BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES,
DOI: 10.24425/bpasts.2024.151960

STAFGCN:A spatial-temporal attention-based fusion graph convolution network for pedestrian trajectory prediction

Guihong LIU[1] , **Chenying PAN**[1]*, **Xiaoyan ZHANG**[1] **and Qiangkui LENG**[1]

[1] School of Electronic and Information Engineering, Liaoning Technical University, Huludao, Liaoning, China

Abstract. Pedestrian trajectory prediction provides crucial data support for the development of smart cities. Existing pedestrian trajectory prediction methods often overlook the different types of pedestrian interactions and the micro-level spatial-temporal relationships when handling the interaction information in spatial dimension and temporal dimension. The model employs a spatial-temporal attention-based fusion graph convolutional framework to predict future pedestrian trajectories. For the different types of local and global relationships between pedestrians, it first employs spatial-temporal attention mechanisms to capture dependencies in pedestrian sequence data, obtaining the social interactions of pedestrians in spatial contexts and the movement trends of pedestrians over time. Subsequently, a fusion graph convolutional module merges the temporal weight matrix and the spatial weight matrix into a spatial-temporal fusion feature map. Finally, a decoder section utilizes Time-Stacked Convolutional Neural Networks to predict future trajectories. The final validation on the ETH and UCY datasets yielded experimental results with an Average Displacement Error(ADE) of 0.34 and an Final Displacement Error(FDE) of 0.55. The visualization results further demonstrated the rationality of the model.

Key words: pedestrian trajectory prediction; micro-level spatial-temporal relationship; spatial-temporal attention; fusion graph convolution; Time-Stacked Convolutional Neural Network

1. INTRODUCTION

Pedestrian trajectory prediction essentially involves analyzing and extracting historical trajectory features of pedestrians, and then predicting their future movement directions and paths. Pedestrian trajectory prediction is crucial in many fields. For example, accurately predicting pedestrian movements in traffic forecasting can improve traffic efficiency and reduce congestion[1][2]. In smart city design, optimizing city layout can be achieved by predicting pedestrian movement trajectories. Additionally, in security monitoring [3], real-time prediction of pedestrian behavior can help promptly detect abnormal behavior or potential safety risks. Therefore, trajectory prediction can advance the development of smart cities, making this research highly significant for practical applications and of great academic value.

The challenge of pedestrian trajectory prediction lies in simultaneously capturing both pedestrian-pedestrian interactions and pedestrian-environment interactions. Before making predictions, it is essential to comprehensively analyze the relationship between time points and spatial positions.As shown in Fig. 1, there is a correlation between the pedestrian's current time point and position, which necessitates assessing the influencing factors within the spatial-temporal context. This increases the complexity of pedestrian trajectory prediction.

Early trajectory prediction methods based on mathematical and physical principles [4][5][6], including Gaussian process regression and kinematic methods, typically focused on short-term predictions of individual trajectories, overlooking pedestrian interactions. Subsequently, researchers turned to deep learning techniques for modeling human trajectories, primarily

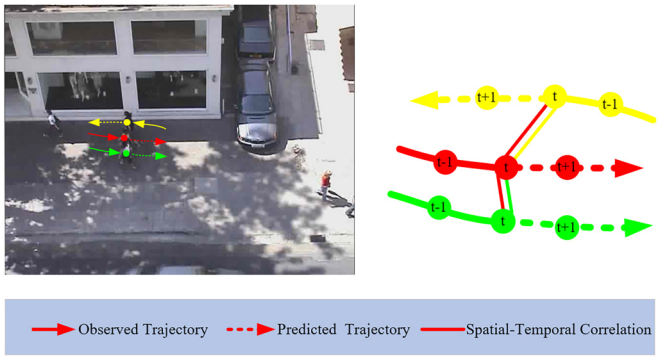

Fig. 1. Pedestrian interaction diagram

using recurrent neural networks[7], long short-term memory networks, and convolutional neural networks [8][9][10][11]. Later, researchers proposed utilizing pooling to gather potential states of pedestrians in the current scene[12]. Many scholars also employed generative adversarial networks to address challenges in behavior inference and uncertainties in future choices[13][14][15]. Early models mostly relied on recurrent structures, which suffered from low training efficiency and high costs[16]. Many models utilizing aggregation layers have been insufficient in intuitively representing physical features among pedestrians. The subsequent article uses graphs to simulate pedestrian movement scenarios[17], which is more fitting for describing pedestrian scenarios than aggregation, but lacks sufficient representation of social aspects. To better leverage graph representations, scholars proposed the Social-STGCNN model[18], which modeled pedestrian scenes as spatial-temporal graphs replacing aggregation layers and used kernel functions to define influences among pedestrians.

*e-mail: 472220529@stu.lntu.edu.cn

The following model builds upon this by using graph attention mechanisms to calculate the weights representing the mutual influence between pedestrians[19], yet accurately representing local and global relationships among pedestrians remains a significant challenge.

In the early trajectory prediction, only the interactions among local pedestrians were considered, while ignoring the motion trends of pedestrians in the distance. The later approach established a holistic model, employing the same mechanism for pedestrians in both spatial and temporal dimensions. However, it failed to fully consider the micro-level changes of individual pedestrians across different dimensions and did not extract interaction weights at multiple levels; Trajectory information between spatial and temporal is often strongly connected; therefore, after obtaining interaction information, we need to analyze pedestrian interactions under the fusion of spatial and temporal; when the model uses Temporal Convolutional Network(TCN) for prediction, there may be cases of insufficient feature extraction; Therefore, this paper proposes the following improvements:

1.After modeling pedestrian trajectories as trajectory graphs, the spatial-temporal attention mechanism applies a convolution operation to compute the weight information of pedestrians across different dimensions. The model dynamically updates the weight information matrix, enabling it to explore micro-level interactions between pedestrians from multiple perspectives.

2. The fusion graph convolution module is used to integrate spatial and temporal features, providing a more comprehensive understanding of the spatial-temporal structure and dynamic changes in the data. This generates more effective interaction feature representations, thereby enhancing the model's ability to understand and represent spatial-temporal data.

3. The decoder utilizes a Time-Stacked Convolutional Neural Network (TSCNN) to recognize and learn long-term dependencies within trajectory data. This enables the model to delve more deeply into learning detailed pedestrian trajectory feature representations and making predictions.

2. RELATED WORK

2.1. Pedestrian interaction model

The earliest model of crowd interaction was proposed by Helbing et al. known as the Social Force Model[20], which represents the attraction and repulsion between pedestrians using Langevin equations. After decades of refinement, experiments in some studies have validated that such models are not sufficiently accurate in representing real-world crowd interactions. Subsequent models such as discrete choice models [21] and continuous dynamics models [22], which integrate mathematical and physical principles, also suffer from insufficient accuracy in trajectory prediction.The integration of deep learning methods with trajectory prediction has improved accuracy[23]. Alahi et al. encode pedestrian interactions as "social" descriptors[24], while Xu et al. use spatial affinity to represent weights between pedestrians[25]. The Behavior-CNN model employs CNNs to model crowd interactions. Zhang et al. simulate neighbors' current intentions using an iteratively updated refinement module [26]. Mohamed et al. utilize kernel functions to extract pedestrian relationships in graph representations. Many studies have shown that graph attention mechanisms can better encode trajectory data, effectively aggregating features of neighboring nodes[27][28]. The AST-GNN model utilizes attention mechanisms to extract agent interactions within spatial-temporal graphs. Subsequent studies have employed self-attention mechanisms to calculate the temporal and spatial interactions among pedestrians[29][30][31]. However, existing models still fail to adaptively consider different-dimensional interactions between pedestrians. To better utilize attention mechanisms in computing interaction matrices, this model employs spatial-temporal attention mechanisms to learn the temporal and spatial relationship weights between nodes. This allows the model to adapt better to different tasks and data distributions, reducing overfitting to specific attention weights. This mechanism flexibly and dynamically extracts information from different dimensions, enhancing understanding of pedestrian behavior in complex environments.

2.2. Graph network in trajectory prediction

In trajectory prediction tasks, the model needs to analyze sequential data in both spatial and temporal dimensions. Sequences can be represented using a graph structure with nodes and edges, where nodes correspond to pedestrians and edges represent interactions between pedestrians. Huang et al. utilized attention mechanisms to extract spatial interactions and employed an LSTM model to capture temporal dimension information[32]. However, different modeling approaches impose limitations on the model's ability to handle different dimensions. Subsequent research has adopted a spatial-temporal graph frameworks for trajectory modeling[33], simulating interactions between pedestrians in the spatial dimension and modeling each pedestrian's historical trajectory in the temporal dimension, as seen in action recognition [34], traffic prediction [35], etc. However, these models do not consider the joint relationships between temporal and spatial dimensions. Our model addresses this by incorporating a fusion module within the spatial-temporal graph framework to analyze the microscopic connections between temporal and spatial dimensions. Additionally, it employs a Time-Stacked Convolutional Neural Network for specific step-length trajectory predictions, thus enhancing its application in complex and dynamic real-world scenarios.

3. PROBLEM FORMULATION

Given the historical observed trajectory positions of N pedestrians from the initial time to time T_o as $tr_o^n = \{\mathrm{P}_t^n = (\mathrm{X}_t^n, \mathrm{Y}_t^n) \,|\, t \in \{1, \ldots, T_o\}\}$, the model needs to predict the position $tr_p^n = \left\{\hat{\mathrm{P}}_t^n = \left(\hat{\mathrm{X}}_t^n, \hat{\mathrm{Y}}_t^n\right) \,|\, t \in \left\{1, \ldots, T_p\right\}\right\}$ at time T_p, the predicted position $(\hat{\mathrm{X}}_t^n, \hat{\mathrm{Y}}_t^n)$ represents the probability distribution random variable of pedestrian N's coordinates at time t. Assuming observed positions follow a bivariate Gaussian distribution $\mathrm{P}_t^n \sim \mathcal{N}(\mu_t^n, \sigma_t^n, \rho_t^n)$, the predicted pedestrian tra-

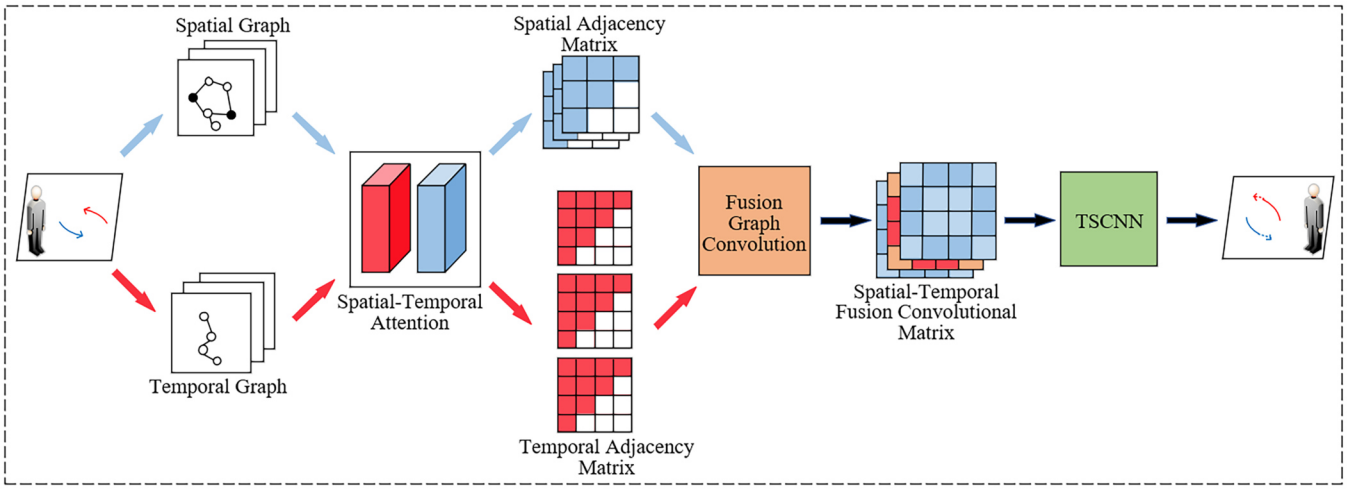

Fig. 2. Spatial-Temporal attention-based fusion graph convolutional model framework diagram

jectories also adhere to this distribution $\hat{\mathrm{P}}_t^n \sim \mathcal{N}(\hat{\mu}_t^n, \hat{\sigma}_t^n, \hat{\rho}_t^n)$. To achieve minimization of the negative log-likelihood loss function for this model:

$$L^n(\mathbf{W}) = -\sum_{t=1}^{T_p} \log(\mathbb{P}(\mathbf{p}_t^n \mid \hat{\mu}_t^n, \hat{\sigma}_t^n, \hat{\rho}_t^n)) \tag{1}$$

Where $\hat{\mu}_t^n$ denotes the mean,$\hat{\sigma}_t^n$ denotes the variance,$\hat{\rho}_t^n$ denotes the correlation of the distribution, and $\mathbf{W}$ represents the learned network parameters.

4. ARCHITECTURE OVERVIEW

The paper introduces the STAFGCN model, which employs an encoder-decoder architecture as depicted in Fig. 2. The encoder consists of two key modules: (1) the spatial-temporal attention module, which extracts interaction weights of pedestrians through spatial-temporal attention mechanisms, encompassing temporal motion trends and spatial social interactions; (2) the fusion graph convolution module, which utilizes fusion graph convolution to capture the spatial-temporal correlations within pedestrians' complex interactions. The decoder utilizes a Time-Stacked Convolutional Neural Network, focusing on predicting long-term future trajectories. Thus, the overall structure of the model is made more complete, enabling it to predict pedestrian movement trajectories with greater accuracy.

4.1. Graph representation of pedestrian trajectories

Due to the sparsity of raw trajectories and the advantageous ability of graph structures to capture complex correlations in sequential information, therefore, pedestrian trajectory data is transformed into graph structures for representation. The original data consists of the coordinate positions of N pedestrians observed in the scene over the past T_o time steps, The size of the input tensor is represented as $(N \times T_o \times 2)$.

1.Spatial graph representation

First, the input pedestrian position information is constructed into a set of spatial graphs $G_s = (V_s, E_s)$, representing the relationships and states among pedestrians at time t. The nodes are denoted as $V_s = \left\{v_s^i | \forall i \in \{1,\ldots,N\}\right\}$, where v_s^i represents the position information (x_t^i, y_t^i) at time t. E_s represents the edge set information of the graph, denoted as $E_s = \left\{e_s^{ij} | \forall i,j \in \{1,\ldots,N\}\right\}$. If there is interaction between two edges, then $e_s^{ij} = 1$; otherwise, $e_s^{ij} = 0$. Then, the spatial-temporal attention mechanism is utilized to obtain the weighted adjacency matrix of the nodes.

2.Temporal graph representation

Modeling pedestrians along the temporal dimension, constructing the temporal graph G_t for the n-th pedestrian. Using the temporal graph $G_t = (V_t, E_t)$, pedestrians' relative positions at different time steps are represented. $V_t = \{v_i^t | \forall i \in \{1,\ldots,T_o\}\}$ denotes the node information of pedestrian positions, where (x_i^t, y_i^t) signifies the coordinates of pedestrian n at time t. The edge set information of the temporal graph is denoted as $E_t = \left\{e_{ij}^t | \forall i,j \in \{1,\ldots,T_o\}\right\}$, while e_{ij}^t indicates the interactions between pedestrians at time t. Subsequently, through spatial-temporal attention mechanisms, more accurate correlations between nodes are captured, providing the model with more precise input features.

4.2. Spatial-temporal attention

The model employs a spatial-temporal attention mechanism to perform feature extraction on the graph structure. Due to the various factors affecting pedestrian movement direction, it's necessary to analyze the diversity of pedestrian motion patterns. This mechanism can model temporal and spatial dependencies from different perspectives, without relying on a fixed weighted adjacency matrix, thereby improving the model's capability to adapt to diverse types of data. The structure is depicted in Fig. 3.

Performing spatial-temporal attention mechanisms on different graph representations, compute the motion trends in the temporal dimension and the social interactions in the spatial dimension. Traditional attention models utilize linear mappings

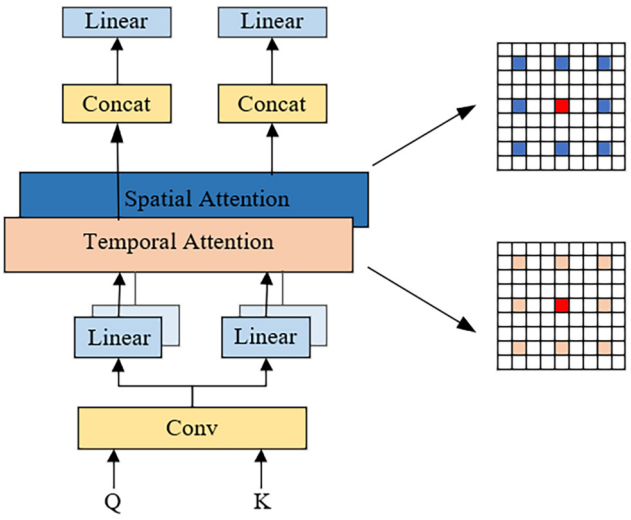

Fig. 3. Spatial-temporal attention architecture diagram

to generate discrete input vectors. However, spatial-temporal attention employs convolutional operations to generate learnable parameters Q and K, producing local evolutionary features. Here, spatial attention and temporal attention are used in parallel to compute pedestrian interaction weights along the temporal and spatial dimensions, respectively. The calculation process of attention coefficients between pedestrians is as follows:

$$E = \varphi(G, W_1) \tag{2}$$

$$Q = conv(E, W_2) \tag{3}$$

$$K = conv(E, W_3) \tag{4}$$

Where G represents the input temporal or spatial graph representation, combined with dynamically changing weight matrix W, we compute the corresponding spatial dimension queries Q_s and keys K_s, as well as temporal dimension Q_t and K_t.

$$a_s = \frac{exp(S(K_s, Q_s))}{\sum_j exp(S(K_s, Q_s))} \tag{5}$$

$$a_t = \frac{exp(S(K_t, Q_t))}{\sum_j exp(S(K_t, Q_t))} \tag{6}$$

Where $S(\cdot)$ denotes the function that computes correlations,j denotes all neighboring nodes, a_s is the normalized spatial attention coefficients, and a_t is the temporal attention coefficients. The attention coefficients at different time points are concatenated to form the temporal adjacency matrix A_t, and use this method to obtain the spatial adjacency matrix A_s for different pedestrians' attention coefficients, , with information propagation and feature updates conducted simultaneously in both spatial and temporal dimensions.

4.3. Fusion graph convolution module

This module consists of two operations. First, perform a fusion operation on the temporal features and spatial features to obtain different fusion matrices. In the next step, the obtained fusion feature matrices are input into a graph convolutional neural network to generate a spatial-temporal fusion convolutional matrix. The specific operations are shown in Fig. 4.

1.Feature fusion

Using the concept of global attention mechanisms, perform feature extraction and fusion operations on temporal and spatial adjacency matrices. First, input weighted adjacency matrices of different dimensions, apply pooling operations to the spatial weighted adjacency matrix to extract features, and then use an activation function to generate the spatial weight matrix.

$$A_{sp} = Sigmoid(Maxpooling(A_s) \cdot W_m + Avgpooling(A_s) \cdot W_a)) \tag{7}$$

Where A_{sp} is the obtained spatial weight matrix, W_m and W_a are the weight during the pooling operation.

The corresponding temporal weighted adjacency matrix is multiplied element-wise with the self-connected spatial weight matrix to obtain the spatial-temporal fusion matrix, while the temporal-spatial fusion matrix is derived by performing a dot product between the spatial weighted adjacency matrix and the self-connected temporal feature matrix.

$$A_{s-t} = A_t \odot (A_{sp} + I) \tag{8}$$

Where A_{s-t} is the spatial-temporal fusion matrix obtained after the dot product $\odot$.

The spatial weighted adjacency matrix is then concatenated with the spatial-temporal fusion matrix, followed by a softmax operation to produce the spatial-temporal fusion aware matrix.

$$R_{s-t} = softmax(cat(A_s, A_{s-t})) \tag{9}$$

Where R_{s-t} is the spatial-temporal fusion perception matrix. The temporal adjacency matrix A_t undergoes similar fusion convolution operations to sequentially produce the temporal feature matrix A_{te} , the temporal-spatial fusion matrix A_{t-s} , and the temporal-spatial fusion aware matrix R_{t-s}.

2.Graph convolutional network

The above-generated fusion aware matrix and the feature map are input into the two-layer graph convolutional network to produce the spatial-temporal fusion convolution matrix, as illustrated in Eq. (10).

$$F_{ST} = \delta(R_{t-s}\delta(R_{s-t}G_sW_s)W_t) + \delta(R_{s-t}\delta(R_{t-s}G_tW_t)W_s) \tag{10}$$

Where F_{st} represents the spatial-temporal fusion convolutional matrix of pedestrians, δ denotes the corresponding activation function, G_s is the spatial graph representation, G_t indicates the temporal graph, and W_s and W_t are trainable linear transformation matrices.The output of graph convolution operations are summed together to obtain the output features, which are the spatial-temporal fusion convolutional matrix.

4.4. Time-Stacked Convolutional Neural Network

In the decoder, a Time-Stacked Convolutional Neural Network is used to predict future trajectories. The spatial-temporal fusion convolutional matrix generated by the encoder serves as the input to the decoder. Future trajectories are generated through a series of feature transformations and time-stacked

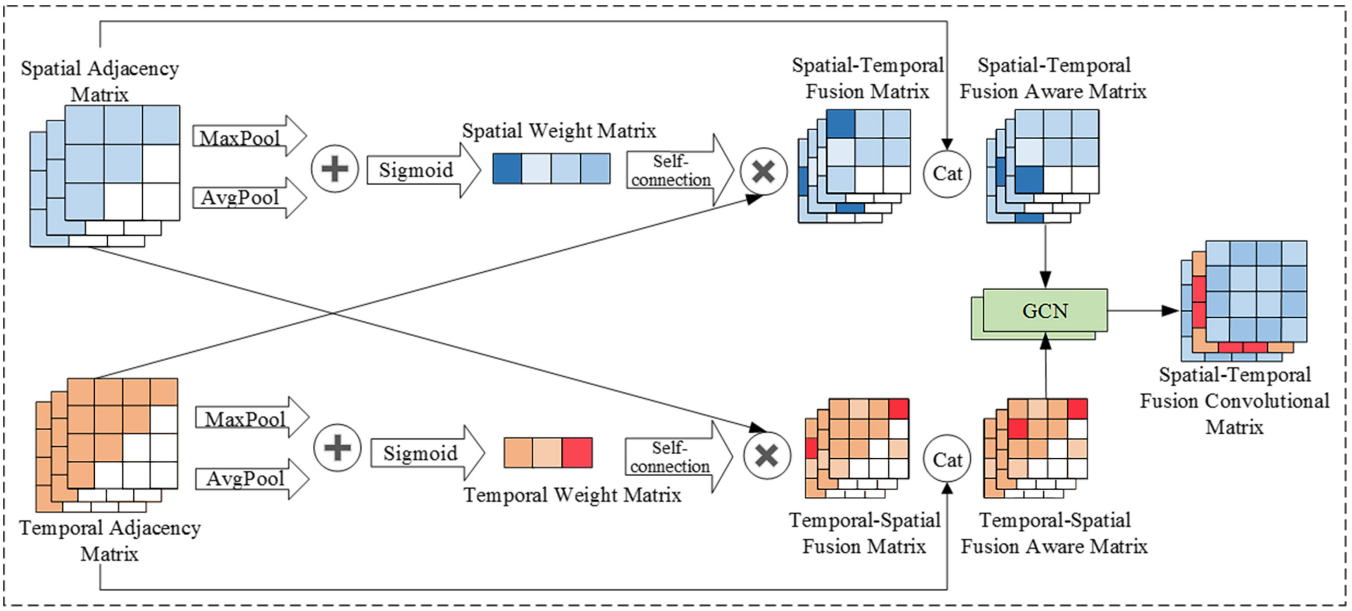

Fig. 4. Fusion graph convolutional structure diagram

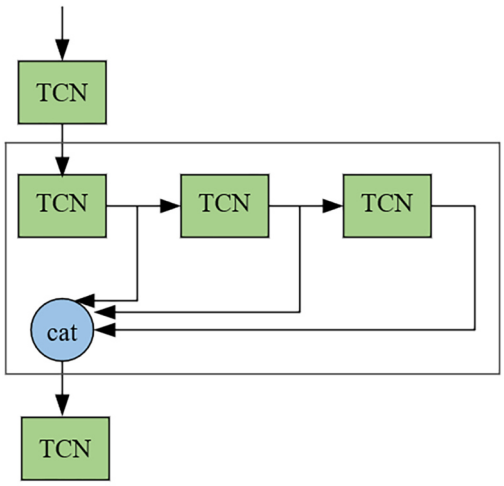

Fig. 5. Time-Stacked Convolutional Network architecture diagram

convolution operations. To fully leverage the trajectory features extracted at each TCN layer and capture feature information at different temporal scales, the features are first input into the initial TCN layer, where they are mapped from 8 dimensions to 12 dimensions. The features are then passed to the subsequent TCN layers to extract trajectory features, and the features learned by the three TCN layers below are stacked together. Finally, they are input into the final TCN layer for more precise feature extraction and prediction. By learning from historical time series data, the network generates position features $\hat{\mathbf{P}}_t^n$ that adhere to a bivariate Gaussian distribution. The features consist of predicted means and covariance matrices. The network architecture is shown in Fig. 5.

The network performs multiple layers of TCN to extract features, The concatenation of features allows the network to better focus on significant regions and more relevant neighboring pedestrians. Compared to previous methods, the network settings aggregated by TSCNN are more conducive to parameter optimization and relationship extraction, enhancing the model's gradient propagation capability and improving deep learning efficiency.

5. EXPERIMENTS AND RESULTS ANALYSIS

The model is implemented on the PyTorch framework, utilizing Adam as the optimizer. Training is configured for 300 epochs, with a batch size of 128 per epoch. The initial learning rate is set to 0.01, with a decay of 0.001 every 100 steps.

5.1. Datasets

ETH[37] and UCY[38] datasets are derived from real street surveillance videos, containing Overhead View and 2D positions of each pedestrian. The ETH dataset comprises two scenes: ETH and HOTEL. The ETH dataset captures pedestrian trajectories from the top floor of the ETH central building, overlooking pedestrian pathways, while the HOTEL dataset records pedestrian trajectories from the fourth floor of a hotel, also overlooking pedestrian pathways. The UCY dataset includes UNIV, ZARA1, and ZARA2. The UNIV dataset depicts scenes from a road within a university campus. ZARA1 and ZARA2 datasets capture pedestrian movements passing by the entrance of ZARA clothing stores. During training and evaluation, similar to other baseline methods, the model employs the preceding 8 frames as observation data to predict pedestrian trajectory information for the subsequent 12 frames.

5.2. Evaluation metrics

The performance of the model is assessed using two trajectory error metrics: Average Displacement Error (ADE) and Final

Table 1. Comparison of trajectory prediction results on ADE/FDE metrics, where ADE and FDE evaluation metrics are measured in meters and world coordinates

Model	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Social-LSTM[12]	1.33/2.94	0.39/0.72	0.82/1.59	0.62/1.21	0.77/1.48	0.79/1.59
SR-LSTM [26]	0.63/1.25	0.37/0.74	0.51/1.10	0.41/0.90	0.32/0.70	0.45/0.94
S-GAN-P[13]	0.87/1.62	0.67/1.37	0.76/1.52	0.35/0.68	0.42/0.84	0.61/1.21
PIF[36]	0.73/1.65	0.30/0.59	0.60/1.27	0.38/0.81	0.31/0.68	0.46/1.00
STGAT[32]	0.68/1.29	0.68/1.40	0.57/1.29	0.29/0.60	0.37/0.75	0.52/1.07
Social-STGCNN[18]	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
AST-GNN[19]	0.66/1.02	0.37/0.61	0.46/0.83	0.32/0.52	0.28/0.45	0.42/0.69
SGCN[29]	0.63/1.03	0.32/0.55	0.37/0.70	0.29/0.53	0.25/0.45	0.37/0.65
STIGCN[30]	0.58/0.96	0.30/0.44	0.38/0.67	0.28/0.47	0.23/0.42	0.35/0.59
RDGCN[31]	0.58/0.94	0.30/0.45	0.35/0.65	0.28/0.48	0.25/0.44	0.35/0.59
STAFGCN(ours)	0.56/0.89	0.31/0.45	0.37/0.62	0.26/0.40	0.22/0.41	0.34/0.55

Displacement Error (FDE). ADE determines the mean distance between each predicted position and the corresponding ground truth position, reflecting the model's average performance over the entire prediction sequence. FDE represents the distance between the predicted final position and the actual final position, with a particular emphasis on evaluating the model's accuracy in predicting the endpoint of the trajectory. By integrating ADE and FDE, a thorough assessment of the model's effectiveness can be achieved. The computation methods are shown in Eq. (11) and Eq. (12):

$$ADE = \frac{\sum_{n\in N}\sum_{t\in T_p} \|\hat{p}_t^n - p_t^n\|_2}{N \times T_p} \tag{11}$$

$$FDE = \frac{\sum_{n\in N} \|\hat{p}_t^n - p_t^n\|_2}{N}, t = T_p \tag{12}$$

5.3. Quantitative Analysis

Table 1 presents the comparative analysis results of errors, indicating that our model demonstrates better performance when compared with various traditional and advanced models, This suggests that our model exhibits higher efficiency and accuracy in handling spatial-temporal interaction information. Regarding the ADE metric, the error outperforms that of the previous best-performing baseline model, showing a 3% improvement. In terms of the FDE metric, the model also shows a significant reduction in error, with a 15% improvement in accuracy compared to the SGCN, outperforming the RDGCN model by 7%, and achieving a 17% improvement in prediction accuracy on the ZARA1 dataset. The results demonstrate that our model achieves the best performance on most datasets. Although it performs slightly lower than other models on some datasets, our model exhibits a higher level of performance in pedestrian trajectory prediction across the majority of datasets. This further proves the effectiveness of the model's predictions, even in densely populated pedestrian movement scenarios, where it maintains high prediction accuracy.

5.4. Ablation study

These experiments systematically remove or modify parts of the model to evaluate the impact of different components on the overall performance.

1.The effectiveness of each module

The experiment validates the contribution of different modules in improving the model's performance. (1) STA represents the model that uses spatial-temporal attention to capture pedestrian interactions; (2) FGCN is the model that incorporates a fusion graph convolution module; (3) TSCNN is the model that utilizes a Time-Stacked Convolutional Neural Network for prediction. Table 2 results indicate that each module enhances the prediction accuracy to varying degrees. After incorporating spatial-temporal attention, the model more accurately captures pedestrian interactions across different dimensions, resulting in errors smaller than those of the original model, with a particularly noticeable improvement in final displacement error. Fusion of temporal and spatial dimensions improves prediction accuracy by 3% and 9%, making predictions more aligned with actual trajectories. In the encoder, TSCNN further enhances prediction accuracy. By simultaneously considering various types of pedestrian interactions, the model more accurately captures the relative importance among pedestrians, adapts to diverse interaction scenarios.

2.The effectiveness of fusion graph convolution

The experiment investigates the impact of fusion graph convolution on model performance. (1) Base refers to the model without the fusion graph convolution module; (2) S-T uses only the spatial-temporal fusion aware matrix in graph convolution for interaction modeling; (3) T-S denotes modeling pedestrian trajectories using only the temporal-spatial fusion feature. From Table 3, it can be seen that the modeling approach is related to the performance of the model. Utilizing spatial features to enhance spatial-temporal interaction modeling improves the model's ability to extract trajectory features that vary with space, modeling pedestrian social trends has led to a 3% improvement in ADE. Incorporating temporal features assists in modeling temporal- spatial interactions, allowing the model to focus on changes in pedestrian movement, and modeling movement trends has reduced the FDE error by 3%. This

Table 2. Ablation study of different modules on the model's performance in ADE/FDE metrics

STA	FGC	TSCNN	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
			0.63/1.03	0.32/0.55	0.37/0.70	0.29/0.53	0.25/0.45	0.37/0.65
✓			0.60/0.99	0.34/0.49	0.39/0.71	0.27/0.46	0.22/0.42	0.36/0.61
	✓		0.59/0.94	0.33/0.47	0.37/0.64	0.29/0.49	0.22/0.43	0.36/0.59
		✓	0.61/1.01	0.33/0.52	0.38/0.69	0.28/0.47	0.23/0.43	0.37/0.62
✓	✓	✓	0.56/0.89	0.31/0.45	0.37/0.62	0.26/0.40	0.22/0.41	0.34/0.55

Table 3. The impact of spatial-temporal fusion operations on model performance in ADE/FDE metrics

Model	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Base	0.58/0.99	0.30/0.45	0.38/0.68	0.28/0.47	0.24/0.43	0.36/0.60
S-T	0.58/0.96	0.31/0.47	0.37/0.64	0.27/0.45	0.23/0.43	0.35/0.59
T-S	0.57/0.93	0.32/0.47	0.38/0.62	0.26/0.44	0.22/0.42	0.35/0.58
STAFGCN	0.56/0.89	0.31/0.45	0.37/0.62	0.26/0.40	0.22/0.41	0.34/0.55

model employs a combination of both fusion modeling methods simultaneously to fully extract the complex relationships within the graph structure, enhancing the diversity of pedestrian interactions. This results in a 6% improvement in ADE and an 8% improvement in FDE. It indicates that fusion graph convolution enables a deeper and multi-faceted exploration of pedestrian interaction relationships, thereby achieving better prediction results.

3.The effectiveness of the TSCNN architecture

The experiment investigates the impact of TCN structure on model performance. Shallow networks struggle to capture subtle changes in trajectories effectively, especially when extracting pedestrian interaction information in high-density scenes. However, an excessive number of layers can lead to overfitting, as the increase in layers also results in increased errors, leading to significant bias in the prediction results. Previous model data suggests that a five-layer TCN yields favorable prediction results, but simply using networks sequentially may impact the effectiveness of information extraction. (1) TSCNN1-4 uses the first TCN layer to process dimensions, sequentially employing four TCN layers to extract features and generate predicted trajectories; (2) TSCNN1-2-2 uses the first layer to process dimensions, stacks the output features from the second and third layers, and then feeds these features into the last two TCN layers; (3) TSCNN1-3-1 uses the first layer to process dimensions, stacks the output features from the middle three layers, and then inputs them into the final layer. As shown in Table 4, stacking the output features from three layers can better reduce prediction errors. Compared to the network structure without stacked TCN, it reduces ADE by 3% and FDE by 8%, and also outperforms the two-layer stacked network structure. Extracting trajectory information at different levels significantly enhances the model's ability to extract features, thereby improving prediction accuracy.

5.5. Model performance comparison

Table 5 provides a comparison of our model with other baseline models regarding parameter count and inference time, it can be seen that these model performance is significantly improved after overcoming the limitations of recurrent architecture and aggregation methods. The inclusion of spatial-temporal attention and fusion graph convolution modules in this model will incur some increase in computational workload and inference time. However, the parallel computation feature of the spatial-temporal attention mechanism results in an increase in inference time only due to the calculation of additional parameters, without significant time cost. The main increase in inference time occurs when fusing features. This indicates that while the model increases in complexity to enhance computational precision, it does not incur an excessive increase in time cost.

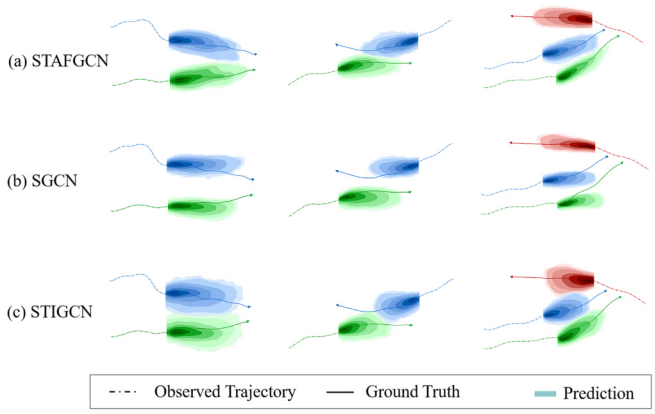

Fig. 6. Visualization of trajectory distribution

5.6. Visualization results

To demonstrate visually the practical performance improvements of the model enhancements, experiments utilized visualization to depict the predicted scenarios at the same moment in time.

Figure 6 shows the visualized trajectory distributions of different models. Each region represents the distribution range of the predicted trajectory mean for an individual pedestrian, with darker colors indicating a higher probability of the trajectory occurring at that location. The first column depicts a

Table 4. The impact of the TSCNN structure on the model's performance in ADE/FDE metrics

Model	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
TSCNN1-4	0.57/0.94	0.32/0.48	0.38/0.68	0.27/0.47	0.22/0.42	0.35/0.60
TSCNN1-2-2	0.58/0.95	0.31/0.47	0.37/0.65	0.27/0.43	0.22/0.42	0.35/0.58
TSCNN1-3-1	0.56/0.89	0.31/0.45	0.37/0.62	0.26/0.40	0.22/0.41	0.34/0.55

Table 5. Model Performance comparison

Model	Inference time(s)	Parameters count
SR-LSTM[26]	0.1758	64.9K
PIF[36]	0.1145	360.3K
Social-STGCNN[18]	0.0020	7.6K
SGCN[29]	0.0040	25.3K
STAFGCN(ours)	0.0045	29.7K

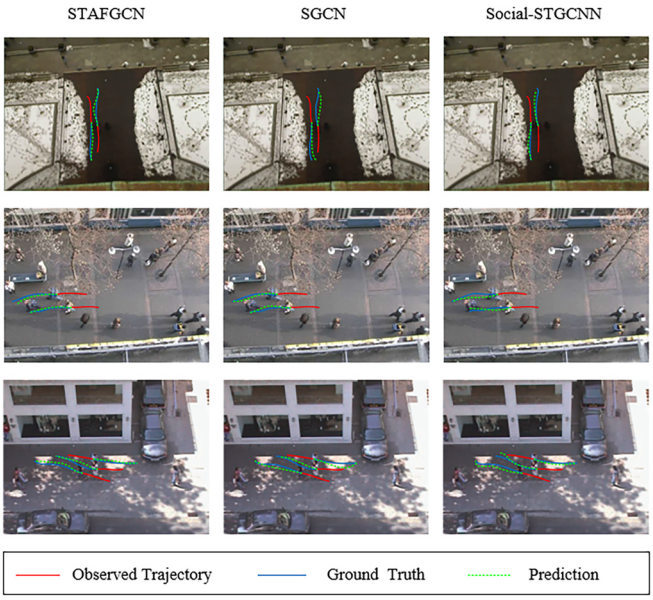

Fig. 7. Visualization of real-world scenarios

scenario with two parallel pedestrians, where our model exhibits minimal deviation and overlap in the trajectory distribution, resulting in a relatively accurate overall prediction effect. The second column features pedestrians walking towards each other, with the STAFGCN model generating the most reasonable avoidance behavior in this scenario. The third column illustrates a scenario with multiple interacting pedestrians. Our model's trajectory distribution closely matches the real-world scenario, while the SGCN model produces an overly sparse distribution. The STIGCN model, influenced by interaction perception, shows overlapping trajectory distributions, resulting in redundant avoidance behaviors. These results demonstrate the feasibility of the model's predictions across various contexts, showing that trajectory distributions are more accurate when dealing with complex interactions.

Figure 7 demonstrates the performance of three models in real-world scenarios. The first row shows a scenario with two pedestrians walking towards each other at an intersection. It can be observed that our model predicts the endpoints closest to the actual trajectories on a micro level, whereas the other two models exhibit inaccuracies in either speed or direction. The second row presents the results of predicting the trajectories of two pedestrians walking one behind the other, with a less influential pedestrian diagonally ahead. It can be seen that the models exhibit varying degrees of avoidance behavior, with our model achieving the smallest average displacement error. The third row depicts a scenario involving multiple interacting pedestrians, where two pedestrians walking together towards a store encounter pedestrians walking in the opposite direction. While SGCN and Social-STGCNN perform well in predicting individual pedestrian trajectories, our model's overall predicted trajectory is the closest to the real trajectory. This indicates that our model excels in capturing the micro-level spatial-temporal interactions between pedestrians, allowing for more accurate predictions of pedestrians' spatial-temporal movement trends. These results further validate the effectiveness and rationality of the model improvements.

6. CONCLUSION

This paper presents a pedestrian trajectory prediction model using a spatial-temporal attention-based fusion graph convolution network. To more accurately extract interaction relationships, the model employs a seq2seq framework. In the encoder, spatial-temporal attention is first used to analyze various types of microscopic relationships between pedestrians and their environment in both temporal and spatial dimensions, and then employs a fusion graph convolution module to extract spatial-temporal correlation information. In the decoder, TSCNN is utilized for a more comprehensive analysis and prediction of trajectories. The model underwent extensive experimental validation on multiple real-world pedestrian trajectory datasets, yielding superior results compared to other mainstream algorithms, with a 3% improvement in ADE and a 7% improvement in FDE. Visualizations comparing different models further confirm the effectiveness of the model improvements. In future research, we will further address the limitations of multi-modal interactions between pedestrians and various types of vehicles, and employ mathematical methods to enhance the safety decisions of our models. We hope to apply the model to complex urban traffic scenarios involving multiple pedestrians and vehicles, aiming to enhance its practicality and reliability.

ACKNOWLEDGEMENTS

This research is financially supported by the National Natural Science Foundation of China (61602056).

REFERENCES

[1] J. Hong, B. Sapp, and J. Philbin, "Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8454–8462, doi:10.1109/cvpr.2019.00865.

[2] M. Golchoubian, M. Ghafurian, K. Dautenhahn, and N. L. Azad, "Pedestrian trajectory prediction in pedestrian-vehicle mixed environments: A systematic review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 11 544–11 567, 2023, doi:10.1109/tits.2023.3291196.

[3] H. Xue, D. Huynh, and M. Reynolds, "Location-velocity attention for pedestrian trajectory prediction," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 2038–2047, doi:10.1109/wacv.2019.00221.

[4] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2, no. 3, doi:10.7551/mitpress/3206.001.0001.

[5] S. Lefèvre, C. Laugier, and J. Ibañez-Guzmán, "Exploiting map information for driver intention estimation at road intersections," in *2011 IEEE Intelligent Vehicles Symposium (IV)*, 2011, pp. 583–588, doi:10.1109/ivs.2011.5940452.

[6] R. Toledo-Moreo and M. A. Zamora-Izquierdo, "Imm-based lane-change prediction in highways with low-cost gps/ins," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 1, pp. 180–185, 2009, doi:10.1109/tits.2008.2011691.

[7] R. Korbmacher and A. Tordeux, "Review of pedestrian trajectory prediction methods: Comparing deep learning and knowledge-based approaches," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24 126–24 144, 2022, doi:10.1109/tits.2022.3205676.

[8] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, doi:10.1109/cvpr.2015.7298878.

[9] R. Emonet, J. Varadarajan, and J.-M. Odobez, "Extracting and locating temporal motifs in video scenes using a hierarchical non parametric bayesian model," in *CVPR 2011*, 2011, pp. 3233–3240, doi:10.1109/cvpr.2011.5995572.

[10] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft + hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *Neural Networks*, vol. 108, pp. 466–478, 2018, doi:10.1016/j.neunet.2018.09.002.

[11] S. Zamboni, Z. T. Kefato, S. Girdzijauskas, C. Norén, and L. Dal Col, "Pedestrian trajectory prediction with convolutional neural networks," *Pattern Recognition*, vol. 121, p. 108252, 2022, doi:10.1016/j.patcog.2021.108252.

[12] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, doi:10.1109/cvpr.2016.110.

[13] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, doi:10.1109/cvpr.2018.00240.

[14] J. Li, H. Ma, and M. Tomizuka, "Conditional generative neural system for probabilistic trajectory prediction," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 6150–6156, doi:10.1109/iros40897.2019.8967822.

[15] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, doi:10.1109/cvpr.2019.00144.

[16] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *CoRR*, vol. abs/1803.01271, 2018, doi:10.48550/arXiv.1803.01271.

[17] 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 *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/d09bf41544a3365a46c9077ebb5e35c3-Paper.pdf

[18] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, doi:10.1109/cvpr42600.2020.01443.

[19] H. Zhou, D. Ren, H. Xia, M. Fan, X. Yang, and H. Huang, "Ast-gnn: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction," *Neurocomputing*, vol. 445, pp. 298–308, 2021, doi:10.1016/j.neucom.2021.03.024.

[20] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E*, vol. 51, pp. 4282–4286, May 1995, doi:10.1103/physreve.51.4282.

[21] G. Antonini, M. Bierlaire, and M. Weber, "Discrete choice models of pedestrian walking behavior," *Transportation Research Part B: Methodological*, vol. 40, no. 8, pp. 667–687, 2006, doi:10.1016/j.trb.2005.09.006.

[22] A. Treuille, S. Cooper, and Z. Popović, "Continuum crowds," *ACM Transactions On Graphics (TOG)*, vol. 25, no. 3, pp. 1160–1168, 2006, doi:10.1145/1179352.1142008.

[23] S. Yi, H. Li, and X. Wang, "Pedestrian behavior understanding and prediction with deep neural networks,"

in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 263–279, doi:10.1007/978-3-319-46448-0$_1$6.

[24] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Socially-aware large-scale crowd forecasting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, doi:10.1109/cvpr.2014.283.

[25] Y. Xu, Z. Piao, and S. Gao, "Encoding crowd interaction with deep neural network for pedestrian trajectory prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, doi:10.1109/cvpr.2018.00553.

[26] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, doi:10.1109/cvpr.2019.01236.

[27] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, 2019, doi:10.1109/tgrs.2018.2864987.

[28] Q. Wang, T. Han, Z. Qin, J. Gao, and X. Li, "Multitask attention network for lane detection and fitting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 3, pp. 1066–1078, 2022, doi:10.1109/tnnls.2020.3039675.

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 8994–9003, doi:10.1109/cvpr46437.2021.00888.

[30] W. Chen, H. Sang, J. Wang, and Z. Zhao, "Stigcn: spatial–temporal interaction-aware graph convolution network for pedestrian trajectory prediction," *The Journal of Supercomputing*, pp. 1–25, 2023, doi:10.21203/rs.3.rs-3170302/v1.

[31] H. Sang, W. Chen, J. Wang, and Z. Zhao, "Rdgcn: Reasonably dense graph convolution network for pedestrian trajectory prediction," *Measurement*, vol. 213, p. 112675, 2023, doi:10.1016/j.measurement.2023.112675.

[32] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "Stgat: Modeling spatial-temporal interactions for human trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019, doi:10.1109/iccv.2019.00637.

[33] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018, doi:10.1609/aaai.v32i1.12328.

[34] L. Wu, C. Zhang, and Y. Zou, "Spatiotemporal focus for skeleton-based action recognition," *Pattern Recognition*, vol. 136, p. 109231, 2023, doi:10.1016/j.patcog.2022.109231.

[35] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "Multi-stage attention spatial-temporal graph networks for traffic prediction," *Neurocomputing*, vol. 428, pp. 42–53, 2021, doi:10.1016/j.neucom.2020.11.038.

[36] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5725–5734, doi:10.1109/CVPRW.2019.00358x.

[37] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 261–268, doi:10.1109/ICCV.2009.5459260.

[38] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Computer graphics forum*, vol. 26, no. 3. Wiley Online Library, 2007, pp. 655–664, doi:10.1111/j.1467-8659.2007.01089.x.