IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 May 2023, accepted 23 May 2023, date of publication 25 May 2023, date of current version 16 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3280068

RESEARCH ARTICLE

Traffic Flow Prediction Based on Two-Channel Multi-Modal Fusion of MCB and Attention

XIAOAN QIN
Anhui Business College of Vocational Technology, Wuhu 241002, China
e-mail: qxa1513@163.com

This work was supported in part by the Quality Engineering Project of Anhui Province under Grant 2021crh014 and Grant 2022jnds024, and in part by the High Level Scientific Research Project Cultivation under Grant 2022ZDG01.

ABSTRACT The accuracy of urban traffic flow prediction is influenced by nearby regional road network, historical traffic flow and seasonal climate, which has complex spatial and temporal dependence. In view of the above factors, we proposed a multi-modal traffic flow prediction model fusing road network, historical traffic flow and weather data. Firstly, the weighted spatio-temporal graph was constructed based on the traffic flow time series data, and the weighted STSGCN model was used to extract the spatio-temporal graph features. Secondly, the image sequence was constructed by road network, vehicle track and sensor position data, and the visual features were extracted by ResNet. Finally, based on the MCB and Attention two-channel multi-modal fusion model, the spatio-temporal graph features and the visual features of the image sequence were fused to obtain the aligned fusion vector. Finally, the aligned fusion vector was combined with the weather feature vector to complete the traffic flow prediction. The experimental results showed that the prediction results of our proposed model were better than those of other baseline models. At the same time, the ablation results also proved the effectiveness of each module in our proposed prediction model.

INDEX TERMS Traffic flow prediction, GCN, MCB, cross-modal attention, multi-modal fusion.

I. INTRODUCTION

Transportation is one of the most important infrastructure in modern cities, providing daily travel services for millions of people. With the acceleration of urbanization and the continuous population increase, the transportation system has become much more complex, including road traffic, rail traffic, pedestrians, and many shared means of transportation such as online car rental. However, there are many problems in the development of cities, such as air pollution and traffic congestion. Early intervention based on traffic flow prediction is considered as an important method to improve the efficiency of the traffic system and alleviate traffic congestion. With the development of smart cities and intelligent transportation systems, sensors such as ring detectors have been set up on the roads to sense the traffic status. The vehicle-mounted GPS system can continuously read the vehicle location information to feedback the traffic status, as well as the road condition monitoring video. By using the smart phones equipped with GPS to realize the data collection of pedestrian travel, it can also indirectly reflect road traffic information. The main process of traffic flow prediction is to predict the future traffic status based on the historical traffic data and environmental factors (weather, holidays, POI, etc.).

Traffic flow prediction is very challenging due to its strong dynamic nature, non-linear data, and unexpected situations such as congestion caused by traffic accidents. The road traffic will be affected both by the traffic of nearby areas and the historical traffic. It is necessary to fully consider the temporal and spatial data dependence. Traditional traffic flow prediction models, such as the ARIMA model [1], mainly use linear time series methods, which cannot deal with complex space-time dependence problems. In order to overcome the defects of the linear time series models, researchers devoloped machine learning methods [2], [3] and deep learning methods [4], [5], [6], such as STSGCN [7] which extracted spatio-temporal dependence features of spatio-temporal map to achieve more accurate prediction. However, the existing methods rarely consider the impact of the spatio-temporal characteristics of multi-modal

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

data on the traffic flow prediction (e.g., the impact of the spatio-temporal characteristics of weather data on the traffic). At the same time, different data have different impact on the traffic flow. Therefore, in this paper we propose a traffic flow prediction method based on MCB (Multimodal Compact Bilinear) and Attention dual-channel multimodal fusion. Our proposed method realizes the effective fusion of multimodal data of spatio-temporal characteristics and visual characteristics through two fusion modules (MCB and Attention) to achieve more accurate traffic flow prediction. Our contributions are summarized as follows:

(1) We propose a spatio-temporal matrix transformation method based on vehicle speed to construct the weighted traffic flow spatio-temporal graph, and use the improved STSGCN model (weighted-STSGCN) to extract the spatio-temporal graph feature.

(2) We propose the image sequence generation method of road network and vehicle track, which can effectively transform road network and track into visual features.

(3) We propose a two-channel multi-modal fusion model of MCB and Attention to fuse the spatio-temporal graph features and visual features, and combined with the weather features to achieve high-precision traffic flow prediction.

The rest of the paper is organized as follows. Section II introduces related work. Section III details the methodology. Section IV conducts the experiments and analyzes the results. Section V concludes the paper.

II. RELATED WORK

A. STATICAL APPROACH

Traffic prediction has a rich research history and there are several statistical models that have been widely used in the time series community to improve the accuracy of traffic prediction. Auto-Regressive Integrated Moving Average (ARIMA) and Kalman Filters (KF) [8] are some examples of statical methods that have been used for traffic prediction. To capture the long-term trends, some methods such as GP model [9] have been proposed to improve prediction accuracy by capturing process dynamics.

Decomposing traffic data into different components has been used as a technique to more accurately track the temporal evolution of traffic flow at different time resolutions and improve traffic prediction accuracy. Literature [10] utilized non-negative matrix factorization (NMF) to learn the latent properties. Literature [11] proposed low-rank decomposition tensor to model the weakly dependency on graph. However, these time series models only take into account the modeling of time series, and seldom take spatial information into account.

Machine learning methods have been used to model nodes such as stations and vehicle locations in Euclidean space, with the purpose of capturing model-pair spatial information. For instance, the ST-KNN [11] model used the KNN model for short-term traffic prediction to identify the current state of the traffic network and integrates generations of similar historical states as prediction results. Literature [12] proposed a unified method in which a linear regression model is built on POI data to extract spatial features.

B. DEEP LEARNING APPROACH

Recently, deep learning has been extensively studied on spatio-temporal tasks, including traffic tasks, due to its powerful ability to extract nonlinear relations. The RNN/LSTM-based model [13] has been used to capture long-term dependencies, and CNN has been used to model the non-linear spatial dependency [14]. Further, Literature [15] proposed a method to jointly model both spatial and temporal dependencies by integrating CNN and LSTM. Conv-LSTM [16] combines convolution and LSTM to fully fuse the relation of traffic flow in the adjacent region of the prediction point. In order to deal with more complex situations, STDN [17] which designed a flow gating mechanism was introduced to learn the dynamic similarity between locations, and a periodically shifted attention mechanism was designed to handle long-term periodic temporal shifting. ST-3DNet [18] adopted 3D convolutions and residual units to effectively extract features from both spatial and temporal dimensions.

To extract spatial relationships from the topological structure of urban network, researchers have proposed models based on GNN network such as STFGNN [19]. Other researchers use GCN technology to convolve the graphical model in the spectral-domain, such as T-GCN [20], which was in combination with the GCN and GRU to learn complex topological structures to capture spatial-temporal dependence. STGCN [21] builded the model with complete convolutional structures on graphs, which enables much faster training speed with fewer parameters. ASTGCN [22] used attention-based spatial-temporal graph convolutional network to model the dynamic spatial-temporal correlations. DCRNN [23] defined a graph convolutional operation in the vertical domain and combined the gating mechanism of RNN, while STSGCN [7] models simultaneously captured the localized spatial-temporal correlations directly. STGODE [24] proposed a tensor-based ordinary differential equation technical to alleviate the over-smoothing problem of ordinary GNNs. Graph WaveNet [25] combined CNN and spatial-based GCN at the same time to deal with very long sequence situations.

As more data related to traffic flow is taken into consideration, more and more researchers propose models that incorporate other data. For example, literature [12] combined weather and event factors into traffic prediction tasks. BuildSenSys [26] predicted traffic volume with dynamic building-traffic correlations.

III. METHOD

A. MODEL STRUCTURE

We use a two-channel and multi-modal fusion model based on MCB and attention mechanism to integrate the spatio-temporal graph features of traffic flow and the visual features of the track road network. Meanwhile, we

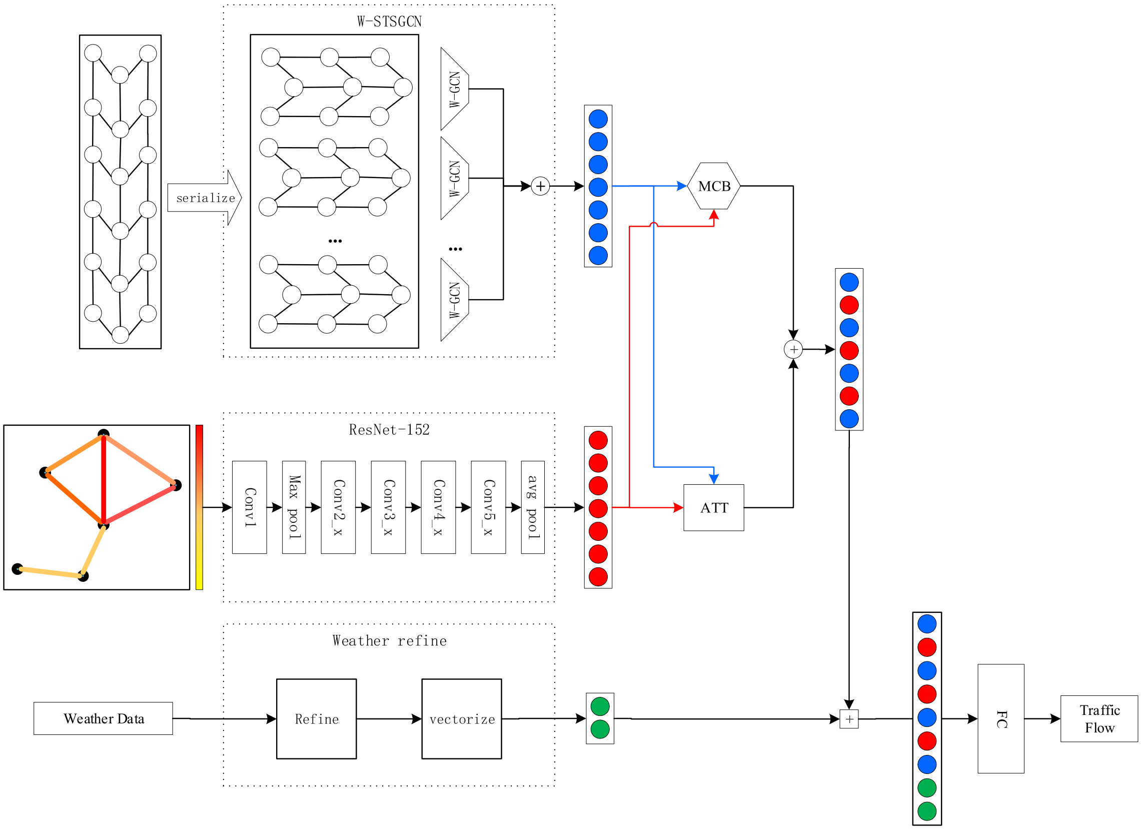

FIGURE 1. Model structure of traffic flow prediction. In the figure, "ATT" represents multi-modal fusion based on cross-modal attention, and "MCB" represents multi-modal fusion based on MCB algorithm. The refine module represents the modification of weather forecast data using polynomial functions. The input of W-STSGCN is the traffic flow spatio-temporal graph constructed from traffic flow sensors data. The input of ResNet is traffic image constructed by GPS trajectory and road network.

combine with weather and environmental factors to predict the future traffic flow. The proposed model structure is shown in Figure 1. Firstly, the traffic flow spatio-temporal graph is constructed according to the temporal data of the traffic flow, and the weighted-STSGCN model is used to extract the features of the spatio-temporal graph, and to obtain the spatio-temporal graph representation vector. Secondly, the image sequence is constructed by road network, vehicle track and sensor position data, and feature extraction is carried out by ResNet to obtain the visual representation vector of the image sequence. Then, the MCB and attention based two-channel multi-modal fusion model is used to fuse the spatio-temporal graph representation vector and the visual representation vector, and obtain the aligned fusion vector of the graph structure and visual features. Finally, the weather factor vector is combined to predict the traffic flow. Specifically, the proposed model includes four parts: feature extraction of spatio-temporal graph, visual feature extraction of road network and track, vectorization of weather features and traffic flow prediction based on multi-modal data.

B. GRAPH FEATURE EXTRACTION

The feature extraction of the traffic flow spatio-temporal Graph includes three parts: the construction of the traffic flow spatio-temporal graph, the spatio-temporal graph encoding and the feature extraction of the spatio-temporal graph based on Graph Convolutional Network (GCN) [27].

1) CONSTRUCTION OF SPATIO-TEMPORAL GRAPH

In this paper, we use traffic flow sensors data to construct a traffic flow spatio-temporal graph. At time step t_1, all sensors are taken as nodes, and the connections of sensors are taken as spatial edges to construct a spatial graph. Similarly, at time step t_2, $t_3\ldots$, the corresponding spatial graph is also constructed. For the spatial graph at time step i, we connect all nodes with themselves at adjacent time steps as temporal edges between sensor nodes, so that the global spatio-temporal graph of sensors can be obtained.

When constructing the traffic flow spatio-temporal graph, the calculation method of the spatial edge weight $\hat{e}^s_{i,j}$ of sensor nodes is shown in formula (1)-(2),

$$e^s_{i,j} = \frac{1}{d_{i,j}} \tag{1}$$

$$\hat{e}^s_{i,j} = \frac{e^s_{i,j} - \min(\{e^s_{i,j}\})}{\max(\{e^s_{i,j}\}) - \min(\{e^s_{i,j}\})} \tag{2}$$

where $d_{i,j}$ is the space distance between sensor i and sensor j. The calculation method of temporal edge weight $\hat{e}^t_{i,i+N}$ of sensor node is shown in formula (3)-(4),

$$e^t_{i,i+N} = \frac{1}{\bar{v} \cdot \Delta t} \tag{3}$$

$$\hat{e}^t_{i,i+N} = \begin{cases} \dfrac{e^t_{i,i+N} - E_{\min}}{E_{\max} - E_{\min}}, & e^t_{i,i+N} < E_{\max} \\ 1, & \text{else} \end{cases} \tag{4}$$

where $E_{\min} = \min(\{e^s_{i,j}\})$ and $E_{\max} = \max(\{e^s_{i,j}\})$, N is the number of sensors, $\bar{v}$ is the average speed of vehicles in the area (can be estimated from sampled data), Δt is the time interval between two time steps.

2) SPATIO-TEMPORAL GRAPH ENCODING

The spatio-temporal correlation can be enhanced by adding spatio-temporal embedding into the spatio-temporal network sequence to make the model consider spatio-temporal information simultaneously. For the constructed spatio-temporal network sequence $X_g \in \mathbb{R}^{N \times C \times T}$, a learnable temporal embedding matrix and a learnable spatial embedding matrix are constructed, so that the two embedding matrices include time information and space information after model training is completed, as shown in formula (5):

$$X_{g,t,s} = X_g \oplus T_e \oplus S_e \in \mathbb{R}^{N \times C \times T} \tag{5}$$

where T_{e} and S_{e} are temporal embedding matrix and spatial embedding matrix respectively, $\oplus$ is broadcast operation, C is sensor node characteristics (longitude, latitude, traffic flow, etc.), T is the length of time series.

3) FEATURE EXTRACTION BASED ON WEIGHTED-STSGCN

In order to extract the features of the traffic flow spatio-temporal graph, an improved STSGCN(named as weighted-STSGCN) is used to extract the features of the spatio-temporal graph structure. It includes two parts: spatio-temporal graph serialization based on sliding window and STSGCM local feature extraction based on weight fusion.

Spatio-temporal graph serialization: Since the STSGCN model needs to serialize and segment the traffic flow global spatio-temporal graph, we adopt the method of sliding window combined with sequential padding to serialize and segment the global spatio-temporal graph to obtain a series of local spatio-temporal graphs. We use a sliding window with a window size of 3 and a step size of 1 to serialize and segment the global spatio-temporal graph along the time direction. Meanwhile, spatial graph randomly initialized at time step t0 and tn+1 were added to the constructed global spatio-temporal graph to reduce information loss during feature extraction. After serialization and segmentation, we get n local spatio-temporal graphs with window size of 3.

Local feature extraction: The local feature extraction process of GCN based on weight fusion includes two parts: feature extraction based on GCN and feature fusion based on weight, the specific process is as follows:

a) feature extraction based on GCN

The feature extraction model of local spatio-temporal graph based on GCN is composed of a set of graph convolution (GCN) operation modules. Each GCN module extracts features from the segmented spatio-temporal graph sequences to obtain corresponding spatio-temporal graph feature sequences. The GCN operation can aggregate the features of each node and its neighboring nodes, and its input is the graph adjacency matrix of the local spatio-temporal graph, so as to realize the aggregation of the features of each node and its neighboring nodes on the adjacent time step. The specific calculation formula is shown in (6) - (7):

$$H^{(l)} = \sigma(A' H^{(l-1)} W + b) \tag{6}$$

$$H_i^{(out)} = \lambda_1 \max(H_i^{(1)}, H_i^{(2)}, \ldots, H_i^{(L)}) + \lambda_2 H_i^{(L)} \tag{7}$$

where A' represents the adjacency matrix of the local spatio-temporal graph with time step 3, $H^{(l-1)}$ is the input of the l-th graph convolutional layer, W and b are learnable parameters, σ is the activation function (such as ReLU, which is the same as standard GCN [27]), $H_i^{(j)}$ represents the j-th graph convolutional layer output vector of the i-th locally spatio-temporal graph sequence, and l is the number of graph convolutional layers, $\lambda_1 + \lambda_2 = 1$.

b) feature fusion based on weight

The weight based fusion process can make use of all the features of the nodes in the previous time step and the next time step, which makes the node feature information of the fusion result more comprehensive and eliminates the feature redundancy information while containing the local temporal and spatial correlation. The specific operation process is shown in Formula (8),

$$\hat{H}_i^{(j)} = \omega_{i-1} H_{t_{i-1}}^{(j)} + \omega_i H_{t_i}^{(j)} + \omega_{i+1} H_{t_{i+1}}^{(j)} \tag{8}$$

where $H_{t_{i-1}}^{(j)}, H_{t_i}^{(j)}, H_{t_{i+1}}^{(j)}$ represents the j-th graph convolution layer output vectors at three time steps of the i-th local spatio-temporal graph sequence, respectively, $\omega_{i-1} + \omega_i + \omega_{i+1} = 1$, substitute formula (8) into formula (7) to get formula (9):

$$\hat{H}_i^{(out)} = \lambda_1 \max(\hat{H}_i^{(1)}, \hat{H}_i^{(2)}, \ldots, \hat{H}_i^{(L)}) + \lambda_2 \hat{H}_i^{(L)} \tag{9}$$

After feature extraction of all local spatiotemporal graph sequences, the output vector sequence $H = \{\hat{H}_i^{(out)}\}, i = 1, 2, \ldots, n$ is finally obtained.

C. VISION FEATURE EXTRACTION

In order to improve the performance of the model, we correlate the road network and vehicle GPS track information to build visual features which can improve the model effect, including the generation of visual features and the extraction of visual features.

1) GENERATION OF VISUAL FEATURES

Based on road network data (including highway, urban expressway, urban main road, urban secondary road, urban branch road, rural road, bicycle path, pedestrian road, internal road and other 10 kinds of road data), same relevant softwares (such as ArcGIS) can be used to visualize road network data and obtain road network visualization diagram, and then the sensor location is also marked on the visual road map. In order to make use of the GPS trajectory information, we calculated the number of vehicles at T_k time on the road network between any two sensors based on the latitude and longitude data of taxis and buses, and calculated the proportion. The calculation formula is shown in (10),

$$R_k^{i,j} = \frac{a \cdot Texi_k^{i,j} + b \cdot Bus_k^{i,j}}{a \cdot Texi_k + b \cdot Bus_k} \tag{10}$$

where $Texi_k^{i,j}$ and $Bus_k^{i,j}$ represent the number of taxis and buses at T_k time between sensor i and sensor j respectively. $Texi_k$ and Bus_k represent the total number of taxis and buses at T_k time respectively. a and b are the traffic flow conversion ratios of taxis and buses respectively.

The calculated results of Formula (10) were taken as the influence index, and road network tracks with different color depths were constructed between sensors according to the influence index, and then the results were marked on the visual road network map. Finally, the road network and track image sequence $P = \{P_{T_1}, P_{T_2}, \ldots, P_{T_m}\}$ at different moments were generated.

2) EXTRACTION OF VISUAL FEATURES

For the generated road network and track image sequence $P = \{P_{T_1}, P_{T_2}, \ldots, P_{T_m}\}$, ResNet [28] was used for feature extraction to obtain visual feature representation vector sequence $V_{all}^{vision} = \{V_{T_1}^{vision}, V_{T_2}^{vision}, \ldots, V_{T_m}^{vision}\}$.

D. VECTORIZATION OF WEATHER FEATURES

In order to further improve the performance of the model, we consider the influence of weather factors (rainfall, temperature, air pressure, wind speed, etc.) after obtaining the graph features and visual features. We extract the features of weather factors by refining weather forecast result and vectorize the weather factors to get the weather representation vector. In the prediction stage, because the future weather is unknown, the model can only use the relevant weather forecast data to predict, and the accuracy of the weather forecast result will affect the accuracy of the prediction model. At the same time, the accuracy of weather forecast will also decay with time, so in this paper, we adopt weather factor extraction algorithm based on time decay to extract weather feature. The specific process of weather factor extraction is as follows:

In the training stage, time sequence feature vectors of four weather factors were constructed according to the meteorological data of historical rainfall, temperature, air pressure and wind speed as the input data of the model.

In the prediction stage, weather forecast data (rainfall, temperature, air pressure and wind speed) at a certain time interval in the future are selected as the initial data, and the polynomial-based method is adopted to refine the weather forecast data. The specific calculation process is shown in Formula (11):

$$w_i^{refine} = \alpha_1 w_1^{gt} + \alpha_2 w_2^{gt} + \ldots + \alpha_{i-1} w_{i-1}^{gt} + \beta w_i^{pred} \tag{11}$$

where $w_j^{gt} = [r_j^{gt}, t_j^{gt}, p_j^{gt}, s_j^{gt}]$ respectively represent the actual rainfall, temperature, air pressure and wind speed at time j, $w_i^{pred} = [r_i^{pred}, t_i^{pred}, p_i^{pred}, s_i^{pred}]$ respectively represent the weather forecast rainfall, temperature, air pressure and wind speed at time i,$w_i^{refine} = [r_i^{refine}, t_i^{refine}, p_i^{refine}, s_i^{refine}]$ respectively represent the refined weather rainfall, temperature, air pressure and wind speed at time i, α_i is the historical weather refine parameter, β is the current weather refine parameter, α_i and β are calculated by the mean of historical statistical results.

E. TRAFFIC FLOW PREDICTION

After the extraction of the three modal features, the dual channel fusion model based on MCB and attention was used to fuse the multi-modal features of the graph structure and vision, and then combined with the weather features to predict the future traffic flow.

1) MCB BASED MULTI-MODAL FUSION

For the feature vector sequence $H = \{\hat{H}_i^{(out)}\}, i = 1, 2, \ldots, n$ of the spatio-temporal graph and the visual vector sequence $V_{all}^{vision} = \{V_{T_1}^{v}, V_{T_2}^{v}, \ldots, V_{T_m}^{v}\}$ of the road network and trajectory, MCB algorithm is used for fusion the multi-modal features. The specific process is as follows:

Step 1. The spatio-temporal graph vector sequence and visual vector sequence were concatenated respectively, and the fully connected network was used to carry out dimension transformation on the concatenated vector, and the representation vectors V^g and V^v with the same dimension were obtained.

Step 2. The Count Sketch mapping function is used to process V^g and V^v respectively, and the corresponding Count Sketch vectors V_{trans}^g and V_{trans}^v are obtained.

Step 3. Vectors V_{trans}^g and V_{trans}^v are transformed into the frequency domain by using the fast Fourier transform (FFT) respectively and multiplied the vectors element-by-element, and then transformed by the inverse fast Fourier transform (IFFT) to obtain the fusion result vector $V_{MCB}^{g,v}$ in the time domain.

2) ATTENTION BASED MULTI-MODAL FUSION

For spatio-temporal graph vector sequence and visual vector sequence, we also used the cross-modal attention mechanism to fuse them, so as to obtain fusion vector based on attention mechanism. The specific operation process is as follows:

Step 1. The spatio-temporal graph vector sequence and visual vector sequence are concatenated and dimensionally transformed to obtain the representation vectors V^g and V^v with the same dimension.

Step 2. Based on the Attention mechanism, the representation vector V^g and V^v of the two modes are fused to obtain the fusion vector $V_{\text{att}}^{g,v}$. For the specific algorithm, see formula (12)-(14),

$$V_{fusion}^{graph} = \text{softmax}(\frac{W_1^g V^g (W_1^v V^v)^T}{\sqrt{d}})W_2^v V^v \tag{12}$$

$$V_{fusion}^{vision} = \text{softmax}(\frac{W_1^v V^v (W_1^g V^g)^T}{\sqrt{d}})W_2^g V^g \tag{13}$$

$$V_{\text{att}}^{g,v} = \text{concat}(V_{fusion}^{graph}, V_{fusion}^{vision}) \tag{14}$$

where W_1^g, W_2^g, W_1^v, W_2^v are all learnable parameters, d is the number of attention heads.

3) TRAFFIC FLOW PREDICTION BASED ON MULTI-MODAL DATA

In this paper, the MCB based multi-modal fusion result vector $V_{MCB}^{g,v}$, the multi-modal fusion vector $V_{\text{att}}^{g,v}$ based on cross-modal attention mechanism and the weather factor vector $V^{weather}$ are concatenate together to get the final fusion vector. Then the traffic flow prediction is carried out to obtain the prediction result Y at the next moment. The specific calculation process is shown in Formula (15) - (17):

$$V_{fusion}^{g,v} = \gamma_1 V_{MCB}^{g,v} + \gamma_2 V_{\text{att}}^{g,v} \tag{15}$$

$$V_{fusion}^{all} = \text{concat}(V_{fusion}^{g,v}, V^{weather}) \tag{16}$$

$$Y = \text{ReLU}(W_1 V_{fusion}^{all} + b_1) \cdot W_2 + b_2 \tag{17}$$

where W_1, b_1, W_2, b_2 are all learnable parameters, $\gamma_1 + \gamma_2 = 1$.

IV. EXPERIMENT

A. EXPERIMENTAL SETTING

1) DATASET

The data set consists of three parts: traffic flow data, GPS track data and weather data. The traffic flow data is the measured flow data of 287 sensor nodes with an interval of 5 minutes from March 1st, 2019 to May 31th, 2019 provided by Wuhu Municipal Bureau of Communications, and the data dimension is (287, 26496). The GPS track data of taxis and buses in this time period are provided by Wuhu Municipal Transportation Bureau, and the dimensions are (2000, 66240) and (150, 66240) respectively. Weather data are provided by Wuhu Meteorological Bureau with the weather data of 143 monitoring points in the region with an interval of one hour and the corresponding weather forecast data, and the data dimension is (143, 2208, 4).

2) HYPERPARAMETER SETTING

We split all datasets with ratio 6:2:2 into training sets, validation sets and test sets. We adopt weighted-STSGCN as the spatio-temporal graph feature extraction model, each weighted-STSGCN contains three graph convolutional operations with 64, 64, 64 filters respectively. The GCN node embedding size is set to 768. We adopt the standard ResNet152 to process the visual images with an output dimension of 2048. We use the Mean Absolute Error (MAE) as the loss function, the Adam optimizer with a learning rate of 10-5 is used as the optimization method, the batch size is set to 16 and the training epochs is set to 30.

3) EVALUATION METRICS

We use three evaluation metrics to evaluate the model performance: MAE, Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). The three are defined as follows:

$$MAE = \frac{1}{m}(\sum_{i=1}^{m} |y_i - \tilde{y}_i|) \tag{18}$$

$$MAPE = \frac{1}{m}(\sum_{i=1}^{m} |\frac{y_i - \tilde{y}_i}{y_i}|) \tag{19}$$

$$RMSE = \sqrt{\frac{1}{m}(\sum_{i=1}^{m} (y_i - \tilde{y}_i)^2)} \tag{20}$$

where m is data size, y_i and $\tilde{y}_i$ represent the predicted value and the actual value (ground truth) respectively. In addition, when calculating the MAPE, in order to avoid the value overflow caused by the denominator being 0, if $y_i = 0$, then $|\frac{y_i - \tilde{y}_i}{y_i}|$ is set to 0.

4) BASELINES

LSTM [29]: The time series features of sensors are extracted by using long and short term memory networks, and the problems of gradient disappearance and long term memory are solved by gating mechanism to predict the traffic flow.

STGCN [21]: The spatio-temporal graph convolutional network is used to encode the spatio-temporal graph structure information formed by sensors to predict the traffic flow.

ASTGCN [22]: The spatio-temporal graph convolutional network is used to encode the spatio-temporal graph structure information composed of sensors. Meanwhile, the spatio-temporal attention module is used to dynamically capture the spatial and temporal correlation to improve the accuracy of traffic flow prediction.

ST-3DNet [18]:3D convolutions and Rc blocks are used to model the two modes in the temporal and spatial, and then combined together in a weighted way to predict the final traffic flow.

STFGNN [19]: The Dynamic Time Warping (DTW) algorithm is used to calculate the similarity of time series to generate "temporal graph", and a novel fusion operation is used to integrate the temporal and traffic spatial graph to obtain hidden spatial-temporal dependencies.

STDEN [30]: A physics-guided deep learning model, which casts the physical mechanism of traffic flow dynamics into a deep neural network framework. This model can bridge the gap between purely data-driven and physics-driven approaches.

LMF [31]: The low-rank multi-modal fusion method is used to decompose the high-order tensor into the low-rank tensor, which greatly reduced the computational complexity, realizing the multi-modal data fusion in the traffic flow prediction model.

MulT [32]: For any target mode, cross-modal Transformer is used to convert the information of other modes to the target mode. It can fully extract the cross-modal feature and model the long distance sequential dependence of the modes, so as to realize the multi-modal traffic flow prediction.

MCGMF [33]: The cross-modal gating mechanism is used to remove inter-modal noise, and complementary information is extracted to enhance the modal representation. The contribution of different modes is paid attention to by weight and similarity constraints. Finally, the multi-level representation of modes is integrated for traffic flow prediction.

B. EXPERIMENT RESULTS

In order to prove the performance of our proposed model in this paper, it is compared with the above baselines experimentally. Table 1 shows the comparison of experimental results of different methods in constructed data sets. Compared with the other 9 benchmark methods, our proposed model shows good performance in each evaluation metric of the data set.

TABLE 1. **Comparison of experimental results.**

Model	MAE	MAPE(%)	RMSE
LSTM	33.15	23.76%	50.78
STGCN	28.54	18.06%	44.45
ASTGCN	27.14	18.05%	41.62
ST-3DNet	31.26	20.15%	48.57
STFGNN	26.56	17.67%	40.78
STDEN	25.08	16.90%	38.85
LMF	25.74	16.74%	40.29
MulT	21.34	14.52%	33.68
MCGMF	19.03	**12.53%**	30.25
ours	**18.32**	12.56%	**29.12**

As shown in Table 1, the LSTM model can effectively utilize the time dependence according to the model characteristics of its time series, but it cannot consider the important spatial dependence in the spatio-temporal prediction. Hence it achieves poor results compared with other baseline models. The STGCN and ASTGCN models process the spatio-temporal graph structure of traffic sensors by using the spatio-temporal graph convolution method, which makes the models add the important dependence of spatial correlation information while considering the temporal correlation. Therefore, compared with the LSTM model, the effect of STGCN and ASTGCN models is greatly improved. At the same time, both STGCN and ASTGCN models use the spatio-temporal graph convolution method to process the spatiotemporal graph structure, but ASTGCN also adds the spatio-temporal attention module to realize effective dynamic capture of spatio-temporal correlation, thus achieving better model effect compared with STGCN model. Similar to ASTGCN and STGCN, STFGNN and STDEN take into account both time and space dependence, but add additional algorithms to enhance the effectiveness of the model while using spatio-temporal information, thus achieving better results. By comparing the experimental results of LSTM, STGCN, ASTGCN, STFGNN and STDEN, we can see that the performance of the spatial-temporal model is better than that of the time series prediction model which only considers the time performance, and the attention mechanism can also improve the performance of the model. Compared with the LSTM, STGCN, ASTGCN, STFGNN and STDEN models, the ST-3DNet model adopts the traffic grid network transformed by road network as the input to predict the traffic flow, so the improved ST-3DNet model is worse than the STGCN and ASTGCN models considering the temporal and spatial characteristics. However, compared with the LSTM model which only considers the time dependence, the ST-3DNet model has a relatively good effect.

LMF, MulT, MCGMF and our proposed model all adopt the multi-modal fusion method to integrate the structure information of spatio-temporal graph, road network information and weather feature information to predict the traffic flow. It can be seen from the table that compared with the model that only considers the single mode characteristics, the multi-modal fusion model has better effect. Compared with the LMF model, MCGMF and our proposed model have improved in all evaluation metrics, indicating that the attention mechanism module can extract significant features of modes, remove irrelevant noise, combine multi-level features to effectively integrate multi-modal features, and achieve more accurate traffic flow prediction. Compared with MulT, MCGMF and our proposed model are both improved, but not by much. This is because MCGMF and our proposed model reduce the use of cross-modal Transformer, and introduce cross-modal attention mechanism and gating mechanism, indicating that cross-modal attention mechanism and gating mechanism can remove inter-mode noise. The complementary information between modes is extracted to enhance the correlation of multimodal features. Compared with MCGMF, our proposed model adopts a dual-channel multi-modal fusion mechanism. The attention mechanism is used as well as the MCB for multi-modal fusion, and the gating mechanism is used to further control the dual-channel fusion results, so a better model effect is achieved.

In order to analyze the multi-step prediction ability of the model, we select 15 minutes, 30 minutes, and 60 minutes as prediction intervals to conduct multi-step prediction experimental analysis, and the experimental results are shown in Table 2.

It can be seen from Table 2 that our proposed model still has advantages in the task of multi-step prediction. Considering that multi-modal information has a long-term impact on traffic conditions, such as weather factors can affect travel throughout the day. LMF, MulT, MCGMF and our proposed model are still superior to the model of single mode in multi-step prediction because long-term influence factors such as weather are taken into account. Compared with the other two models, the self-attention module in MCGMF and our proposed model is sensitive to dynamic weights and can capture the importance change of multi-modal features in the time dimension. Hence, it has better results in traffic flow prediction.

TABLE 2. Multi-step prediction experimental results.

Time	Metric	LSTM	STGCN	ASTGCN	ST-3DNet	STFGNN	STDEN	LMF	MulT	MCGMF	ours
15 min	**MAE**	34.64	29.26	28.35	32.65	27.36	25.99	26.77	22.37	19.95	**19.14**
	MAPE(%)	25.04%	19.05%	18.97%	21.48%	18.26%	17.18%	17.48%	15.42%	13.64%	**13.41%**
	RMSE	52.85	45.61	43.74	50.18	42.36	39.97	41.74	35.10	31.65	**30.19**
30 min	**MAE**	37.72	30.46	29.52	34.19	28.38	27.37	28.00	24.17	21.13	**20.95**
	MAPE(%)	27.18%	19.90%	20.23%	22.11%	19.31%	17.91%	18.26%	16.96%	14.56%	**14.33%**
	RMSE	57.23	47.37	45.80	52.78	44.14	42.02	43.70	37.60	33.35	**33.11**
60 min	**MAE**	41.08	32.65	30.90	36.67	30.68	29.80	30.08	26.67	22.52	**22.16**
	MAPE(%)	29.96%	21.48%	21.32%	23.57%	20.64%	19.23%	19.73%	19.24%	**15.40%**	15.50%
	RMSE	61.79	50.18	47.37	57.10	46.92	46.42	46.56	41.01	35.33	**34.72**

C. ABLATION ANALYSIS

In order to further analyze the role of different modules in our proposed model, we conduct the following three types of ablation experiments on the model:

① **ablation of road network feature**: The road network information features are removed, and only the sensor spatio-temporal graph features and weather features are used for traffic flow prediction.

② **ablation of weather feature**: The weather features are removed, and only the sensor spatio-temporal graph features and road network information features are used to predict the traffic flow.

③ **ablation of spatio-temporal graph feature**: The sensor spatio-temporal graph features are removed, and only weather features and road network information features are used for traffic flow prediction.

As can be seen from the results of evaluation metrics in Table 2, the characteristics of the three modules in our proposed model are all very important for the traffic flow prediction. However, due to the significant nonlinear temporal dynamic variation of traffic flow prediction, model ③ without the structural feature module of the spatio-temporal graph has the worst prediction performance due to the uncaptured spatio-temporal variation. However, model ① is inferior in performance because it does not consider potential road network traffic grid information and ignores road network vehicle information factors that have a great influence on traffic flow prediction. However, model ② only considers the influence of weather factors, and the weather factors use the revised data of weather forecast in the model prediction, which reduces the importance of weather factor features, so the model performance is less reduced than our proposed model. In general, all the modules can effectively improve the prediction effect of the model, but there are differences among them.

As can be seen from Table 4, in the multi-step prediction task, the model effect decreases to varying degrees after the ablation of each module. As multi-modal transport information has a long-term impact on traffic conditions, compared with other ablation models, our proposed model fused more modal information, so the robustness of our proposed model in multi-step prediction is better than other ablation models. At the same time, it can be seen from the multi-step prediction process that the model ③ without the feature module of the spatio-temporal graph structure still has the worst multi-step prediction effect. The model ① without considering the potential traffic grid information of the road network is only higher than the model ③, and the model ② only considering the influence of weather factors is second only to our proposed model, which is consistent with the experimental results in Table 3. This shows the effectiveness of each proposed module, and the multi-modal fusion model has higher stability for different time step prediction.

TABLE 3. Ablation experiment results.

Model	MAE	MAPE(%)	RMSE
①	25.05	17.82%	38.85
②	21.56	14.62%	33.99
③	27.39	18.42%	41.95
Ours	**18.32**	**12.56%**	**29.12**

TABLE 4. Multi-step prediction results of ablation experimental.

Time	Metric	①	②	③	Ours
15 min	**MAE**	25.98	22.55	28.24	**19.14**
	MAPE(%)	18.11%	15.04%	19.32%	**13.41%**
	RMSE	39.98	35.43	43.90	**30.19**
30 min	**MAE**	27.24	24.13	29.34	**20.95**
	MAPE(%)	19.01%	15.97%	20.39%	**14.33%**
	RMSE	42.05	37.59	44.90	**33.11**
60 min	**MAE**	29.65	25.85	31.70	**22.16**
	MAPE(%)	20.45%	17.43%	20.80%	**15.50%**
	RMSE	45.63	40.18	49.18	**34.72**

V. CONCLUSION

Based on the traditional time series prediction model and the time series data of traffic sensors, we construct the traffic flow spatio-temporal graph, and adopts the weighted STSGCN algorithm to extract the structure information of the traffic flow spatio-temporal graph to obtain the features. At the same time, road network and track images are generated based on road network and track data, and then ResNet is used to extract image features to get the visual features of road network and track. Finally, the spatio-temporal graph features and visual features were fused based on the MCB and attention dual-channel multi-modal fusion model, and the traffic flow was predicted by combining weather features. The effectiveness of our proposed model is proved by comparison of experimental results, and the effectiveness of each module of our proposed model is also proved by the results of ablation experiments. The spatio-temporal matrix transformation algorithm of vehicle speed proposed in this paper has defects in the weighted traffic flow spatio-temporal graph. In the future, we will improve the spatio-temporal matrix transformation algorithm to further enhance the effect of the traffic flow prediction model.

REFERENCES

[1] A. Sarker, H. Shen, and J. A. Stankovic, "MORP: Data-driven multi-objective route planning and optimization for electric vehicles," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 1–35, Jan. 2018.

[2] V. Alarcon-Aquino and J. A. Barria, "Multiresolution FIR neural-network-based learning algorithm applied to network traffic prediction," *IEEE Trans. Syst., Man Cybern., C Appl. Rev.*, vol. 36, no. 2, pp. 208–220, Mar. 2006.

[3] M. Castro-Neto, Y.-S. Jeong, M.-K. Jeong, and L. D. Han, "Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions," *Exp. Syst. Appl.*, vol. 36, no. 3, pp. 6164–6173, Apr. 2009.

[4] Z. Duan, Y. Yang, K. Zhang, Y. Ni, and S. Bajgain, "Improved deep hybrid networks for urban traffic flow prediction using trajectory data," *IEEE Access*, vol. 6, pp. 31820–31827, 2018.

[5] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2191–2201, Oct. 2014.

[6] S. George and A. K. Santra, "Deep learning techniques for traffic flow prediction in intelligent transportation system: A survey," *Test Eng. Manag.*, vol. 82, pp. 9773–9789, Jan./Feb. 2020.

[7] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial–temporal network data forecasting," in *Proc. 34th AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 1, pp. 914–921.

[8] Y. Xie, Y. Zhang, and Z. Ye, "Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 22, no. 5, pp. 326–334, Jul. 2007.

[9] J. Zhou and A. K. H. Tung, "SMiLer: A semi-lazy time series prediction system for sensors," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, May 2015, pp. 1871–1886.

[10] D. Deng, C. Shahabi, U. Demiryurek, L. Zhu, R. Yu, and Y. Liu, "Latent space model for road networks to predict time-varying traffic," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1525–1534.

[11] P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, and J. Sun, "A spatiotemporal correlative K-nearest neighbor model for short-term traffic multistep forecasting," *Transp. Res. C, Emerg. Technol.*, vol. 62, pp. 21–34, Jan. 2016.

[12] Y. Tong, Y. Chen, Z. Zhou, L. Chen, J. Wang, Q. Yang, J. Ye, and W. Lv, "The simpler the better: A unified approach to predicting original taxi demands based on large-scale online platforms," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1653–1662.

[13] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: A deep learning approach for short–term traffic forecast," *IET Intell. Transp. Syst.*, vol. 11, no. 2, pp. 68–75, Mar. 2017.

[14] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 1–7.

[15] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, "Deep multi-view spatial–temporal network for taxi demand prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–12.

[16] Y. Liu, H. Zheng, X. Feng, and Z. Chen, "Short-term traffic flow prediction with conv-LSTM," in *Proc. 9th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2017, pp. 1–6.

[17] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, "Revisiting spatialtemporal similarity: A deep learning framework for traffic prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5668–5675.

[18] S. Guo, Y. Lin, S. Li, Z. Chen, and H. Wan, "Deep spatial–temporal 3D convolutional neural networks for traffic data forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3913–3926, Oct. 2019.

[19] S. Li, L. Ge, Y. Lin, and B. Zeng, "Adaptive spatial–temporal fusion graph convolutional networks for traffic flow forecasting," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, vol. 35, Jul. 2022, pp. 4189–4196.

[20] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-GCN: A temporal graph convolutional network for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3848–3858, Sep. 2020.

[21] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," 2017, *arXiv:1709.04875*.

[22] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial–temporal graph convolutional networks for traffic flow forecasting," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 922–929.

[23] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," 2017, *arXiv:1707.01926*.

[24] Z. Fang, Q. Long, G. Song, and K. Xie, "Spatial-temporal graph ODE networks for traffic flow forecasting," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 364–373.

[25] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph WaveNet for deep spatial–temporal graph modeling," 2019, *arXiv:1906.00121*.

[26] X. Fan, C. Xiang, C. Chen, P. Yang, L. Gong, X. Song, P. Nanda, and X. He, "BuildSenSys: Reusing building sensing data for traffic prediction with cross-domain learning," *IEEE Trans. Mobile Comput.*, vol. 20, no. 6, pp. 2154–2171, Jun. 2021.

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction," 2018, *arXiv:1801.02143*.

[30] J. Ji, J. Wang, Z. Jiang, J. Jiang, and H. Zhang, "STDEN: Towards physics-guided neural networks for traffic flow prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 4, pp. 4048–4056.

[31] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Stroudsburg, PA, USA: ACL, vol. 1, 2018, pp. 2247–2256.

[32] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics* Stroudsburg, PA, USA: ACL, 2019, pp. 6558–6569.

[33] Y. Miao, S. Yang, T. Liu, W. Zhang, L. Zhu, and M. Zhou, "Multimodal sentiment analysis based on cross-modal gating mechanism and improved fusion method," *Appl. Res. Comput.*, vol. 40, no. 7, pp. 1–8, 2023.

XIAOAN QIN was born in 1982. He received the master's degree from Anhui University, in 2010. He is currently a Vice Professor with the School of Information and Artificial Intelligence, Anhui Business College of Vocational Technology. His research interest includes big data technology.