IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 June 2024, accepted 7 July 2024, date of publication 11 July 2024, date of current version 22 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3426958

RESEARCH ARTICLE

METF: Modeling Macro-Micro Human Intention by Multi-Encoder Transformer for Human Trajectory Prediction

XINCHENG HU[1], BO YANG[1,2], JIXING YANG[1], AND TENG ZHANG[1]

[1]College of Computer Science, South-Central Minzu University, Wuhan 430074, China

[2]Key Laboratory of Cyber-Physical Fusion Intelligent Computing (South-Central Minzu University), State Ethnic Affairs Commission, Wuhan 430074, China

Corresponding author: Bo Yang (yangbo@mail.scuec.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 72104254, and in part by Hubei Provincial Natural Science Foundation of China under Grant 2022CFB469.

ABSTRACT Human trajectory prediction tasks find applications in many fields, like autonomous driving and social robots. The main challenge arises from the fact that pedestrians, while walking, consider their own route and constantly account for their spatial and temporal interactions with other pedestrians to avoid collisions. However, most existing state-of-the-art models either overlook the balance between a pedestrian's own path and their interactions with others, or they focus solely on either one of these aspects. We posit that an effective pedestrian trajectory prediction should incorporate both macro and micro perspectives. In this paper, We propose a Multi-Encoder-TransFormer-network (METF), which can balancing the information between micro and macro. First, we propose a multi-encoder architecture to simultaneously encode macroscopic and microscopic information and allocate different degrees of importance for macroscopic and microscopic information. Then, we introduce a graph attention mechanism to capture the interactions between pedestrians at each moment, and also introduce an attention module to learn the time dependence of the interaction in different moments within a long time range. We also redesigned the input, output and computational methods of the transformer decoder for the trajectory prediction problem, and reduced the computational cost while maintaining the accuracy. Upon comparing with a wide range of methods, we found that METF achieved superior performance on two publicly available datasets (ETH and UCY), producing trajectories that align more closely with pedestrian social walking patterns. Ablation experiments illustrate the effectiveness of the designs for various parts in the METF.

INDEX TERMS Trajectory prediction, graph attention networks, social interactions.

I. INTRODUCTION

This section introduces the importance and challenges of human trajectory prediction, as well as the limitations of existing methods in balancing the pedestrian's own path and interactions with others.

Human trajectory prediction pedestrian trajectory prediction is a complex but highly valuable task, it has received considerable attention in the fields of robotics [1], [2], autonomous driving [3], [4], and computer vision [5], [6].

In trajectory prediction, modeling the complex and diverse interaction between humans is crucial and challenging. Early work used handcrafted energy functions, but they failed to establish crowd interaction between pedestrians in crowded spaces. In recent years, some methods have proposed modeling pedestrian interactions based on LSTM and Attention mechanisms. Alahi et al. [7] use LSTM's hidden state to represent the states of pedestrians, and propose a ''pool'' scheme to model pedestrian interactions. This method employs a grid of fixed size to partition the neighborhood, and aggregates the hidden states within each grid, thereby integrating the influence of surrounding pedestrians. Unlike the ''pool'' scheme, another method uses

The associate editor coordinating the review of this manuscript and approving it for publication was Anandakumar Haldorai.

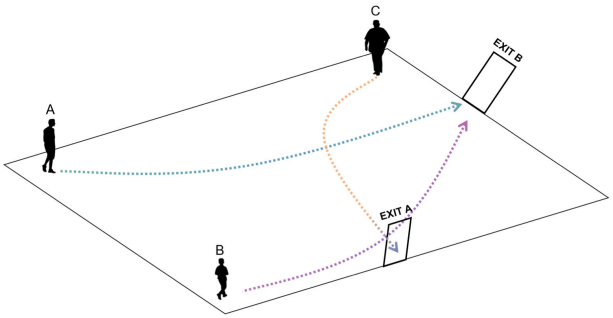

FIGURE 1. Illustration of the macro direction and micro interaction of pedestrians. The three dashed lines represent the macro travel paths of three pedestrians, A, B, and C. They intend to walk along these lines. However, when encountering other pedestrians on their route, they will temporarily adjust their path to avoid collisions.

the Attention mechanism to capture dynamic interactions between pedestrians [8], [9]. Compared with the ''pool'' scheme, the Attention mechanism is not limited to the manually set attention range and can adaptively allocate the impact to adjacent pedestrians. Therefore, attention based models can better model the interaction behavior between pedestrians.

However, despite the extensive research conducted on various methods of capturing pedestrian interactions, most previous work has overlooked a crucial factor. In pedestrian trajectory prediction, we posit that personal walking habits and goals play a significant role. Furthermore, striking a balance between pedestrian interactions and their individual walking habits and destinations is of paramount importance. Individuals have their own destinations, which largely dictate the approximate direction of their travel route. As pedestrians proceed in their intended direction, they must take measures to avoid collisions with others.

As shown in Figure.1, pedestrians A, B, and C all have their destination:exits A and exits B. We reckon that the pedestrian's own forward goal can be seen as the ''macro'' direction of the pedestrian's trajectory, which highlights the dotted line under pedestrians's feet, and they may encounter others along their own ''macro'' route. To avoid collisions with other pedestrians, pedestrians may change their short-term forward path, but their main forward direction does not change significantly. We believe that this is an ''micro'' adjustment made by pedestrians in their ''macro'' direction.

Numerous existing methods have been thoroughly examined in the context of modeling pedestrian interactions, yet they frequently neglect the equilibrium between ''macro'' and ''micro'' information. In pedestrian interaction modeling, most models focus on a single moment. Huang et al. [10] argue that both immediate and over-time interactions are key. Thus, STGAT employs ''M-LSTM'' and GAT for momentary interactions, and ''G-LSTM'' for interactions over significant time periods. However, numerous studies [11], [12] have demonstrated that the attention mechanism surpasses LSTM in predicting long sequences. This is because, in the prediction of long sequences, it becomes challenging for hidden states to retain sufficient information. Moreover, STGAT did not consider the significant impact of interactions over multiple periods, whereas the structure of LSTM may only capture crucial interactions within a single period, potentially leading to congestion and information loss in long-term sequence prediction. The attention mechanism can concentrate on global information and autonomously assign varying degrees of importance to different pieces of information, unlike LSTM, which tends to focus solely on a specific period of interaction, potentially leading to the loss or excessive compression of certain crucial information.

To address the aforementioned issues: 1. the limitations of existing methods in balancing the pedestrian's own path and interactions with others, 2. the existing model does not fully consider the temporal correlation of pedestrian interaction or cannot achieve a better effect due to methodological limitations, we propose a novel model: the Multi-Encoder Transformer. This model is designed to harmonize the macro-level direction of pedestrian movement with the micro-level obstacle avoidance behavior. First, we designed a novel architecture named Multi-Encoder Transformer, with the aim that the model not only possesses the ability to simultaneously focus on diverse kinds of information, but also can fully balance the significance among different information. Then, in the modeling of pedestrian interaction, we integrate the Graph Attention Network (GAT) with the Attention module to more effectively capture the temporal correlation of pedestrian interaction, with the intention of attaining a superior modeling effect for pedestrian interaction. Finally, aiming at the problem of pedestrian trajectory prediction, we have carried out a new design for the input and output of the Transformer Decoder, while reducing the computational load and maintaining the accuracy of the model at the same time. The main contributions of this paper are as follows:

1)This paper proposes a novel architecture named Multi-Encoder Transformer which can simultaneously focus on macro-micro information, and balance the significance among different information.

2)In the interaction stage, a gat module is introduced to model the pedestrian interaction at each individual time point, adaptively allocating the importance degree of the influence of different pedestrians on the current pedestrian. Moreover, we use an additional transformer encoder to capture the temporal correlation of pedestrian interactions at different time points. This module allows the model to adaptively capture the pedestrian interaction at the time points that are more important for the pedestrian's future trajectory.

3)In the Decoder, we redesigned the input, output and calculation methods of this module for the pedestrian trajectory prediction problem, and reduced the amount of computation compared to the original transformer decoder while maintaining the prediction accuracy of the model. The experimental results are presented in Section IV.

The rest of this paper is arranged as follows. In Section II, we analyze recent work on pedestrian trajectory prediction.

In Section III, we explain the principle of the METF model in detail. In Section IV, we do comparative experiments with other models on the open data sets and analyze the experimental results, and some ablation experiments are introduced to prove the effectiveness of the module design. We also visualized the prediction results of the model in order to observe the prediction results obviously. In Section V, we summarize the work of this paper.

II. RELATED WORK

This section reviews the related work in crowd interaction modeling, recurrent neural networks for sequence prediction, models based on attention mechanism, and graph neural networks. It provides the background and foundation for the proposed method.

A. CROWD INTERACTION MODELING

The initial work on modeling pedestrian interaction was spearheaded by Yu et al. [13]. Their primary focus was on pedestrian dynamics and social force models. The social force model is a methodology that describes pedestrian movement, predicated on the interaction between a pedestrian's internal motivation and the external environment. The methodology of the social force model aligns with our approach of considering pedestrian trajectory prediction from both macro and micro perspectives. It posits that pedestrian movement is primarily influenced by the following forces: 1. The force propelling pedestrians in the expected direction; 2. The force compelling pedestrians to maintain a certain distance from other pedestrians and boundaries; 3. The force of attraction between pedestrians. However, other studies [7], [14] have demonstrated that social force models struggle to accurately model the interactions of complex populations in intricate scenarios. In recent years, many deep learning-based models have tried to model the interactions among pedestrians. Alahi et al. [7] use the LSTM hidden state to represent the state of the pedestrian and use the ''pool scheme'' to model the pedestrian interaction with the hidden state. Vemula et al. [9] propose an attention-based approach to model pedestrian interactions. They use a soft attention model to capture the relative importance of each person in the crowd, regardless of their proximity. Mohamed et al. [15] proposes a trajectory prediction framework based on Spatio-Temporal Graph Transformer (STAR), which decomposes spatio-temporal attention modeling into temporal and spatial modeling. In the past years, the seq2seq-based models have achieved great success in modeling human-human interactions. The models mentioned above all model human-human interactions by different method.

B. RECURRENT NEURAL NETWORKS FOR SEQUENCE PREDICTION

Sequence prediction encompasses multiple sub-problems, one of which involves using historical sequences to forecast future sequences. For instance, the LSTM model in neural networks [16] is a type of neural network specifically designed to predict future sequences based on historical data. Pedestrian trajectory prediction can also be framed as the task of using past sequences to anticipate future sequences. However, merely employing LSTM for pedestrian trajectory prediction proves challenging when it comes to modeling pedestrian interaction. To tackle this issue, numerous researchers have embarked on various explorations. For instance, one approach [7] utilizes the ''social pool'' mechanism to amalgamate information from multiple pedestrians to predict their trajectories. Another method [10] employs the Graph Attention Network (GAT [17]) to model the hidden state of pedestrians in LSTM output, thereby modeling the interaction between pedestrians at various times, and subsequently using ''G-LSTM'' to model the temporal correlation of pedestrians at each moment. However, each approach has its own set of challenges: the ''social pool'' mechanism in [7] heavily depends on artificially set attention ranges, and in [10], it is challenging for LSTM to adaptively allocate the importance of all pedestrians.

C. MODELS BASED ON TRANSFORMER

The Transformer network, has demonstrated exceptional performance in fields such as natural language processing and machine translation. According to research [11], [12], its performance in machine translation, speech recognition, and natural language understanding surpasses that of LSTM [18]. In [19], the Transformer model was employed to predict pedestrian trajectories, demonstrating superior effectiveness compared to the LSTM model, even without considering pedestrian interaction. Yu et al. [20] proposes a multi-agent framework based on the Transformer Network, to predict people future trajectories. Experimental results show that the Transformer Network performs well in the task of pedestrian trajectory prediction, and has advantages in dealing with missing data and long-term prediction. In [21], TGConv, a Transformer-based graph convolution mechanism, was utilized to model crowd interactions at various time points. The time dependence at each point was modeled by an independent time transformer, which subsequently became the state-of-the-art model at that time. Yu et al. [20] designed the temporal transformer and the spatial transformer based on the transformer for the pedestrian trajectory prediction problem. It uses the temporal transformer to capture the state information of the pedestrian at different time points of itself, and uses the spatial transformer to capture the interaction information among pedestrians at the same time point.

D. GRAPH NEURAL NETWORK

Graph Neural Networks (GNNs) have demonstrated superior performance in machine learning tasks that involve graph data types. For instance, the Graph Convolution Network (GCN) proposed by Kipf and colleagues has yielded impressive results in tasks such as social network analysis, node classification, and community discovery. In [15], Graph

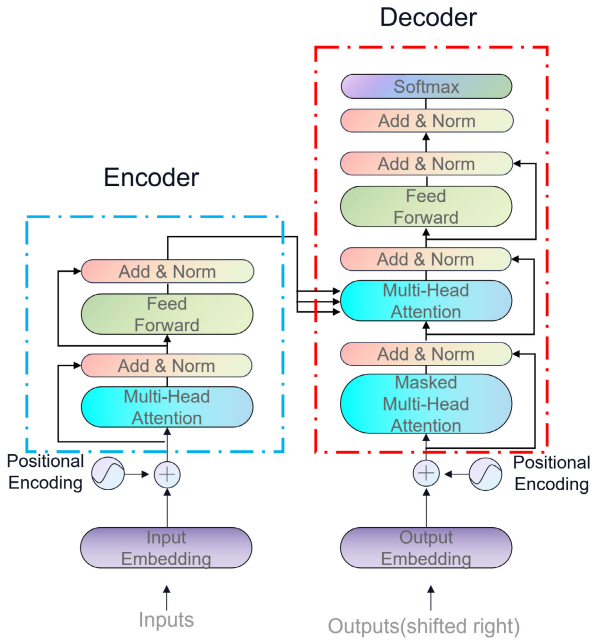

FIGURE 2. Illustration of Transformer. Transformer makes use of the encoder representation of the observed input positions and the previously predicted outputs. Shown in blue and are the self-attention and encoder-decoder attention modules, which enable Transformer to learn on which past position it is necessary to focus in order to predict an accurate result.

Convolution Networks are utilized to extract spatial and temporal information of pedestrians from the graph, thereby generating suitable embeddings and leveraging them to predict pedestrian trajectories. In [17], Graph Attention Networks (GAT) also exhibit strong performance in tasks such as facial recognition, pose estimation, and social network analysis. In these tasks, GAT is capable of learning the relationships and significance between nodes, thereby significantly enhancing the model's comprehension ability.

In addressing the problem of pedestrian trajectory prediction, when modeling the simultaneous interaction between pedestrians, we represent each pedestrian as a node with edges connecting them. This approach transforms the data of all pedestrians in a scene at a given moment into graph data, where the edges symbolize the intensity of the interaction between two pedestrians. Likewise, other research [10] has employed GAT to model pedestrian interaction and has attained optimal results.

We incorporated GAT and attention mechanisms into the ''macro encoder'' and ''micro encoder'' of our model, leveraging their respective strengths, thereby enabling their application to tasks in which they excel. Ultimately, the model yielded optimal test results.

III. METHOD

This section introduces the problem formulation and the various components of the model proposed in this paper, as well as their design ideas.

A. PROBLEM FORMULATION

The problem of human trajectory prediction is as follows: We assume that there are N pedestrians in a scene, represented as $p_1, p_2, \ldots, p_N$,the position of pedestrianp_i $(i \in [1, N])$ represented as $S_i^t = (x_i^t, y_i^t)$.. Pedestrians 's position S_i^t is known for a continuous period of time $t = 1, \ldots, T_{obs}$,our goal is to predict the future continuous time, S_i^t in $t = T_{obs+1}, \ldots, T_{pred}$.

B. ENCODER-DECODER TRANSFORMER

Transformer is a kind of modular architecture which consists of an encoder and a decoder. As shown in Figure 2, both the encoder and the decoder in the Transformer are six layers, and each layer contains three main modules: 1. an attention module, 2. a feed-forward fully-connected module, 3. two residual connections after each of the previous blocks.

The ability of the network to capture sequence non-linearities mainly resides in the attention modules. Inside each attention module, an entry of a sequence, termed ''query'' (Q), is compared with all the other sequence entries, named ''keys'' (K) through a scaled dot product, which is scaled by the equal query and key dk embedding dimensionality.Then, the output is utilized to weight the same sequence entries which are now named ''values'' (V). Hence, attention is provided by the equation:

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{1}$$

The aim of the encoding stage is to generate a representation for the observation sequence that enables the model to have memory.For this purpose, after the encoding of the input, encoder produces two vectors of keys, K_{enc}, and values V_{enc}, which would be passed on to the decoder. The decoder predicts the future track positions in an auto-regressive manner. At each new prediction step, a new decoder query Q_{dec} is compared with the encoder keys K_{enc} and values Venc in accordance with Eq. (1).

In the model we proposed, both the macro encoder and the micro encoder as well as the decoder adopt the same design as the Transformer, but the designs in aspects such as the output of the encoder and the calculation method of the decoder are completely different, and the details will be elaborated in the DECODER chapter and the PREDICTION chapter of this section.

C. MULTI ENCODER ARCHITECTURE

As depicted in Fig.3, our proposed architecture comprises two sub-encoders: the micro-encoder and the macro-encoder, each with distinct responsibilities. We partition the problem of human trajectory prediction into two components: macro and micro. Correspondingly, each encoder is tasked with one of these components.

Specifically, the macro-encoder models pedestrian direction, inferring future plans from historical trajectories, i.e., the pedestrian's planned path to their destination. We assume that deviations from a planned route are inevitable during most

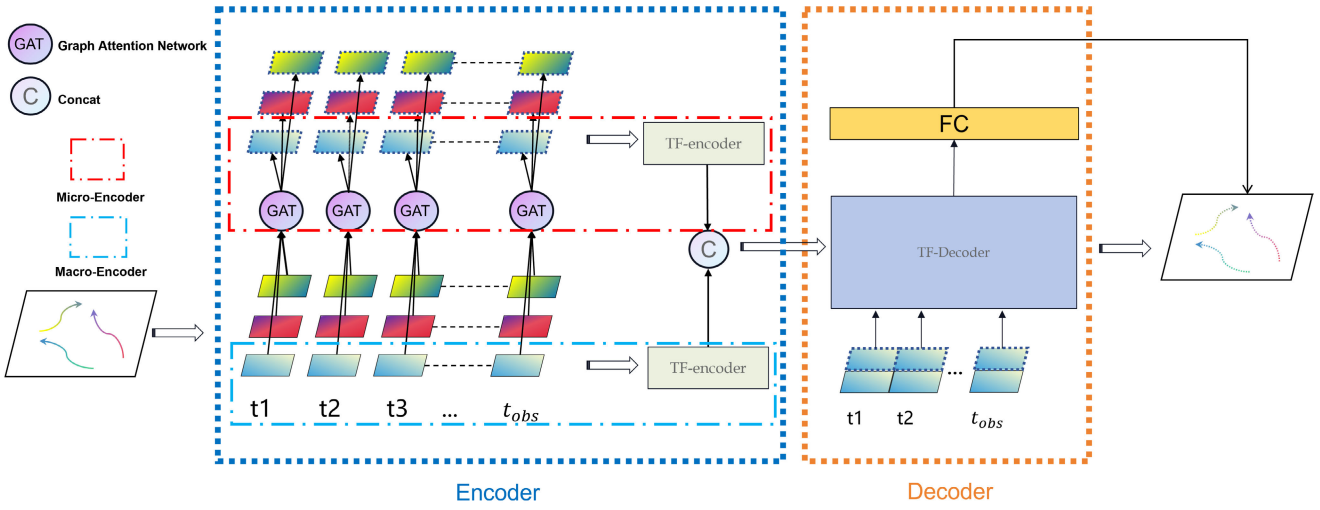

FIGURE 3. The architecture of our proposed METF model. The entire model is based on seq2seq and consists of two parts:Encoder and Decoder. The Encoder contains two sub encoders:micro encoder and macro encoder. 1.Macro-encoder,The blue dashed box along the way represents a single pedestrians's input of each time position to the macro encoder,which is a single pedestrian attention encoder; 2.Micro encoder, including GAT and attention module, the input of the GAT is the information of all pedestrians at each moment, and the output is the pedestrian interaction information and serves as the input of the attention module to model the temporal correlation of pedestrian interactions. The Memory contains both micro and macro information for each pedestrian. Decoder generates future trajectories based on Memory.

walks. Frequently, we encounter pedestrians on our planned path and choose to avoid them, but this does not imply a change in our initial route. In reality, we alter our path to avoid collisions on short sections of the planned route. Thus, from a macro perspective, there is little difference between the planned and actual walking routes. However, from a micro perspective, locations where we encounter other pedestrians differ from our planned route. Occasionally, these micro-level route changes can prompt us to replan the macro-level route. The micro-encoder addresses these micro-level pedestrian path changes.

D. MACRO ENCODER

Each pedestrian exhibits unique behavioral patterns, encompassing walking speed, acceleration mode, and macroscopic destination path. The Attention mechanism has been demonstrated to effectively capture a single pedestrian's behavior patterns based on their information, without considering the influence of other pedestrians [10]. Consequently, we employ the attention mechanism in modeling pedestrian movement intentions. It solely observes the pedestrian's information and can efficiently allocate the importance of the pedestrian's state at different times, thereby effectively encoding pedestrian intention. Ultimately, its output serves as the output of the ''macro'' encoder. This part is the macro-encoder.

In our deployment, our raw data is pedestrians 's coordinate S_i^t at time $t = 1, \ldots, T_{obs}$. In order to enable the attention module to process input data, we embed it into a higher D-dimensional space through a linear projection with a weight matrix W_x, as shown in equation 1.

$$e_{obs}^{(i,t)} = S_i^t W_x \tag{2}$$

Due to the fact that the attention module receives inputs from various time points simultaneously, it cannot distinguish the input time information.Therefore, we adopt the same approach as Transformer [12], after embedding the input into dimension D, we use ''position encoding'' to encode the data of all pedestrians at all times.The specific approach is to add the input embedding $e_{obs}^{(i,t)}$ and a position encoding vector p^t, which has the same dimension D with $e_{obs}^{(i,t)}$. We use sine and cosine functions to define p^t.

$$\mathbf{p}^t = \{p_{t,d}\}_{d=1}^{D} \tag{3}$$

$$\text{where} \quad p_{t,d} = \begin{cases} \sin\left(\dfrac{t}{10000^{d/D}}\right) & \text{for} d \text{even} \\ \cos\left(\dfrac{t}{10000^{d/D}}\right) & \text{for} d \text{odd} \end{cases} \tag{4}$$

Each dimension of the positional encoding corresponds to a sinusoid. The wavelengths form a geometric progression from 2π to $10000 \cdot 2\pi$.

After obtaining $\mathbf{p}^t$,add $\mathbf{p}^t$ and $e_{obs}^{(i,t)}$ to obtain an embedding with time order information.

$$D{:}\xi_{obs}^{(i,t)} = \mathbf{p}^t + \mathbf{e}_{obs}^{(i,t)} \tag{5}$$

Macro encoder employs the standard encoder architecture of a naive transformer on $\xi_{obs}^{(i,t)}$. The objective of employing macro encoding is to encapsulate the pedestrian's walking status, inclusive of travel speed and path planning, via observable pedestrian trajectories. This approach aims to furnish the decoder with effective macro route and speed information for prediction. This rationale underpins our choice of the standard encoder architecture of a naive transformer over LSTM to encapsulate the macro direction of pedestrians.

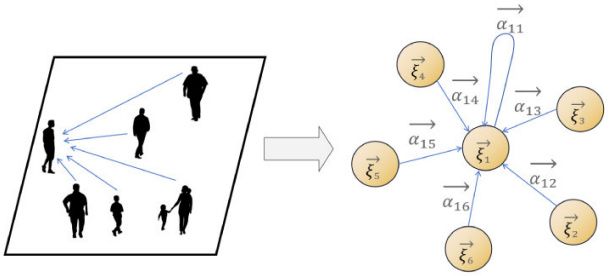

FIGURE 4. Illustration of modeling the interaction of pedestrians at the same time point using GAT. In a given scene, all pedestrians are represented as points on a graph. The influence of each pedestrian on others is adaptively assigned importance using GAT. Here, $\alpha_{i,j}$ represents the degree of influence of point j on point i. This model is used to simulate the interaction between pedestrians at a specific moment.

Contrary to LSTM, the naive transformer's standard encoder architecture can parallelly observe pedestrian status at all time points and adaptively assign importance to pedestrian status at different time points, instead of sequentially receiving information at each time point and opting to forget or remember specific information. Therefore, we posit that attention encoders, compared to LSTM, offer the following advantages: 1. Parallel computing allows attention encoders to achieve faster computation speeds during training and inference stages; 2. The attention encoder observes global information simultaneously, rather than step-by-step. We contend that the proximity of the historical trajectory to the prediction point does not necessarily denote its importance. Instead, for different pedestrians, the significance of historical trajectory points may vary, regardless of their proximity to the prediction point. LSTM's limitations can result in over-compression or forgetting of historical trajectory information that is farther from the prediction point, leading to the loss of crucial information. This enables attention encoders to more efficiently concentrate on useful information and better encapsulate pedestrian walking status.

E. MICRO ENCODER

We hypothesize that a pedestrian's true route can be bifurcated into their macroscopic route and their microscopic adjustments on this macroscopic route. The micro encoder's role is to discern the mutual influence among pedestrians and equip the decoder with information on the micro adjustments that pedestrians might make on their macro route in the future due to the influence of other pedestrians. Numerous prior studies have underscored the importance and efficacy of modeling pedestrian interactions at various time points. However, we contend that not every interaction moment carries significant impact. It could be the interaction over a duration that holds more importance, or the interaction at a specific point in time.

Consequently, we employ a combination of GAT and attention mechanisms to model pedestrian interaction. GAT is utilized to capture the interaction between pedestrians at each moment, while attention is used to identify significant time periods of pedestrian interaction. Ultimately, its output serves as the input for the "micro" encoder.

Micro encoder handles raw data in the same way as macro encoder. It is worth noting that when the micro encoder embeds the original data, we do not use a new linear layer for embedding, but instead use the same embedding $\xi_{obs}^{(i,t)}$ as the macro encoder. This is because we have empirically found that using the same embedding as the macro encoder and using a new linear layer embedding has little impact on the accuracy of the model.

To obtain the interaction between pedestrians at each moment, we consider all pedestrians in a scene as multiple points in a graph, and consider this graph as a fully connected graph, assuming that each point in this graph may affect each other(Recent research findings indicate that when considering the influence of other pedestrians, it is necessary to consider every pedestrian [8], [22]). We use the latest progress of GNNs (Graph Neural Networks), GAT (Graph Attention Network), to model the mutual influence between pedestrians at different time points, because GAT allows for the fusion of information between points and can adaptively assign weights to each point, simulating the actual situation where the degree of influence between each pedestrian varies in real scenes. Therefore, we employs the GAT on $\xi_{obs}^{(i,t)}$. Shown as Fig.4.

The input of GAT is ξ at the same time, $\xi = \{\vec{\xi}_1^t, \vec{\xi}_2^t, \ldots, \vec{\xi}_N^t\}$, N is the number of people going down at that moment, that is, the number of points in the graph. $\alpha_{i,j}$ is the correlation coefficient between nodes (i, j), indicating the degree of influence of node j on node i.

$$\alpha_{ij}^t = \frac{\exp(\text{LeakyReLU}(a^T[W\xi_i^t \| W\xi_j^t]))}{\sum_{k\in\mathcal{N}_i}\exp(\text{LeakyReLU}(a^T[W\xi_i^t \| W\xi_k^t]))} \tag{6}$$

where $\|$ represents the connection operation, α_{ij}^t is the attention coefficient of two points (i, j) at the same time. N_i represents the number of nodes at that time. The final output of GAT is also determined by α_{ij}^t and ξ are calculated. $\mathbf{W}(\mathbf{W} \in \mathbf{R}^{d' \times d})$ is the weight matrix of a shared linear transformation which is applied to each node (d is the dimension of ξ_i^t, d' is the dimension of $\hat{\xi}_i^t$). x^T represents transpose operation to x, a is the weight matrix of a single-layer feedforward neural network. It finally normalized by LeakyReLU.

The final output of GAT is also determined by α_{ij}^t and ξ are calculated.

$$\hat{\xi}_i^t = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^t \mathbf{W}\xi_j^t\right) \tag{7}$$

σ is a nonlinear function.W is the weight matrix of a shared linear transformation. $\hat{\xi}_i^t$ is the output of GAT.

The objective of the micro encoder is to convey the influence of other pedestrians on the current pedestrian to the decoder, enabling the decoder to model the micro adjustments of pedestrians on their macro routes. As stated at the outset of this section, not every interaction moment is significant.

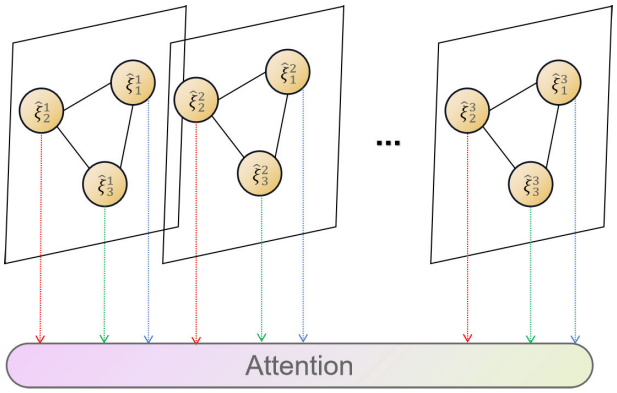

FIGURE 5. Illustration of use of attention mechanisms to capture significant group interactions at a specific or continuous time point. The Attention encoder takes as input the information regarding a pedestrian's interactions with other pedestrians at various time points. In the figure, dashed lines of the same color represent the inputs of a pedestrian's interactions at different times into the Attention encoder.

It could be a specific moment or a few moments of interaction that play a pivotal role in the micro adjustment of pedestrians on the macro route. We posit that in certain scenarios, it is not a single moment of interaction that holds importance, but rather a continuous or several consecutive interactions.

Therefore, we are like macro encoders, deploying the standard encoder architecture of a naive transformer on $\hat{\xi}_i^t$, use it to capture certain individual or continuous important interactions.Shown as Fig.5.

The input for the Attention encoder is $\hat{\xi} = \{\hat{\xi}_i^1, \hat{\xi}_i^2, \ldots, \hat{\xi}_i^{t_{obs}}\}$, that is, the interaction information between the current pedestrian and other pedestrians at different times.We treat the output of the attention encoder as the output of the micro encoder.

F. DECODER

The role of the decoder is to forecast the macro route of a pedestrian and the micro adjustments made by the pedestrian on the macro route, drawing upon the macro and micro information encapsulated by the macro encoder and micro encoder. In our implementation, the decoder employs the standard decoder architecture of a naive transformer. However, it is noteworthy that our design for memory and input diverges significantly from theirs.

Memory consists of outputs from macro encoders and micro encoders, the outputs of the macro encoder and micro encoder for a pedestrian are u, $u = \{u_i^1, u_i^2, \ldots, u_i^{T_{obs}}\}$; n, $n = \{n_i^1, n_i^2, \ldots, n_i^{T_{obs}}\}$. The memory is calculated as:

$$m_i = u_i \parallel n_i \tag{8}$$

The input also consists of two parts: ξ and $\hat{\xi}$. $\xi = \{\vec{\xi}_i^1, \vec{\xi}_i^2, \ldots, \vec{\xi}_i^{t_{obs}}\}$, $\hat{\xi} = \{\hat{\xi}_i^1, \hat{\xi}_i^2, \ldots, \hat{\xi}_i^{t_{obs}}\}$, the input is calculated as:

$$q_i = \hat{\xi}_i \parallel \xi_i \tag{9}$$

The memory and the input are concatenated in the feature dimension by u, n, and $\xi, \hat{\xi}$, respectively. Special attention should be paid to the u, n and $\xi, \hat{\xi}$'s concatenate order:Align u with $\hat{\xi}$ and n with ξ in the memory and the input. That is to say,in the second attention layer of the decoder, the $\hat{\xi}$ (containing pedestrian interaction information at every moment) after being processed by the self attention layer in the decoder, then it will calculate attention score with the output n of the micro encoder(which includes crowd interaction information); The ξ (which only including information about pedestrians at various time points) after being processed by the self attention layer in the decoder, then it will calculate attention score with the output u of the micro encoder (which only have pedestrian's own information). The purpose of our design is to integrate macro and micro information, enabling more accurate prediction of pedestrian macro routes and micro adjustments of pedestrians in macro routes.

G. PREDICTION

During the final prediction phase, numerous models opt to predict the Gaussian distribution of future pedestrian trajectories, or introduce random vectors during the prediction phase and generate 20 predictions, with the optimal result among the 20 predictions serving as the model's evaluation result. We contend that this method can assess a model's upper limit, but gauging the model's standard level proves challenging, hence we refrained from incorporating randomness at this stage. Ultimately, we employ a linear layer to output the future trajectory coordinates of the concluding pedestrian.

The output of the Decoder is recorded as $o_i^1, o_i^2, \ldots, o_i^{T_{obs}}$. The superscript is the time point, and the subscript is the pedestrian number. We concatenate the decoder output of a pedestrian in the feature dimension.

$$o = o_i^1 \parallel o_i^2 \parallel \ldots \parallel o_i^{T_{obs}} \tag{10}$$

Then, we will input the o into the linear layer.

$$(x_i^{T_{obs}+1}, y_i^{T_{obs}+1}, x_i^{T_{obs}+2}, y_i^{T_{obs}+2}, \ldots, x_i^{T_{pred}}, y_i^{T_{pred}}) = \delta(o) \tag{11}$$

Our loss design is very simple.

$$Loss = ||(x_i, y_i) - (\hat{x}_i, \hat{y}_i)||_2 \tag{12}$$

The (x_i, y_i) is prediction, and the $(\hat{x}_i, \hat{y}_i)$ is ground truth.

IV. EXPERIMENTS

This section introduces the datasets, evaluation metrics, comparison with existing methods, quantitative and qualitative evaluation of the proposed METF method.

A. DATASETS

The ETH and UCY datasets [32] are two common datasets used to train and evaluate pedestrian trajectory prediction methods. They are all real pedestrian trajectories captured from a bird's-eye perspective, including 1536 pedestrians and rich multiplayer interaction scenes. Their sampling frequency

TABLE 1. Comparison with baselines models on public benchmarks ETH and UCY for ADE/FDE. The method name followed by an "*" indicates that the results of the method in the table are based on the Best-of-20 Samples, the others are deterministic result. Lower is better.

Method	ETH-UNIV	ETH-HOTEL	UCY-UNIV	UCY-ZARA1	UCY-ZARA2	AVERAGE
Social LSTM*[7]	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
Social GAN*[22]	0.81/1.52	0.72/1.61	0.60/1.26	0.34/0.69	0.42/0.84	0.58/1.18
SoPhie*[24]	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.51/1.15
Social-BiGAT*[25]	0.69/1.29	0.49/1.01	0.55/1.32	0.30/0.62	0.36/0.75	0.48/1.00
STGAT*[10]	0.65/1.12	0.35/0.66	0.52/1.10	0.34/0.69	0.29/0.60	0.43/0.83
Social-STGCNN* [15]	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
SGCN*[26]	0.63/1.03	0.32/0.55	0.37/0.70	0.29/0.53	0.25/0.45	0.37/0.65
Social TAG*[27]	0.61/1.00	0.37/0.56	0.51/0.87	0.33/0.50	0.30/0.49	0.42/0.68
D-STGCN*[28]	0.63/1.03	0.40/0.65	0.50/0.89	0.37/0.60	0.32/0.50	0.44/0.73
HSTGA*[28]	0.53/1.03	0.31/0.52	0.44/0.98	0.31/0.62	0.27/0.61	0.37/0.75
FlowChain*[29]	0.55/0.99	0.20/0.35	0.29/0.54	0.22/0.40	0.20/0.34	0.29/0.52
STAGP*[30]	0.65/1.21	0.41/0.73	0.38/0.68	0.28/0.46	0.25/0.44	0.40/0.70
Social GAN[22]	1.03/2.02	0.90/1.97	0.58/1.22	0.38/0.84	0.47/1.01	0.67/1.41
STGAT[10]	0.80/1.53	0.52/1.08	0.51/1.12	0.39/0.87	0.30/0.64	0.50/1.05
SR-LSTM[31]	0.63/1.25	0.37/0.74	0.51/1.10	0.41/0.90	0.32/0.70	0.45/0.94
STAR[20]	0.56/1.11	0.26/0.50	0.52/1.13	0.40/0.89	0.31/0.71	0.41/0.87
CSCNet[32]	0.51/1.05	0.22/0.42	0.47/1.02	0.36/0.81	0.31/0.68	0.37/0.79
METF(ours)	**0.65/1.31**	**0.19/0.30**	**0.53/1.12**	**0.38/0.76**	**0.32/0.67**	**0.41/0.83**

TABLE 2. Comparison with different design of decoder. We compute the MACs of each model by simultaneously predict the trajectories of 10 agents. Params is the parameter quantity of the entire model. The Lower the better.

Method	ETH-UNIV	ETH-HOTEL	UCY-UNIV	UCY-ZARA1	UCY-ZARA2	AVERAGE	MACs	Params
METF(Naive TF Design)	0.72/1.43	0.18/0.31	0.62/1.21	0.43/0.82	0.34/0.67	0.46/0.89	2472.21M	2.86M
METF(Our Design)	0.65/1.31	0.19/0.30	0.53/1.12	0.38/0.76	0.32/0.67	0.41/0.83	227.88M	2.96M

TABLE 3. Ablation Study on METF. We removed the macro encoder and micro encoder from METF separately. WITHOUT MACRO denotes removed the macro encoder from METF; WITHOUT MICRO denotes removed the micro encoder from METF.

Method	ETH-UNIV	ETH-HOTEL	UCY-UNIV	UCY-ZARA1	UCY-ZARA2	AVERAGE
METF(WITHOUT MACRO)	1.40/2.09	1.01/1.19	1.25/1.62	1.02/1.08	1.37/1.61	1.21/1.51
METF(WITHOUT MICRO)	0.76/1.55	0.37/0.56	0.65/1.32	0.46/0.89	0.35/0.72	0.51/1.00
METF	0.65/1.31	0.19/0.30	0.53/1.12	0.38/0.76	0.32/0.67	0.41/0.83

is ($\Delta t = 0.4$s). The ETH and UCY have five scenes in total:ETH, HOTEL, UNIV, ZARA1 and ZARA2.

We following previous studies [7], [10], [22], we set $T_{obs} = 8$, $T_{pred} = 12$, that is take 8 frames (3.2s) as observation and the next 12 frames (4.8s) as prediction. This setting has been recognized and adopted by the vast majority of studies.And besides pedestrian coordinates, we did not use any additional data, such as map information, etc.

B. METRICS

In model evaluation, We are also following previous studies [7], [10], using ADE (Average Displacement Error) and FDE (Final Displacement Error). The ADE evaluate average error between model prediction trajectories and ground truth trajectories in every time steps. The FDE evaluate error between model prediction trajectories and ground truth trajectories in the last time steps. The smaller the evaluation numbers, the better the model results.

$$\text{ADE} = \frac{1}{T_{pred}} \sum_{t=T_{obs}}^{T} ||(x_t - \hat{x}_t) + (y_t - \hat{y}_t)||_2 \tag{13}$$

$$\text{FDE} = ||(x_{t_{pred}} - \hat{x}_{t_{pred}}) + (y_{t_{pred}} - \hat{y}_{t_{pred}})||_2 \tag{14}$$

C. COMPARISON WITH STATE-OF-ART METHODS

We compared our proposed METF with a wide range methods, they are all highly representative methods, including Social LSTM, Social GAN [22], SoPhie [23], STGAT, Social-BiGAT [24], SGCN [25], CSCNet [31], Social-STGCNN [15], STAR, social TAG [26] and STAGP [29]. Table 1 presents the ADE and FDE results of each model. The performance of METF is noteworthy, particularly given that our prediction results are derived from a single sample, in contrast to over half of the models which require 20 samples to achieve their best results. As discussed in Section III-F, while the best-20 protocol may provide a better examination of a model's upper limit, it poses challenges in determining a model's standard level.

METF has yielded results that are relatively ideal. With the exception of CSCNet as shown in Table 1, METF performs best among other deterministic results. It is worth noting that STGAT uses GAT in modeling pedestrian interaction, just like our method of modeling micro route changes for pedestrians, and STGAT also only models macro routes for pedestrians based on their own walking information. However, Compared to STGAT deterministic result, the ADE and FDE of our method has decreased by 18.0%

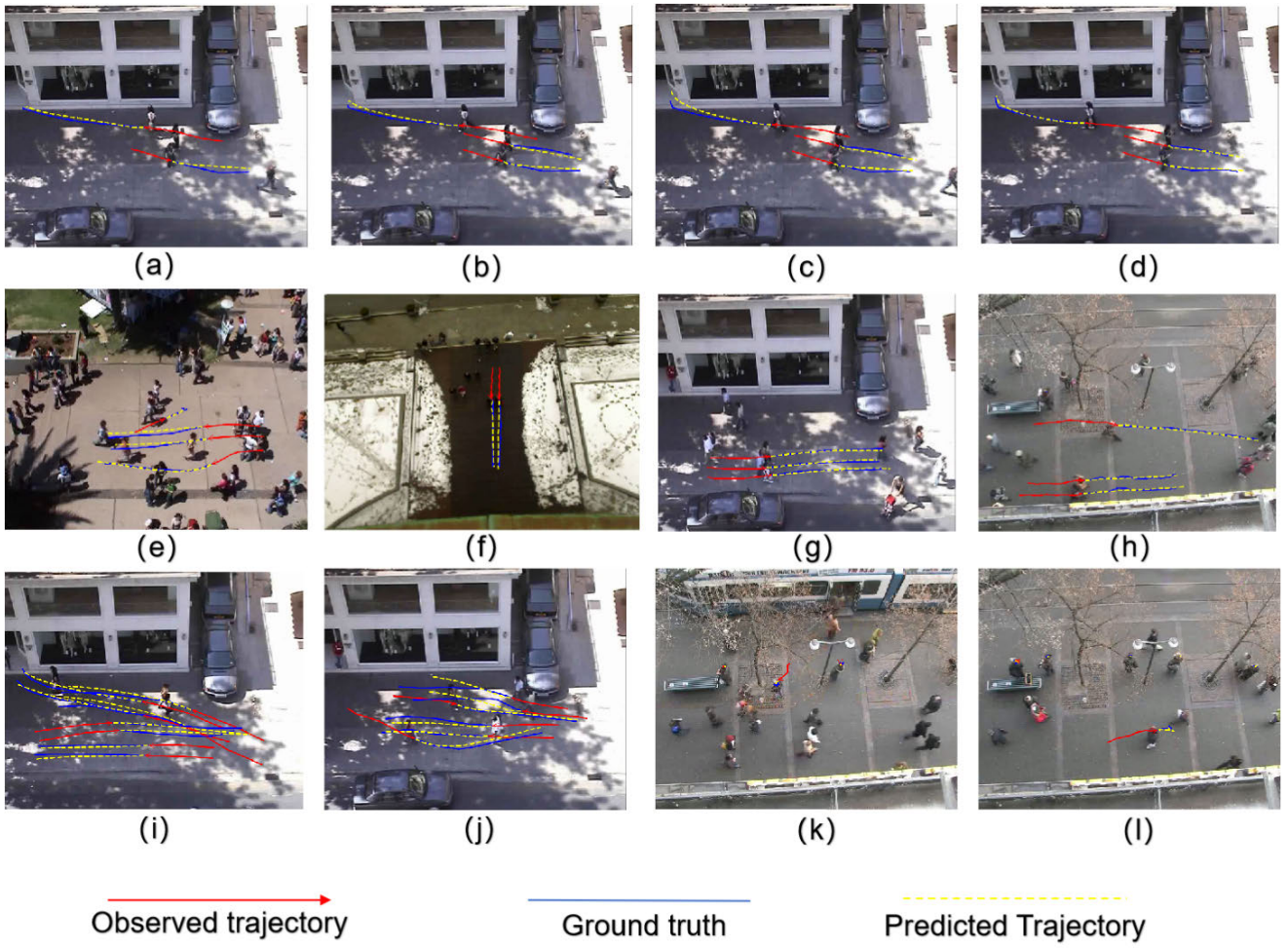

FIGURE 6. Visualization Results. The solid red line in the figure are observed trajectory, solid blue line are ground truth, the yellow dashed line is the predicted trajectory of our METF. The first row displays the model's predicted trajectory of pedestrians over a continuous period of time. The second row displays the behavior of pedestrians walking together in different scenarios. The third row shows the situation of complex scenes with multiple pedestrians, as well as situations where pedestrians suddenly stop and slow down.

and 21.1% respectively. METF has achieved a 5% decrease in ADE compared to STGAT's Best-of-20 Samples result. We believe that is to be due to STGAT's use of LSTM to summarize the macro and micro states of pedestrians. LSTM can only linearly observe the embedded information of pedestrians at each time point, potentially leading to excessive compression or neglect of embedded information at earlier time points. However, in pedestrian prediction, the information at each time point is crucial. METF employs an attention mechanism to observe the information of pedestrians at each time point simultaneously and adaptively allocate their importance. The Decoder structure of METF further enables it to adaptively allocate the importance of macro and micro information for pedestrians in subsequent routes. Whether in the modeling of macro or micro information, or in the process of outputting future pedestrian trajectories through macro and micro information, METF demonstrates a high degree of adaptability in capturing more useful information for predicting future pedestrian trajectories. This is also believed to be the primary reason for METF's significant decrease in ADE and FDE compared to STGAT.

D. QUANTITATIVE EVALUATION

1) EFFECT OF THE SPECIAL DESIGN OF THE INPUT OF DECODER

In the naive Transformer, the input of the Decoder is the output of the previous time step Decoder, as predicted in [19] using this method. To test the effectiveness of our redesign of the decoder input, we trained and evaluated the decoder design using the original Transformer while keeping the remaining parameters of the model unchanged, in order to compare the effectiveness of the two methods. The results are shown in Table 2.

It can be seen that our design has reduced by 11.9% and 7.8% in ADE and FDE, respectively, compared to the naive transformer's design.Moreover, we also listed the number of model parameters and MACs of these two methods. Calculation of MACs are based on the assumption of predicting ten pedestrians simultaneously. It can be seen

that our design has increased by 3.4% in params, compared to the naive transformer's design. But, this design makes our model is only 9.2% of the designed for naive Transformers in MACs. This is a fairly cost-effective deal, by spending only an additional 3.4% of memory usage, the model's error was reduced by neraly 10% and the MACs were reduced tenfold.

Our design has significantly reduced the MACs, as it allows the model to predict the position of pedestrians simultaneously at all times when forecasting their future trajectories, unlike the original Transformer. The naive Transformer predicts the position of pedestrians sequentially, as it requires the pedestrian position predicted by the model at the previous time step when predicting the position at the current time. This results in the naive Transformer recalculating the encoder's output and all the calculations required in the previous time step predictions for each prediction, leading to a substantial amount of duplicate calculations.

2) EVALUATION OF MICRO-ENCODER AND MACRO-ENCODER

To investigate the distinct roles and contributions of the macro and micro encoders, we separately retained only the macro and micro encoders in METF for testing on the ETH and UCY datasets. The test results are presented in Table 3.

In comparison to METF with only macro encoders, the complete METF has achieved a reduction of 19.7% and 17.0% in ADE and FDE, respectively. This implies that the micro encoder contributes to a nearly 20.0% reduction in the model's error. As previously mentioned, the problem of pedestrian trajectory prediction can be decomposed into the macroscopic route intention of pedestrians. However, for more accurate predictions, the model needs to capture the changes in pedestrians' microscopic routes. The role of the micro encoder is to model these microscopic route changes influenced by other pedestrians, such as following or avoiding behaviors. In the following section, we will visualize the prediction results of two models for a clearer observation of this outcome.

As shown in Table 3, the performance of METF with only the micro encoder is less than ideal. This outcome aligns with our previous hypothesis that the pedestrian trajectory prediction problem can be segmented into macro and micro components. By modeling the macro aspect of the pedestrian trajectory, we can approximate the path planning of pedestrians. Coupled with the capability to model micro changes in pedestrians, the model can more precisely predict the micro alterations made by pedestrians in their own macro path due to other pedestrians. Modeling the micro level of pedestrians significantly enhances the model's accuracy, given that we initially have a model of the macro level route of pedestrians for it to function effectively.

E. QUALITATIVE EVALUATION

Fig.6 presents the visualization of METF prediction results on the ETH and UCY datasets, representing common scenarios and situations encountered in daily life. In the figure, the solid red line denotes the observed trajectory, the solid blue line represents the ground truth, and the yellow dashed line illustrates the predicted trajectory of our METF.

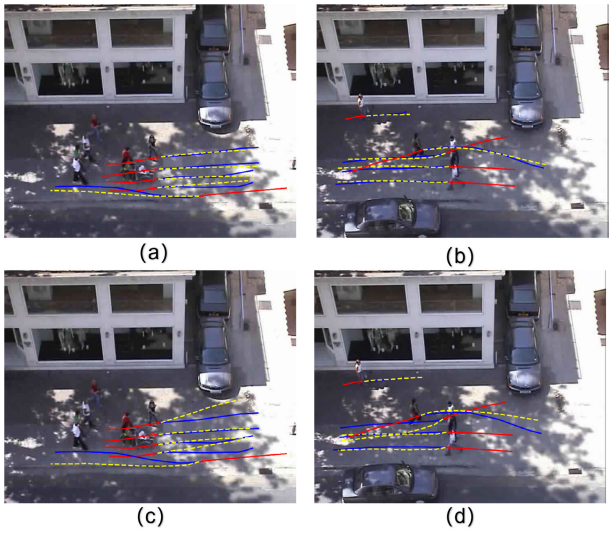

FIGURE 7. Comparison of METF with and without micro encoders. (a) and (b) are the results of the complete METF for predicting crowd trajectories, while (c) and (d) are the crowd trajectory prediction results of the METF without a micro encoder.

The four images (a) - (d) in the first row depict the continuous movement of pedestrians within the same scene, encompassing behaviors such as walking together and turning. The METF's prediction results demonstrate commendable performance in this continuous prediction, accurately forecasting the pedestrians' turning and accompanying behaviors. This suggests that METF has effectively learned specific behavioral patterns of pedestrians, thereby ensuring a high degree of accuracy.

In the second row of Fig.6, images (e) - (h) illustrate the METF's modeling of pedestrian walking behavior across various scenarios, a frequent occurrence in our daily lives. When people walk together, they keep a certain distance, preserving a walking posture while evading collisions with each other. Images (e) - (h) underscore the effectiveness of METF in micro modeling, given its ability to model pedestrian walking behavior in diverse scenarios.

Scenes (i) and (j) depict a higher density of pedestrians, leading to more intricate interactions. These include behaviors such as detouring, group walking, and counter-directional movement to prevent collisions. The METF model has demonstrated its proficiency in accurately simulating these diverse pedestrian behaviors, thereby validating its effectiveness in complex scenarios. Scene (k) and (j) illustrate another prevalent pedestrian behavior: abrupt stopping or deceleration during movement. The METF model's successful prediction of this behavior underscores its robust macro-modeling capabilities and its ability to learn and replicate pedestrian behavior patterns.

Fig.7 provides a visual comparison of the METF model with only a macro encoder and the complete METF model in a scenario involving multiple pedestrian interactions. The predicted results of the complete METF model are depicted in (a) and (b), while (c) and (d) represent the outcomes of the METF model with just a macro encoder. The complete METF model, as shown in (a) and (b), effectively models pedestrian walking behavior due to its micro encoder, yielding accurate predictions. Conversely, in (c) and (d), the METF model with only a macro encoder makes generally correct predictions in the macro direction. However, the absence of a micro encoder prevents the model from acquiring information about all other pedestrians. This limitation hinders the model's ability to adjust its route based on other pedestrians' influence and to effectively model behaviors such as group walking and avoidance. Consequently, it struggles to produce accurate predictions in scenarios with a high density of pedestrians and complex interactions.

V. CONCLUSION

This section summarizes the main contributions and findings of the paper, and discusses the potential and limitation for further development of the proposed Multi-Encoder Transformer.

We introduce a novel framework, the Multi-encoder Transformer, designed specifically for pedestrian trajectory prediction tasks. We employ the attention mechanism to capture the macroscopic route of pedestrians and integrate it with the Graph Attention Network (GAT) to model the microscopic pedestrian information, facilitated by a novel decoder design. We put forth a unique solution that bifurcates pedestrian trajectory prediction into macroscopic and microscopic components. Our model yielded promising results, with qualitative experiments illustrating that the Multi-encoder Transformer Framework (METF) produces socially acceptable human trajectories across diverse scenarios. This underscores the rationale behind decomposing the problem into macroscopic and microscopic levels.

However, our model has not yet taken into account map information, obstacle information, etc. Therefore, when modeling the microscopic path change of pedestrians in some scenarios, the model fails to obtain sufficient information, resulting in a decrease in the prediction effect. But this does not affect the superiority of our approach. Moreover, our Multi-encoder Transformer harbors significant potential for further development. In future study, we aim to leverage the capabilities of multi-encoders to incorporate the hitherto unused information, excluding other data such as map information, obstacle information, scene context and pedestrian speed. We will continue to explore how to exploit the advantages of the multi-encoder structure and utilize more information to model the problem of pedestrian trajectories. And we will also attempt to extend the model proposed in this paper to other researches with commonalities, such as vehicle trajectory prediction and ship trajectory prediction, in order to test the universality of this model.

REFERENCES

[1] Z. Chen, C. Song, Y. Yang, B. Zhao, Y. Hu, S. Liu, and J. Zhang, "Robot navigation based on human trajectory prediction and multiple travel modes," *Appl. Sci.*, vol. 8, no. 11, p. 2205, Nov. 2018.

[2] A. J. Sathyamoorthy, J. Liang, U. Patel, T. Guan, R. Chandra, and D. Manocha, "Densecavoid: Real-time navigation in dense crowds using anticipatory behaviors," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2020, pp. 11345–11352.

[3] Y. Cai, L. Dai, H. Wang, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, "Pedestrian motion trajectory prediction in intelligent driving from far shot first-person perspective video," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5298–5313, Jun. 2022.

[4] F. Marchetti, F. Becattini, L. Seidenari, and A. D. Bimbo, "Multiple trajectory prediction of moving agents with memory augmented networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6688–6702, Jun. 2023.

[5] L. Rossi, M. Paolanti, R. Pierdicca, and E. Frontoni, "Human trajectory prediction and generation using LSTM models and GANs," *Pattern Recognit.*, vol. 120, Dec. 2021, Art. no. 108136.

[6] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2016, pp. 549–565.

[7] 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.

[8] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft + hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection," *Neural Netw.*, vol. 108, pp. 466–478, Dec. 2018.

[9] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4601–4607.

[10] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: Modeling spatial–temporal interactions for human trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6271–6280.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, Jul. 2017, pp. 5998–6008.

[13] W. Yu, R. Chen, L.-Y. Dong, and S. Dai, "Centrifugal force model for pedestrian dynamics," *Phys. Rev. E*, vol. 72, no. 2, pp. 1–19, Jan. 2005.

[14] P. Yadav, A. Mishra, and S. Kim, "A comprehensive survey on multi-agent reinforcement learning for connected and automated vehicles," *Sensors*, vol. 23, no. 10, p. 4710, May 2023.

[15] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14412–14420.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[17] P. Velič ković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[18] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5884–5888.

[19] 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.

[20] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2020, pp. 507–523.

[21] X. Zhao, Y. Chen, J. Guo, and D. Zhao, "A spatial–temporal attention model for human trajectory prediction," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 4, pp. 965–974, Jul. 2020.

[22] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.

[23] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1349–1358.
[24] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting using bicycle-GAN and graph attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.
[25] L. Shi, L. Wang, C. Long, S. Zhou, M. Zhou, Z. Niu, and G. Hua, "SGCN: Sparse graph convolution network for pedestrian trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8990–8999.
[26] X. Zhang, P. Angeloudis, and Y. Demiris, "Dual-branch spatio-temporal graph neural networks for pedestrian trajectory prediction," *Pattern Recognit.*, vol. 142, Oct. 2023, Art. no. 109633.
[27] B. I. Sighencea, I. R. Stanciu, and C. D. Caleanu, "D-STGCN: Dynamic pedestrian trajectory prediction using spatio-temporal graph convolutional networks," *Electronics*, vol. 12, no. 3, p. 611, Jan. 2023.
[28] T. Maeda and N. Ukita, "Fast inference and update of probabilistic density estimation on trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 9761–9771.
[29] Z. Liu, L. He, L. Yuan, K. Lv, R. Zhong, and Y. Chen, "STAGP: Spatio-temporal adaptive graph pooling network for pedestrian trajectory prediction," *IEEE Robot. Autom. Lett.*, vol. 9, no. 3, pp. 2001–2007, Mar. 2024.
[30] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12077–12086.
[31] B. Xia, C. Wong, Q. Peng, W. Yuan, and X. You, "CSCNet: Contextual semantic consistency network for trajectory prediction in crowded spaces," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108552.
[32] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 261–268.

XINCHENG HU received the B.S. degree in software engineering from the Wenhua College (WHC), Wuhan, China, in 2022. He is currently pursuing the master's degree in computer technology with South-Central Minzu University (SCMU). He has published one articles. His research interests include deep learning, video image processing, and people trajectory prediction.

BO YANG received the B.S. degree in computer science and technology from the School of Computer Science and Technology, Huazhong University of Science and Technology (HUST), in 2001, the M.S. degree in water conservancy and hydroelectric engineering and the Ph.D. degree in spatial information science and technology from HUST, in 2004 and 2008, respectively. From 2008 to 2011, she worked as a Postdoctoral Research Fellow at HUST. Since 2012, she has been affiliated with South-Central Minzu University (SCMU), where her research interests span computer modeling and simulation, GIS, computer vision, and sign language recognition.

JIXING YANG received the B.S. degree in network engineering from South-Central Minzu University (SCMU), Wuhan, China, in 2022, where he is currently pursuing the master's degree in computer technology. He has published two articles. His research interests include deep learning and image and video processing.

TENG ZHANG received the B.S. degree in software engineering from Hubei University of Economics, Wuhan, China, in 2023. He is currently pursuing the master's degree in computer technology with South-Central Minzu University (SCMU). His research interests include deep learning, video image processing, and people trajectory prediction.