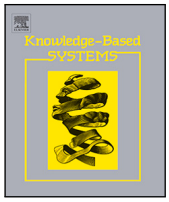
Knowledge-Based Systems 299 (2024) 112038

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



A multimodal stepwise-coordinating framework for pedestrian trajectory prediction

Yijun Wang [a], Zekun Guo [a], Chang Xu [b], Jianxin Lin [a,*]

[a] College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410082, China
[b] Microsoft Research Asia, Beijing, 100080, China

ARTICLE INFO

Keywords:
Pedestrian trajectory prediction
Multimodal fusion
Transformer network

ABSTRACT

Pedestrian trajectory prediction from the first-person view has still been considered one of the challenging problems in automatic driving due to the difficulty of understanding and predicting pedestrian actions. Observing that pedestrian motion naturally contains the repetitive pattern of the gait cycle and global intention information, we design a Multimodal Stepwise-Coordinating Network, namely MSCN, to sufficiently leverage the underlying human motion properties. Specifically, we first design a multimodal spatial-frequency encoder, which encodes the periodicity of pedestrian motion with a frequency-domain enhanced Transformer and other visual information with a spatial-domain Transformer. Then, we propose a stepwise-coordinating decoder structure, which leverages both local and global information in sequence decoding through a two-stage decoding process. After generating a coarse sequence from the stepwise trajectory predictor, we design a coordinator to aggregate the corresponding representations used to generate the coarse sequence. Subsequently, the coordinator learns to output a refined sequence through a knowledge distillation process based on the aggregated representations. In this way, MSCN can adequately capture the representations of short-term motion behaviors, thus modeling better long-term sequence prediction. Extensive experiments show that the proposed model can achieve significant improvements over state-of-the-art approaches on the PIE and JAAD datasets by 16.1% and 16.4% respectively.

1. Introduction

First-person view pedestrian trajectory prediction [1–5], which forecasts the future locations of pedestrians in an ego-centric view of a moving vehicle, is crucial for automatic driving systems since it helps to avoid collisions with pedestrians. Such a task requires not only a high-level understanding of pedestrian historical behaviors but also a clear perception of environments to accurately predict the future trajectory of pedestrians. Therefore, existing works [2,4–10] have widely explored leveraging additional modalities, such as ego-vehicle motion data, optical flow data, etc., to improve the performance on first-person view pedestrian trajectory prediction tasks compared to traditional trajectory-based methods [4,11]. However, there are few analyses on the fundamental and core point of view of pedestrian motion in existing works.

In fact, pedestrian motion is a process containing both fine-grained periodicity and global intention information. Specifically, pedestrian motion naturally forms a series of gait cycles. For example, when a pedestrian is walking, we can take the right foot of the pedestrian as a reference. The right foot first touches the ground, then the left foot touches the ground, and finally, the right foot touches the ground again. Therefore, we can find that the entire pedestrian walking process is composed of a repetitive pattern of several gait cycles step by step, indicating a periodicity [12–14]. Moreover, existing works [14–16] have shown that human walking trajectory can be well modeled and simulated based on human gait motion properties. All of the studies indicate that the inherent property of short-term *periodic gait cycle* should be taken into consideration for better pedestrian trajectory prediction.

On the other hand, it is worth noting that pedestrian motion is mostly driven by some intention or goal as global information along with the fine-grained step-by-step process. For example, in the context of traffic, pedestrians usually reach a certain destination within a given time frame [17]. In existing works, incorporating goal estimation as an auxiliary optimization objective has been shown to improve trajectory prediction [3,17,18]. However, not much attention has been paid to reviewing trajectory prediction results and refining the results based on the global context understanding.

* Corresponding author.
E-mail addresses: wyjun@hnu.edu.cn (Y. Wang), guozekun@hnu.edu.cn (Z. Guo), chanx@microsoft.com (C. Xu), linjianxin@hnu.edu.cn (J. Lin).

https://doi.org/10.1016/j.knosys.2024.112038
Received 14 March 2024; Received in revised form 20 May 2024; Accepted 29 May 2024
Available online 1 June 2024

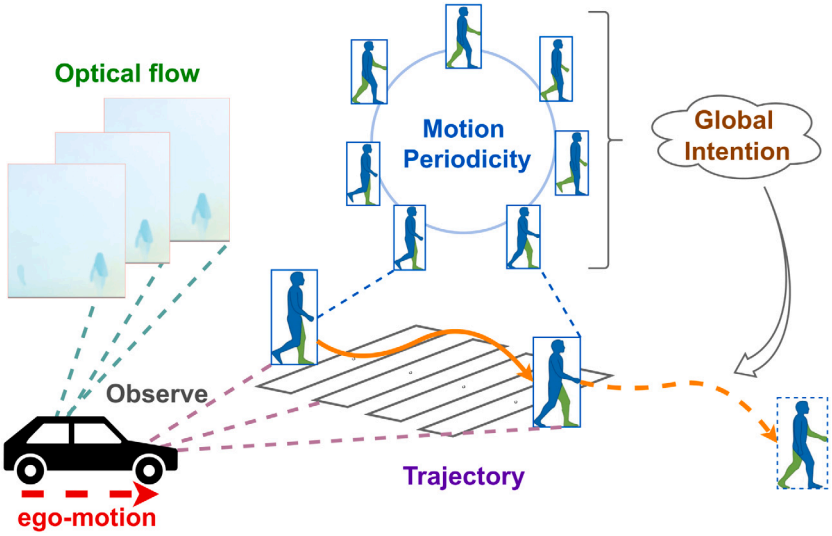


Fig. 1. Illustration of pedestrian motion from the first-person view. It involves multiple modalities, including the ego-motion, the observed pedestrian trajectory, and the optical flow information. The pedestrian motion inherently takes on the property of motion periodicity and global intention information.

Inspired by these observations in pedestrian motion, in this paper, we propose a Multimodal Stepwise-Coordinating Network (MSCN), leveraging the implicit property of periodic gait cycles and the global information of pedestrians' intention for better trajectory prediction (the design tenet is depicted in Fig. 1). To be specific, we first design a multimodal spatial-frequency encoder in which frequency-domain enhanced Transformer and spatial Transformer are utilized to model the periodicity of pedestrian motion and other visual information respectively. Then, using a two-stage decoding process, we present a stepwise-coordinating decoder that makes use of the global information in sequence decoding. We create a *coordinator* to aggregate the corresponding representations from the stepwise trajectory predictor which generates the coarse sequence result. Afterward, we propose to make the coordinator output a refined sequence through a knowledge distillation process based on the aggregated representations. In this way, the global information of the target sequence can be utilized to refine the generation process. Extensive evaluations on two pedestrian trajectory prediction datasets PIE [1] and JAAD [19] demonstrate the effectiveness of MSCN and its advantages over state-of-the-art baselines.

The main contributions of this paper can be summarized as:

1. We highlight a new direction for pedestrian trajectory prediction by modeling local periodic gait cycle and global context information during decoding.
2. We design a multimodal spatial-frequency encoder to effectively model the periodicity of pedestrian motion with frequency-domain enhanced Transformer fused with other spatial information.
3. We propose a stepwise-coordinating decoder, which adopts a two-stage decoding process, i.e., the stepwise trajectory predictor and the coordinator, to more effectively capture both short-term and long-term motion interaction.

2. Related works

Trajectory prediction in first-person view circumstance. For modeling first-person view trajectory prediction tasks, three modalities are usually considered, i.e., pedestrian trajectory, ego-motion, and image. For image modality, Rasouli et al. [1] proposed to incorporate pedestrian intention estimation and vehicle speed prediction for future Trajectory Prediction by combining multiple LSTMs (Long Short-Term Memory). Quan et al. [20] proposed a holistic LSTM to incorporate multiple sources of information from pedestrians and vehicles adaptively. Yang et al. [21] further proposed to fuse local patch features with global semantic segmentation information for better scene understanding. Rasouli et al. [7] designed a categorical interaction module to generate interaction latent representation, thus capturing the relationship between target pedestrians and surroundings. Yin et al. [6] used the historical video to obtain optical flow information and utilized Transformer architecture [22] to realize the coarse-grained fusion and the fine-grained fusion for multimodal data. For pedestrian trajectory and ego-motion modalities, Rasouli et al. [7] and Yang et al. [21] proposed to model the trajectory and ego-motion through LSTM [23] and multi-level GRU (Gate Recurrent Unit) [24] respectively. Other works [5,6] applied the cross Transformer to model trajectories and ego-motions to extract their potential relationships. In this paper, to model various modality information in trajectory prediction, we propose a multimodal spatial-frequency encoder, in which frequency-domain enhanced Transformer is used to encode the periodicity of pedestrian motion and spatial-domain Transformer is used to extract other spatial information.

Decoders in trajectory prediction. In trajectory prediction, it is simple and efficient to apply MLP (Multilayer Perceptron) as a decoder [18,25]. However, this kind of decoder cannot guarantee the smoothness of the prediction trajectory. The problem of trajectory smoothness can be solved by fitting the trajectory with a fixed mathematical function, such as cubic polynomial curve [26] or Bézier curve [5], which is called curve-based decoder, while prediction flexibility is still limited due to the pre-set fitting function. Pang et al. [27] introduced Bayesian fully connected layers to handle uncertainty in trajectory data, while others [17,28–30] have adopted GRU/LSTM as decoders to better capture temporal dependencies. Wu et al. [29] utilized a stacked RNN that maps motion vectors from source views to target views in multiview trajectory prediction. Additionally, Li et al. [31] proposed a diffusion model-based decoder to capture overall trajectory characteristics. Other works [32,33] employed decoders for multi-agent trajectory prediction, demonstrating their ability to consider distinct characteristics of individual agents when inferring future positions. The high computational cost of the autoregressive Transformer decoder is a significant problem that makes it unsuitable for real-time applications [34,35], particularly for tasks like predicting pedestrian movements. To tackle this issue, inspired by streaming/real-time speech recognition methods [36,37], we design a stepwise-coordinating decoder based on LSTM. The coordinator aggregates the corresponding representations from the stepwise trajectory predictor and refines

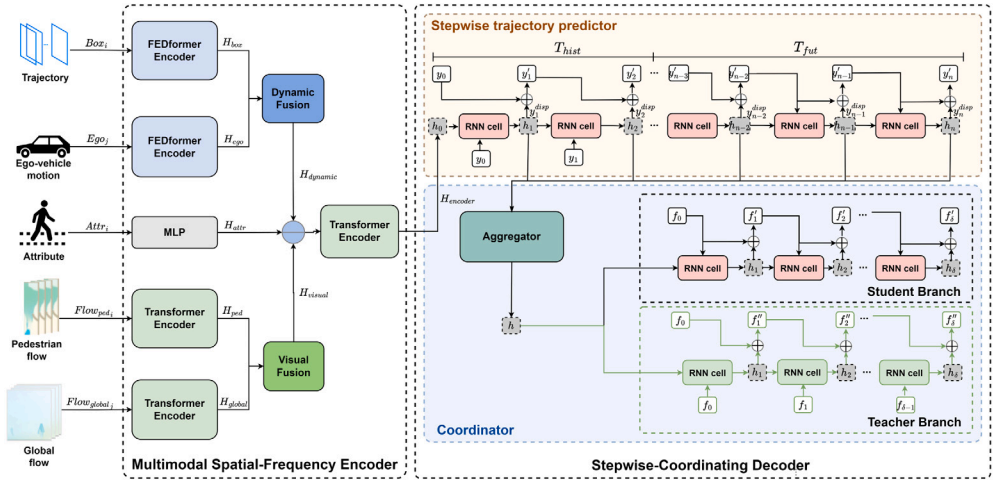


Fig. 2. The overall structure of MSCN. The MSCN structure mainly includes two parts: a multimodal spatial-frequency encoder and a stepwise-coordinating decoder. Green operations occur during training only.

the prediction result through a knowledge distillation process, which avoids error accumulation and obtains a more reasonable trajectory.

3. Method

In this section, we describe the details of our method which includes the multimodal spatial-frequency encoder, the stepwise-coordinating decoder, and the loss functions.

3.1. Problem formulation

Assuming that pedestrian i appears in the field of view from a driving vehicle j, our goal is to optimize a generative model to predict the future trajectory of pedestrian i in future time $T_{fut} = \{t^* + 1, t^* + 2, \dots, t^* + \delta\}$ where δ and t^* are the number of time steps in the future and in historical period, respectively. The given information includes the pedestrian attribute vector $Attr_i$ and the historical information $\{Ego_j^t, Box_i^t, Flow_{global}{}_j^t, Flow_{ped}{}_i^t | t \in T_{hist}\}$, where the symbols Ego, Box, $Flow_{global}$, and $Flow_{ped}$ represent the vehicle state, pedestrian trajectory, global optical flow information, and pedestrian optical flow information respectively. To be specific, $T_{hist} = \{t^* - \tau + 1, t^* - \tau + 2, \dots, t^*\}$ denotes the historical time steps where τ represents the number of accessible historical time steps. Ego_j^t represents the ego-motion information of the driving vehicle. $Box_i^t = \{x1_i^t, y1_i^t, x2_i^t, y2_i^t\}$ is the coordinates of the upper left and lower right corners of the bounding box. $Flow_{global}{}_j^t \in R^{(\tau-1)\times 2\times h_{glb}\times w_{glb}}$ and $Flow_{ped}{}_i^t \in R^{(\tau-1)\times 2\times h_{ped}\times w_{ped}}$ are the motion information of the optical flow, where h_{glb}, w_{glb}, h_{ped} and w_{ped} are height and width of two kinds of optical flow. The global optical flow and pedestrian optical flow are divided into M and N blocks as well as spatially averaged and pooled.

3.2. Framework overview

Fig. 2 shows the overall framework of our method, including a multimodal spatial-frequency encoder and a stepwise-coordinating decoder. The encoder integrates multimodal information including observed trajectory, ego-vehicle speed, optical flows, and pedestrian attributes. Owing to the periodicity of pedestrian motion, a frequency-domain enhanced Transformer processes trajectory and ego-vehicle speed to produce a hybrid representation. The optical flow representations of ego-vehicle and pedestrians are extracted through a spatial-domain Transformer. The attribute feature is extracted by an MLP. Then, the representations of different modalities are fused hierarchically to obtain a feature vector as an initial hidden state of the decoder. The decoder is a two-stage structure including a stepwise trajectory predictor which generates a coarse sequence result and a coordinator which aggregates the corresponding representation of the coarse result and outputs a refined sequence through a knowledge distillation process.

3.3. Multimodal spatial-frequency encoder

The proposed multimodal spatial-frequency encoder captures the pedestrian motion pattern as a latent vector by integrating multiple modalities with Transformer-based architecture. As shown in Fig. 3, it mainly consists of frequency-domain enhanced Transformer, spatial-domain Transformer, and hierarchical multimodal fusion stages.

Frequency-domain enhanced Transformer. Since pedestrian motion is composed of periodic gait cycles, we can take the pedestrian trajectory as a periodic time series. Therefore, we propose to extract the features of the whole trajectory from the perspective of frequency-domain combining with Transformer to capture both the global profile of the trajectory and more detailed structures. As shown in Fig. 3(a), a FEDformer encoder block [38] is composed of a Frequency Enhanced Block (FEB) and a feedforward network which is both followed by a Mixture Of Experts Decomposition block (MOEDecomp).

Since the pedestrian trajectory is obtained relative to vehicle driving, it is necessary to consider both vehicle speed Ego_j and pedestrian trajectory Box_i. To capture the overall characteristic of pedestrian trajectory, we utilize a Transformer-based structure which is incorporated with a seasonal-trend decomposition approach and Fourier analysis, i.e., FEDformer encoder [38], to encode vehicle speed and pedestrian trajectory.

Specifically, Box_i or Ego_j is coded by a linear network and then added with a positional code to get the embedding X_{box} or X_{ego}. Then, a Frequency Enhanced Block (FEB) with Fourier transform is used to capture important structures of $X_{box/ego}$ through frequency domain mapping. To be specific, the embedding $X_{box/ego}$ is first linearly projected, getting $Q_{box/ego}$. Then, $Q_{box/ego}$ is converted from the time domain to the frequency domain with Fourier transform. In frequency domain, we use a select operator to randomly select modes as

$$\tilde{Q}_{box/ego} = \text{Select}(\mathcal{F}(Q_{box/ego})), \tag{1}$$

where $\mathcal{F}$ denotes the Fourier transform. Afterward, the FEB performs a reconstruction stage as

$$\text{FEB}(Q_{box/ego}) = \mathcal{F}^{-1}(\text{Padding}(\tilde{Q}_{box/ego} \odot R)), \tag{2}$$

where R is a randomly initialized parameterized kernel while $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. $\tilde{Q} \odot R$ is defined as $\tilde{Q} \odot R =$

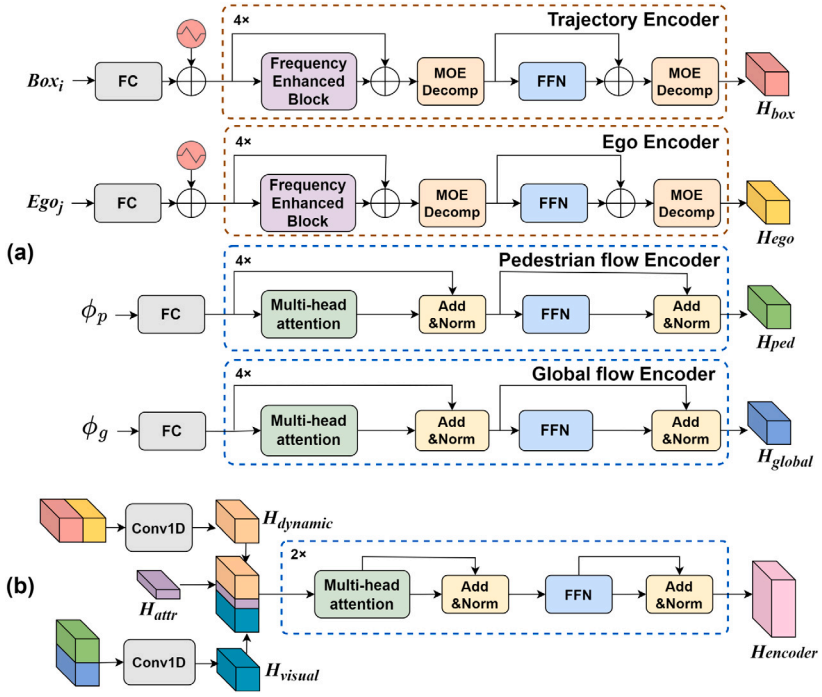


Fig. 3. Details of the Multimodal Spatial-Frequency Encoder. (a) is the encoding stage of each modality. (b) is the hierarchical fusion stage.

$\sum_{d_i=0}^{D} \tilde{Q}_{m,d_i} \cdot R_{d_i,d_o,m}$ where d_i and d_o are the input and output channels respectively. The result of $\tilde{Q}_{box/ego} \odot R$ is padded with zero and converted back into the time domain through inverse Fourier transform.

Then, the trend information is decoupled by the Mixture Of Experts Decomposition block (MOEDecomp) as follows:

$$X_{\text{trend}} = \text{Softmax}(L(x)) * (F(x)), \tag{3}$$

where $F(.)$ represents an average pooling filter, $Softmax(L(x))$ is the weights for trends mixing. The features extracted by FEDformer from pedestrian trajectories Box_i and vehicle speeds Ego_j are $H_{box} \in R^{L\times d}$ and $H_{ego} \in R^{L\times d}$ respectively.

Spatial-domain Transformer.

Apart from trajectory and ego-motion modalities, we also consider information in the spatial domain. We utilize optical flow between frames as trajectory and ego-motion data rather than image data. This choice is motivated by the fact that optical flow represents the instantaneous velocity of moving objects on the imaging plane, where the central region and target boxes denote the motion of the ego-vehicle and pedestrians, respectively [6].

For optical flow representation, we first divide the global optical flow $Flow_{global}$ and pedestrian optical flow $Flow_{ped}$ into N and M patches along the spatial dimension, resulting in spatial motion representations of ego-vehicle ϕ_g and target pedestrian ϕ_p. Then, as shown in Fig. 3(a), ϕ_g or ϕ_p are projected into a C-dimensional space through a fully connected layer. Afterward, we adopt multi-head attention in Transformer blocks to jointly focus on information from different representation subspaces at different locations and increase feature representation ability through feed-forward network [22]. Finally, we obtain the global optical flow spatial dynamic feature representation $H_{global} \in \mathbb{R}^{N\times d}$ and the pedestrian optical flow spatial dynamic feature representation $H_{ped} \in \mathbb{R}^{M\times d}$.

Hierarchical multimodal fusion.

Apart from obtaining representations of trajectory, ego-motion and optical flows, for pedestrian attributes, we apply an MLP to obtain the attribute feature H_{attr}. Then, as shown in Fig. 3(b), we perform hierarchical fusion to integrate multiple modalities at distinct stages to more effectively capture the highly dynamic motion information. Specifically, the first level of fusion is as follows:

$$H_{visual} = Conv1D(cat([H_{global}, H_{ped}])), \tag{4}$$

$$H_{dynamic} = Conv1D(cat([H_{ego}, H_{box}])), \tag{5}$$

The matrices $H_{visual} \in \mathbb{R}^{K\times d}$ and $H_{dynamic} \in \mathbb{R}^{L\times d}$ represent the fused visual and motion features, respectively, where K and L denote the number of visual and motion features that are combined, respectively. It is noteworthy that temporal modality features (H_{ego} and H_{box}) are aligned along the time dimension to emphasize their temporal characteristics, whereas optical flow modality features (H_{ped} and H_{global}) are concatenated along the spatial dimension to highlight their spatial characteristics.

While there is a strong correlation among vehicle speed, pedestrian trajectory sequences, and optical flow features, direct concatenation operation may cause such relationships to be ignored. Therefore, in the second level fusion process, we employ a multi-head attention mechanism to capture the inherent correlations among temporal, spatial, and attribute features, as depicted in the following equation:

$$H_{encoder} = TF(cat([H_{dynamic}, H_{attr}, H_{visual}])), \tag{6}$$

where $H_{encoder} \in R^{(K+L+1)\times d}$, and TF is a Transformer encoder with multi-head attention to finally fuse multimodal information.

3.4. Stepwise-coordinating decoder

After the encoder transforms multimodal historical information into a vector, the proposed stepwise-coordinating decoder generates the final prediction using a two-stage process. As shown in Fig. 2, the stepwise trajectory predictor first generates a coarse result. Then, the coordinator aggregates the corresponding representation of the coarse result and generates refined results through a knowledge distillation process. Although Transformer has been demonstrated in high efficiency for multimodality data feature extraction and fusion, the heavy computational cost of the autoregressive Transformer decoder is a key issue to prevent it in real-time applications [34,35], especially the pedestrian

trajectory prediction task. Therefore, inspired by streaming/real-time speech recognition methods [36,37] using Transformer as encoder and RNN as predictor, we use RNN as the autoregressive decoder due to the consideration of speed and memory.

Stepwise trajectory predictor.

The initial hidden state h_0 of RNN is obtained by flattening the encoded feature $H_{encoder}$. The stepwise trajectory predictor will generate a coarse result and its corresponding hidden states $[h_1 \dots h_n]$, where $n = \delta + \tau - 1$. Specifically, the stepwise trajectory predictor consists of a historical trajectory reconstruction stage and a future trajectory prediction stage. In each time step t of RNN, we estimate the *motion displacement* y_t^{disp} of the pedestrian trajectory. The decoding process is formalized as

$$Box_i^t = \mathrm{RNN}(Box_i^{t-1}, h_{t-1}) + Box_i^{t-1}; t \in [T_{hist}, T_{fut}], \tag{7}$$

where the RNN cell estimates box coordinates' motion displacement $y_t^{disp} = Box_i^t - Box_i^{t-1}$.

In the training phase, as shown in Fig. 2, for the historical trajectory period, we take the ground-truth label of the last step y_{t-1} as the input of the decoder. For future trajectory periods, the input of the decoder is the prediction result y'_{t-1} of the last step from itself.

Coordinator. Inspired by the fact that pedestrian motion is mostly driven by some intention or goal as global information along with the fine-grained step-by-step process, we design a coordinator consisting of an aggregator and a knowledge distillation strategy to take advantage of global information. Specifically, the aggregator receives all the hidden states generated by the stepwise trajectory $[h_1 \dots h_n]$ including historical reconstruction and future prediction, then coordinates the global information as

$$h = Hardswish\left(\omega \times [h_1 \dots h_n] + \beta\right), \tag{8}$$

where ω and β are learnable parameters while the activation function $Hardswish$ [39] is used to enhance its expression ability.

Then, in order to make the generation process more comprehensive, we introduce a knowledge distillation strategy in the coordinator as shown in the right part of Fig. 2. It is divided into two branches, including a teacher branch that takes the ground-truth label f_{t-1} as input in each step t of the RNN decoder and a student branch that takes the prediction result of the last step f'_{t-1} as the input of the RNN decoder. The teacher branch and the student branch both take the aggregated representation h as the initial hidden state in order to incorporate global information. Meanwhile, the teacher branch corrects the step-by-step prediction of the student branch in time, guiding the step-by-step prediction process to capture short-term motion information.

3.5. Loss functions

The proposed model is trained end-to-end using multiple losses.

Historical trajectory reconstruction loss. For historical trajectory reconstruction, we utilize the mean square error (MSE) loss,

$$L_{hist} = \left\| P - \hat{P} \right\|, \tag{9}$$

where P and $\hat{P}$ represent the ground-truth historical trajectory and the reconstructed trajectory respectively.

Future trajectory prediction loss. For future trajectory prediction, inspired by the exponential L2 loss [40], we propose a loss function to emphasize the importance of the prediction result in early steps,

$$L_{fut} = \left\| F - \check{F} \right\| \times e^{\frac{(\delta - T_{fut})}{\gamma}}, \tag{10}$$

where F and $\check{F}$ are the ground-truth future trajectory and the predicted trajectory of the student branch respectively, while γ is a hyperparameter used to control the decreasing trend of importance in the sequence.

Distillation loss. For knowledge distillation training, we also adopt the exponential L2 loss,

$$L_{dist} = \left\| \hat{F} - \check{F} \right\| \times e^{\frac{(\delta - T_{fut})}{\gamma}}, \tag{11}$$

where $\hat{F}$ is the output of teacher branch.

Then, the final loss is given by

$$L = L_{hist} + L_{fut} + L_{dist}. \tag{12}$$

4. Experiments

4.1. Datasets and metrics

In this paper, we mainly adopt two widely used public datasets to verify the performance of our proposed method, i.e., JAAD [19] and PIE [1]. JAAD and PIE are sampled from 2200 and 1842 pedestrians (30 Hz) respectively and provide a large number of trajectories in the first-person perspective traffic environment. Both datasets provide image information and annotations of pedestrian crossing intentions. For ego-motion, PIE provides detailed information on the velocity, while JAAD only provides annotations of vehicle activities (e.g., moving slowly, stopping, accelerating), which serve as a proxy for ego-motion. For a fair comparison, we adopt the same ego vehicle sensor information as in [6], such as vehicle speed and train–test split. We apply 15 frames (0.5 s) for historical scenes and 45 frames (1.5 s) for future scenes.

There are three evaluation metrics widely used, including MSE (mean squared error), C_{MSE} (box center mean squared error), and CF_{MSE} (box center final mean squared error). All predictions are given in pixels with lower errors indicating better performance.

4.2. Implementation details

We follow MTN [6] for all dataset preparation to have a fair comparison. For example, the hyperparameters associated with representing optical flow data, such as the height h_{glb}, width w_{glb}, **M** and **P**, mirror those settings in MTN. As for training-related hyperparameters, including learning rate, these are selected by following common deep-learning methods' practice. We have accordingly detailed these descriptions.

For the JAAD and PIE datasets, we followed the data preparation procedures outlined in MTN [6] to ensure consistency and fairness in comparison. For each sample, the optical flow data comes from [6], where the height h_{glb} and width w_{glb} of the global optical flow information are set to be 160 pixels, while the number of patches M and P are 64 and 9 respectively, as settings in MTN. We set the observation sequence length τ to 15 frames (0.5 s), and the prediction sequence length δ to 45 frames (1.5 s) according to the PIE dataset's setting. For this task, we use attribute information including pedestrian ID, age (represented as child (0), young (1), adult (2) or senior (3)), and gender (represented as n/a (0), female (1) or male (2)). Our coordinator coordinates and examines the hidden state and the gate state, respectively, when it compiles the global data. Finally, the hidden state and the gate state are transmitted to the next stage for subsequent decoding. The sequence of trajectories is randomly flipped with a probability of 0.1 during training. We set the random seed to 42, and the total number of training epochs to 80. The hyperparameter γ of the exponential L2 loss is set to 10. The batch size is set to 32. Following common deep-learning methods' practice, the Adam optimizer [42] is used, with the learning rate initialized to 0.001. All experiments were performed on a single GTX 3090.

Fig. 4. Qualitative comparison results. The white color indicates the observed trajectory, and the future trajectory display corresponds to ground truth, SGNet-ED [17], Context-Aware [41], MTN [6], and MSCN(ours).

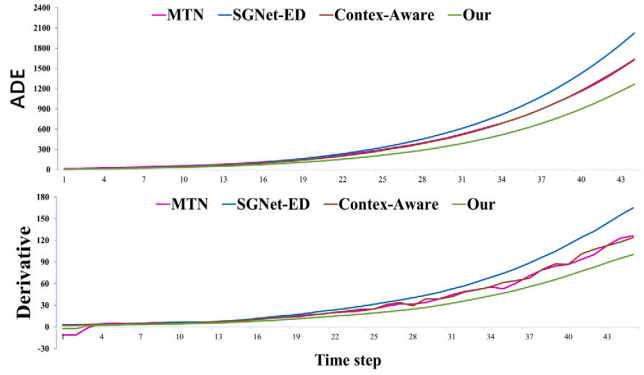


Fig. 5. The variation of box error over time. (top) is the box error value changing with time, and (bottom) is the box error derivative changing with time.

Table 1
Quantitative comparison on PIE dataset and JAAD datasets.

Method	PIE			JAAD		
	MSE (0.5 s/1.0 s/1.5 s)	C_{MSE} (1.5 s)	CF_{MSE} (1.5 s)	MSE (0.5 s/1.0 s/1.5 s)	C_{MSE} (1.5 s)	CF_{MSE} (1.5 s)
B-LSTM [4]	101/296/855	811	3259	159/539/1535	1447	5615
TDP-MOF [43]	-/-/665	566	2373	-/-/1158	1014	4143
PIE_{full} [1]	-/-/559	520	2162	-/-/-	–	–
PIE_{traj} [1]	58/200/636	596	2477	110/399/1248	1183	4780
MTN [6]	57/161/444	414	1627	95/325/1005	951	4010
SGNet-ED [17]	34/133/442	413	1761	82/328/1049	996	4076
Context-Aware [41]	33/127/398	372	1519	78/324/1020	974	3937
Ours	**30/110/334**	**309**	**1269**	**72/289/853**	**808**	**3209**

4.3. Comparisons with state-of-the-art methods

We mainly compared B-LSTM [4], TDP-MOF [43], PIE_{full} [1], PIE_{traj} [1], MTN [6], SGNet-ED [17], and Context-Aware [41] with our proposed method. For the sake of fairness, for TDP-MOF, the experimental results were taken from [6], in which the input and output length of TDP-MOF were modified. Table 1 shows the deterministic prediction results on both PIE and JAAD datasets, where our method achieves significant performance improvement over existing methods. For example, considering the metric CF_{MSE} (1.5 s), which evaluates long-term prediction accuracy, our method improves 16.5% and 18.5% over the best baseline [41] on the JAAD and PIE datasets, respectively. In addition, our method also exhibits strong trajectory prediction ability with a large reduction in the metrics MSE (1.5 s) and C_{MSE} (1.5 s).

Table 2
Time Complexity Comparison.

Method	Para. (M)	FLOPs (M)	Inference time (ms)
B-LSTM	0.86	4.1	177.1
TDP-MOF	11.27	2413.63	17.4
PIE_{full}	2.52	54.1	1328.4
PIE_{traj}	0.62*	**2.1**	254.1
MTN	**0.13**	3.3*	7.8*
SGNet-ED	4.36	1559.1	310.7
Context-Aware	2.97	70.1	**3.3**
Ours	0.64	40.4	10.6

As illustrated in Fig. 5 (top), we conducted a frame-by-frame analysis of the model's predicted positions and compared them with other models such as MTN [6], SGNet-ED [17], and Context-Aware [41]. Additionally, in Fig. 5 (bottom), we presented the derivative of prediction errors over time. It is notable that as the prediction horizon extends, particularly beyond 20 frames, the derivative of prediction errors for our model gradually diverges from those of other models. SGNet-ED, focusing solely on a single modality, lacks environmental analysis, leading to poorer long-term prediction robustness. While MTN and Context-Aware are designed in Transformer-based encoder–decoder architecture, their prediction error derivatives do not exhibit smooth curves, indicative of suboptimal architectural designs. In contrast, our approach incorporates various modalities and boasts a more interpretable architectural design, thereby enhancing model performance and rendering it more robust for long-term prediction, effectively attenuating the exponential growth of prediction errors over time and thus ensuring greater reliability.

Time complexity plays a critical role in pedestrian trajectory prediction. Through time complexity analysis with prominent models in this field (refer to Table 2), we ascertain that our model demonstrates moderate levels of parameter count, computational workload, and inference time on identical hardware. The objective of the pedestrian trajectory prediction task is to forecast the pedestrian trajectory for the subsequent 1.5 s (45 frames) based on a 0.5 s input (15 frames) [1,19]. Compared to the other three SOTA methods, despite higher complexity than MTN, our method still satisfies real-time requirements while significantly enhancing prediction accuracy.

Fig. 6. Visualization of the efficacy of Multiple Modalities. The white color indicates the observed trajectory, while the future trajectory display corresponds to ground truth, using all modal information, without using trajectory, without using vehicle speed, without using pedestrian attributes, without using global optical flow, without using pedestrian optical flow. (Please zoom in for better viewing.).

Fig. 7. Visualization of the efficacy of Frequency-Domain Enhanced Transformer. The white color indicates the observed trajectory, while the future trajectory display corresponds to ground truth, using FEDformer, without using FEDformer. (Please zoom in for better viewing.).

Table 3
Ablation study on multimodal information.

Box	Ego	$Attr$	$Flow_{global}$	$Flow_{ped}$	MSE	C_{MSE}	CF_{MSE}
–	√	√	√	√	352	327	1353
√	–	√	√	√	436	409	1618
√	√	–	√	√	336	311	1283
√	√	√	–	√	360	335	1371
√	√	√	√	–	356	330	1352
√	√	√	√	√	**334**	**309**	**1269**

Table 4
Ablation study on stepwise-coordinating strategy.

Stepwise		Coordinator		MSE	C_{MSE}	CF_{MSE}
DE	HTR	Aggregator	Distill			
–	√	√	√	347	321	1302
√	–	√	√	399	372	1472
√	√	–	√	359	332	1385
√	√	Weighted Sum	√	395	368	1529
√	√	√	–	354	328	1347
√	√	√	√	**334**	**309**	**1269**

Table 5
Ablation study on frequency domain enhancement.

Type	DE	MSE	C_{MSE}	CF_{MSE}
Transformer	–	357	330	1338
FEDformer	–	347	321	1302
Autoformer	√	380	353	1468
Transformer	√	356	330	1335
FEDformer	√	**334**	**309**	**1269**

Table 6
Ablation study on historical trajectory reconstruction length.

HTR Length	MSE	C_{MSE}	CF_{MSE}
0	399	372	1472
5	384	357	1421
10	350	324	1321
14	**334**	**309**	**1269**

4.4. Visualizations

The visualization of trajectory prediction results is shown in Fig. 4. We can find that our prediction results are significantly closer to the ground truth trajectory compared with other methods regardless of simple scenarios or scenarios with complicated relations, which is consistent with our quantitative evaluations as reported in Table 1. As shown in Fig. 4(a), our approach can provide a precise forecast of the pedestrian's intent to cross in front of the ego-vehicle, which avoids a collision. In Fig. 4(b), our method predicts that the pedestrian walks along the sidewalk, demonstrating the ability to understand global context information. Furthermore, it is observed that in challenging scenarios, such as when pedestrian trajectories exhibit significant spatial variations (e.g., as shown in subfigures a, g, and h of Fig. 4), our model demonstrates pronounced advantages. This underscores the adaptability and robustness of our model in handling complex scenarios.

4.5. Ablation study

Multimodal information. In this part, we conduct an ablation experiment to assess the impact of multiple-modal information. As demonstrated in Table 3, each modality serves a distinct purpose, and effective fusion of information maximizes their respective contributions. Particularly, as shown in the second row of Table 3, the Ego modality, i.e., car movement, was found to significantly influence the pedestrian's trajectory, as it determines the alteration of the pedestrian's reference coordinate system.

We also give result visualizations of our method with ablated modality as depicted in Fig. 6. The observations revealed that the absence of specific modal information for different pedestrians has a certain effect on the prediction accuracy of their trajectories. Nonetheless, the overall prediction accuracy improved significantly when all modal information was available.

Stepwise-coordinating. In order to verify the effectiveness of the stepwise-coordinating strategy, we conduct several experiments with different settings, including no displacement estimation (DE), no historical trajectory prediction (HTR), no aggregator, replacing the aggregator with a weighted summation, and no knowledge distillation. When

replacing the aggregator, we use a weighted summation instead:

$$h = \sum_{k=1}^{n} Softmax(\epsilon) \times h_k, \tag{13}$$

where ϵ is a learnable parameter. The experimental results are shown in Table 4. The results verify that our full model can achieve the best performance and all proposed components contribute to the full model. It should be noted that without an aggregator, a weighted summation operation can even bring counterproductive results. In addition, the historical trajectory prediction strategy brings a huge performance gain, showing its importance in trajectory prediction task. Overall, the results demonstrate that the stepwise predictor and the coordinator are complementary to each other, verifying that our MSCN can sufficiently leverage the underlying human movement properties with a special encoder–decoder design.

Frequency-domain enhancement. In order to investigate the effectiveness of frequency-domain enhancement in the encoder, we conducted a comparative ablation study by replacing FEDformer [38] with different Transformers, including vanilla Transformer [22] and Autoformer [44], while keeping other components fixed. In order to verify that pedestrian motion periodicity is effectively encoded and decoded by our model, we also examine the impact of displacement estimation strategy on different Transformers. The results are shown in Table 5, where the performance of FEDformer with displacement estimation has the best performance. We can observe whether using displacement estimation has a great influence on the FEDformer and has almost no effect on the vanilla Transformer, which further validates the effectiveness of the proposed network architecture.

We also give result visualizations of our method with different Transformers as shown in Fig. 7. The visualizations revealed that FEDformer is more able to deal with the frequency patterns of pedestrian movements, leading to more accurate predictions of their future trajectories. In contrast, methods not utilizing FEDformer may lead to overestimation (e.g., case b) or underestimation (e.g., cases a and c) of pedestrian movements. These findings demonstrate the advantages of FEDformer in enhancing prediction accuracy for pedestrian trajectory prediction.

Historical trajectory reconstruction length. Historical sequence prediction is commonly used in sequence generation tasks, such as text generation, to enhance the feature representation ability. We found it is also useful in our task as an ablation study in Table 4 and Table 6. As shown in Table 6, increasing the length of historical trajectories further improves the model performance.

Knowledge distillation. We also conduct an additional experiment to investigate the impact of the distillation strategy on coordinator learning. Here we add a teacher network loss to minimize the difference between the output of the teacher network and the ground-truth trajectory:

$$L_{teacher} = \left\| F - \hat{F} \right\| \times e^{\frac{(\delta - T_{fut})}{\gamma}}. \tag{14}$$

Then we set the ablation experiment including (1) no distillation strategy; (2) distillation strategy with both $L_{distill}$ and $L_{teacher}$; (3) original distillation strategy $L_{distill}$ in our model. We report the performance of both student and teacher branches with different settings as shown in Table 7. It should be noted that the teacher branch takes the ground-truth labels as input in each step of the RNN. From the results, we can see that with or without $L_{teacher}$, our model (student branch) can still produce accurate prediction results. Although adding $L_{teacher}$ can bring significant performance gain for the teacher branch, we notice a little drop in performance of the student branch, which indicates that a softer regularization between teacher and student networks may bring better performance as [45].

Table 7
Ablation study on knowledge distillation strategy for coordinator learning.

Branch	$L_{distill}$	$L_{teacher}$	MSE	C_{MSE}	CF_{MSE}
Student	–	–	354	328	1347
Teacher	√	√	**72**	**65**	**241**
Student	√	√	338	314	1287
Teacher	√	–	325	302	1242
Student	√	–	**334**	**309**	**1269**

5. Conclusion

In this paper, we present a Multimodal Stepwise-Coordinating Network, namely MSCN, for first-person view pedestrian trajectory prediction, which aims to make sufficient use of human motion properties. We design a multimodal spatial-frequency encoder to effectively model the periodicity of pedestrian motion and spatial information. In addition, we introduce a stepwise-coordinating decoder structure to effectively capture both short-term and long-term motion interaction. We demonstrate that our proposed method greatly outperforms existing methods on first-person view trajectory prediction tasks using publicly available benchmarks. By undertaking ablation studies, we are able to further demonstrate the overall contributions of our proposed modules. In future work, it will be interesting to apply our method to additional computer vision and robotics problems, like action prediction, interaction prediction, and human motion simulation.

One of our method's limitations is the lack of leveraging contextual information from the environment, like other pedestrians' and vehicles' states or the semantic segmentation map. We believe that improving upon these limitations in future work will result in a better accuracy rate for predicting pedestrian trajectories.

CRediT authorship contribution statement

Yijun Wang: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Zekun Guo:** Writing – review & editing, Validation, Methodology, Formal analysis. **Chang Xu:** Validation, Investigation, Data curation. **Jianxin Lin:** Supervision, Resources, Project administration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jianxin Lin reports financial support was provided by National Natural Science Foundation of China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments. This research was partially supported by grants from the National Natural Science Foundation of China (Grants No. 62202158, 62206089), the Natural Science Foundation of Hunan Province, China (Grants No. 2023JJ40167, 2023JJ40178), the science and technology innovation Program of Hunan Province, China (Grants No. 2023RC3098) and the Fundamental Research Funds for the Central Universities, China (Grants HNU: 531118010668, 531118010786).

References

[1] A. Rasouli, I. Kotseruba, T. Kunic, J.K. Tsotsos, Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6262–6271.

[2] T. Yagi, K. Mangalam, R. Yonetani, Y. Sato, Future person localization in first-person videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7593–7602.

[3] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, X. Du, Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation, IEEE Robot. Autom. Lett. 6 (2) (2021) 1463–1470.

[4] A. Bhattacharyya, M. Fritz, B. Schiele, Long-term on-board prediction of people in traffic scenes under uncertainty, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4194–4202.

[5] Z. Su, G. Huang, S. Zhang, W. Hua, Crossmodal transformer based generative framework for pedestrian trajectory prediction, in: 2022 International Conference on Robotics and Automation, ICRA, IEEE, 2022, pp. 2337–2343.

[6] Z. Yin, R. Liu, Z. Xiong, Z. Yuan, Multimodal transformer networks for pedestrian trajectory prediction., in: IJCAI, 2021, pp. 1259–1265.

[7] A. Rasouli, M. Rohani, J. Luo, Bifold and semantic reasoning for pedestrian behavior prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15600–15610.

[8] Z. Sui, Y. Zhou, X. Zhao, A. Chen, Y. Ni, Joint intention and trajectory prediction based on transformer, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2021, pp. 7082–7088.

[9] Y. Yao, M. Xu, C. Choi, D.J. Crandall, E.M. Atkins, B. Dariush, Egocentric vision-based future vehicle localization for intelligent driving assistance systems, in: 2019 International Conference on Robotics and Automation, ICRA, IEEE, 2019, pp. 9711–9717.

[10] S. Malla, B. Dariush, C. Choi, Titan: Future forecast using action priors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11186–11196.

[11] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese, Social lstm: Human trajectory prediction in crowded spaces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 961–971.

[12] A. Kharb, V. Saini, Y. Jain, S. Dhiman, A review of gait cycle and its parameters, IJCEM Int. J. Comput. Eng. Manag. 13 (2011) 78–83.

[13] K. Jordan, J.H. Challis, K.M. Newell, Walking speed influences on gait cycle variability, Gait Posture 26 (1) (2007) 128–134.

[14] L. Ren, R.K. Jones, D. Howard, Predictive modelling of human walking over a complete gait cycle, J. Biomech. 40 (7) (2007) 1567–1574.

[15] Y. Xiang, J.S. Arora, S. Rahmatalla, K. Abdel-Malek, Optimization-based dynamic human walking prediction: One step formulation, Internat. J. Numer. Methods Engrg. 79 (6) (2009) 667–695.

[16] A.E. Martin, J.P. Schmiedeler, Predicting human walking gaits with a simple planar model, J. Biomech. 47 (6) (2014) 1416–1421.

[17] C. Wang, Y. Wang, M. Xu, D.J. Crandall, Stepwise goal-driven networks for trajectory prediction, IEEE Robot. Autom. Lett. 7 (2) (2022) 2716–2723.

[18] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, A. Gaidon, It is not the journey but the destination: Endpoint conditioned trajectory prediction, in: European Conference on Computer Vision, Springer, 2020, pp. 759–776.

[19] A. Rasouli, I. Kotseruba, J.K. Tsotsos, Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 206–213.

[20] R. Quan, L. Zhu, Y. Wu, Y. Yang, Holistic LSTM for pedestrian trajectory prediction, IEEE Trans. Image Process. 30 (2021) 3229–3239.

[21] D. Yang, H. Zhang, E. Yurtsever, K.A. Redmill, Ü. Özgüner, Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention, IEEE Trans. Intell. Veh. 7 (2) (2022) 221–230.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, 2017.

[23] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[24] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, 2014, arXiv preprint arXiv:1406.1078.

[25] Y. Liu, J. Zhang, L. Fang, Q. Jiang, B. Zhou, Multimodal motion prediction with stacked transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7577–7586.

[26] L. Fang, Q. Jiang, J. Shi, B. Zhou, Tpnet: Trajectory proposal network for motion prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6797–6806.

[27] Y. Pang, X. Zhao, J. Hu, H. Yan, Y. Liu, Bayesian spatio-temporal graph transformer network (b-star) for multi-aircraft trajectory prediction, Knowl.-Based Syst. 249 (2022) 108998.

[28] H. Zhou, X. Yang, M. Fan, H. Huang, D. Ren, H. Xia, Static-dynamic global graph representation for pedestrian trajectory prediction, Knowl.-Based Syst. 277 (2023) 110775.

[29] M. Wu, H. Ling, N. Bi, S. Gao, Q. Hu, H. Sheng, J. Yu, Visual tracking with multiview trajectory prediction, IEEE Trans. Image Process. 29 (2020) 8355–8367.

[30] B. Yang, F. Fan, R. Ni, J. Li, L. Kiong, X. Liu, Continual learning-based trajectory prediction with memory augmented networks, Knowl.-Based Syst. 258 (2022) 110022.

[31] Z. Li, H. Liang, H. Wang, X. Zheng, J. Wang, P. Zhou, A multi-modal vehicle trajectory prediction framework via conditional diffusion model: A coarse-to-fine approach, Knowl.-Based Syst. 280 (2023) 110990.

[32] Y. Yuan, X. Weng, Y. Ou, K.M. Kitani, Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9813–9823.

[33] K.-I. Na, U.-H. Kim, J.-H. Kim, SPU-BERT: Faster human multi-trajectory prediction from socio-physical understanding of BERT, Knowl.-Based Syst. 274 (2023) 110637.

[34] M. Ghazvininejad, O. Levy, Y. Liu, L. Zettlemoyer, Mask-predict: Parallel decoding of conditional masked language models, 2019, arXiv preprint arXiv:1904.09324.

[35] X. Song, Z. Wu, Y. Huang, C. Weng, D. Su, H. Meng, Non-autoregressive transformer asr with ctc-enhanced decoder input, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 5894–5898.

[36] X. Chen, Y. Wu, Z. Wang, S. Liu, J. Li, Developing real-time streaming transformer transducer for speech recognition on large-scale dataset, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 5904–5908.

[37] C.-F. Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, M.L. Seltzer, Transformer-transducer: End-to-end speech recognition with self-attention, 2019, arXiv preprint arXiv:1910.12977.

[38] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, R. Jin, Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting, 2022, arXiv preprint arXiv:2201.12740.

[39] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1314–1324.

[40] J. Sun, Q. Jiang, C. Lu, Recursive social behavior graph for trajectory prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 660–669.

[41] H. Damirchi, M. Greenspan, A. Etemad, Context-aware pedestrian trajectory prediction with multimodal transformer, in: 2023 IEEE International Conference on Image Processing, ICIP, IEEE, 2023, pp. 2535–2539.

[42] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[43] O. Styles, A. Ross, V. Sanchez, Forecasting pedestrian trajectory with machine-annotated training data, in: 2019 IEEE Intelligent Vehicles Symposium, IV, IEEE, 2019, pp. 716–721.

[44] H. Wu, J. Xu, J. Wang, M. Long, Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, Adv. Neural Inf. Process. Syst. 34 (2021) 22419–22430.

[45] J. Gou, B. Yu, S.J. Maybank, D. Tao, Knowledge distillation: A survey, Int. J. Comput. Vis. 129 (6) (2021) 1789–1819.