Fusion of Multi-scale Spatio-Temporal Features for Traffic Accident Risk Prediction

Qingrong Wang, Bingyan Huang, Changfeng Zhu, Runtian He

ABSTRACT—**Traffic accident risk prediction has significant implications for urban emergency response and traffic safety management. However, most existing prediction methods primarily focus on spatio-temporal characteristics within specific regions or time frames, often overlooking the complex interactions between global and local regions. To address this gap, we propose the Multi-Head Flow Attention Mechanism based Multi-Scale Spatio-Temporal Feature Map Convolutional Network Model (MFA-MSSTGCN). This model incorporates multi-scale spatio-temporal features of traffic accidents, considering both global and local scales. Firstly, the model uses a multi-view GCN to capture global spatial features. These are combined with a multi-head flow attention mechanism-based bidirectional gated recurrent unit (MFA-Bi GRU) to capture global temporal features. The multi-head attention mechanism dynamically selects important features while filtering out irrelevant information. This approach also avoids excessive computational complexity by employing a source competition and sink allocation mechanism. At the local scale, the model extracts spatial features from both the original grid data and high-resolution regions formed by aggregating neighboring areas. Temporal features are captured using gated temporal convolutions (gated-TCN). Additionally, We propose a sample-weighted loss function to tackle the zero-inflation problem arising from data sparsity. In the end, we carried out thorough testing on two datasets—NYC， Chicago. The outcomes reveal that MFA-MSSTGCN surpasses current models in RMSE, MAP, Recall, highlighting its superior performance.**

Index Terms—**traffic accidents; risk prediction; graph convolutional networks; attention mechanisms**

I. Introduction

With the cities grow and vehicle numbers rise, the likelihood of traffic accidents also increases. Predicting traffic accident risks is of significant practical value for enhancing traffic safety, reducing accidents, helping drivers choose optimal routes, and improving traffic network planning. However, predicting traffic accidents is affected by numerous complex factors, both internal and external, exhibiting clear spatio-temporal characteristics. Furthermore, accident occurrence patterns show clear spatio-temporal characteristics. A major difficulty lies in thoroughly exploring the relationships between these factors and creating a scientifically sound predictive model.

Traffic accident risk prediction involves various complex factors and requires the simultaneous consideration of spatio-temporal relationships between regions. Early studies primarily including HA [1], ARIMA [2], etc. However, these studies require significant computational resources and rely heavily on specialized parameter settings, which limit their prediction accuracy. As machine learning emerges in traffic prediction, researchers have turned to methods such as SVM [3] and KNN [4] to address these limitations, enabling better handling of complex data. Nonetheless, these models still depend on manual feature extraction, which restricts their performance when dealing with large-scale and complex spatio-temporal data.

Recently, Deep Learning (DL) has gained popularity in traffic prediction, greatly improving the accuracy of spatio-temporal data forecasting. LSTM [5] and GRU[6], frequently integrated with other algorithms, are commonly applied to capture spatial and temporal patterns in traffic data [7] . Chang et al. [8] introduced the POA-BiGRU-CNN model for ship traffic prediction. This model employs adaptive hyper-parameter optimization through POA, enhancing both global search capability and convergence speed. Despite these advancements, these models still encounter gradient issues and weak generalization with long sequences and limited data, and they often show poor generalization when the available data is limited. To address these issues, researchers have introduced the Temporal Convolutional Network (TCN) [9][10], which captures medium- and long-term dependencies in data more effectively.

Existing studies primarily focus on temporal correlations, often overlooking spatial features, which significantly limits prediction performance. Although Convolutional Neural Networks (CNN) [11] effectively extract spatio-temporal features through convolutional operations, they are generally more suited for Euclidean spatial data (e.g., images and videos) and face limitations when dealing with graph-structured data with dynamic nodes[12]. In contrast, Graph Convolutional Networks (GCN) [13], [14] excel in learning node features and capturing neighbor information through multi-layer graph convolution, enabling effective representation of graph nodes.

Chen et al. [15] proposed Seq2Seq-TGCN, which effectively captures and predicts spatio-temporal traffic flow

Manuscript received February 14, 2025; revised August 16, 2025.

This work was supported in part by the National Natural Science Foundation of China (72161024); "Double First-Class" Major Research Project of the Education Department of Gansu Province (GSSYLXM - 04).

Qingrong Wang is a professor at School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China (e-mail: 329046272@qq.com).

Bingyan Huang is a postgraduate student at School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China (Corresponding author, e-mail:3303687948@qq.com).

Changfeng Zhu is a professor at School of Traffic and Transportation. Lanzhou Jiaotong University, Lanzhou 730070, China (e-mail: cfzhu003@163.com).

Runtian He is a doctoral student at School of Traffic and Transportation. Lanzhou Jiaotong University, Lanzhou 730070, China (e-mail: 1939700273@qq.com).

patterns. This framework dynamically extracts and integrates spatio-temporal sequence data from strongly correlated road sections, thus facilitating robust traffic pattern modeling. Ye et al. [16] proposed DSTHGCN, which combines GCN with hypergraph co-convolution (HGCN) and introduces a dynamic pruning mechanism for hyperedge outliers (HOR). By suppressing noise and optimizing structure, DSTHGCN effectively captures node-hyperedge interactions, enhancing traffic flow prediction accuracy. Gao et al. [17] developed the DS-STGCN for traffic flow prediction. which models traffic flow by integrating node features, topological correlations, and time-slot dynamics within a triple graph structure. This approach enables collaborative modeling and dynamic sensing of global-local dependencies. Yao et al. [18] proposed an Auto-STCN model that captures spatio-temporal dependencies in traffic flows. Ali et al. [19] and Zhao et al. [20] introduced a model combining LSTM, GCN, and Bi-GRU to capture the periodic patterns of traffic flow. Wu et al. [21] enhanced the long-term time series modeling capacity of their model by combining dilated causal convolution with graph convolution to construct Graph WaveNet (GWN). However, the influence of various times and spaces on regional traffic accidents varies significantly. To improve prediction accuracy, it is crucial to analyze how spatio-temporal features at different scales impact the prediction[22][23][24].

The attention mechanism in deep learning models helps the system focus on key features by assigning different weights to sample features. This enhances training efficiency and the ability to handle long time series. Fares Alhaek et al. [25] used global semantic and local geospatial temporal relationships to predict urban traffic accident risk, creating multiple maps to capture both static and dynamic spatial connections, while using an attention mechanism to selectively emphasize key spatio-temporal information. Wang et al. [26] employed a multi-view attention approach to identify spatial relationships within a knowledge graph, facilitating traffic accident risk prediction at both broad and detailed levels. Li et al. [27] developed the developed the Direction-Distance Sensitive Graph Transformer (DDGformer), which integrates a direction distance bi-domain self-attention mechanism and a dynamically augmented graph convolution module for dynamic spatio-temporal correlation modeling.

Yu et al. [28] proposed a fusion of the multi-head flow attention mechanism and spatio-temporal traffic flow features learned by a GCN; Patara Trirat et al. [29] modeled a dynamic and static graph in a heterogeneous environment. Guo et al. [30] use a self-attention approach to capture local context and a dynamic graph convolution module for spatio-temporal traffic flow forecasting. Fang et al. [31] applied a multi-head spatio-temporal attention method to improve the interaction of traffic state data across different time scales, combining location and semantic features for forecasting traffic flow. Chen et al. [32] combined the attention mechanism and graph convolution by fusing node dynamics and spatio-temporal features; Geng et al. [33] investigated the traffic flow based on the distance self-attention mechanism by considering the gated temporal self-attention Prediction theory system. Chen et al. [34] propose a multi-granular hierarchical spatio-temporal network that integrates remotely sensed data, captures spatial closeness and semantic relevance.

Building on this, this paper proposes a multi-scale spatio-temporal feature map convolutional network model (MFA-MSSTGCN) using a Multi-Head Flow Attention (MFA). The primary contributions of this framework are outlined below:

(1) At the global scale, we introduced the MFA mechanism, which replaces the traditional attention weight calculation method with a source competition and sink allocation mechanism, thereby reducing the complexity of the model. At the same time, we combined the Bi-GRU network to capture global temporal features and used multi-view GCN to learn global spatial features, further improving prediction performance.

(2) At the local scale, low-scale spatial features and high-scale spatial features aggregated from adjacent regions are extracted. The gated-TCN model is used to capture local temporal similarity, thereby enhancing the model's local spatio-temporal modeling capabilities.

(3) It dynamically fuses spatio-temporal features from both global and local scales and designs a sample-weighted loss function to address data sparsity issues.

II. Traffic Accident Influencing Factors and Spatio-Temporal Characterization

A. Analysis of Factors Affecting Traffic Accidents

A.1 Intrinsic factors

Intrinsic factors influencing traffic accidents mainly include road structure and functional layouts. Specifically, Specifically, the design of road intersections has a significant impact on the occurrence of accidents. In addition, spatial differences exist across regions. The distribution of Points of Interest (POIs) in different functional zones directly influences accident patterns. Accident patterns within the same time period differ notably among various functional zones. Also, the spatial clustering of multiple POIs complicates regional accident patterns further. Therefore, carefully analyzing how internal factors relate to traffic accidents is important. The interaction of POIs is demonstrated in Fig 1.

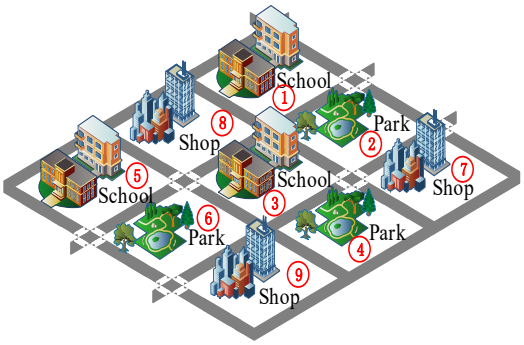
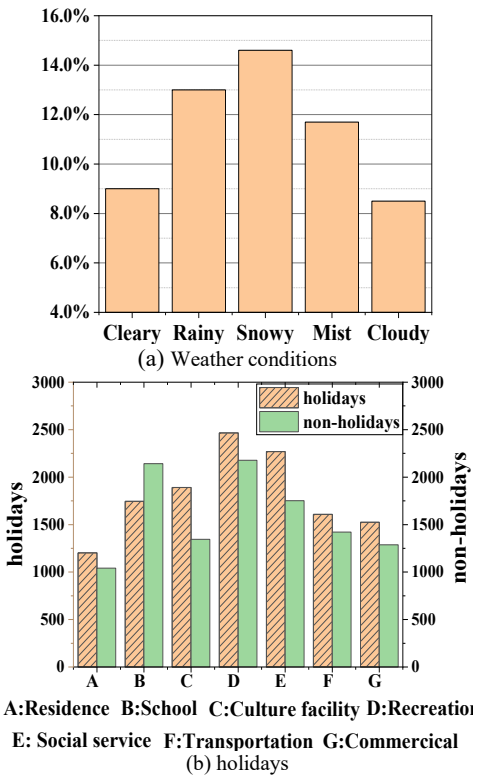


Fig. 1. The interaction of POIs

A.2 External factors

Weather conditions and holidays also play a crucial role in influencing the occurrence of road incidents. Rainy or snowy weather can lead to deteriorated road conditions or limited visibility for drivers; during holidays, the surge in traffic volume and increased traffic density in different

functional zones can also result in a higher frequency of accidents. These factors exhibit varying degrees of influence over time and space, leading to complex nonlinear and time-varying characteristics in accident occurrence patterns. By analyzing data from selected areas in Chicago, examining the proportion of traffic accidents under different weather conditions, and assessing the impact of holidays on accident volumes across various functional zones, we can more clearly elucidate the role of these external factors. The effects of weather conditions and holidays on traffic accidents are illustrated in Fig 2.



(a) Weather conditions

(b) holidays

Fig. 2. Impact of weather conditions and holidays on traffic accidents

B. Spatial- temporal characterization

B.1 Temporal characterization

Assuming the prediction of traffic accident risk at time t. the temporal correlation refers to the temporal auto-correlation of the variable at time t with the past k moments, i.e. $x_t = f(x_{t-1}, x_{t-2}, ..., x_{t-k})$. x_t is associated with the period from x_{t-1} to x_{t-k}, while also exhibiting similarities with historical data at different scales.

The neighboring time series exhibit low-scale proximity, with notable variations in traffic patterns on weekdays and weekends. Additionally, the accident pattern demonstrates a repetitive weekly regularity, indicating high-scale weekly periodicity. Therefore, the historical data are divided into two categories: the neighboring period of the predicted time and the weekly cycle period. That is

$$X_h = \left[X_h^{t-p}, X_h^{t-(p-1)}, ..., X_h^{t-1} \right] \quad (1)$$

$$X_w = \left[X_w^{t-q \cdot w_{weekly}}, X_w^{t-(q-1) \cdot w_{weekly}}, ..., X_w^{t-w_{weekly}} \right] \quad (2)$$

Where p represents the data of the time interval adjacent to the forecast period; q denotes the number of weekly cycle time segments, and w_{weekly} indicates the length of the weekly cycle series. The low-scale sequence data captures local variations, while the high-scale sequence data better reflects the overall trend of data changes, making it more suitable for mid- to long-term forecasts. Based on these, the time feature matrix of the region at time t is constructed. That is

$$X_{n \times d_t} = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^{d_t} \\ x_2^1 & x_2^2 & \cdots & x_2^{d_t} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \cdots & x_n^{d_t} \end{bmatrix} \quad (3)$$

Where $x_n^{d_t}$ represents the d_t-dimensional traffic accident time characteristics of node n at time t. Assuming the prediction period is from 8:00 to 9:00 am on March 30, 2023, the corresponding time series segment is constructed as illustrated in Fig 3.

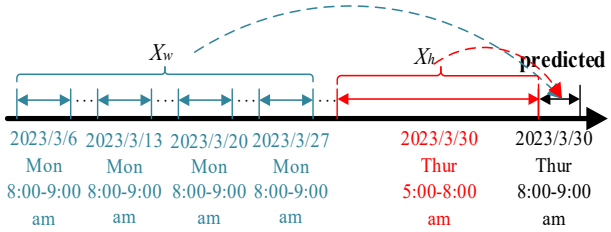


Fig. 3. Time series segment construction diagram

B.2 Spatial characterization

Spatial characteristics of traffic accidents include road network structures and implicit spatial relationships formed by factors such as Points of Interest (POIs) and road attributes. Local spatial dependencies occur in neighboring areas connected by roads. Global spatial dependencies result from similar road attributes and POIs, meaning distant areas with similar road features can have comparable accident patterns. Additionally, there is clear spatial heterogeneity across regions, so accident patterns within the same time period often differ greatly. However, areas with similar functional zones typically share highly similar accident patterns. By considering both broad and localized spatial-temporal relationships, more precise traffic accident risk predictions can be made. Fig 4 displays the global and local interaction diagrams.

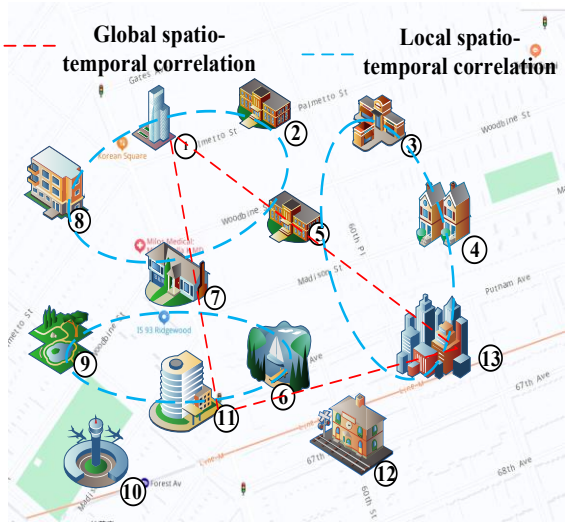


Fig. 4. The global and local interaction diagrams.

III. Model Construction

A. Description of the problem

City R is partitioned into $A \times B$ grid zones according to geographic coordinates, with each uniform section labeled as $R = \{r_{1,1},...,r_{i,j},...,r_{a,b}\}$. Among these, only N regions contain road and traffic accident data; the data for other regions are set to 0. Based on the historically observed eigenvalues of region $(X_1, X_2, ..., X_T)$, the graph signal matrices $(S_1, S_2, ..., S_T)$ and Z_{T+1} are used to predict the risk of traffic accidents in the next time interval $\hat{Y}_{T+1} \in R^{A \times B}$.

Where $\boldsymbol{X_t} \in R^{A \times B \times d_g}$ denotes the grid features of all regions at time interval t, and d_g denotes the feature dimensions of the nodes, including external factors such as traffic accident risk value, POI, and weather. Let $\boldsymbol{S_t} \in R^{N \times d_g}$ denote the graph signal matrix of the three graphs at time interval t. $Z_t \in R^{d_t}$ denotes the temporal features at moment t, including 24 hours a day, every day in a week, and whether it is a holiday or not. Owing to the varying scale influences on traffic accidents, the input data $X_1,...,X_q,...,X_T\,(T = p+q)$ is obtained by selecting the most recent p time and the same time intervals in the previous q weeks.

B. Traffic accident risk prediction model based on MFA-MSSTGCN

Fig. 5 presents the structure of the MFA-MSSTGCN model, comprising three main components: global spatio-temporal feature modeling with a multi-head flow attention mechanism, local spatio-temporal feature modeling through multi-scale fusion, and fusion of global and local spatio-temporal features. A multi-view Graph Convolutional Network (GCN) captures global spatial correlations. MFA-BiGRU models global temporal dependencies. For local spatio-temporal modeling, we begin by applying a multi-channel convolutional network to derive spatial information from the original grid data. We apply multi-layer deconvolution to extract spatial patterns from aggregated high-scale regions. Next, these spatial features are passed through a gated temporal convolutional network (gated-TCN) to model local temporal patterns. Since global and local scales contribute differently, we fuse them with a weighted mechanism. Additionally, A weighted loss function is introduced to handle the zero-inflation problem resulting from sparse data.

C. MFA-based full-domain spatio-temporal feature modeling

C.1 Full-domain spatial feature extraction based on multi-view GCN

1) Multi-view construction

Considering that the non-linear relationship between road conditions and POIs across regions introduces additional potential factors influencing traffic accidents, we construct the POI, the road , the accident risk similarity graphs: $\mathcal{G}_p = (V, E_p, A_p)$, $\mathcal{G}_R = (V, E_R, A_R)$, $\mathcal{G}_K = (V, E_K, A_K)$,to capture the latent dynamic information in traffic accidents. Here, V denotes the collection of nodes., where each node corresponds to a region containing roads. Where $|V| = N$, E represents the edges that establish connectivity among the nodes, $\boldsymbol{A} \in \boldsymbol{R}^{N \times N}$ represents the adjacency matrix of the graph $\mathcal{G}$, which a_{ij} denotes the spatial connectivity state between nodes v_i, v_j.We compute the similarity between two nodes using the $Jensen-Shannon(JS)$ divergence and subsequently construct the similarity graph. That is

$$Sim_p(i,j) = 1 - JS\left(R_P^i - R_P^j\right) \tag{4}$$

$$JS\left(R_P^i, R_P^j\right) = \frac{1}{2}\left(\sum_{k=1}^{K} R_p^i(k) log \frac{2R_p^i(k)}{R_p^i(k) + R_p^j(k)} + \sum_{k=1}^{K} R_p^j(k) log \frac{2R_p^j(k)}{R_p^i(k) + R_p^j(k)}\right) \tag{5}$$

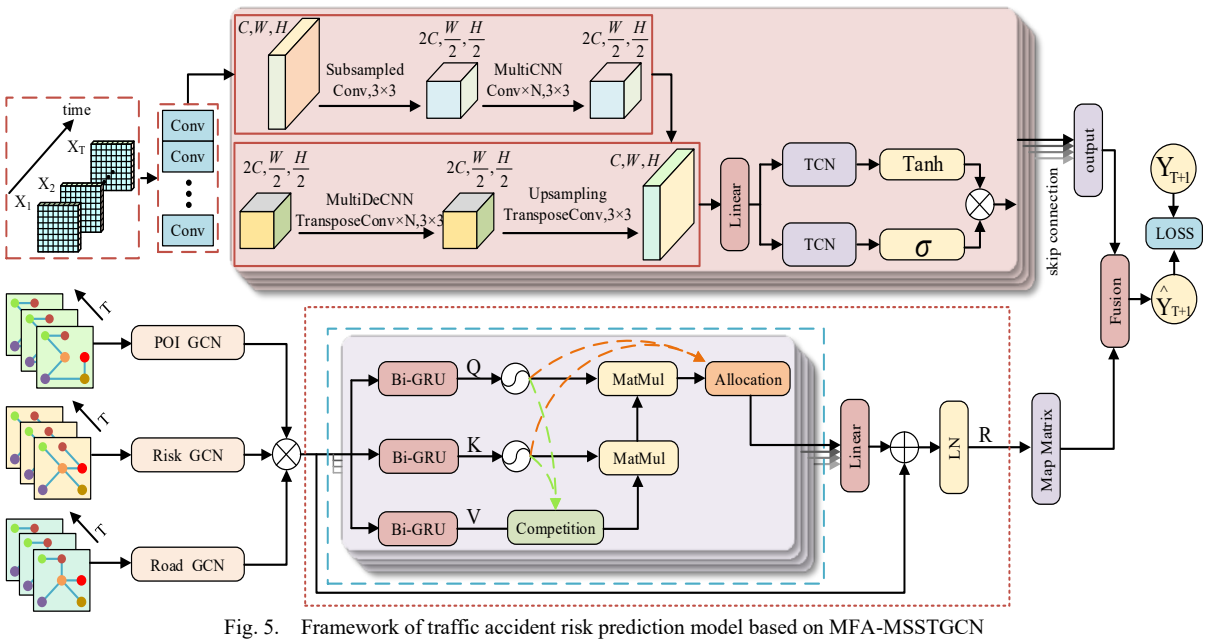


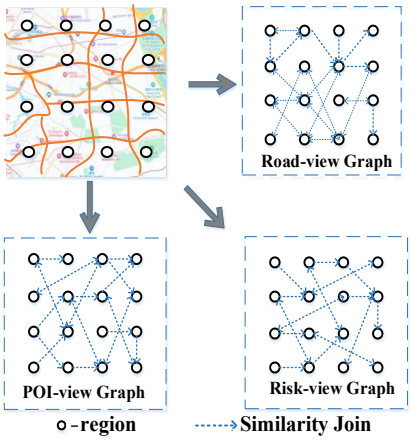
Fig. 5. Framework of traffic accident risk prediction model based on MFA-MSSTGCN

where R_P^i and R_P^j denote the POI distributions of region i and region j; $R_p^i(k)$ denotes the kth dimension of R_P^i. $Sim(i,j)$ denotes the degree of POI similarity for regions v_i and v_j. Road similarity graphs and accident risk similarity graphs are constructed following the same methodology. For each graph, we choose the L nodes with the highest similarity scores to build the corresponding adjacency matrix. Specifically, $A \in (A_P, A_R, A_K)$ is outlined below:

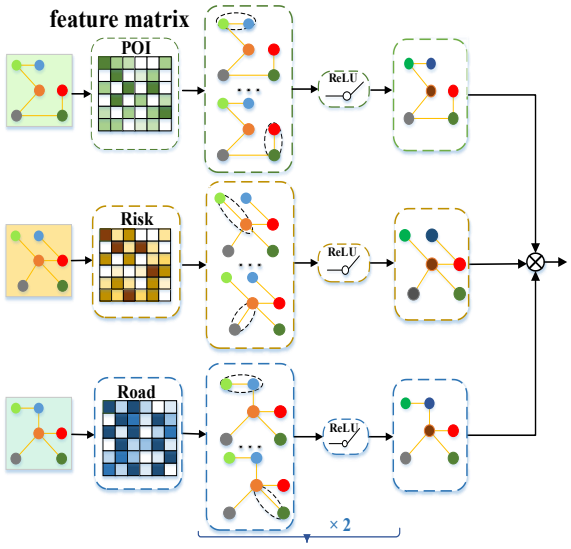
$$A_{i,j} = \begin{Bmatrix} Sim(i,j), e_{i,j} \in E \\ 0, else \end{Bmatrix} \tag{6}$$

2) Global Spatial GCN

The previously constructed views reveal that regions with similar features in the global domain may not be directly connected. The global domain representation learning process of the GCN is illustrated in Fig 6.



(a) Multi-view architecture based on road network topology



(b) Multi-view GCN

Fig. 6. Multi-view GCN full domain representation learning

In other words, higher-order structural relationships exist between these regions. To effectively capture these potential higher-order relationships in the global road network, we utilize a multi-view GCN. The input consists of three graph signal matrices $(S_1, S_2, ..., S_T)$ corresponding to different historical time periods. At each layer, we apply a graph convolution operation followed by an activation function to extract non-linear features from the data. Specifically, this process is outlined below:

$$S' = ReLu\left(AReLu\left(ASW^{(0)} + b^{(0)}\right)W^{(1)} + b^{(1)}\right) \tag{7}$$

Where A refers to the collection of three similarity matrices constructed. The global domain representation learning process of the GCN is illustrated in Fig 6.

C.2 *Full-domain temporal feature extraction based on MFA-Bi GRU*

1) *Bi GRU-based temporal feature extraction*

Traffic accidents exhibit strong temporal auto-correlation. However, their spatio-temporal interactions are dynamic, changing due to the spatial heterogeneity of urban functional areas. To capture the temporal correlations across the entire domain, We employ the Bi-GRU to model accident dynamics by utilizing both forward and backward information flows. It also adaptively learns temporal features, including high-frequency variations and long-term dependencies. The Bi-GRU is depicted in Fig. 7.

From Fig 7, we can see that the Bi GRU hidden output h_t at time t combines the forward GRU hidden output $\vec{h}_t$ and the backward GRU hidden output $\overleftarrow{h}_t$. That is:

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right) \tag{8}$$

$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right) \tag{9}$$

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

$$\tilde{h}_t = tanh\left(W_h \cdot [r_t \times h_{t-1}, x_t]\right) \tag{11}$$

$$\vec{h}_t = GRU_{forward}\left(x_t, \vec{h}_{t-1}\right) \tag{12}$$

$$\overleftarrow{h}_t = GRU_{backward}\left(x_t, \overleftarrow{h}_{t-1}\right) \tag{13}$$

$$h_t = \vec{h}_t \oplus \overleftarrow{h}_t \tag{14}$$

Where x denotes the continuous sequence information of the region; r_t、z_t denote the reset, update gate; $\sigma()$ and $tanh()$ are the activation functions; h_t is the hidden state; $\tilde{h}_t$ is the candidate hidden state; $\vec{h}_t$, $\overleftarrow{h}_t$ is the hidden state of the forward、backward GRU; $\oplus$ is the splicing of $\vec{h}_t$ and $\overleftarrow{h}_t$, to obtain the output of the final bidirectional GRU h_t.

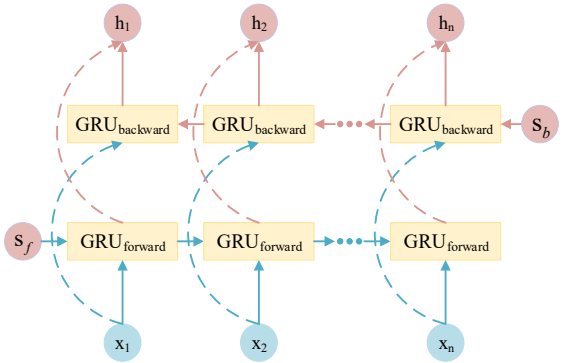


Fig. 7. Bi GRU structure

2) *Temporal MFA*

In Section C.1, we concatenate graph signal matrices from different time steps along the time dimension to create a higher-dimensional graph signal matrix $R^{N\times(T\times d_g)}$ as input. Although this approach captures temporal correlations, it fails to distinguish which specific time steps the features belong to. Consequently, it becomes challenging to dynamically and effectively model the influence of historical data on the prediction of future traffic accidents. To overcome this limitation, we introduce MFA mechanism. The MFA employs multiple parallel attention heads, assigning different weights to input sequences from various perspectives and focusing on the time-step information most relevant to traffic accidents. Furthermore, the MFA's source competition and sink allocation mechanisms reduce computational complexity, enhance feature screening, and minimize information redundancy.

Considering that the multi-layer nesting of the model may increase computational complexity, we replace the fully connected layer with Bi GRU for linear transformation. This generates the corresponding query Q, key K, and value V matrices. That is

$$\boldsymbol{Q}=\boldsymbol{W}_{\boldsymbol{Q}}\mathrm{h}_t,\boldsymbol{K}=\boldsymbol{W}_K h_t,\boldsymbol{V}=\boldsymbol{W}_V h_t \tag{15}$$

Where h_t is the Bi GRU output. Let there be a total of i sinks and j sources, and normalize the obtained $\boldsymbol{Q},\boldsymbol{K}$. That is

$$\frac{\varphi(\boldsymbol{Q})}{I} \tag{16}$$

$$\frac{\varphi(\boldsymbol{K})}{O} \tag{17}$$

Where $\varphi()$ is a non-negative non-linear transformation; $I\in R^{i\times 1}$ and $O\in R^{j\times 1}$ denote the inflow information flow of the sink and the outflow information flow of the source, respectively; $\frac{\varphi(\boldsymbol{Q})}{I}$ denotes the conservation of the sink and $\frac{\varphi(\boldsymbol{K})}{O}$ denotes the conservation of the source. The preservation of the outflow information flow of the source and the conservation of the inflow information flow of the sink are realized through normalization. That is

$$I'=\varphi(\boldsymbol{Q})\sum_{m=1}^{j}\frac{\varphi(\boldsymbol{K}_m)^T}{O_m} \tag{18}$$

$$O'=\varphi(\boldsymbol{K})\sum_{n=1}^{i}\frac{\varphi(\boldsymbol{Q}_n)^T}{I_n} \tag{19}$$

where $n\in\{1,2,...,i\}$, $m\in\{1,2,...,j\}$, $I'\in R^{i\times 1}$, $O'\in R^{j\times 1}$ denote the amount of conserved information flowing into and out of the information flow, respectively.

To implement information flow conservation in the source and sink of the MFA, we introduce a mechanism for information flow competition. We set the inflow information flow of each sink to 1, ensuring that the output information flow from the source competes for a unique position (Competition). Next, we apply maximum pooling to the GCN output in the time dimension to aggregate the information flow (Aggregation). By setting the output information flow of the source to 1, we enable the sinks to compete for a unique flow. Finally, we filter the inflow information flow of each sink to achieve sink allocation. That is

$$Competition: V'=soft\max(\boldsymbol{O}')\odot V \tag{20}$$

$$Aggregation: A=\frac{\varphi(\boldsymbol{Q})}{I}\left(\varphi(\boldsymbol{K})^T V'\right) \tag{21}$$

$$Allocation: R=LN\left(sigmoid(I')\odot A+H\right) \tag{22}$$

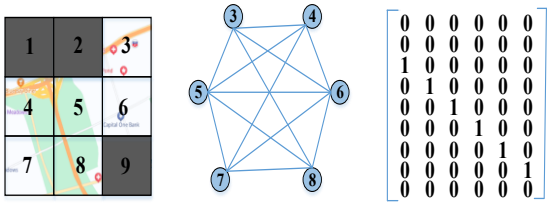
Where $\odot$ denotes element-wise multiplication, $softmax()$ and $sigmoid()$ are activation functions, and $LN()$ denotes layer normalization. Additionally, the residual mechanism in the MFA prevents network performance degradation and mitigates the gradient vanishing problem.

To be consistent with the output dimensions of the local spatio-temporal features, the dimensional transformation is performed via a mapping matrix $\boldsymbol{M}\in R^{A\times B\times N}$. That is

$$\hat{Y}_G=\boldsymbol{M}\boldsymbol{R} \tag{23}$$

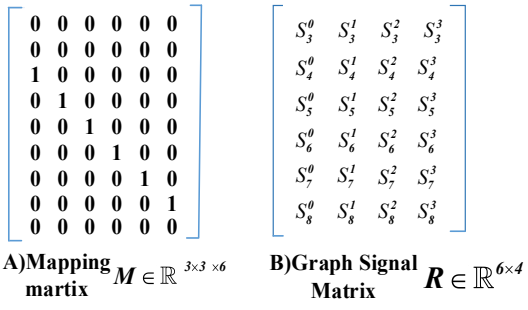
where $\boldsymbol{M}$ is the mapping matrix; $\boldsymbol{R}$ is the output of the multicurrent temporal attention; and $\hat{Y}_G$ is the output of the global spatio-temporal correlation module.

The primary process of mapping matrix transformation proceeds as follows. We divide the city into 3×3 grid regions numbered from 1 to 9, as shown in Fig 8, A). Regions 1, 2, and 9 contain no roads, while regions 3 to 8 include only a few nodes in the similarity map, as show in Fig 8, B).

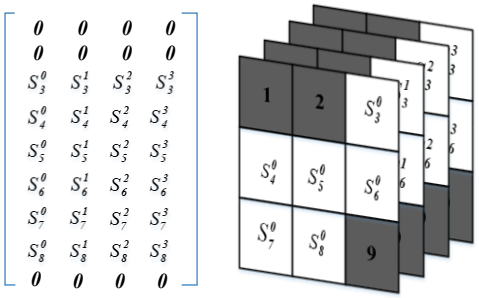


A) Grid data B) Graph data C) Mapping matrix $M\in R^{3\times3\times6}$

Fig. 8. Example of diagram-grid mapping matrix



A)Mapping martix $M\in\mathbb{R}^{3\times3\times6}$ **B)Graph Signal Matrix** $R\in\mathbb{R}^{6\times4}$



C) $\hat{Y}_G\in\mathbb{R}^{3\times3\times4}$ **D)** $\hat{Y}_G\in\mathbb{R}^{3\times3\times4}$

Fig. 9. Graph data-grid data conversion

The mapping matrix has a dimension of $\boldsymbol{M} \in R^{A\times B\times N}$, each row signifies a region, while each column signifies a node. If node n corresponds to region A, the matrix assigns the value $M_{A,n}=1$; otherwise, it assigns the value $M_{A,n}=0$.The graph signal matrix $R\in\mathbb{R}^{6\times 4}$ assigns each node a feature dimension of 4. We obtain $\hat{Y}_G \in \mathbb{R}^{3\times 3\times 4}$ by multiplying the feature matrix with the mapping matrix and then transforming the dimensions to produce the final output. The graph data-grid data conversion is shown in Fig 9.

D. Local spatio-temporal feature modeling based on multi-scale fusion

D.1 Multi-scale local spatial feature extraction

Geographically neighboring regions tend to exhibit similar time-lagged accident patterns due to the spatio-temporal diffusion effects of traffic flow. Therefore, we employ a strategy for learning features at multiple scales. Specifically, we preserve the original low-scale regional division while constructing a high-scale representation space through neighborhood aggregation. This approach identifies the spatio-temporal patterns of traffic accidents while resolving data refinement and local bias problems typically found in low-scale analyses. The transformation of regional scales is illustrated in Fig. 10.

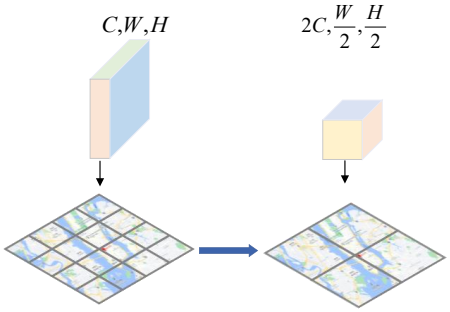


Fig. 10. a transformation of a scale

In this case, we extract simple spatial features of traffic accidents between low-scale cell regions using multi-channel convolution. That is

$$C_t^k = \varphi(\boldsymbol{W}_t^k * C_t^{k-1} + b_t^k) \tag{24}$$

Subsequently, we apply a 3×3 convolutional filter to downsize the input feature map and enhance the feature channel count. This minimizes information loss during the feature fusion process. Next, we aggregate spatial features in high-scale regions through multi-layer convolution. That is

$$x_\tau = MultiCNN_\tau\left(x_{\tau-1}\right) \tag{25}$$

Where τ is the current number of convolutional layers; $x_{\tau-1}$ is the input to the τ layer. *MultiCNN()* for the multi-layer convolution operation. Multi-layer deconvolution is applied to capture spatial correlations in high-scale regions. That is

$$x_s = MultiDeCNN_s\left(x_{s-1}\right) \tag{26}$$

Where s is the number of current deconvolution layer layers; x_{s-1} serves as the input for the s th layer; $MultiDeCNN()$ is the multi-layer deconvolution operation. Finally, the feature map dimension is reduced by 3*3 deconvolution.

D.2 Gated temporal convolutional layer

Adjacent areas may belong to different functional zones and exhibit spatial differences in accident patterns. However, they often exhibit comparable time-lagged behaviors over time, owing to the spatio-temporal diffusion effects of traffic flow. To model these features, we use TCN to capture local temporal dependencies. The TCN effectively models temporal correlations across functional areas. To improve feature selection and reduce redundancy, we introduce a gating mechanism, creating a gated-TCN. This mechanism governs the flow of information, strengthening the model's capability to identify complex time-based patterns. The TCN network's design is depicted in Fig. 11.

Dilated causal convolution combines the benefits of dilated convolution and causal convolution. It effectively captures distant relationships by enlarging the receptive field. At the same time, it preserves the causal relationship in the time series. This ensures that predictions at each step only depend on current and past information. Thus, future information does not affect current predictions.

$$X *_d f(t) = \sum_{i=0}^{k-1} f(t)_i \, X_{t-d\times s} \tag{27}$$

Where $X\in R^T$ denotes the feature vector, and $f(t)\in R^K$ represents the convolution filter. The dilation factor d increases as the network layers deepen. The padding size helps maintain the output dimensions and smooths boundary effects. The expanded causal convolution is illustrated in Fig 12.

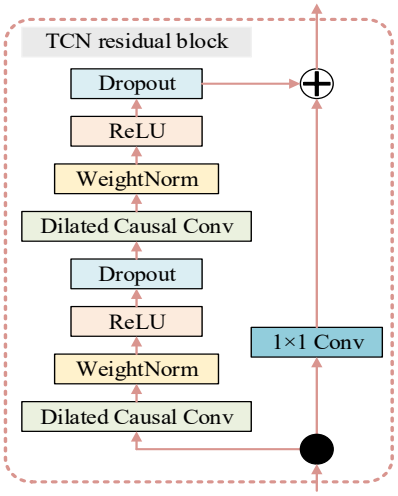


Fig. 11. TCN Network Architecture

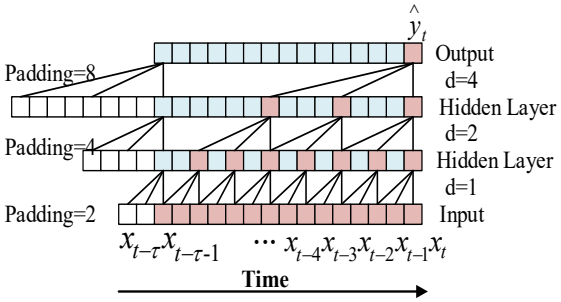


Fig. 12. Expanded causal convolution structure diagram

The gated-TCN contains two independent parallel TCNs that regulate the intensity of the information flow transmission through the Tanh function and the Sigmoid function regulates the flow of information. That is

$$h = tanh\left(\theta_1 *_d X + b\right) \odot \sigma\left(\theta_2 *_d X + c\right) \tag{28}$$

where: θ_1, θ_2, b, c is the convolution operation parameter; $*_d$ denotes the inflated causal convolution; $\odot$ is the Hadamard product.

E. Dynamic Fusion Layer

We dynamically fuse the global domain and local domain spatio-temporal features. Subsequently, the predicted values are obtained through a fully connected layer. That is

$$\hat{Y} = FC\left(\boldsymbol{W}_1 * \hat{Y}_G + \boldsymbol{W}_2 * \hat{Y}_L\right) \tag{29}$$

Where the outputs $\hat{Y}_G$ and $\hat{Y}_L$ represent the global domain spatio-temporal feature modeling based on MFA mechanisms and the local domain feature modeling based on multi-scale fusion, respectively. The feature fusion prediction result shows the likelihood of traffic accidents occurring in the area during the next time step.

Based on the number of accident casualties, the accident severity is classified into three levels: minor accident (risk value = 1), general accident (risk value = 2), and major accident (risk value = 3). Consequently, the accident risk level is categorized into four levels $L = \{0,1,2,3\}$, where a risk value of 0 indicates no accident. Let Y_t^i represent the total traffic accident risk in region *i* at time ***t***. In order to tackle the problem of zero-inflation in the data, we design a weighted loss function to assign higher weights to samples with elevated traffic accident risks. The weights corresponding to the four risk levels are 0.05, 0.2, 0.25, and 0.5, respectively. The Loss is expressed as:

$$Loss\left(Y, \hat{Y}\right) = \frac{1}{2}\sum_{l \in L} \eta_l \left(Y(l) - \hat{Y}(l)\right)^2 \tag{30}$$

Where Y is the real data; $\hat{Y}$ is the predicted value; $Y(l)$ to indicate the risk level of traffic accidents for *l*. samples; η_l is the traffic accident risk level of l weight.

IV. Calculated Case Analysis

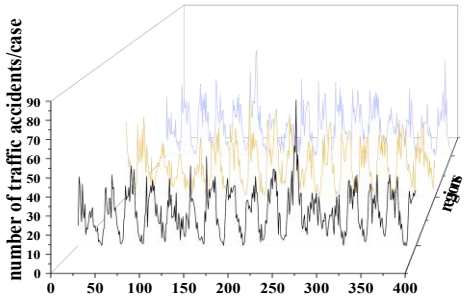
A. Introduction to the dataset

We evaluate the effectiveness of MFA-MSSTGCN using multi-source spatio-temporal data from two U.S. cities: NYC and Chicago. Each dataset includes time, location, and the number of injuries and fatalities in traffic accidents. The NYC dataset also provides point-of-interest (POI) information around accident locations, Divided into seven categories: residential zones, educational institutions, cultural sites, recreational spaces, social services, transportation, commercial zones. Weather information encompasses temperature and various atmospheric conditions. Roadway characteristics cover type, length, width, and snow removal priority. Dataset Overview and Metrics are presented in Table I.

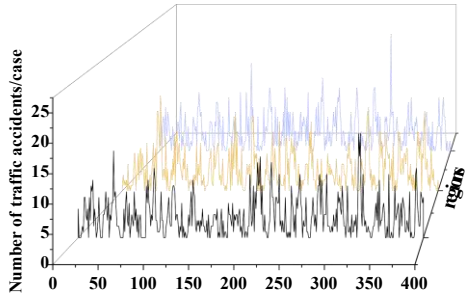
TABLE I
Dataset Overview and Metrics

Dataset	NYC	Chicago
time span	2013.01-2013.12	2016.02-2016.09
Accidents	147000	44000
Road Network	103000	56000
Weather	8760	5832
POI	15625	---

Fig. 13 presents a subset of traffic accident volumes from the NYC(June 1 to December 1, 2013) and Chicago(October 2 to October 16, 2016.) datasets for specific time periods.



(a) NYC Dataset



(b) Chicago Dataset

Fig. 13. Changes in Traffic Accident Volume for Selected Areas on the NYC/Chicago Dataset

B. Experimental Setup

We implement this experiment using the PyTorch framework. The dataset is split into training, validation, and test sets in a 6:2:2 ratio. The road network is partitioned into $2\text{km} \times 2\text{km}$ small regions, and the data is normalized to the range [0, 1] to enhance model stability. The proximity period p=3, and the weekly cycle length q=4. A 2-layer GCN with 64 filters is employed to learn full-domain spatial features. In the MFA mechanism, the quantity of heads is configured to 8, the quantity of layers is configured to 4, and each layer contains 256 hidden units. The Bi-GRU have 128 hidden units in each direction. The Gated-TCN consists of 128 hidden units per layer. During training, the batch size is configured to at 32, the learning rate is 0.001.

C. Evaluation metrics

Derived from the literature [22], this work uses RMSE, Recall, and MAP to evaluate the model performance. That is

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \hat{Y}_t\right)^2} \tag{31}$$

$$\text{Recall} = \frac{1}{T}\sum_{t=1}^{T}\frac{S_t \cap R_t}{|R_t|} \tag{32}$$

$$\text{MAP} = \frac{1}{T}\sum_{t=1}^{T}\frac{\sum_{j=1}^{|R_t|} pre(j) \times rel(j)}{|R_t|} \tag{33}$$

Where Y_t is the true accident risk of all regions at moment *t*, and $\hat{Y}_t$ is the predicted accident risk of all regions at moment *t*. The modeling performance for high-risk regions is summarized as follows R_t the set of regions with real

accidents at time t. S_t denotes the areas that rank at the top $|R_t|$ regarding predicted accident risk. $pre(j)$ denotes the prediction accuracy of all regions ranked top j in S_t; $rel(j)$ is whether region j is involved in an accident at time t, and its value is 1 if it is, and 0 if it is not. The model evaluates prediction performance for peak accident hours, specifically 7:00-9:00 and 16:00-19:00, using RMSE*, Recall*, and MAP* as metrics.

D. Comparative analysis of prediction results

To evaluate the effectiveness of the suggested model, a comparison will be conducted against the following baseline models:

- SDCAE [10]: It learns spatial correlations across regions using denoising convolution layers.
- Conv LSTM [5]: ConvLSTM integrates convolution and LSTM into a unified framework, enabling efficient processing of spatio-temporal data.
- T-GCN [13]: T-GCN is new temporal graph convolutional network model.
- ST-RiskNet [21]: This model combines local and global spatial-temporal features, considering multiple factors to predict traffic accident risk.
- ASTGCN [31]: A network model that uses an attention mechanism to dynamically capture spatio-temporal features in traffic data.
- MVMT-STN[23]: A recent multi-task learning approach predicts both detailed and broad citywide traffic accident risks at the same time.
- GSNet [22]: A model designed to extract spatio-temporal relationships by integrating geographic and semantic perspectives.
- MGHSTN [33]: MGHSTN integrates remote sensing to evaluate urban traffic accident risks.

Table 2 presents a comparison of the MFA-MSSTGCN model's outcomes with reference models on the NYC and Chicago datasets. SDCAE performs slightly worse than other baseline models, because it captures spatial correlations solely through multi-layer CNNs, neglecting temporal features. Although ConvLSTM incorporates both temporal and spatial correlations, it extracts only basic features, resulting in limited performance improvement. T-GCN captures spatio-temporal correlations in the road network structure using GCN and GRU; however, none of these three models account for external factors influencing traffic accidents.

ASTGNN improves prediction accuracy by dividing input data into multiple scales, refining temporal information, and using a spatio-temporal dynamic attention mechanism. GSNet captures global semantic and local geographic correlations by applying attention-based GRU, modeling semantic, geographic, and temporal dependencies. Building on GSNet, ST-RiskNet simultaneously considers global and local spatio-temporal correlations. MVMT-STN splits the road network into coarse-grained and fine-grained regions based on latitude and longitude. These three models also include external factors such as weather and temperature when making predictions for different regional scales. MGHSTN performs better than other multi-view models, highlighting the advantages of attention mechanisms in learning spatio-temporal features.

Compared to these baseline models, our proposed method significantly improves prediction performance. This demonstrates the benefit of considering multi-scale regions. Table 2 shows that MFA-MSSTGCN reduces RMSE by 1.2211 compared to GSNet on the NYC dataset. Recall increases by 1.29%, and MAP improves by 0.0126.

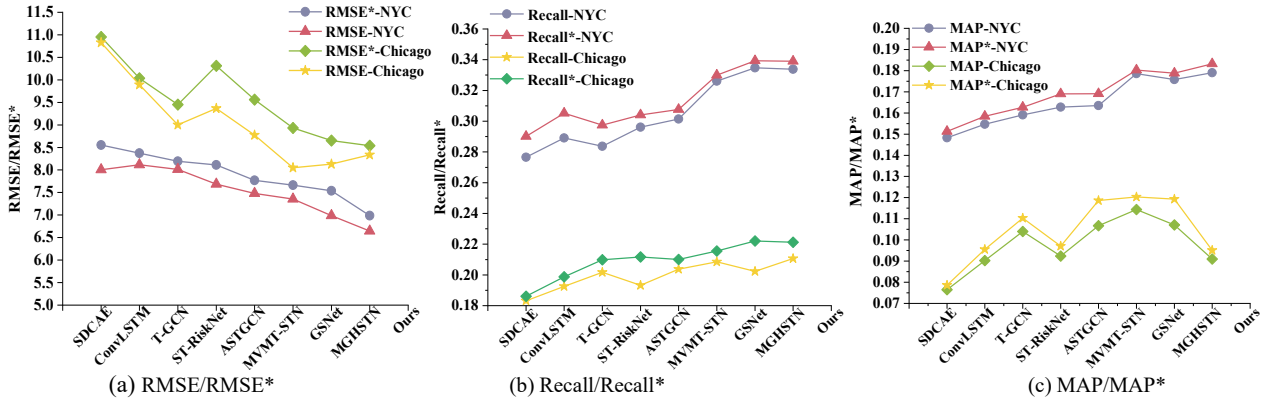
Because the Chicago dataset lacks POI data, prediction accuracy during both general periods and peak accident periods is lower compared to the NYC dataset. This emphasizes the importance of including regional functional similarity to improve accuracy. Fig.14 compares RMSE/RMSE*, Recall/Recall*, and MAP/MAP* across various models for the NYC and Chicago datasets.

E. Analysis of the impact of weather factors

This study examines the impact of weather on traffic accidents by performing a comparative assessment of prediction outcomes with the NYC dataset and chosen baseline models, with and without the inclusion of weather factors.

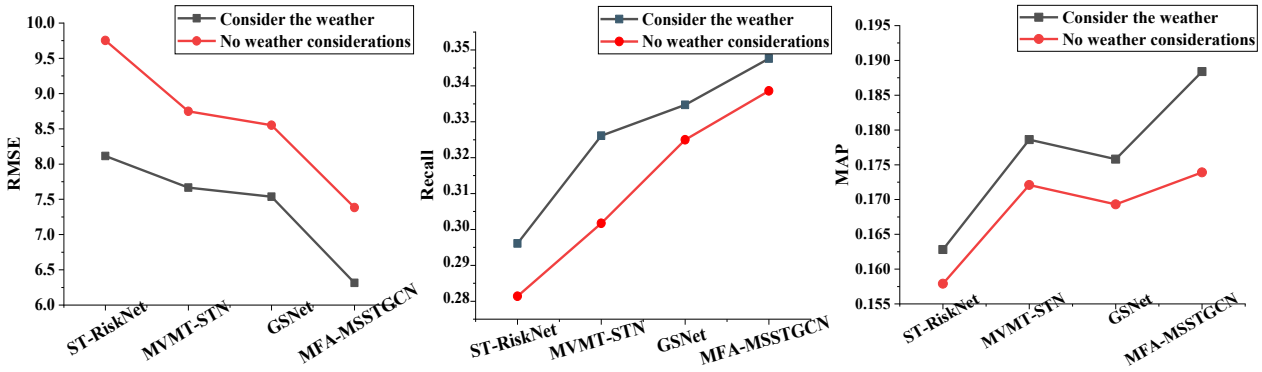
TABLE II
COMPARISON OF MODEL PERFORMANCE ON THE NYC/CHICAGO DATASET

Models	NYC			Chicago		
	RMSE/RMSE*	Recall/Recall*	MAP/MAP*	RMSE/RMSE*	Recall/Recall*	MAP/MAP*
SDCAE	8.5542/8.0063	27.66%/29.01%	0.1483/0.1513	10.9520/10.8255	18.32%/18.61%	0.0765/0.0786
ConvLSTM	8.3741/8.1149	28.91%/31.52%	0.1547/0.1584	10.0361/9.8931	19.25%/19.87%	0.0902/0.0955
T-GCN	8.1945/8.0126	28.37%/29.75%	0.1591/0.1627	9.4523/9.0029	20.17%/20.98%	0.1040/0.1103
ST-RiskNet	8.1132/7.6854	29.61%/30.41%	0.1628/0.1690	10.3120/9.3651	19.32%/21.17%	0.0924/0.0971
ASTGNN	7.7698/7.4783	30.14%/30.76%	0.1635/0.1691	9.5625/8.7736	20.37%/21.01%	0.1067/0.1186
MVMT-STN	7.6650/7.3536	32.61%/32.99%	0.1786/0.1803	8.9310/8.0479	20.85%/21.56%	0.1144/0.1203
GSNet	7.5366/6.9848	33.47%/33.93%	0.1758/0.1788	8.6525/8.1294	20.23%/22.21%	0.1071/0.1193
MGHSTN	6.9871/6.6430	33.38%/33.90%	0.1790/0.1832	8.5400/8.3381	21.07%/22.13%	0.0910/0.0951
Ours	**6.3155/5.4961**	**34.76%/35.48%**	**0.1884/0.1915**	**8.0231/7.7785**	**22.25%/23.91%**	**0.1254/0.1305**



(a) RMSE/RMSE* (b) Recall/Recall* (c) MAP/MAP*

Fig. 14. Performance comparison of different models on NYC and Chicago dataset



(a) RMSE (b) Recall (c) MAP

Fig. 15. Comparison of Whether Weather is Considered on the NYC Dataset

As illustrated in Fig 15, the comparison of forecasting results between several baseline models and the proposed MFA-MSSTGCN model, both with and without weather factors. The data shows clear differences in prediction accuracy among models when weather conditions are considered. ST-RiskNet performs the worst among all tested models, with or without weather factors. Without weather conditions, ST-RiskNet has an RMSE of 9.7531, a Recall of 28.14%, and an MAP of 0.1579. MVMT-STN and GSNet perform better under the same conditions. MVMT-STN achieves an RMSE of 8.7481, a Recall of 30.17%, and an MAP of 0.1721. GSNet achieves an RMSE of 8.5513, a Recall of 32.50%, and an MAP of 0.1693.

In comparison, our proposed MFA-MSSTGCN model outperforms all baseline models. Specifically, without weather factors, MFA-MSSTGCN reduces RMSE by 16.20%, increases Recall by 3.85%, and improves MAP by 7.16% compared to GSNet. These results highlight the robustness and stability of MFA-MSSTGCN. The model maintains strong performance whether or not weather factors are included. This demonstrates its effectiveness in improving prediction accuracy and reliability across various environmental conditions, confirming its advantage over existing baseline models.

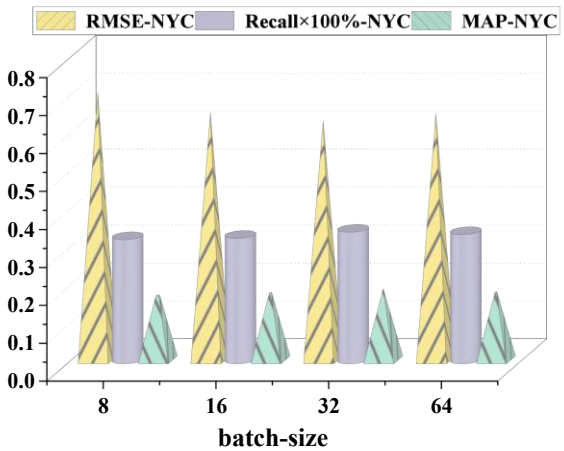
F. Ablation Experiment

To investigate the model's performance under varying hyper parameters, we examine and assess the implications of batch-size, learning rate with the NYC and Chicago datasets. Moreover, we investigate the effect of GCN layer count, the quantity of heads in the MFA mechanism, and hidden unit count in the Bi-GRU on forecasting results using the NYC dataset. To facilitate a unified comparison of RMSE, Recall, and MAP within the same coordinate system, the overall RMSE values are scaled by a factor of 0.1 for visualization purposes.

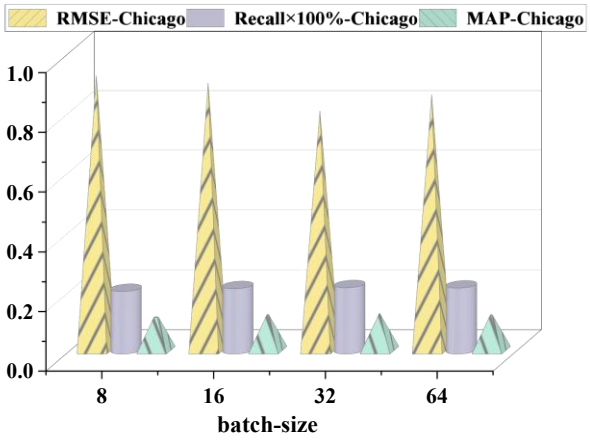
F.1 Influence of batch size on the model

Considering the implications of batch size on model efficiency, training speed and memory usage, we set the batch size to {8, 16, 32, 64}. The comparison of prediction performance with different batch sizes is depicted in Fig 16.

With a batch size of 32, the model yields the highest performance on both the NYC and Chicago datasets. While a larger batch size can enhance model performance and reduce gradient noise, it may compromise the model's generalization ability. Conversely, a smaller batch size can lead to unstable training, thereby adversely affecting model performance.



(a)batch-size NYC



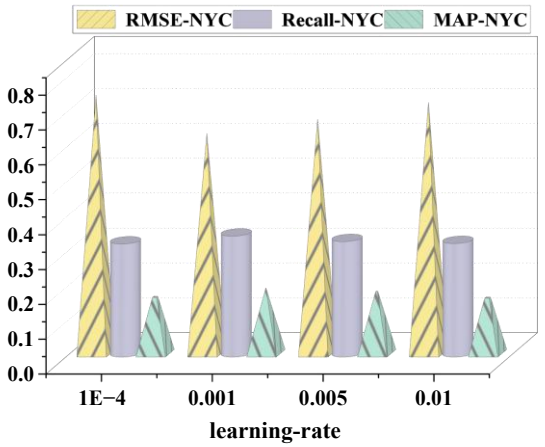
(b) batch-size Chicago

Fig. 16. Comparison of Prediction Performance under Different Batch Processing Sizes

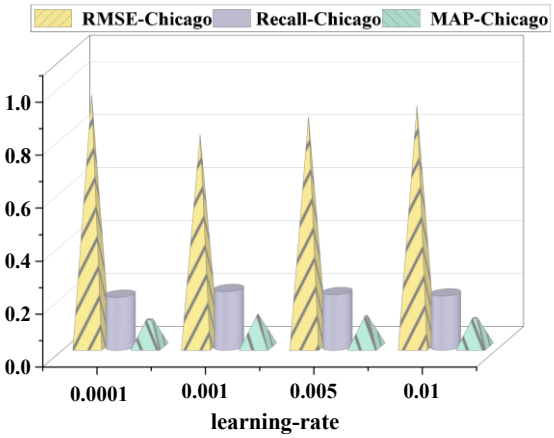
F.2 Learning rate influence.

The learning rate is set to {0.0001, 0.001, 0.005, 0.01} to test the effect on the prediction results. The comparison of prediction performance under different learning rate sizes is depicted in Fig 17.

The model reaches its best performance with a learning rate of 0.001, as evidenced by the minimal RMSE value and the maximal Recall and MAP values. A larger learning rate may result in unstable or divergent training, whereas a smaller learning rate can slow the training process and increase the risk of converging to suboptimal local solutions.



(a)learning-rate NYC



(b)learning-rate Chicago

Fig. 17. Comparison of Prediction Performance under Different Learning Rates

F.3 Impact of GCN layer quantity.

Given the varying effects of different GCN layers on feature learning, the layers are set to {1, 2, 3, 4} to assess their impact on forecasting accuracy. The comparison of prediction performance with different GCN layer numbers is depicted in Fig 18. (a).

The best results are obtained when the GCN layers are configured to 2. With just a solo layer, the model is unable to effectively extract the features of graph nodes. While increasing the quantity of GCN layers enhances the aggregation of node features and improves the model's representational capacity, it also leads to higher parameter complexity and computational costs. Furthermore, an excessively deep network may suffer from gradient explosion or vanishing gradients, compromising the stability of model training.

F.4 Effect of MFA head count on performance

The head count in the MFA mechanism is configured as {2, 4, 6, 8, 10} to evaluate its impact on the prediction outcomes. The performance with different numbers of heads in the MFA is depicted in Fig 18. (b).

The model performs best with 8 heads. Each head independently focuses on distinct aspects or hierarchical levels of features, enabling finer-grained learning of correlations between graph nodes. While too few heads limit the model's representational capacity, an excessive number of heads increases computational resource consumption, ultimately degrading model performance.

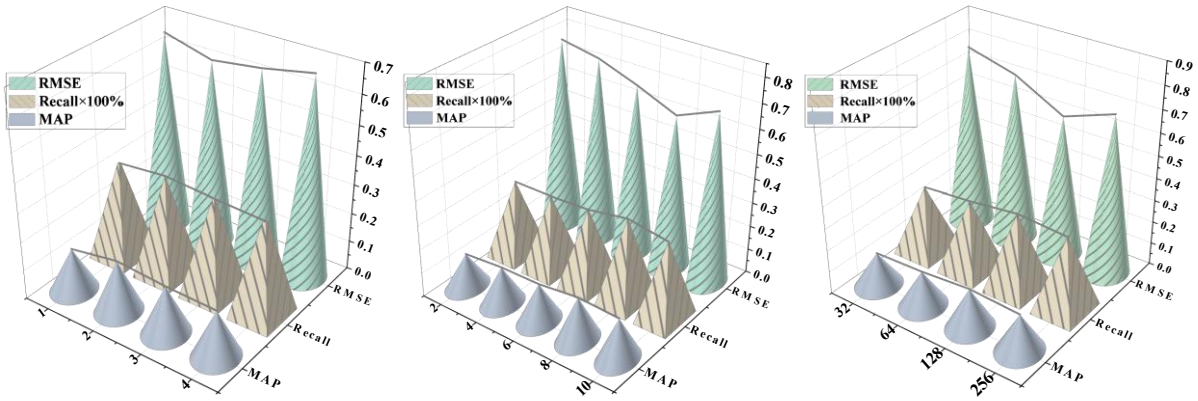
F.5 Effect of hidden layer unit count in Bi GRU on model performance.

Taking into account the varying impacts of hidden layer unit count on feature learning, the hidden layer unit count in the Bi GRU for both directions is configured as {32, 64, 128, 256} to assess its influence on prediction performance. The comparison of the prediction performance for different Bi GRU hidden layer unit numbers is depicted in Fig 18. (c).

The model reaches peak performance with 128 hidden units. Having fewer than 128 hidden units restricts the model's capacity to learn long-term relationships, leading to underfitting and a decline in all three evaluation metrics. Conversely, a higher number of hidden units induces overfitting, which similarly degrades model performance.

G. Ablation Experiment

The ablation experiment was conducted by removing the following four components. (1) no-Bi GRU: Replace the Bi GRU in the MFA mechanism with a fully connected layer. (2) no flow-attention: Replace the MFA mechanism with a traditional attention mechanism to validate the merits of MFA. (3) no gated-TCN: Remove the gated-TCN module from the local temporal correlation module, i.e., only consider global temporal correlation. (4) no LS-FE: Remove the high-scale region aggregation part, capturing only the spatiotemporal features of the initial partitioned grid regions. The comparisons of RMSE/RMSE*, Recall/Recall*, and MAP/MAP* for different components on the NYC、Chicago dataset are shown in Fig 19, 20.



(a) different GCN Layers (b) different head numbers of MFA (c) different Bi-GRU hidden layer unit numbers

Fig 18. Predictive performance of the model with different hyperparameters

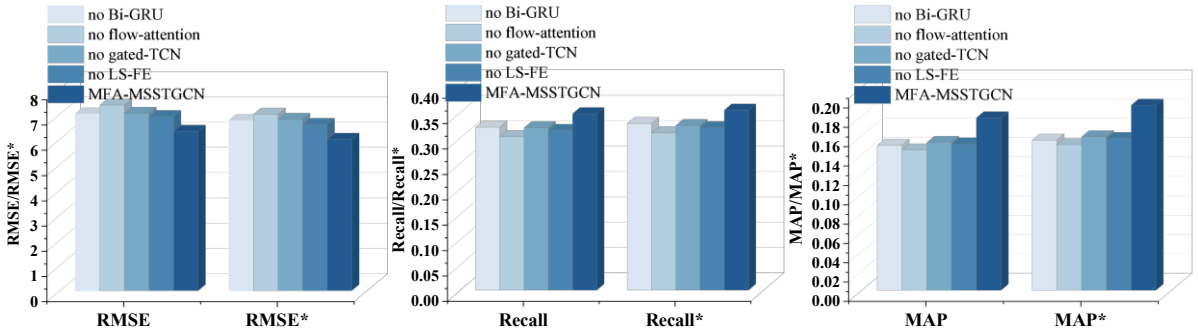


Fig. 19. Comparison of RMSE/RMSE*、Recall/Recall*、MAP/MAP* for different component on NYC

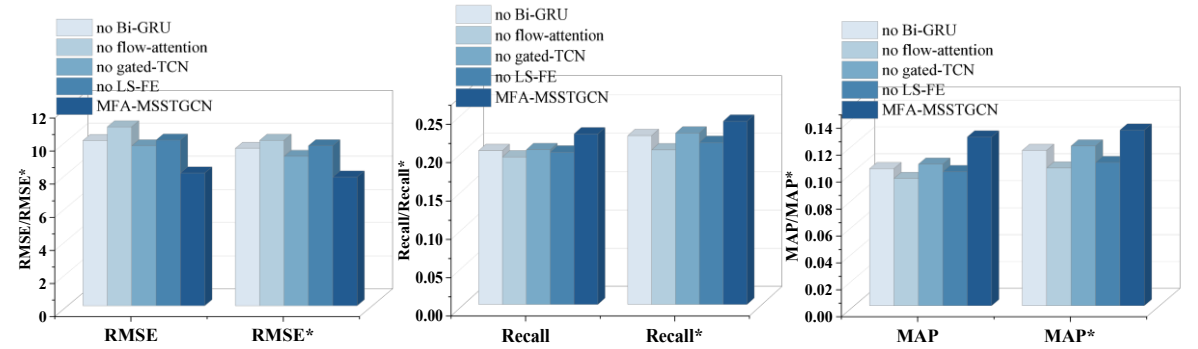


Fig. 20. Comparison of RMSE/RMSE*、Recall/Recall*、MAP/MAP* for different component on Chicago

As shown in Fig 19, reducing different components in the NYC dataset leads to a decrease in the final prediction results, indicating that each component plays a positive role in prediction performance. As depicted in Fig 20, removing the MFA mechanism results in the worst prediction performance, with the number of computational parameters being 9,914,763. However, when the MFA mechanism is added, the number of parameters decreases to 8,827,448, indicating that the multi-head attention mechanism addresses the quadratic complexity issue in traditional attention mechanisms, thereby improving model performance. Removing the Bi-GRU and using a fully connected layer for the MFA mechanism's matrix mapping —i.e., not considering global temporal similarity—yields prediction results similar to those of the model without the gated-TCN, which does not consider local temporal similarity. This indicates that the temporal features influencing accident occurrence exhibit multi-scale dependencies. After removing high-scale region aggregation, the lack of consideration for local spatio-temporal interactions also leads to a decline in performance. This demonstrates that MFA-MSSTGCN employs different prediction modules for global and local spatio-temporal feature capture, and each module effectively enhances traffic accident risk prediction performance.

H. Contribution of the weighted loss function

To evaluate the performance of the weighted loss function, trials were carried out using the NYC dataset. The comparison data for the weighted loss function is shown in Table III.

Table III
ALGORITHM COMPARISON DATA

Model	RMSE/RMSE*	Recall/Recall*	MAP/MAP*
un-weighted	9.1426/8.6382	34.21%/34.53%	0.1891/0.1857
weighted	6.3155/5.4961	34.76%/35.48%	0.1884/0.1915

As illustrated in Table III, the weighted loss function has a relatively minor impact on Recall and MAP but significantly reduces RMSE and RMSE* by 2.8271 and 3.1421, respectively. This suggests that adding the weighted loss function improves the model's overall forecasting ability across all regions and helps reduce the problem of data zero-inflation.

I. *Visualization of partial region prediction*

In the following analysis, the predicted traffic accident values are contrasted with the real data for selected areas in the NYC and Chicago datasets at 30-minute intervals on a specific day. The visualization of the traffic accident risk prediction comparison is presented in Fig 21 and Fig 22.

The MFA-MSSTGCN prediction results for the selected areas exhibit strong alignment with the actual accident conditions. Due to higher traffic flow and frequent crowd gatherings in NYC, its accident risk consistently remains elevated. Furthermore, the addition of POI data in the NYC dataset leads to better forecasting results than those observed in the Chicago dataset. Although the Chicago dataset records significantly fewer accidents than NYC, the introduction of a weighted loss function partially mitigates the zero-inflation issue, thereby enhancing prediction accuracy for regions with higher accident risks.

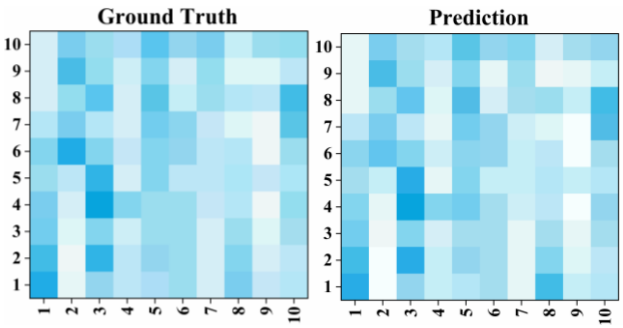


Fig. 21. Visualization of NYC dataset traffic accident risk prediction

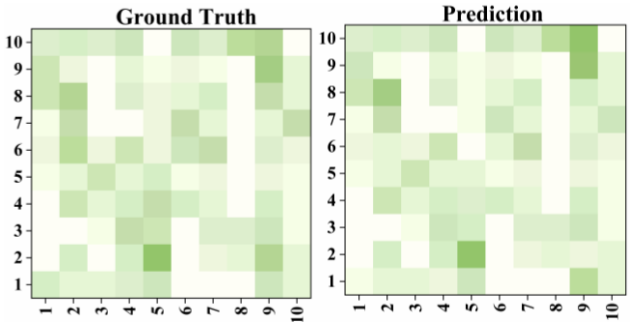


Fig. 22. Visualization of Chicago dataset traffic accident risk prediction

V. Conclusion

To overcome the limitations of current traffic accident prediction models, which typically focus only on global or local spatial features or rely on single-scale temporal modeling, this paper proposes the MFA-MSSTGCN based on multi-head flow attention. The main findings are:

(1) Combining Bi-GRU with multi-head attention improves modeling of global spatio-temporal features. This method considers POIs, road features, and accident-risk similarities. Additionally, the source competition and sink allocation strategies in multi-head attention significantly reduce computational complexity. Experimental results show that MFA strongly improves the model's spatio-temporal modeling ability and prediction accuracy.

(2) Local spatio-temporal convolutions allow the model to reconstruct the original spatio-temporal feature maps by extracting features from different spatial scales. This helps capture and predict spatio-temporal dependencies across areas with varying scales.

(3) Dynamically combining global and local spatio-temporal features enables the model to better represent complex accident patterns within regions. The weighted loss function also effectively addresses the zero-inflation problem caused by sparse data, further improving model performance.

Although MFA-MSSTGCN shows improved prediction accuracy, this study has two main constraints. Primarily, the model does not yet provide fine-grained predictions at the road segment level. Second, it does not fully consider how multiple external factors affect accident patterns. Future research will focus on fine-grained predictions and integrating multiple external factors to improve the model's performance.

References

[1] B. Pan, U. Demiryurek and C. Shahabi, "Utilizing Real-World Transportation Data for Accurate Traffic Prediction," 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, pp. 595-604, 2012.

[2] Williams, B. M. Multivariate Vehicular Traffic Flow Prediction: Evaluation of ARIMAX Modeling. Transportation Research Record, vol. 1776, no. 1, pp. 194-200, 2001.

[3] Jeong Y S, Byon Y J, Castro-Neto M M, et al. Supervised weighting-online learning algorithm for short-term traffic flow prediction. IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 4, pp. 1700-1707, 2013.

[4] Van Lint J W C, van Hinsergen C. Short-term traffic and travel time prediction models. Artificial Intelligence Applications to Critical Transportation Issues, vol. 22, no. 1, pp. 22-41, 2012.

[5] SHI, Xingjian, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems, pp. 28, 2015.

[6] Lu S, Chen H, Teng Y. Multi-Scale Non-Local Spatio-Temporal Information Fusion Networks for Multi-Step Traffic Flow Forecasting. ISPRS International Journal of Geo-Information, vol. 13, no. 3, pp 71, 2024.

[7] Ma, C., Gu, K., Zhao, Y. et al. Research on highway traffic flow prediction based on a hybrid model of ARIMA–GWO–LSTM. Neural Comput & Applic, vol. 37, pp. 14703–14722, 2025.

[8] Chang, Y.; Ma, J.; Sun, L.; Ma, Z.; Zhou, Y. Vessel Traffic Flow Prediction in Port Waterways Based on POA-CNN-BiGRU Model. J. Mar. Sci. Eng, vol. 12, pp. 2091, 2024.

[9] Chang, Yumiao, Jianwen Ma, Long Sun. et al. "Vessel Traffic Flow Prediction in Port Waterways Based on POA-CNN-BiGRU Model". Journal of Marine Science and Engineering. vol. 12, no.11, pp. 2091, 2024.

[10] W. Zhao, Y. Gao, T. Ji, X. Wan, F. Ye and G. Bai, "Deep Temporal Convolutional Networks for Short-Term Traffic Flow Forecasting," in IEEE Access, vol. 7, pp. 114496-114507, 2019.

[11] C. Chen, X. Fan, C. Zheng, L. Xiao, et al. "SDCAE: Stack Denoising Convolutional Auto encoder Model for Accident Risk Prediction Via Traffic Big Data," 2018 Sixth International Conference on Advanced Cloud and Big Data, pp. 328-333, 2018.

[12] Méndez M, Merayo M G, Núñez M. Long-term traffic flow forecasting using a hybrid CNN-BiLSTM model. Engineering Applications of Artificial Intelligence, vol. 121, pp. 106041, 2023.

[13] Hu Z, Sun R, Shao F, et al. An efficient short-term traffic speed prediction model based on improved TCN and GCN. Sensors, vol. 21, no. 20, pp. 6735, 2021.

[14] Zhao L, Song Y, Zhang C, et al. T-GCN: A temporal graph convolutional network for traffic prediction. IEEE transactions on intelligent transportation systems, vol. 21, no 9, pp. 3848-3858, 2019.
[15] Chen G, Wei Y, Peng J, et al. Spatio-temporal graph neural network based on time series periodic feature fusion for traffic flow prediction. The Journal of Supercomputing, vol. 81, pp. 129, 2025.
[16] Ye, Z., Wang, H., Przystupa K, et al. Dynamic Spatio-Temporal Hypergraph Convolutional Network for Traffic Flow Forecasting. Electronics. vol. 13, pp. 4435, 2024.
[17] Gao M, Du Z Qin H, et al. Dynamic multi-scale spatial-temporal graph convolutional network for traffic flow prediction. Knowledge-Based Systems, vol. 305, pp. 112586, 2024.
[18] Yao Y, Chen L, Wang X, et al. Research on Traffic Flow Forecasting of Spatio-temporal Convolutional Networks with Auto-correlation. International Journal of Automotive Technology, pp. 1-10, 2024.
[19] Ali A, Zhu Y M, Zakarya M. Exploiting dynamic spatio-temporal graph convolutional neural networks for city wide traffic flows prediction. Neural Networks, vol. 145, pp. 233-247, 2022.
[20] Zhao W, Zhang S, Zhou B, et al. Multi-spatio-temporal fusion graph recurrent network for traffic forecasting. Engineering Applications of Artificial Intelligence, vol. 124, pp. 106615, 2023.
[21] Wu Z H, Pan S R, Long G D, et al. Graph Wave Net for deep spatial-temporal graph modeling. arXiv preprintarXiv, pp.1906-00121, 2019.
[22] WANG Beibei, WAN Huaiyu, GUO Shengnan, et al. Local and Global Spatial-Temporal Networks for Traffic Accident Risk Forecasting. Journal of Frontiers of Computer Science and Technology, vol. 15, no. 9, pp.1694-1702, 2021.
[23] Wang B, Lin Y, Guo S, et al. GSNet: Learning spatial-temporal correlations from geographical and semantic aspects for traffic accident risk forecasting. Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 5, pp. 4402-4409, 2021.
[24] Wang S, Zhang J, Li J, et al. Traffic accident risk prediction vi a multi-view multi-task spatio-temporal networks. IEEE Transacti ons on Knowledge and Data Engineering, vol. 35, no.12, pp. 12 323-12336, 2021.
[25] Alhaek F, LiT, Rajeh T M, et al. Encoding global semantic and localized geographic spatial-temporal relations for traffic accident risk prediction. Information Sciences, vol. 697, pp. 121767, 2025.
[26] Wang Q ,Zhang K, Zhu C, et al. Traffic Accident Risk Prediction for Multi-factor Spatio-temporal Networks. IAENG International Journal of Applied Mathematics, vol. 53, no.4, pp. 1315-1331, 2023.
[27] Li Y, Xu H, Zhang T, et al. DDGformer: Direction-and distance-aware graph transformer for traffic flow prediction. Knowledge-Based Systems, vol. 302, pp.112381, 2024.
[28] Yu L, Liu W, Wu D, et al. Spatial–temporal combination and multi-head flow-attention network for traffic flow prediction. Scientific Reports, vol. 14(1), pp.9604, 2024.
[29] P. Trirat, S. Yoon and J. -G. Lee, "MG-TAR: Multi-View Graph Convolutional Networks for Traffic Accident Risk Prediction," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 4, pp. 3779-3794, April 2023.
[30] GUO S, LIN Y, WAN H, et al. Learning Dynamics and Heterogeneity of Spatial-Temporal Graph Data for Traffic Forecasting. IEEE Transactions on Knowledge and Data Engineering, pp. 5415-5428, 2022.
[31] Fang J, Wu Z, Xu M, et al. Multistep traffic speed prediction from multiple time-scale spatiotemporal features using graph attention network. Applied Intelligence, pp.1-14, 2024.
[32] Chen J, Zheng L, Hu Y, et al. Traffic flow matrix-based graph neural network with attention mechanism for traffic flow prediction. Information Fusion, vol. 104, pp. 102146, 2024.
[33] Geng Z, Xu J, Wu R, et al. STGAFormer: Spatial–temporal Gated Attention Transformer based Graph Neural Network for traffic flow forecasting. Information Fusion, vol. 105, pp. 102228, 2024.
[34] Chen M, Yuan H, Jiang N, et al. Urban Traffic Accident Risk Prediction Revisited: Regionality, Proximity, Similarity and Sparsity. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, pp. 281-290, 2024.