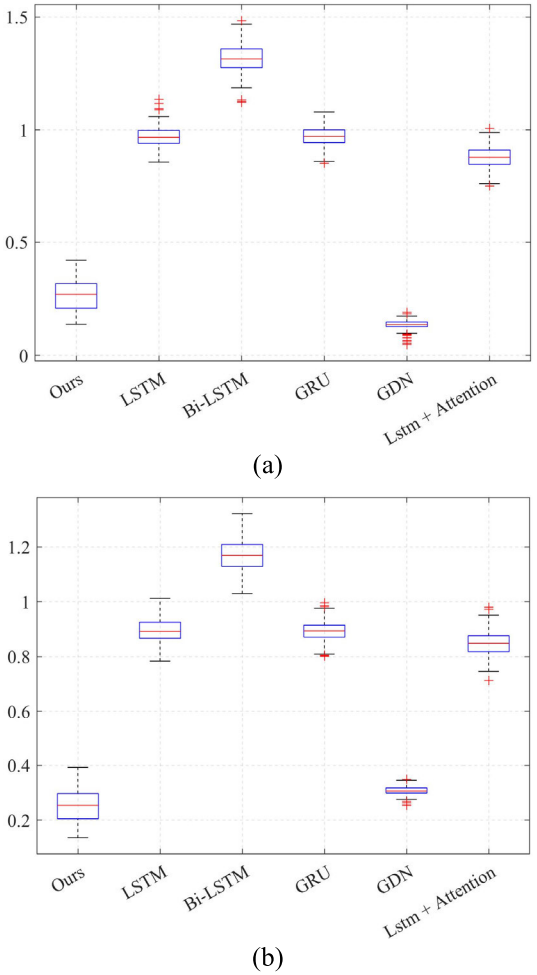


FIGURE 5. Box diagrams of head pose angles for different algorithms. Where (a) denotes the sway angle, (b) denotes the pitch angle.

control groups to address this perturbed data.

$$Channel(j) = \sqrt{\frac{1}{M}\sum_{n=1,\cdots,M}\left(x_j^n - \hat{x}_j^n\right)^2} \tag{15}$$

where $j = 1, 2, 3$ denote the head pitch angle, the transverse swing angle, and the lateral inclination angle, respectively. x_j^n and $\hat{x}_i^n$ denote the predicted and actual values of different head pose angles at the moment n, respectively.

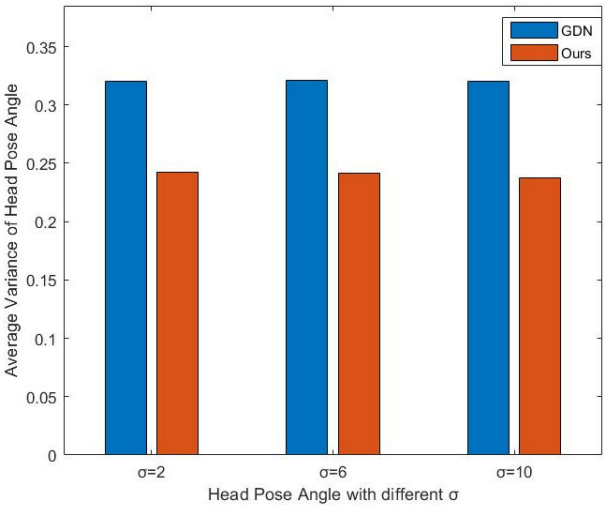


FIGURE 6. The average variance of head pose angle for different algorithms.

As our method is derived from modifications to GDN and LSTM, we initially designated GDN and LSTM as two control groups. Additionally, we included several other widely-used prediction methods based on time-series data (Bi-LSTM, GRU, LSTM+Attention) as control groups for comparison. The results are presented in Fig. 5. To expedite the selection of methods most closely aligned with the effectiveness of our approach for further comparison, we initially conducted preliminary comparisons among various algorithms when $\sigma = 1$, as illustrated in Figure 5.

As depicted in Fig. 5, our method exhibits higher accuracy and smaller fluctuations in the prediction interval compared to the other three methods (LSTM, Bi-LSTM, GRU, and LSTM+Attention). Our algorithm surpasses the GDN algorithm in predicting the head pitch angle. Furthermore, our method generates fewer outliers, indicating its greater ability to eliminate noise interference. It is evident that the GDN algorithm exhibits performance similar to our method when interference is minimal, and even slightly outperforms our method in pitch angle. Therefore, we have chosen the GDN algorithm as a further benchmark for comparison, we further compared them by introducing additional control groups with increased noise. The comparison is based on the total mean values of the sway angle and the pitch angle.

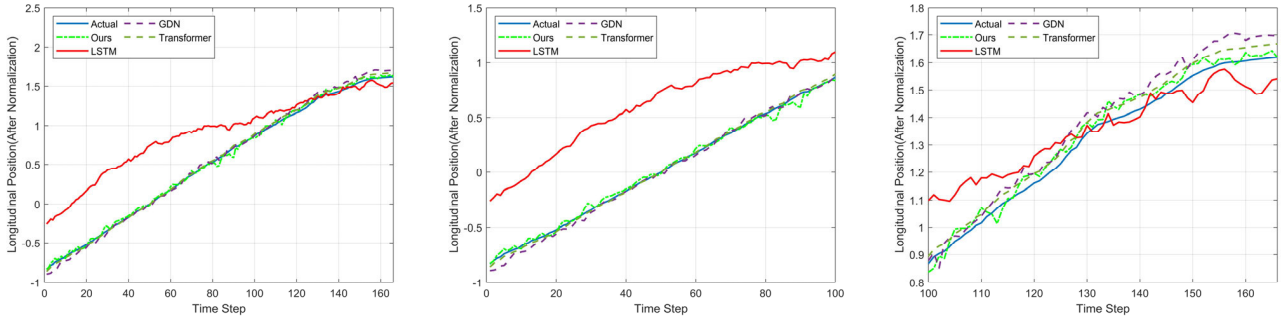


FIGURE 7. Results of partial pedestrian trajectory prediction. Where (a) represents the entire trajectory prediction, (b) represents the prediction results during the first half of the period, and (c) represents the prediction results during the second half of the period.

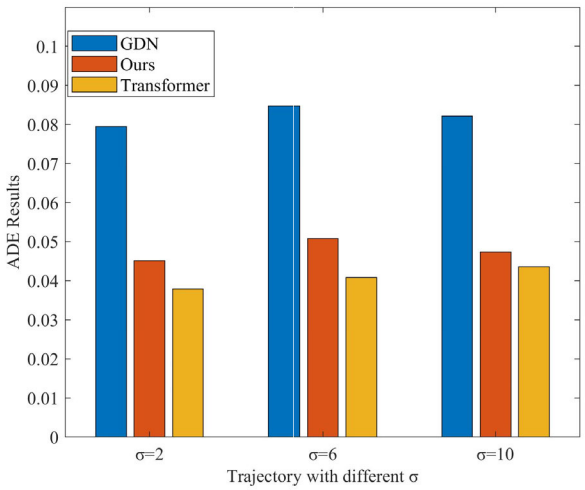


FIGURE 8. ADE result.

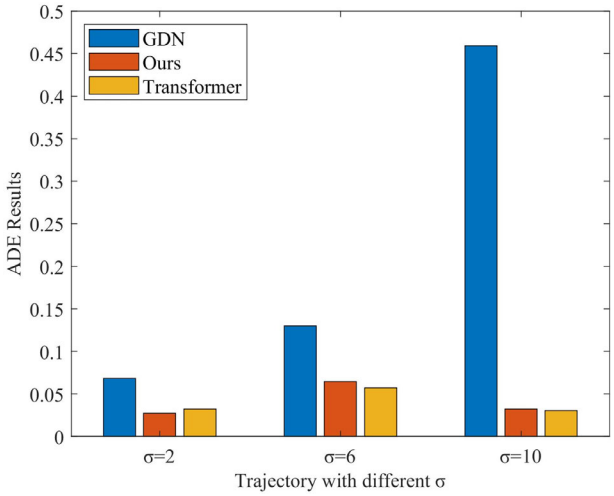


FIGURE 9. FDE result.

The results, depicted in Fig. 6, clearly show that our prediction results are closer to the actual values.

C. TRAJECTORY PREDICTION COMPARISON TEST

In pedestrian trajectory prediction experiments, to identify methods that closely match the effectiveness of our approach, we initially visually compared the overall performance of different methods ($\sigma = 1$). As our method is an improvement upon GDN and LSTM, we first selected GDN and LSTM as control groups. Additionally, we chose the transformer-based method proposed by Giuliari et al. [24] as another control group.

Demonstrated in Fig. 7 are some of the prediction results of the three methods in pedestrian longitudinal trajectories. From Fig. 7(a), it is evident that our algorithm and Giuliari's algorithm consistently approximate the actual value throughout the longitudinal trajectory prediction. Observing Fig. 7(b) and Fig. 7(c), it becomes apparent that in the early stage of the prediction, the LSTM algorithm exhibits a larger bias. Over time, it gradually converges towards the actual value in the subsequent stages of the prediction. In contrast, the GDN algorithm is closer to the actual value at the beginning of the prediction but gradually deviates from the actual value in the subsequent prediction process. We believe that the phenomenon described above arises from the fact that the LSTM algorithm fails to consider the inherent relationships among input features. Consequently, it is highly susceptible to noise during the early stages. However, over time, it gradually suppresses some of the noise based on observations within certain time steps. In contrast, the GDN algorithm, benefiting from the interrelationships among features, can suppress the influence of noise during the early stages. Nevertheless, the GDN algorithm itself lacks the capability for noise suppression. Therefore, in the later stages, due to the accumulation of noise, the predicted values gradually deviate from the actual values.

In the above experiments, the differences between the remaining control groups are not significant enough, except for the LSTM control group, which is less effective. To further compare the differences among the remaining control groups, we introduced additional interference into the original trajectories to create control groups. We then evaluated them using commonly used metrics in pedestrian trajectory prediction, namely Average Displacement Error (*ADE*) and

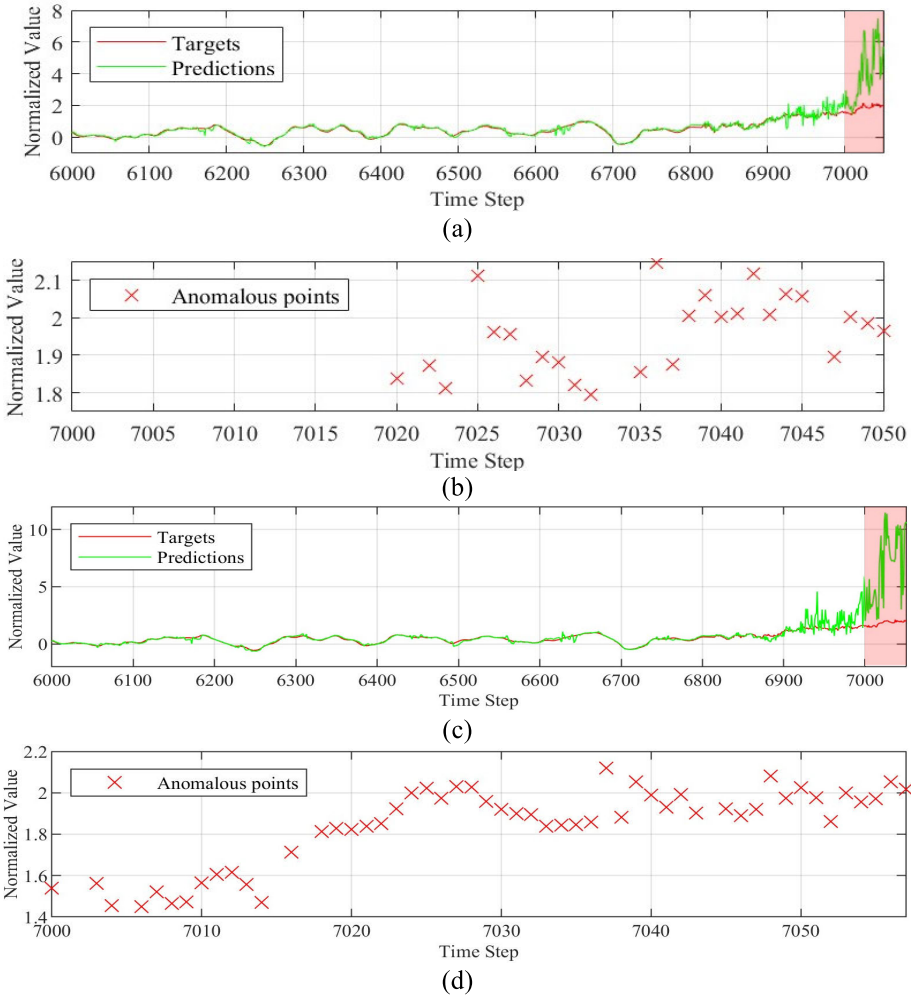


FIGURE 10. Anomaly detection test using pedestrian longitudinal trajectory as an example. Gaussian noise with with mean 0, variance 4 added in (a), and Gaussian noise with with mean 0, variance 8 added in (c). Figs. 10(a) and 10(b) represent the actual and predicted values, while Figs. 10(b) and 10(d) display the anomaly detection results in the red interval in (a) and (c), respectively.

Final Displacement Error (*FDE*). *ADE* represents the mean square error between the actual trajectory coordinates and the model-predicted trajectory coordinates for all trajectories. *FDE* is the error between the actual trajectory end coordinates and the model-predicted trajectory end coordinates for all trajectories. We employed these two metrics to evaluate the predictive capability of the model. The equations for the two metrics are shown in (16) and (19):

$$ADE(k) = \frac{1}{T} \sum_{t=1,2\cdots,T} \sqrt{\left(x_k^t - \hat{x}_k^t\right)^2 + \left(y_k^t - \hat{y}_k^t\right)^2} \quad (16)$$

$$ADE = \frac{1}{N} \sum_{k=1,2\cdots,N} ADE(k) \quad (17)$$

$$FDE(k) = \sqrt{\left(x_k^T - \hat{x}_k^T\right)^2 + \left(y_k^T - \hat{y}_k^T\right)^2} \quad (18)$$

$$FDE = \frac{1}{N} \sum_{k=1,2\cdots,N} FDE(k) \quad (19)$$

where $\left(x_k^t, y_k^t\right)$ denoted the model-predicted trajectory coordinate points for all pedestrian trajectories at all moments; $(\hat{x}_k^T, \hat{y}_k^T)$ denoted the final pedestrian trajectory coordinate points predicted by the model; $(\hat{x}_k, \hat{y}_k)$ denoted the actual trajectory coordinate points corresponding to the predicted trajectory coordinate points; T denoted the predicted sequence length; N denoted the number of all pedestrian trajectories.

Some of the experimental results are shown in Figures 8 and 9. From Figures 8 and 9, we can see that in the three groups of experimental results with mean 0, variance 2, 6 and 10, our method outperforms the GDN algorithm in all the indexes, and basically has a similar performance with the Transformer-based method, which occurs because, compared with the GDN algorithm, we inhibit the effect of the high-noise moment data on the whole time series.

D. ANOMALY DETECTION TEST

Although our algorithm slightly underperforms compared to the Transformer in terms of ADE and FDE evaluation

TABLE 1. Main variables and descriptions.

Variables	*Variable Description*
$\mathbf{X}_e$	pre-collected pedestrian trajectory and pose data used as empirical dataset
$\mathbf{X}$	real-time data (including pedestrian trajectory and pose)
$\mathbf{X}_t$	actual inputs to the GDN and AI-LSTM modules
$\mathbf{P}_t'$	the set of predicted values $\hat{\mathbf{s}}^{(t)}$ of each channel parameter obtained after GDN (containing the predicted values of $\mathbf{X}_e$)
$\mathbf{z}_i^{(t)}$	denotes the representations for all N nodes
$\mathbf{X}_i^{(t)}$	input features of node i
$\alpha_{i,j}$	attention coefficients(After normalized)
$\pi(i,j)$	attention coefficients(Before normalized)
$\mathbf{g}_i^{(t)}$	concatenate the embedded data $\mathbf{v}_i$ with the corresponding transformed features $\mathbf{W}\mathbf{X}_i^{(t)}$
$\mathbf{P}_t$	The set of predicted values $\hat{\mathbf{s}}^{(t)}$ of each channel parameter obtained after GDN(without the predicted values of $\mathbf{X}_e$)
c_t	cells' state at the moment t (Used in the standard LSTM unit)
f_t	control the forgetting status of the state at the previous moment (Used in the standard LSTM unit)
i_t	retaining status of the current state to obtain the output of the cells' state at the moment t
$\mathbf{E}_t$	difference between actual and predicted values for each channel at time t
$\boldsymbol{\alpha}_t$	control the forgetting status of the state at the previous moment with f_t (Used in the AI-LSTM unit)
$\boldsymbol{\beta}_t$	retaining status of the current state to obtain the output of the cells' state at the moment t with i_t (Used in the AI-LSTM unit)
$\mathbf{W}_\alpha, \mathbf{W}_\beta, \mathbf{W}$	trainable weight matrix
t_{ij}	target value of channel j at moment i
y_{ij}	predicted value of channel j at moment i

metrics, our method possesses interpretability. Specifically, for our model, if it can identify outlier data and their corresponding time points throughout the entire time series, its overall good performance can be attributed to the suppression of these outlier data. This indirectly suggests that the model's convergence process possesses a certain degree of interpretability. To assess this capability, we introduced different noises in the data interval from 6850 to 7050 steps. As depicted in Figs. 10(b) and 10(d), our method effectively identifies anomalous data nodes.

In summary, our method is effective not only in detecting anomalous data but also in suppressing the effect of the anomalous data on pose estimation throughout the entire time series in the subsequent AI-LSTM module. Our method is not only noise-suppressive but also interpretable compared to the method of Giuliari et al.

V. CONCLUSION

This paper proposes a method for PPE and PTP under noise interference, achieved by combining a GDN network and a novel LSTM unit. Initially, we obtained the predicted value by considering the relationship between the pedestrian's position and pose through the GDN module. Subsequently, we utilized this value and the actual value as inputs to the AI-LSTM network, ultimately obtaining prediction results with noise suppression. The effectiveness of our method was also verified through experiments from both pose estimation and trajectory prediction perspectives. The shortcomings of our study are mainly reflected as follows:

(1). The premise of our method was based on pedestrians crossing the road rationally. Our method may not effectively handle irrational crossing patterns (e.g., crossing the road directly without observation) or data anomalies caused by special behaviors (e.g., making a phone call while crossing the road). In the future, we could explore introducing quantum characterization to enhance uncertainty in data processing, addressing the aforementioned challenges [25].

(2). Our method does not address the issue of head pose changes due to interference behavior between different pedestrians. We may consider introducing a GAP (Group Activity Recognition) algorithm in later studies [26].

(3). Our study did not consider the impact of inter-pedestrian influence on trajectory prediction results. In future research, we could explore the incorporation of an inhibition module into the social-LSTM algorithm [27] for a more comprehensive investigation.

APPENDIX

See the Table 1.

REFERENCES

[1] R. Quan, L. Zhu, Y. Wu, and Y. Yang, "Holistic LSTM for pedestrian trajectory prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 3229–3239, 2021.

[2] Z. Fang and A. M. López, "Intention recognition of pedestrians and cyclists by 2D pose estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4773–4783, Nov. 2020.

[3] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2008.

[4] A. Deng and B. Hooi, "Graph neural network-based anomaly detection in multivariate time series," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, pp. 4027–4035.

[5] G. Li and J. J. Jung, "Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges," *Inf. Fusion*, vol. 91, pp. 93–102, Mar. 2023.

[6] R. Korbmacher and A. Tordeux, "Review of pedestrian trajectory prediction methods: Comparing deep learning and knowledge-based approaches," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24126–24144, Dec. 2022.

[7] Z. Li, J. Ye, M. Song, Y. Huang, and Z. Pan, "Online knowledge distillation for efficient pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11720–11730.

[8] K. Wang, L. Lin, C. Jiang, C. Qian, and P. Wei, "3D human pose machines with self-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1069–1082, May 2020.

[9] H. Liu, T. Liu, Y. Chen, Z. Zhang, and Y.-F. Li, "EHPE: Skeleton cues-based Gaussian coordinate encoding for efficient human pose estimation," *IEEE Trans. Multimedia*, early access, Aug. 8, 2022, doi: 10.1109/TMM.2022.3197364.

[10] S. Li, H. Zhang, H. Ma, J. Feng, and M. Jiang, "CSIT: Channel spatial integrated transformer for human pose estimation," *IET Image Process.*, vol. 17, no. 10, pp. 3002–3011, Aug. 2023.

[11] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "ViTAEv2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," *Int. J. Comput. Vis.*, vol. 131, no. 5, pp. 1141–1162, Jan. 2023.

[12] Y. Wu, G. Chen, Z. Li, L. Zhang, L. Xiong, Z. Liu, and A. Knoll, "HSTA: A hierarchical spatio-temporal attention model for trajectory prediction," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 11295–11307, Nov. 2021.

[13] S. Zhang, M. Abdel-Aty, Y. Wu, and O. Zheng, "Pedestrian crossing intention prediction at red-light using pose estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2331–2339, Mar. 2022.

[14] J. Zhou, X. Bai, and W. Hu, "Recognition and prediction of pedestrian hazardous crossing intentions in visual field obstruction areas based on IPVO-LSTM," *Appl. Sci.*, vol. 13, no. 5, p. 2999, Feb. 2023.

[15] P. Kothari, S. Kreiss, and A. Alahi, "Human trajectory forecasting in crowds: A deep learning perspective," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7386–7400, Jul. 2022.

[16] S. Ahmed, A. A. Bazi, C. Saha, S. Rajbhandari, and M. N. Huda, "Multi-scale pedestrian intent prediction using 3D joint information as spatio-temporal representation," *Expert Syst. Appl.*, vol. 225, Sep. 2023, Art. no. 120077.

[17] X. Zhang, P. Angeloudis, and Y. Demiris, "ST CrossingPose: A spatial–temporal graph convolutional network for skeleton-based pedestrian crossing intention prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20773–20782, Nov. 2022.

[18] P. Czech, M. Braun, U. Kreßel, and B. Yang, "On-board pedestrian trajectory prediction using behavioral features," in *Proc. 21st IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2022, pp. 437–443.

[19] V. Rajinikanth, S. Kadry, D. Taniar, K. Kamalanand, M. A. Elaziz, and K. P. Thanaraj, "Detecting epilepsy in EEG signals using synchro-extracting-transform (SET) supported classification technique," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 8, pp. 10123–10141, Aug. 2023.

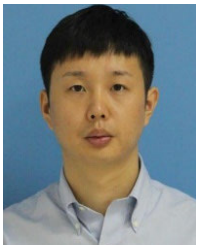
[20] S. You, B. H. Cho, Y.-M. Shon, D.-W. Seo, and I. Y. Kim, "Semi-supervised automatic seizure detection using personalized anomaly detecting variational autoencoder with behind-the-ear EEG," *Comput. Methods Programs Biomed.*, vol. 213, Jan. 2022, Art. no. 106542.

[21] M.-S. Cheong, M.-C. Wu, and S.-H. Huang, "Interpretable stock anomaly detection based on spatio-temporal relation networks with genetic algorithm," *IEEE Access*, vol. 9, pp. 68302–68319, 2021.

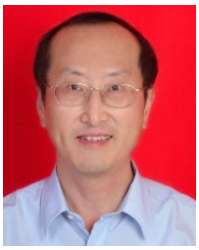
[22] R. N. Madurawe, B. K. D. I. Jayaweera, T. D. Jayawickrama, I. Perera, and R. Withanawasam, "Collusion set detection within the stock market using graph clustering & anomaly detection," in *Proc. Moratuwa Eng. Res. Conf. (MERCon)*, Jul. 2021, pp. 450–455.

[23] D. K. Sharma, T. Dhankhar, G. Agrawal, S. K. Singh, D. Gupta, J. Nebhen, and I. Razzak, "Anomaly detection framework to prevent DDoS attack in fog empowered IoT networks," *Ad Hoc Netw.*, vol. 121, Oct. 2021, Art. no. 102603.

[24] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 10335–10342.
[25] Q. Song, W. Fu, W. Wang, Y. Sun, D. Wang, and J. Zhou, "Quantum decision making in automatic driving," *Sci. Rep.*, vol. 12, no. 1, p. 11042, Jun. 2022.
[26] M. Han, D. J. Zhang, Y. Wang, R. Yan, L. Yao, X. Chang, and Y. Qiao, "Dual-AI: Dual-path actor interaction learning for group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2980–2989.
[27] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.



JINCAO ZHOU received the M.S. and Ph.D. degrees in vehicle engineering from Chang'an University, Xi'an, China, in 2014 and 2018, respectively. He is currently a Lecturer with the College of Mechanical and Precision Instrument Engineering, Xi'an University of Technology. His research interests include autopilot systems and its environmental perception and driving behavior decision theory, and ADAS and its theory and technology.



WEIPING FU received the Ph.D. degree in mechanical design and theory from Xi'an Jiaotong University, Xi'an, China, in 1996. He is currently a Professor with the College of Mechanical and Precision Instrument Engineering, Xi'an University of Technology. His research interests include intelligent robot, modern logistics system engineering and technology, intelligent vehicle and its control theory and technology, electromechanical systems, and manufacturing system dynamics and control.

BENYU NING received the B.S. degree in mechanical engineering from Shaanxi University of Technology, China, in 2022. He is currently pursuing the master's degree with Xi'an University of Technology. His current research interests include ADAS systems and computer vision.

SIYUAN HE received the Bachelor of Engineering degree from Xi'an University of Technology, in 2022, where he is currently pursuing the master's degree. His current research interests include robot autonomous assembly, point cloud visual recognition, and counterfactual reasoning.