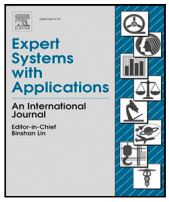
Expert Systems With Applications 237 (2024) 121548

Contents lists available at ScienceDirect

Expert Systems With Applications

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



Multi-task-based spatiotemporal generative inference network: A novel framework for predicting the highway traffic speed

Guojian Zou [a,b], Ziliang Lai [a,b], Ting Wang [a,b], Zongshi Liu [a,b], Jingjue Bao [a,b], Changxi Ma [c], Ye Li [a,b,*], Jing Fan [a,b,d]

[a] The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 201804, PR China
[b] College of Transportation Engineering, Tongji University, Shanghai 201804, PR China
[c] School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou 730070, PR China
[d] China Railway First Survey and Design Institute Group Co.,LTD, Xi'an, PR China

ARTICLE INFO

Keywords:
Long-term highway traffic speed prediction
Dynamic spatiotemporal correlation
Graph neural networks
Fusion gate mechanism
Generative inference
Multi-task learning

ABSTRACT

Accurately predicting the highway traffic speed can reduce traffic accidents and transit time, and it also provides valuable reference data for traffic control in advance. Three essential elements should be considered in highway traffic speed prediction: (1) dynamic spatiotemporal correlation, (2) prediction error propagation elimination, and (3) traffic speed heterogeneity on the highway network. A multi-task-based spatiotemporal generative inference network (MT-STGIN) is proposed to address the above challenges. [1] MT-STGIN consists of three parts: an encoder based on spatiotemporal correlation extraction block (ST-Block), a decoder based on masked ST-Block and generative inference, and multi-task learning. The encoder is first used to extract the dynamic spatiotemporal correlation of the highway network. The decoder concentrates on correlating historical and - target sequences and generates the target hidden outputs rather than a dynamic step-by-step decoding way. Finally, a multi-task learning method is used to predict traffic speed on different types of road segments because of heterogeneity and shares the underlying network parameters. The evaluation experiments use the monitoring data of the highway in Yinchuan City, Ningxia Province, China. The experimental results demonstrate that the performance of the proposed prediction model is better than that of the baselines, and it can efficiently solve the problem of long-term highway speed prediction.

1. Introduction

The highway network plays a key role in the normal operation of the city owing to its higher speed, capacity, and safety (Magazzino & Mele, 2021). Recently, various intelligent algorithms have been applied in transportation engineering, such as the stability analysis (Cheng, Lyu, Zheng, & Ge, 2022; Wang, Cheng, & Wu, 2022) and forecasting (Chen et al., 2020) of heterogeneous traffic flow, travel time estimation (Zou, Lai, Ma, Tu et al., 2023), and traffic speed perdition (Chen et al., 2022; Zou, Lai, Ma, Li and Wang, 2023). Traffic speed prediction is the main function of the highway intelligent monitoring system, which provides helpful information for travelers and traffic management departments to ensure the safe and smooth operation of the highway (James, Markos, & Zhang, 2021; Yu, Lee, & Sohn, 2020). And traffic speed is one of the essential factors in measuring highway traffic state, i.e., freely or congested.

Highway traffic speed prediction is a dynamic spatiotemporal problem affected by both temporal and spatial dimensions, as shown in Fig. 1. The classic traditional methods include the historical average model (HA), autoregressive integrated moving average (ARIMA) (Ahmed & Cook, 1979), and their improved methods are widely used in traffic-related time series prediction tasks (Duan, Mao, Zhang, & Wang, 2016; Wang, Liu, Dong, Qian, & Wei, 2016). The advantage of these methods is that only a small batch of samples can roughly predict the target data. However, they face the challenge of not being able to extract the nonlinear correlation of input data. Methods based on machine learning, such as support vector machine (SVM) (Vanajakshi & Rilett, 2004), have been widely studied for traffic prediction tasks. These

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: 2010768@tongji.edu.cn (G. Zou), 2033402@tongji.edu.cn (Z. Lai), 2110763@tongji.edu.cn (T. Wang), chuochuoliu@tongji.edu.cn (Z. Liu), baojingjue@tongji.edu.cn (J. Bao), machangxi@mail.lzjtu.cn (C. Ma), JamesLI@tongji.edu.cn (Y. Li), jing.fan@tongji.edu.cn (J. Fan).

[1] https://doi.org/10.24433/CO.8500259.v1

https://doi.org/10.1016/j.eswa.2023.121548
Received 4 April 2023; Received in revised form 6 September 2023; Accepted 9 September 2023
Available online 15 September 2023

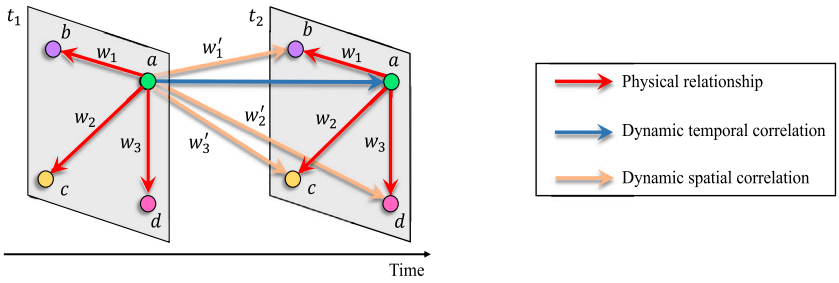


Fig. 1. The four nodes *a*, *b*, *c* and *d* represent four different road segments. The red arrow indicates the physical relationship between segment *a* and adjacent segments *b*, *c* and *d*. The blue arrow represents the impact of road segment *a* on itself in the next time step. The orange arrow indicates the influence of segment *a* on the adjacent segments *b*, *c* and *d* in the next time step.

methods take into account the nonlinear correlations of the data from the perspective of big data and have a more remarkable accuracy improvement than the traditional classical methods. However, traffic speed data has dynamic spatiotemporal correlation, and these methods lack the flexibility to implement and process multi-dimensional traffic data. Therefore, the accuracy and application value is difficult to be further improved.

In recent years, deep learning has brought revolutions in computer vision (Lu et al., 2023) and natural language processing (Eloundou, Manning, Mishkin, & Rock, 2023). Subsequently, some deep learning methods have been deployed in traffic speed prediction and achieved satisfactory performance (Jia et al., 2021; Lee, Jeon, & Sohn, 2020; Qu, Lyu, Li, Ma, & Fan, 2021). For instance, Chen et al. (2022) propose an ensemble framework to estimate the vehicular moving speed via trajectory extracted from video. Since traffic speed data has the characteristics of spatial and temporal two dimensions, existing deep learning methods are modeled from these two dimensions. Spatial correlation is mainly extracted using convolutional neural networks (CNNs) (Rempe, Franeck, & Bogenberger, 2022), while temporal correlation is extracted using recurrent neural networks (RNNs) (Afrin & Yodo, 2022). The highway traffic data conforms to discrete distribution, that is, non-Euclidean structure data. However, CNNs are mainly used to extract the spatial correlation of the Euclidean structure data, such as pictures, which is unsuitable for non-Euclidean structural data (Zafeiriou et al., 2022). Some current methods force the sampling of traffic data into a standard form and feed it to CNNs (Yang, Liu, Zhu, Ban, & Wang, 2021), which may lose important spatial correlation information of the traffic data. The ideal way to express traffic data consists in maintaining the original spatial structure, and expressing it in a graph network. The latest researches have extended CNNs to graph neural networks (GNNs) (Li, Chen et al., 2021; Li, Lu, Yi and Gong, 2021), which can process data with arbitrary non-Euclidean structural data (Wu et al., 2020). GNNs have been successfully used in traffic prediction tasks (Zou, Lai, Ma, Tu et al., 2023), including traffic speed prediction (Jin et al., 2023; Zou, Lai, Ma, Li et al., 2023).

The accuracy of highway traffic speed forecasts is affected by numerous elements within the traffic system, especially the dynamic spatial and temporal correlations asserted by previous research (Fang, Zhao, Qin, Luo, & Wang, 2022; Feng & Tassiulas, 2022; Zheng, Fan, Wang, & Qi, 2020). Absorbing the proven experiences, we proposed new questions hidden in long-term traffic speed forecasting. For example, traffic incidents like road maintenance easily affects traffic speeds, making velocities fluctuate and forecasting more challenging. In addition, traffic speed on the target road segment is affected directly by the speed in the upstream and downstream directions, defined as the physical relationship, and easily neglected in dynamic spatial correlations modeling. Moreover, for long-term traffic speed prediction, existing studies have difficulty solving the prediction error propagation in the spatial and temporal dimensions, resulting in the accumulation of errors and limited model prediction accuracy. Furthermore, the highway network has tremendous traffic speed heterogeneity. There are two root causes: (1) the difference in speed limitation between ramps and main roads; (2) high traffic flow complexity owing to various factors, including vehicle type differences and changing travel behavior. Therefore, the highway segments are divided into three categories: entrance toll to the gantry (i.e., leaving the entrance toll), gantry to gantry (i.e., main roads), and gantry to exit ramp toll (i.e., nearing the exit toll), regarded as multi-task learning.

In this paper, a multi-task-based spatiotemporal generative inference network (MT-STGIN) for highway traffic speed prediction is proposed. MT-STGIN, employing multi-task learning to handle velocity heterogeneity, accurately forecasts various types of road traffic speeds; in addition, the designed model effectively addresses speed fluctuations in complex traffic environments, such as changes in traffic conditions following incidents; moreover, the essential core function is to model dynamic traffic spatiotemporal correlation; furthermore, it effectively avoids error propagation by utilizing one-step decoding instead of the autoregressive model, which involves step-by-step decoding. Therefore, a spatiotemporal correlation extraction block (ST-block) is first designed: a semantic layer based on 1-D CNNs is constructed to enhance the contextual semantic with n-grams (Cao, Lu, Zhou, & Li, 2018); a fusion gate mechanism (F-Gate) is proposed to incorporate the physical relationship extracted by multi-head graph convolutional network (multi-head GCN) into dynamic spatial correlation extracted by multi-head graph attention network (multi-head GAT); a temporal attention network models the dynamic temporal dependency, which is later combined with spatial correlation through F-Gate. To avoid long-term prediction error propagation in the spatial and temporal dimensions, a specific transformer generates the long-term target hidden outputs in a single step by bridging the relationship between the historical and target sequences, called bridge transformer (BridgeTrans), and this architecture is defined as generative inference. Finally, a multi-task learning method shares the underlying network parameters for different tasks (Gao et al., 2023). The main contributions of this paper are summarized as follows:

1. A novel multifunctional feature extractor ST-Block is proposed for spatiotemporal correlation modeling. The semantic enhancement module that employs semantic layers to perceive local contextual semantics to overcome speed fluctuation is first designed. Then, the physical relationship is molded for traffic diffusion simulation and subsequently embedded adaptively into dynamic spatial correlation via the fusion gate mechanism, achieving syncretism. Finally, the dynamic temporal dependency and spatial correlation are fused automatically without additional parameters rather than concatenation or addition.
2. For long-term highway traffic speed prediction processing, to avoid error propagation in spatial and temporal dimensions, a generative inference is developed and then applied to pay attention to the correlation between historical and -target sequences to generate the hidden target outputs rather than dynamic step-by-step decoding like ST-GRAT. Moreover, residual connection (Huang, Ye, Ding, Yang, & Xiong, 2022) and batch

normalization (BN) (Huang et al., 2023) are added to each network layer.
3. Highway segments express the velocity differences combined with preceding explanations, i.e., speed limitation and traffic flow complexity. We partition traffic speed prediction on highway networks into three subtasks due to speed heterogeneity on various types of roads. Therefore, multi-task learning for highway speed heterogeneity is devised to correlate relative tasks, share the underlying network parameters, and improve the performance of each subtask.
4. Several experiments are conducted on the highway traffic dataset. The experimental results show that the proposed MT-STGIN model outperforms the state-of-the-art baseline methods.

The remainder of this paper is organized as follows. In Section 2, the previous related studies are summarized. Section 3 describes the relative definition and problem statement. Section 4 details the proposed MT-STGIN model. Section 5 presents the experiments and the results. Finally, the conclusion and future work are drawn in Section 6.

2. Related work

The existing traffic speed prediction techniques can be divided into three categories: statistical methods, machine learning methods, and deep learning methods.

2.1. Statistical methods

Statistical methods have been successfully applied to traffic speed prediction tasks, including the HA and ARIMA model (Ahmed & Cook, 1979; Duan et al., 2016; Wang et al., 2016). HA uses the average value of the historical data at the same time every day as the predicted value at the same time instant in the future prediction task (Liu et al., 2019). ARIMA is a traditional time series prediction method that combines moving average and autoregressive components to model historical time series data (Ahmed & Cook, 1979; Duan et al., 2016; Wang et al., 2016). However, traffic speed is nonlinear, and the parametric approaches are based on prior knowledge, theoretical assumptions, and simple mathematical statistics. Therefore, they face difficulties in accurately predicting traffic speed.

2.2. Machine learning approaches

Traditional machine learning methods alleviate the problems encountered by statistical methods because they can efficiently extract the nonlinear features of the traffic data (Hong, 2011; Lin, Li, Chen, Ye, & Huai, 2017). For instance, Vanajakshi and Rilett (2004) propose a regression technique, referred to as a SVM, for short-term traffic speed prediction. Jiang and Fei (2016) propose a novel vehicle speed forecast model in the context of vehicular networks, in which hidden Markov models (HMMs) are then used to present the statistical relationship between individual vehicle velocities and traffic speeds. Shin and Sunwoo (2018) propose a vehicle speed prediction algorithm based on a random model, which uses a Markov chain with speed constraints as the basis. Shin and Sunwoo (2018) propose a prediction algorithm for vehicle speed based on a stochastic model using a Markov chain with speed constraints estimated by an empirical model. Significantly, the Markov chain, which forms the basis of the proposed algorithm, generates the velocity trajectory stochastically within speed constraints. Zhang, Feng, Lu, Song, and Wang (2020) propose a traffic factor state network model (TFSN) based on high-order multivariate Markov models, in order to establish the relationship between speed and related factors. However, these machine learning methods concentrate on modeling shallow nonlinear features and limited extraction of the complex spatiotemporal correlation of the traffic variables.

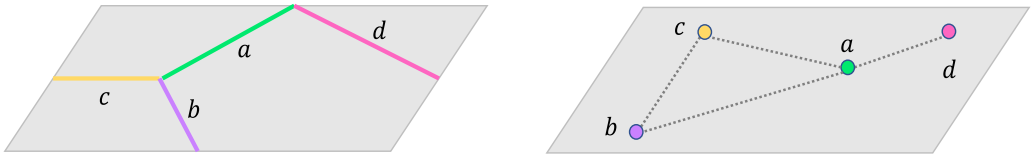
2.3. Deep learning algorithms

In recent years, deep learning has achieved a high performance when dealing with regression problems. In the early studies on traffic speed prediction using neural networks, the researchers found that neural network algorithms are more suitable for processing nonlinear big traffic data than machine learning methods (Tang, Yiu, Chan, & Zhang, 2023). For example, Jia, Wu, and Du (2016) propose a deep belief network (DBN) model based on deep learning for short-term traffic speed information prediction. Tang, Liu, Zou, Zhang, and Wang (2017) propose a traffic speed prediction model based on an improved fuzzy neural network (FNN). Tang et al. (2023) propose a novel neighbor subset deep neural network (NSDNN) for short-term traffic speed prediction, and the approach conjoins deep neural network (DNN) and the subset selection method to extract valuable variables from nearby roads. However, traffic speed is a long-term dependency series, traditional neural networks are challenging to model this temporal correlation and more complex deep learning methods are required.

Recurrent neural network (RNN) is a typical temporal correlation extractor, and its variants have been used in traffic speed prediction (Gu, Lu, Qin, Li, & Shao, 2019; Yi & Bui, 2020). For example, Ma, Tao, Wang, Yu, and Wang (2015) propose a novel neural network architecture, referred to as the long short-term neural network (LSTM NN), to efficiently capture the nonlinear traffic dynamic characteristics. Wang, Chen, and He (2019) use a bidirectional long short-term memory neural network (Bi-LSTM NN) to model each critical path. They then use multiple Bi-LSTM layers stacked together to merge time information. Meng et al. (2020) propose a long short-term memory with a dynamic time warping (D-LSTM) model for time series processing, which integrates a dynamic time warping algorithm, can fine-tune the time feature, thus adjusting the current data distribution to be close to the historical data. Qu et al. (2021) propose the features injected recurrent neural networks (FI-RNNs) that combine time series data and use a stacked RNN and encoder to learn sequential features of traffic data. However, if only RNNs are used to process the traffic data's temporal correlation, the spatial correlation's impact on the prediction may be ignored.

To solve the problems encountered by RNNs, the spatiotemporal prediction models based on CNN have been designed (Jia et al., 2021; Lv et al., 2018; Song et al., 2017; Zang, Ling, Wei, Tang, & Cheng, 2018). For example, Zhou, Zhang, Yu, and Chen (2019) propose a novel speed prediction method, referred to as the spatiotemporal and deep tensor neural networks (ST-DTNN), which is mainly used for large-scale urban networks with mixed road types. Yang et al. (2021) propose a path-based speed prediction neural network (PSPNN) composed of CNN and a bidirectional LSTM (Bi-LSTM) network, which extracts the temporal and spatial correlations of historical data to perform path-based speed prediction. In order to predict the lane-level short-term traffic speed, Lu, Rui, and Ran (2020) propose a novel mixed deep learning (MDL), which consists of a convolutional long short-term memory (Conv-LSTM) layer, a convolutional layer, and a fully connected layer. To model the spatiotemporal features of traffic speeds, a hybrid model combining convolutional neural networks with the gated recurrent unit (CNN-GRU) is developed by Ma, Zhao, Dai, Xu, and Wong (2022). CNN is designed to extract deep spatial features, and the bidirectional GRU is used to model long-range dependence. The common problem with these methods is that they extract the spatial correlation using the CNNs, while the traditional CNNs deal with this issue in the context of Euclidean space, and therefore they are not suitable for non-Euclidean space.

Obtaining complex spatial dependence is a critical issue in traffic speed prediction. In recent years, GNNs have been widely used for traffic speed prediction (Jiang & Luo, 2022), since they can process non-Euclidean structure data (Fang et al., 2022; James et al., 2021; Zhang, Li, Song, & Dong, 2021). For example, Zhao et al. (2019) propose a novel temporal graph convolutional network (T-GCN) for



(a) Highway Network (b) Graph Network

Fig. 2. (a) Example of highway network with four road segments. Each road segment is represented by different characters and colors. (b) The highway network can be represented as a graph, while the road segments can be represented as nodes.

traffic prediction, which is combined with the graph convolutional network (GCN) and gate recurrent unit (GRU). Li, Yu, Shahabi, and Liu (2018) propose a diffusion convolutional recurrent neural network (DCRNN), a deep learning framework incorporating spatial and temporal dependency into traffic prediction. These approaches extend the previous works to GCN; however, over-depend on handling pre-defined directed graphs. Therefore, Zheng et al. (2020) proposed a graph multi-attention network (GMAN) to predict long-term traffic speed at different locations on the traffic network. Significantly, the graph attention mechanism is first used to extract dynamic spatial correlation without pre-defined topology architecture of road network, temporal attention based on Transformer (Vaswani et al., 2017) is then applied to model the time series dependency, and an adaptive feature fusion method is finally proposed to combine them. Park et al. (2020) propose a novel spatiotemporal graph attention (ST-GRAT) method based on the self-attention mechanism. It efficiently captures the spatiotemporal dynamic features in the road network and improves prediction accuracy. Li and Lasenby (2021) propose an attention-based spatiotemporal graph attention network (AST-GAT) for segment-level traffic speed prediction, consisting of a self-attention-based GAT network and an attention-based LSTM network. We recently proposed a novel data-driven spatiotemporal generative inference network (STGIN) for long-term highway traffic speed prediction (Zou, Lai, Ma, Li et al., 2023), which consists of semantic enhancement, spatiotemporal correlation extractor, and generative architecture, but cannot adaptively fuse physical- and dynamic correlations, does not extract deep target spatiotemporal dependencies, and ignores speed heterogeneity.

In spatiotemporal correlation extraction, these popular traffic prediction methods are not conscious of the traffic speed fluctuation for semantic continuity, not considering the necessity of traffic diffusion embedded into dynamic spatial correlation, neglecting adaptive fuse spatial- and temporal dependencies of intra and inter without additional parameters. In addition, the error propagation in spatial and temporal dimensions is not resolved via a decoder framework for long-term traffic speed prediction, such as GMAN (Zheng et al., 2020) and EGAF-Net (Qiu, Zhu, Jin, Sun, & Du, 2023). Moreover, traffic speed heterogeneity is not distinguished, such as in our recently proposed speed prediction model (Zou, Lai, Ma, Li et al., 2023); multi-task learning is a positive way to correlate the diversified speeds because they have a close relation, such as ramps and main roads. In this paper, inspired by the recent studies on graph neural networks in traffic speed prediction, a novel long-term highway speed prediction model, referred to as MT-STGIN, is proposed.

3. Preliminary

In this section, we first present several preliminaries and define our problem formally.

Definition 1 (*Road Network*)**.** The highway network can be abstracted as a graph network, as shown in Fig. 2, while each road segment can be mapped to the graph network node (cf. Fig. 2(b)). The graph network is defined as G = (V,E,A), where V represents the nodes, E denotes the edges, $A \in R^{N\times N}$ is the adjacency matrix, and N represents the number of nodes. The adjacency matrix A indicates whether there is a directed connection between the road segments. More precisely, 'one' indicates a connection between two road segments (and vice versa), while 'zero' indicates no connection.

Definition 2 (*Embedding Statement*)**.** Three types of input data information are included for traffic speed prediction: traffic speed, road segment embedding, and timestamp embeddings. The traffic speed input to MT-STGIN at time step t is $XS_t \in R^{N\times d_X}$, road segment embedding at time step t is $XP_t \in R^{N\times d}$, timestamp embeddings at time step t including minute embedding $XM_t \in R^{N\times d}$, hour embedding $XH_t \in R^{N\times d}$, day embedding $XD_t \in R^{N\times d}$, and day of week embedding $XW_t \in R^{N\times d}$. The timestamp and road segment embedded methods are similar to the embedded method in the BERT (Kenton & Toutanova, 2019), which is mapped to the dense matrix through one-hot (Zou, Lai, Ma, Tu et al., 2023). $d_X = 1$ denotes the input speed dimension at time step t; to consider computation cost, the dimension of d is set to 64 based on our prior knowledge.

Unlike GMAN (Zheng et al., 2020), the road segment embeddings in this paper trained in the training phase are the same as the BERT (Kenton & Toutanova, 2019). We also fed timestamp embeddings XM + XH + XD + XW and road segment embedding XP to a two-layer fully connected neural network and obtained spatiotemporal embeddings (STE) as additional information input to our proposed model (Zou, Lai, Ma, Tu et al., 2023).

Definition 3 (*Problem Statement*)**.** MT-STGIN aims at predicting the long-term traffic speed on each highway segment. Assume that the input time steps length is P and the prediction time steps length is Q. Given the historical sequence of observations $XS = \left\{XS_{t_1}, \ldots, XS_{t_P}\right\} \in R^{P\times N\times d_X}$ of N nodes in P time steps and spatiotemporal embeddings $STE \in R^{(P+Q)\times N\times d}$ of N nodes in P + Q time steps, we aim to predict the target sequence values of Q time steps for N nodes, expressed as $\hat{Y} = \left\{\hat{Y}_{t_{P+1}}, \ldots, \hat{Y}_{t_{P+Q}}\right\} \in R^{Q\times N\times 3}$.

4. Proposed approach

4.1. Framework overview

The proposed MT-STGIN for long-term highway traffic speed prediction, an encoder–decoder architecture, is illustrated in Fig. 3.

ST-Block, as the core of the encoder–decoder, is first designed that consist of four components, semantic enhancement, spatial correlation extractor, fusion gate mechanism, and temporal correlation extractor: semantic enhancement concentrates on the contextual semantics of each time step with n-gram; spatial correlation extractor is used to extract dynamic spatial correlation and physical relationship; temporal correlation extractor is used to model the dynamic temporal correlation; fusion gate mechanism is used to adaptively incorporate the physical relationship into dynamic spatial correlation and automatically fuse the dynamic spatial and - temporal correlations. In order to avoid error propagation in long-term prediction, a generative inference architecture based on the BridgeTrans is proposed to generate the target hidden outputs in one step. Finally, a multi-task learning method is applied and shares the underlying network parameters for different tasks. Each part of the proposed method is detailed in the sequel.

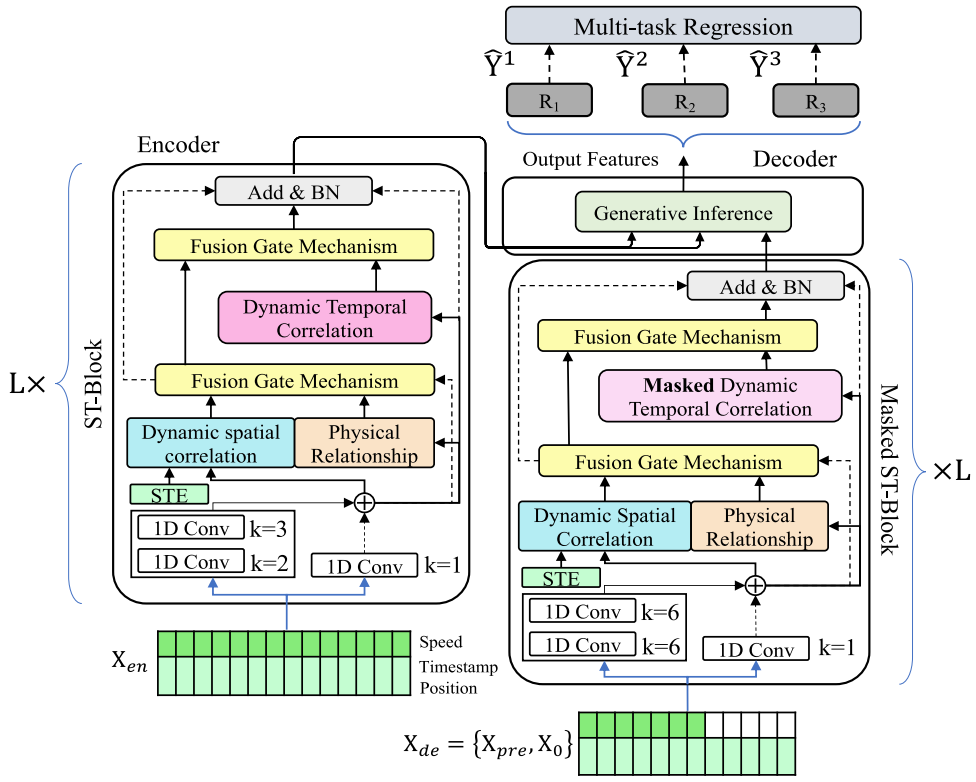


Fig. 3. The framework of the proposed MT-STGIN. L layers of ST-Block can be stacked to extract spatiotemporal dependency, with each layer taking inputs derived from the layer below it.

4.2. ST-block architecture

The working process of the ST-Block is shown in Fig. 3. In the proposed MT-STGIN, like Transformer architecture (Vaswani et al., 2017), we stack several layers ST-Block to extract the dynamic spatiotemporal correlation of the highway network. Assume that the initial input of ST-Block is STE $\in R^{U\times N\times d}$ and X $\in R^{U\times N\times d}$, and the output is HST $\in R^{U\times N\times d}$, where U denotes the input series length of ST-Block.

4.2.1. Semantic enhancement

Neural language processing (NLP) always employs contextual semantics to characterize the logical context relationship within a sentence (Liu et al., 2023). As in the sentence 'I like playing basketball', the context conforms to the grammatical constraint (I-Subject, like-Verb, and playing basketball-Object). The relationship between words in the sentence from left to right and vice versa is substantial. For the traffic speeds, as with a sentence, the observed values are continuous in both directions, also referred to as contextual semantics in this paper.

The traffic speed changes of the highway network are complex. For example, during the morning rush hours, the traffic speed fluctuations sharply lead to observed values not continuing, and this property can be defined as semantic information sparsity. The sparsity problem reflects the traffic speed at a specific time step that lacks perception for neighbors (lefts and rights), especially for irregular speed changes, which increases the prediction difficulty. Fortunately, 1-D convolutional neural networks (1-D CNNs) are employed with an NLP-inspired n-gram concept to capture local context semantics (Zhou, Li, Chi, Tang, & Zheng, 2022). Therefore, we use 1-D CNNs with kernel size k to model the contextual semantics of traffic speed at each time step, as shown in Fig. 3. The semantic enhancement component, such as in the encoder, contains two 1-D CNNs with kernel size $k = 2$ and $k = 3$, respectively, and a 1-D CNN with kernel size $k = 1$ used for a residual connection. For example, when kernel size $k = 3$, the contextual semantic information of three neighbor time steps is mapped into one central time step, as shown in Fig. 4.

As a result, semantic enhancement works can be defined as,

$$\begin{cases} H_1^l = f\left(reverse\left(X\right) * \omega_1\right),\ H_2^l = f\left(reverse\left(X\right) * \omega_2\right) \\ H^l = f\left(reverse\left(H_1^l\right) + reverse\left(H_2^l\right) + X * \omega_3\right) \end{cases} \tag{1}$$

where $*$ denotes convolution calculation, ω represents convolution kernel, and f is a nonlinear transformation function as ReLU at high frequencies used in this study. A reverse function transposes raw input semantic order; avoiding zero-speed padding affects semantic enhancement in the decoder phase. The initial input of semantic enhancement component is X = XS $\in R^{U\times N\times d_X}$ when $l = 1$, and the output is $H^l \in R^{U\times N\times d}$.

4.2.2. Dynamic spatial correlation

The traffic speed of the target road segment is affected by global road segments, and the influence weight changes dynamically with time, called dynamic spatial correlation. For example, a congestion road segment is influenced by other road segments in the traffic network during the congestion period, and the influence weight may be weakened as the congestion is relieved. To model the dynamic spatial correlation, we design a spatial attention approach base on multi-head GAT to adaptively model the correlation between different road segments of the highway network, as shown in Fig. 5(a). For node v_i, at time step t_j, the correlation coefficient between nodes v_i and v in the lth spatial attention layer is,

$$\alpha_{v_i,v}^{l,m} = \frac{exp\left(s_{v_i,v}^{l,m}\right)}{\sum_{v_r\in V} exp\left(s_{v_i,v_r}^{l,m}\right)} \tag{2}$$

where $s_{v_i,v}^{l,m}$ denotes the relevance between v_i and v in the lth layer, V represents the input nodes of ST-Block; r represents a subscript, and the subscript range is $1 \le r \le N$.

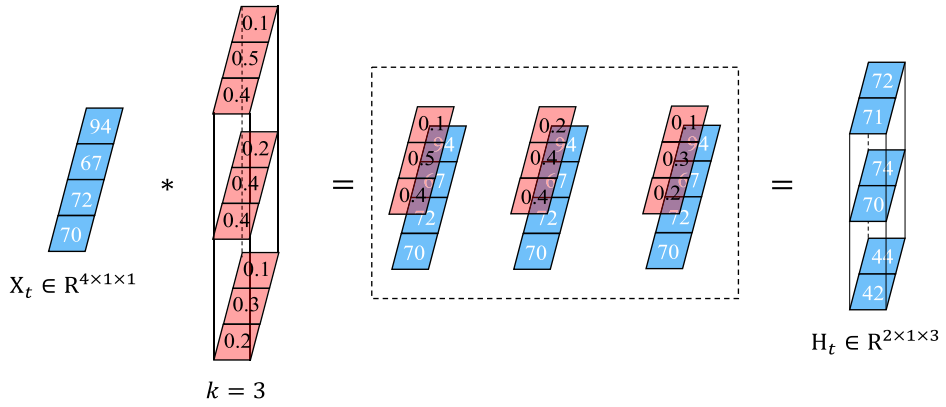
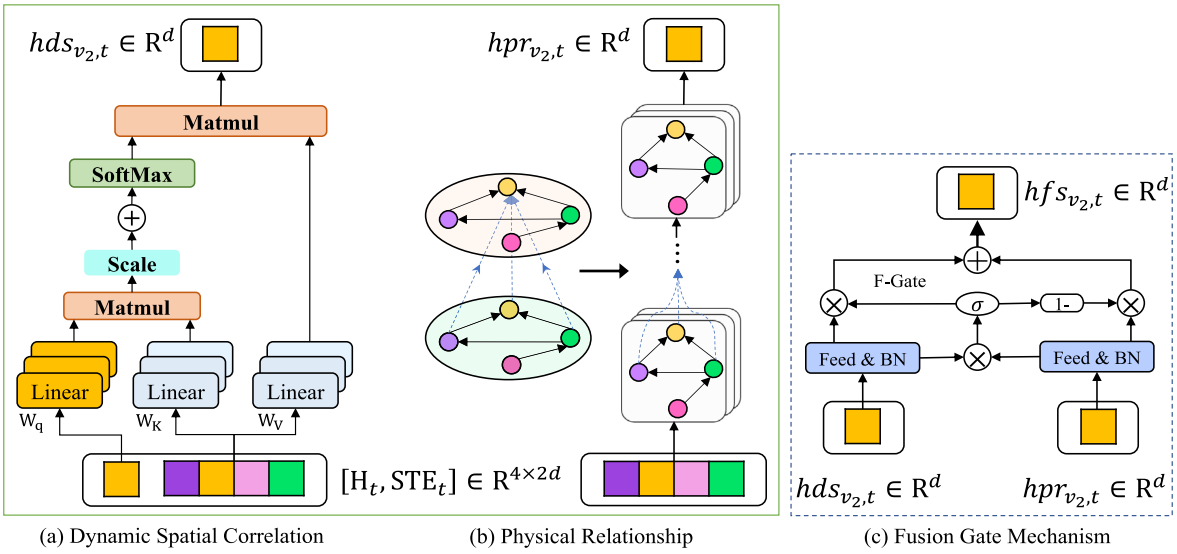


Fig. 4. 1-D CNN.



(a) Dynamic Spatial Correlation (b) Physical Relationship (c) Fusion Gate Mechanism

Fig. 5. Example of spatial correlation processing. (a) dynamic spatial correlation extractor, (b) physical relationship extractor, and (c) fusion gate mechanism.

The relevance $s^{l,m}_{v_i,v}$ can be obtained by the inner product of the query vector of node v_i and the key vector of node v,

$$s^{l,m}_{v_i,v} = \frac{\left\langle f^m_q\left(hds^{l-1}_{v_i,t_j}\right), f^m_k\left(hds^{l-1}_{v,t_j}\right)\right\rangle}{\sqrt{d}} \tag{3}$$

where f^m_q and f^m_k are respectively the nonlinear transformation functions in the mth head attention of the query vector and the key vector, and $\langle *, * \rangle$ represents the inner product operator.

After obtaining the correlation coefficient $\alpha^{l,m}_{v_i,v}$ between nodes v_i and v in the mth head attention, the lth layer dynamic spatial correlation $hds^l_{v_i,t_j}$ of node v_i at time step t_j can be formulated as,

$$hds^{l,m}_{v_i,t_j} = \sum_{v_r \in V} \alpha^{l,m}_{v_i,v_r} f^m_v\left(hds^{l-1}_{v_r,t_j}\right) \tag{4}$$

$$hds^l_{v_i,t_j} = \mathrm{BN}\left(\|^{\mathrm{M}}_{m=1} hds^{l,m}_{v_i,t_j} \mathrm{W}_{ds}\right) + hds^{l-1}_{v_i,t_j} \tag{5}$$

where f^m_v is the nonlinear transformation function in the mth head attention of the value vector, and $\|$ represents the concatenation operation. The final dynamic spatial correlation $hds^l_{v_i,t_j} \in \mathrm{R}^d$ of node v_i can be calculated using Eqs. (2)–(5) at time step t_j. The initial input of the lth spatial attention layer is $\left[\mathrm{H}^l, \mathrm{STE}\right] \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times2d}$, and the output is $\mathrm{HDS}^l \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times d}$.

4.2.3. Physical relationship

In the dynamic spatial correlation extraction processing, the traffic road network's physical property is easily neglected by us, defined as a physical relationship in this paper. The traffic network is a directed graph, so the traffic speed study should consider the highway network architecture, that is, the upstream and downstream traffic speed directly related to the target road segment, which is caused by traffic diffusion. For example, as shown in Fig. 5(b), at a yellow target road segment in the highway network, its traffic speed is directly affected by two upstream road segments, purple and green, and may be affected by the pink in the following time steps. The spatial attention measures the influences of global road segments on the target, concentrating on road sections with huge weights. However, spatial attention could not percept the traffic flow direction, which caused the connectivity and traffic diffusion direction to be neglected. To model these psychical relationship properties, we design a multi-head GCN to focus on the traffic speed from different subspaces of different road segments. The physical relationship aggregation is used as input to a standard nonlinear transformation layer, in order to generate the lth layer embedding of the graph nodes, as shown in Eqs. (6) and (7). For node v_i, at time step t_j, the correlations between nodes v_i and v_i's first-order neighbors V_{v_i} in the mth head GCN is,

$$hpr^{l,m}_{v_i,t_j} = f\left(\widetilde{\mathrm{D}}^{-0.5}\widetilde{\mathrm{A}}_{v_i}\widetilde{\mathrm{D}}^{-0.5}\mathrm{HPR}^{l-1,m}_{t_j}\mathrm{W}^{l-1,m}\right) \tag{6}$$

$$hpr^l_{v_i,t_j} = \mathrm{BN}\left(\|^{\mathrm{M}}_{m=1} hps^{l,m}_{v_i,t_j} \mathrm{W}_{pr}\right) + hpr^{l-1}_{v_i,t_j} \tag{7}$$

where $\widetilde{\mathrm{D}}^{-0.5}\widetilde{\mathrm{A}}_{v_i}\widetilde{\mathrm{D}}^{-0.5}$ denotes the normalized adjacency matrix with added self-connections, $\widetilde{\mathrm{A}}_{v_i} = \mathrm{A}_{v_i} + \mathrm{I}_{v_i}$ is the adjacency matrix of the

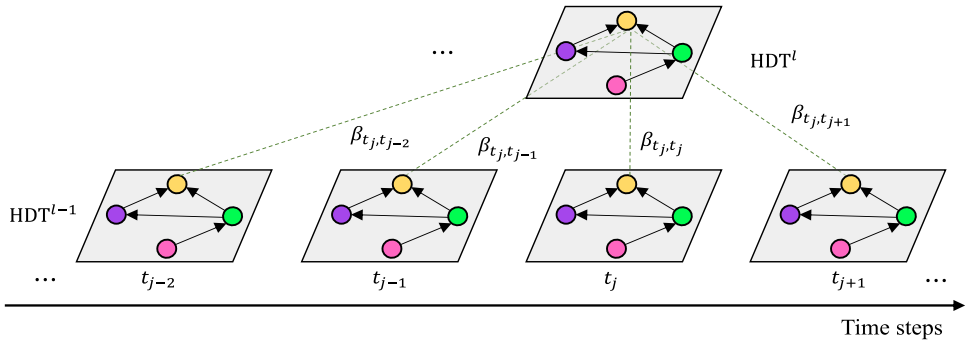


Fig. 6. Temporal attention model of the temporal correlation between different time steps.

graph with added self-connections, $\mathrm{I} \in \mathrm{R}^{\mathrm{N}\times\mathrm{N}}$ represents the identity matrix, and $\widetilde{\mathrm{D}} \in \mathrm{R}^{\mathrm{N}\times\mathrm{N}}$ is the degree matrix of $\widetilde{\mathrm{A}} \in \mathrm{R}^{\mathrm{N}\times\mathrm{N}}$; BN represents the batch normalization. Multi-head GCN can be built by stacking multi-convolutional layers in parallel; as Eqs. (6) and (7), the topological relationship between nodes v_i and v_i's first-order neighbors V_{v_i} can be obtained, and the topological architecture of the highway network and the attributes of the road segment are encoded in order to obtain the physical relationship $hpr^{l}_{v_i,t_j} \in \mathrm{R}^{d}$ of node v_i at time step t_j. The initial input of the lth multi-head GCN layer is $\mathrm{H}^{l} \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times d}$, and the output is $\mathrm{HPR}^{l} \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times d}$.

4.2.4. Fusion gate mechanism (F-Gate)

The traffic speed of a road segment at a certain time step is correlated with directed neighbors and global traffic conditions. As shown in Fig. 5(c), we design a new simple network to incorporate the physical relationship into dynamic spatial correlation adaptively, do not add additional parameters, and the fused spatial dependency $\mathrm{HFS}^{l} \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times d}$ is formed via F-Gate. The working process of F-Gate is,

$$\mathrm{HFS}^{l} = \mathcal{Z} \odot \mathrm{HDS}^{l} + (1-\mathcal{Z}) \odot \mathrm{HPR}^{l} \tag{8}$$

with

$$\mathcal{Z} = \sigma\left(\mathrm{HDS}^{l} \odot \mathrm{HPR}^{l}\right) \tag{9}$$

where σ represents *sigmoid* activation, $\odot$ denotes Hadamard product, and $\mathcal{Z} \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times d}$ is weight vector that controls the flow of physical and dynamic spatial representations at each time step.

4.2.5. Dynamic temporal correlation

The traffic speed at each road segment in the highway network is affected by previous traffic speed and impacts on the future, and the influence weight changes dynamically with time, called dynamic temporal correlation. For example, the traffic speed during morning rush hour is negatively affected by the previous traffic speed, and the influence may gradually increase the occurrence of congestion and then gradually release. In this study, we design a temporal attention approach to effectively model the correlations between different time steps, as shown in Fig. 6.

For node v_i, at time step t_j, the correlation coefficient between time steps t_j and t in the lth temporal attention layer is,

$$\beta^{l,m}_{t_j,t} = \frac{exp\left(\mu^{l,m}_{t_j,t}\right)}{\sum_{t_r\in\mathcal{N}_{t_\mathrm{U}}} exp\left(\mu^{l,m}_{t_j,t_r}\right)} \tag{10}$$

where $\mu^{l,m}_{t_j,t}$ denotes the relevance between t_j and t, $\mathcal{N}_{t_\mathrm{U}}$ denotes a set of time steps before t_U, and the range of subscript r is $1 \le r \le \mathrm{U}$.

The relevance $\mu^{l,m}_{t_j,t}$ can be obtained by the inner product of the query vector of node v_i at time step t_j and the key vector of node v_i at time step t,

$$\mu^{l,m}_{t_j,t} = \frac{\left\langle f^{m}_{q}\left(hdt^{l-1}_{v_i,t_j}\right), f^{m}_{k}\left(hdt^{l-1}_{v_i,t}\right)\right\rangle}{\sqrt{d}} \tag{11}$$

Once the correlation coefficient $\beta^{l,m}_{t_j,t}$ in the mth head attention is obtained, the lth layer temporal correlation $hdt^{l}_{v_i,t_j}$ of node v_i at time step t_j can be formulated as,

$$hdt^{l,m}_{v_i,t_j} = \sum_{t_r\in\mathcal{N}_{t_\mathrm{U}}} \beta^{l,m}_{t_j,t_r} f^{m}_{v}\left(hdt^{l-1}_{v_i,t_r}\right) \tag{12}$$

$$hdt^{l}_{v_i,t_j} = \mathrm{BN}\left(\|^{\mathrm{M}}_{m=1}\, hdt^{l,m}_{v_i,t_j}\mathrm{W}_{d_t}\right) + hdt^{l-1}_{v_i,t_j} \tag{13}$$

The final temporal correlation $hdt^{l}_{v_i,t_j} \in \mathrm{R}^{d}$ of node v_i can be calculated using Eqs. (10)–(13) at time step t_j. The initial input of the lth temporal attention layer is $\mathrm{H}^{l} \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times d}$, and the output is $\mathrm{HDT}^{l} \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times d}$.

The fused spatial dependency $\mathrm{HFS}^{l} \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times d}$ and temporal correlation $\mathrm{HDT}^{l} \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times d}$ of highway network can be fused automatically using F-Gate, dynamic spatiotemporal correlation $\mathrm{HST}^{l} \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times d}$ formed. The final dynamic spatiotemporal correlation $\mathrm{HST}^{l} \in \mathrm{R}^{\mathrm{U}\times\mathrm{N}\times d}$ extracted by lth layer ST-Block can be formulated as,

$$\begin{cases} \mathrm{HST}^{l} = \mathrm{BN}\left(\mathcal{Z}' \odot \mathrm{HFS}^{l} + \left(1-\mathcal{Z}'\right) \odot \mathrm{HDT}^{l} + \mathrm{HFS}^{l} + \mathrm{H}^{l}\right) \\ \mathcal{Z}' = \sigma\left(\mathrm{HFS}^{l} \odot \mathrm{HDT}^{l}\right) \end{cases} \tag{14}$$

4.3. Encoder

Given the input $\mathrm{X}_{en} = \left\{\mathrm{X}_{en,t_1}, \ldots, \mathrm{X}_{en,t_\mathrm{P}}\right\} \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d_{model}}$, $\mathrm{X}_{en,t_j} = \left[\mathrm{XS}_{t_j}, \mathrm{STE}_{t_j}\right] \in \mathrm{R}^{\mathrm{N}\times d_{model}}$, an encoder, stacking several layers of ST-Block, is used to convert the raw highway network data into dynamic spatiotemporal representation $\mathrm{HST} = \left\{\mathrm{HST}_{t_1}, \ldots, \mathrm{HST}_{t_\mathrm{P}}\right\}$, $\mathrm{HST} \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$. The latter is then used in the generative inference layer. Note, $d_{model} = d_{\mathrm{X}} + d$.

4.4. Decoder

We designed a special decoder structure in Fig. 3, and it composed two components, including the Masked ST-Block and generative inference.

4.4.1. Masked ST-block

In the case where the target traffic speed is not known, we need to model the correlations of target time steps. Therefore, the ST-Block is applied in the decoder to model spatiotemporal correlation in order to initialize the target time steps X_0. We feed the following embeddings to the decoder,

$$\mathrm{X}_{de} = \mathrm{Concat}\left(\mathrm{X}_{pre}, \mathrm{X}_0\right) \in \mathrm{R}^{(\mathrm{S}+\mathrm{Q})\times\mathrm{N}\times d_{model}} \tag{15}$$

where $\mathrm{X}_{pre} \in \mathrm{R}^{\mathrm{S}\times\mathrm{N}\times d_{model}}$ is the part of the encoder historical input sequence $\mathrm{X}_{en} \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d_{model}}$ from time step $t_{\mathrm{P}-\mathrm{S}}$ to t_P, and S is the length of the last known time steps. As Fig. 3 shows, we take the last 7 known time steps before the target sequence, and feed the decoder

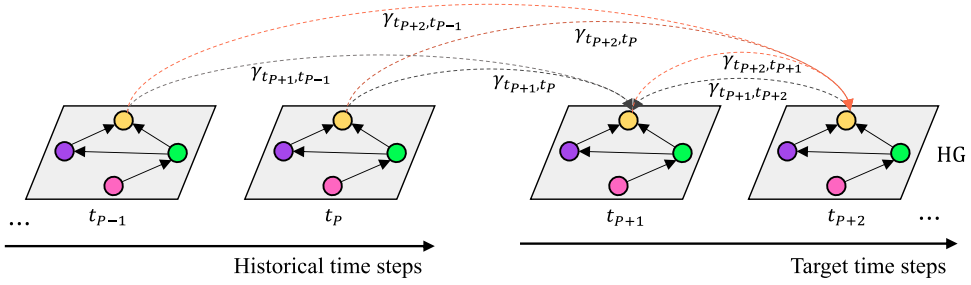


Fig. 7. Bridge transformer models the correlations between historical time steps and target time steps directly.

with $X_{de} = \text{Concat}\left(X_{pre}, X_0\right), X_{pre} \in R^{7\times N\times d_{model}}$; $X_0 \in R^{Q\times N\times d_{model}}$ is the target input sequence (setting the traffic speed to 0), and X_0 also contains the spatiotemporal embeddings $STE \in R^{Q\times N\times d}$.

However, we modify the temporal attention in the ST-Block to prevent time step t_{P+j} from attending to subsequent time steps. Because future speeds are unknown, this masking ensures that the initializations for time step t_{P+j} can depend only on the known historical time steps less than t_P. After the masked ST-Block of the decoder, the initialized $HST' \in R^{Q\times N\times d}$ is obtained from feed input X_{de}.

4.4.2. Generative inference

For the predictive inference process, we need to avoid the propagation of errors and the inference speed. In this paper, we design a particular generative inference architecture that directly combines the correlations between historical time steps and target time steps to generate the final target hidden output $HG \in R^{Q\times N\times d}$. For example, to infer the output $HG_{t_{P+j}} \in R^{N\times d}$ at time step t_{P+j}, we calculate it through the bridge transformer (BridgeTrans) based on the historical sequence $HST \in R^{P\times N\times d}$ and the target initialized sequence $HST' \in R^{Q\times N\times d}$; this does away with the time-consuming "dynamic decoding" operation, as shown in Fig. 7.

For node v_i, the correlation coefficient between the target time step t_{P+j} $(t_{P+j} = t_{P+1}, \ldots, t_{P+Q})$ and the input time step t $(t = t_1, \ldots, t_{P+Q})$ is measured,

$$\gamma^m_{t_{P+j},t} = \frac{exp\left(\lambda^m_{t_{P+j},t}\right)}{\sum_{t_r\in\mathcal{N}_{t_{P+Q}}} exp\left(\lambda^m_{t_{P+j},t_r}\right)} \tag{16}$$

$$\lambda^m_{t_{P+j},t} = \frac{\left\langle f^m_q\left(hst'_{v_i,t_{P+j}}\right), f^m_k\left(hst'_{v_i,t}\ if\ \left(t > t_P\right)\ else\ hst_{v_i,t}\right)\right\rangle}{\sqrt{d}} \tag{17}$$

where $\lambda^m_{t_{P+j},t}$ denotes the relevance between t_{P+j} and t, $\mathcal{N}_{t_{P+Q}}$ represents a set of time steps before t_{P+Q}, and the range of subscript r is $1 \le r \le P + Q$.

Once the correlation coefficient $\gamma^m_{t_{P+j},t}$ in the mth head attention is obtained, the correlation $hg_{v_i,t_{P+j}}$ of node v_i at time step t_{P+j} can be formulated as,

$$hg^m_{v_i,t_{P+j}} = \sum_{t_r\in\mathcal{N}_{t_{P+Q}}} \gamma^m_{t_{P+j},t_r} f^m_v\left(hst'_{v_i,t_r}\ if\ \left(t > t_P\right)\ else\ hst_{v_i,t_r}\right) \tag{18}$$

$$hg_{v_i,t_{P+j}} = \text{BN}\left(\|^M_{m=1}\ hg^m_{v_i,t_{P+j}} W_{hg}\right) + hst'_{v_i,t_{P+j}} \tag{19}$$

The final correlation $hg_{v_i,t_{P+j}}$ of node v_i can be calculated using Eqs. (16)–(19) at time step t_{P+j}. The initial input of BridgeTrans is $HST \in R^{P\times N\times d}$ and $HST' \in R^{Q\times N\times d}$, and the output is $HG \in R^{Q\times N\times d}$.

4.5. Multi-task learning

There are huge traffic condition differences between the ramps and main roads in the real world, and caused this phenomenon mainly by two aspects, (1) vehicle speed limitation and (2) the traffic flow complexity. These properties can be summarized as traffic speed heterogeneity. We divide speed prediction into three categories: entrance toll to the gantry, gantry to gantry, and gantry to exit ramp toll. In addition, there is a high relationship between ramps and main roads, as described in Section 4.2; for example, the vehicle on the in-ramp will directly affect the traffic speed on the main road. Therefore, we design a multi-task learning component to share input features and underlying model parameters, adding three separate feed-forward layers for three different tasks,

$$\hat{Y} = \begin{cases} \hat{Y}^1 = HG \bullet W_1 \\ \hat{Y}^2 = HG \bullet W_2, \hat{Y} \in R^{Q\times N\times 3} \\ \hat{Y}^3 = HG \bullet W_3 \end{cases} \tag{20}$$

where $W_1 \in R^{d\times 1}$, $W_2 \in R^{d\times 1}$, and $W_3 \in R^{d\times 1}$ represent the weight matrices of the three different fully connected layers, and $\bullet$ is a matrix multiplication operation.

The loss function of MT-STGIN corresponding to the multi-task layer is defined as the mean absolute error (MAE) between observed values Y and predicted values $\hat{Y}$,

$$L(\theta) = \frac{1}{Q\times N}\sum_{t=t_{P+1}}^{t_{P+Q}}\sum_{i=1}^{3}\left|Y^i_t - \hat{Y}^i_t\right| + \frac{\lambda}{2}\|\theta\|^2 \tag{21}$$

where λ is the regularization parameter, and θ denotes all the learnable parameters in MT-STGIN.

5. Experiments

5.1. Data description

The traffic speed data used in this study is provided by the ETC intelligent monitoring sensors at the gantries and the toll stations of the highway in Yinchuan City, Ningxia Province, China. The 66 ETC intelligent monitoring sensors record the vehicle driving data in real-time, including 13 highway toll stations (each toll station contains an entrance and exit) and 40 highway gantries. Therefore, these monitoring sensors divide the highway network into 108 road segments, as shown in Fig. 8. The traffic speed of each road segment is measured at a certain frequency, such that one sample is measured every 15 min, and therefore the time series form of the traffic speed is obtained. In addition, the traffic speed data also includes the other two factors, timestamps and road segment index. Because of traffic speed heterogeneity on different types of road segments, the traffic speed is divided into three types, from entrance toll to gantry, called **ETTG**; gantry to gantry, called **GTG**; and gantry to exit toll, called **GTET**. The time span is from June 1, 2021, to August 31, 2021. The road segment index does not change over time, and there are 108 road segments in total, that is, 108 indexes. In the experiment, 70% of the data are used as the training set, 10% of the data are used as the validation set, and the remaining 20% are considered as the test set.[2]

[2] https://github.com/zouguojian/Traffic-speed-prediction/tree/main/MT-STGIN/data.

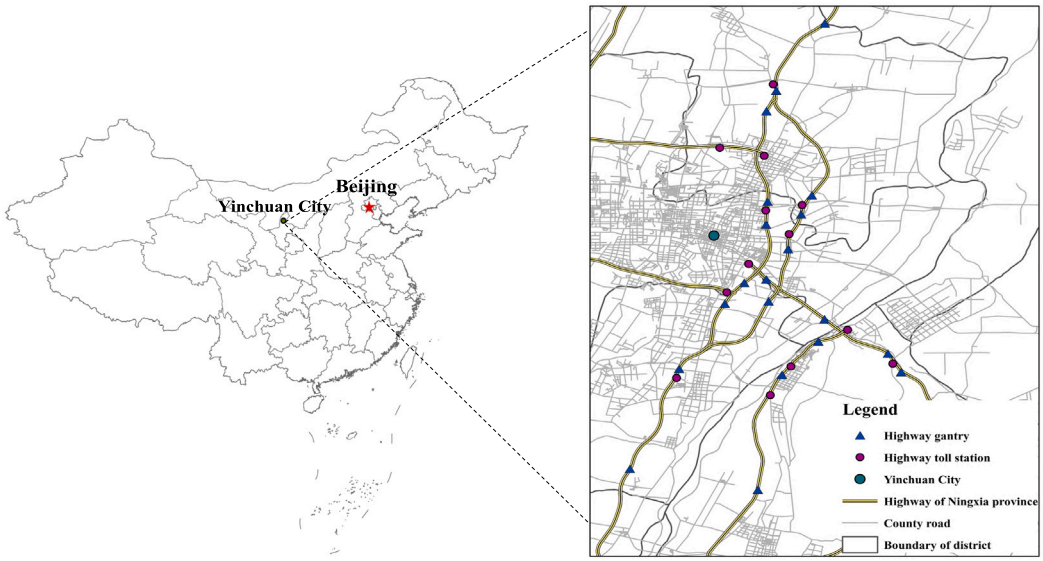


Fig. 8. Study area.

5.2. Baselines and metrics

The proposed model for highway traffic speed prediction is compared with the following prediction methods:

HA, the HA model uses the average value of the historical data, at the same time every day, as the predicted value at the same time in the future prediction task (Liu et al., 2019).

ARIMA, this is a traditional time series prediction method that combines the moving average and autoregressive components in order to model the historical time series data (Duan et al., 2016).

SVM, it refers to support vector machine, is a regression technique for short-term prediction of traffic speed (Vanajakshi & Rilett, 2004).

LSTM NN, it is used to capture the nonlinear traffic dynamic characteristics (Ma et al., 2015).

Bi-LSTM NN, it refers to bidirectional long short-term memory neural network. It models each critical path, and then uses the multiple Bi-LSTM layers stacked together in order to merge the time information (Wang et al., 2019).

FI-RNNs, it refers to features injected recurrent neural networks. It combines the time series data and uses a stacked RNN and encoder, in order to learn the sequential features of the traffic data (Qu et al., 2021).

PSPNN, it refers to path-based speed prediction neural network. It is composed of a CNN and a Bi-LSTM network, that extract the temporal and spatial correlations of the historical data, in order to perform the path-based speed prediction (Yang et al., 2021).

MDL, it refers to novel mixed deep learning. This method is used to predict the lane-level short-term traffic speed. It consists of a Conv-LSTM layer, a convolutional layer and a fully connected layer (Lu et al., 2020).

T-GCN, it combines the GCN and GRU to model the spatiotemporal correlation (Zhao et al., 2019).

DCRNN, it is a diffusion convolutional recurrent neural network, a deep learning framework incorporating spatial and temporal dependency into traffic prediction (Li et al., 2018).

GMAN, it refers to graph multi-attention network. This network is based on spatial and temporal attention. It predicts the traffic speed at different locations on the road network (Zheng et al., 2020).

AST-GAT, it refers to attention-based spatiotemporal graph attention network. It consists of a self-attention-based GAT network and an attention-based LSTM network, for segment-level traffic speed prediction (Li & Lasenby, 2021).

ST-GRAT, it is a novel spatiotemporal graph attention model based on the self-attention mechanism that effectively captures the dynamic spatiotemporal correlation of the road network (Park et al., 2020).

In order to evaluate the prediction performance of the MT-STGIN model, three metrics are used to determine the difference between the observed values Y and the predicted values $\hat{Y}$: the mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). Note that low MAE, RMSE, and MAPE values indicate a more accurate prediction performance. These metrics are presented in Eqs. (22), (23) and (24), respectively.

$$\mathrm{MAE} = \frac{1}{\mathrm{D}\times\mathrm{Q}}\sum_{j=1}^{\mathrm{Q}}\sum_{i=1}^{\mathrm{D}}\left|\mathrm{Y}_{i,j}-\hat{\mathrm{Y}}_{i,j}\right| \tag{22}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{D}\times\mathrm{Q}}\sum_{j=1}^{\mathrm{Q}}\sum_{i=1}^{\mathrm{D}}\left(\mathrm{Y}_{i,j}-\hat{\mathrm{Y}}_{i,j}\right)^2} \tag{23}$$

$$\mathrm{MAPE} = \frac{100\%}{\mathrm{D}\times\mathrm{Q}}\sum_{j=1}^{\mathrm{Q}}\sum_{i=1}^{\mathrm{D}}\frac{\left|\mathrm{Y}_{i,j}-\hat{\mathrm{Y}}_{i,j}\right|}{\mathrm{Y}_{i,j}} \tag{24}$$

where D is the number of samples in test set.

5.3. Experimental settings

We applied the grid search in the proposed MT-STGIN method to find the optimal model on the validation dataset. Especially among all candidate hyperparameter selections, every possibility is tried through loop traversal, and the hyperparameter group with the best performance on the validation dataset is selected as the final result. Note for these continuous hyperparameter values, sample at equal intervals. For each hyperparameter group, the optimal parameters of the proposed MT-STGIN model and baseline techniques are determined during the training process with minimal MAE on the validation set, and specific processing follows,

In the experiment, the maximum number of epochs is 200; and the batch size is 64, which divides the training set into 92 iterations in a single epoch. Updating the model's parameters via backpropagation with a batch of data is called one iteration. Specifically, we evaluate the prediction model on the validation set after one epoch. If the MAE on the validation set is improved, the model parameters are updated and recorded to replace the last one saved. In addition, when the forecasting performance of the prediction model on the validation set is optimal, the training process ends after many parameter adjustments

Table 1
Model hyperparameters.

Module name	Hyperparameter	Values	Input dimension	Output dimension
Semantic enhancement	Filter size	$[2,3,1]_{en}$, $[6,6,1]_{de}$	$[64, 12, 108, 1]_{en,de}$	$[64, 12, 108, 64]_{en,de}$
	Stride	1		
	Is padding	True		
	In channel	1		
	Out channel	64		
Multi-head GCN	Hidden nodes	64	$[64, 12, 108, 64]_{en}$	$[64\ 12, 108, 64]_{en}$
	Number of heads (M)	1	$[64, 6, 108, 64]_{de}$	$[64, 6, 108, 64]_{de}$
Multi-head GAT	Hidden nodes	64	$[64, 12, 108, 64]_{en}$	$[64, 12, 108, 64]_{en}$
	Number of heads (M)	4	$[64, 6, 108, 64]_{de}$	$[64, 6, 108, 64]_{de}$
Temporal attention	Hidden nodes	64	$[64, 12, 108, 64]_{en}$	$[64, 12, 108, 64]_{en}$
	Number of heads (M)	4	$[64, 6, 108, 64]_{de}$	$[64, 6, 108, 64]_{de}$
Generative inference	Hidden nodes	64	$[64, 12, 108, 64]_{de}$	$[64, 6, 108, 64]_{de}$
	Number of heads (M)	4	$[64, 6, 108, 64]_{de}$	
Multi-task layer	Hidden nodes	[64,32,1]	$[64, 6, 108, 64]_{de}$	$[64, 6, 108, 1]_{de}$
		[64,32,1]	$[64, 6, 108, 64]_{de}$	$[64, 6, 108, 1]_{de}$
		[64,32,1]	$[64, 6, 108, 64]_{de}$	$[64, 6, 108, 1]_{de}$
–	Number of blocks (L)	$[1]_{en}$ and $[1]_{de}$	–	–
–	Batch size	64	–	–
–	Dropout	0.3	–	–
–	Decay rate	0.99	–	–
–	Learning rate	0.0005	–	–
–	λ	0.001		
–	Epochs	200	–	–
–	Training method	Adam	–	–

and experiments. We use an early-stop mechanism in all experiments, and the number of early-stop epochs is set to 50, defined as patience. The early-stop mechanism means the training stops early if the MAE on the validation set is not decreased under the patience before the maximum number of epochs. Finally, the prediction result is obtained by iterating all the samples in the test set. To consist with existing studies, we set the target time steps Q and historical time steps P to 6 and 12, respectively, representing the time span is 270 min.

After multiple training steps, the final model framework parameters are determined. Table 1 presents the number of layers, nodes, output size and related hyperparameters of the MT-STGIN model. We implement the MT-STGIN and baselines in TensorFlow and PyTorch. The server's one NVIDIA Tesla V100S-PCIE-32 GB GPUs and 24 CPU cores are used for model training and testing. Note that the **implementation codes** of the proposed MT-STGIN model and baseline models are open source, and are available at the personal **GitHub homepage**.[3]

5.4. Experimental results

5.4.1. Predicting performance comparison

Long-term highway traffic speed prediction is a challenging task, and it is related to the precise control of highway traffic in the future. This experiment uses 12-time steps of historical data to predict the traffic speed in the next six time steps. For example, 7:00–10:00 am is used as the input period, and 10:00–11:30 am is considered as the predicted period. Tables 2, 3, and 4 show the highway traffic speed prediction comparison results on different tasks.

The performance of HA and ARIMA is lower than that of other baseline models, demonstrating the difficulty of long-term highway traffic speed prediction. Temporal models based on RNNs perform better and more steadily than statistical methods because they can extract the complex nonlinear temporal correlation of traffic speed. For example, for third-time step traffic speed forecasting compared with ARIMA, LSTM NN, Bi-LSTM NN, and FI-RNNs reduced by 2.419%, 2.966%, and 2.482% regarding MAE in ETTG; by 4.791%, 6.113%, and 3.211% in GTG; by 0.285%, 1.399%, and 0.611% in GTET. In addition, for the next six time steps, LSTM NN, Bi-LSTM NN, and FI-RNNs improved MAE by 2.185%, 2.970%, and 2.524% in ETTG compared with ARIMA; by 4.391%, 5.831%, and 4.289% in GTG; by 0.378%, 1.378%, and 0.797% in GTET. The experimental results demonstrate that RNN and its variants outperform statistical methods such as ARIMA in temporal correlation extraction and long-term highway traffic speed forecasting. Note that the performance of SVM high than RNNs in GTG because the SVM predicts traffic speed in different time steps separately without error propagation compared with RNNs.

Spatial correlation is another essential factor in traffic speed prediction tasks. Unlike RNNs, most spatiotemporal models, such as PSPNN and MDL, utilize CNNs as spatial feature extractors. The results show that PSPNN and MDL outperform temporal models based on RNNs regarding the three metrics for all tasks, especially in GTG and GTET. Compared with LSTM NN for the next six-time steps prediction, PSPNN and MDL reduced MAE by 5.479% and 7.447% in GTG; reduced by 3.289% and 4.939% in terms of RMSE; reduced by 1.732% and 4.294% regarding MAPE. Moreover, for the next six time steps, PSPNN and MDL lowered MAE by 3.403% and 1.030% in GTET compared with LSTM NN; reduced RMSE by 2.972% and 1.973%; improved MAPE by 2.911% and 4.449%. The above comparisons evaluate that embedding the spatial factor through CNNs into temporal correlation extraction processing help improve the prediction performance extended. However, with the non-Euclidean finding in the speed prediction task, the weakness of CNNs on spatial correlation extraction is exemplified by existing studies.

Graph convolutional networks (GCNs) are first used in non-Euclidean space, such as T-GCN and DCRNN, and showed some effects for speed prediction compared with spatiotemporal methods based on CNNs. For instance, T-GCN and DCRNN outperform MDL in GTG for sixth-time step traffic speed prediction regarding the two metrics, lowering the MAE by 4.905% and 3.490%, reducing the MAPE by 3.986% and 10.182%. However, the prediction performance on the other two tasks is slightly insufficient compared with PSPNN and MDL. These findings demonstrate that GCNs can handle non-Euclidean structure traffic data, but they deem that the spatial dependency from neighbors is equal in default; moreover, such an example of T-GCN is challenging to converge and needs training of more than 5000 epochs due to its combination limitation. In contrast, recently, some existing studies thought spatial correlation is dynamic and not immutable.

[3] https://github.com/zouguojian/Traffic-speed-prediction/tree/main/MT-STGIN.

Table 2
Highway traffic speed prediction results of different methods in different target time steps for ETTG.

Model	15 min (first-time step)			45 min (third-time step)			90 min (sixth-time step)			Six-time steps (avg)		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
HA	6.598	11.027	15.065%	6.598	11.027	15.065%	6.598	11.027	15.065%	6.598	11.027	15.065%
ARIMA	4.627	7.633	17.171%	4.754	7.789	13.484%	4.727	7.512	10.519%	4.714	7.701	12.997%
SVM	4.816	8.735	11.859%	4.883	8.672	**10.858%**	5.278	9.244	14.478%	5.059	9.026	13.703%
LSTM NN	4.404	7.405	17.158%	4.639	7.662	13.287%	4.690	7.488	10.770%	4.611	7.600	13.060%
Bi-LSTM NN	4.405	7.404	16.811%	4.613	7.682	13.114%	4.629	7.442	10.584%	4.574	7.588	12.809%
FI-RNNs	4.398	7.403	16.993%	4.636	7.659	13.289%	4.657	7.460	10.637%	4.595	7.592	12.944%
PSPNN	4.354	7.325	17.071%	4.534	7.555	13.078%	4.571	7.302	10.355%	4.516	7.482	12.841%
MDL	4.568	7.633	17.906%	4.803	8.050	14.583%	4.777	7.767	10.751%	4.771	7.973	13.910%
T-GCN	4.868	8.008	18.330%	4.992	8.126	13.873%	4.880	7.867	10.853%	4.938	8.082	13.712%
DCRNN	4.730	7.732	13.142%	4.700	7.725	13.387%	4.794	7.859	13.630%	4.737	7.765	13.369%
GMAN	4.335	7.332	17.288%	4.360	**7.374**	12.940%	4.214	7.016	9.592%	4.314	7.300	12.571%
AST-GAT	4.497	7.535	16.840%	4.650	7.653	12.966%	4.548	7.460	10.030%	4.615	7.678	12.636%
ST-GRAT	4.234	7.280	**12.462%**	4.394	7.503	13.119%	4.676	7.914	13.520%	4.449	7.592	13.115%
MT-STGIN	**4.178**	**7.233**	16.141%	**4.225**	7.390	11.873%	**4.132**	**6.995**	**9.329%**	**4.192**	**7.277**	**11.938%**

Table 3
Highway traffic speed prediction results of different methods in different target time steps for GTG.

Model	15 min (first-time step)			45 min (third-time step)			90 min (sixth-time step)			Six-time steps (avg)		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
HA	5.030	8.456	7.520%	5.030	8.456	7.520%	5.030	8.456	7.520%	5.030	8.456	7.520%
ARIMA	5.604	8.915	8.244%	5.824	9.261	7.517%	6.106	9.290	9.817%	5.899	9.218	8.724%
SVM	5.176	8.955	**7.203%**	5.125	8.519	9.061%	5.442	9.063	8.804%	5.301	8.804	8.321%
LSTM NN	5.274	8.497	7.814%	5.545	8.971	7.152%	5.903	9.068	9.425%	5.640	8.969	8.430%
Bi-LSTM NN	5.182	8.403	7.800%	5.468	8.922	7.152%	5.800	8.936	9.406%	5.555	8.879	8.414%
FI-RNNs	5.267	8.484	7.854%	5.637	9.018	7.233%	5.857	9.019	9.370%	5.646	8.955	8.450%
PSPNN	5.018	8.252	7.683%	5.258	8.760	7.002%	5.481	8.648	9.315%	5.331	8.674	8.284%
MDL	4.922	8.099	7.520%	5.200	8.641	6.929%	5.301	8.452	8.957%	5.220	8.526	8.068%
T-GCN	4.881	8.011	7.409%	4.991	8.431	6.689%	5.041	8.145	8.600%	5.023	8.318	7.819%
DCRNN	5.219	8.668	8.279%	5.105	8.559	8.156%	5.116	8.547	**8.045%**	5.130	8.573	8.134%
GMAN	4.740	7.907	7.286%	4.747	8.241	6.375%	4.728	7.890	8.102%	4.780	8.134	7.527%
AST-GAT	**4.680**	**7.823**	7.096%	4.718	**8.138**	**6.306%**	4.673	**7.833**	8.280%	4.754	**8.063**	**7.495%**
ST-GRAT	4.772	8.264	7.682%	4.995	8.491	8.027%	5.485	9.039	8.679%	5.102	8.620	8.164%
MT-STGIN	4.683	7.956	7.275%	**4.672**	8.377	6.409%	**4.614**	7.933	8.268%	**4.708**	8.217	7.592%

Table 4
Highway traffic speed prediction results of different methods in different target time steps for GTET.

Model	15 min (first-time step)			45 min (third-time step)			90 min (sixth-time step)			Six-time steps (avg)		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
HA	7.918	11.924	19.643%	7.918	11.924	19.643%	7.918	11.9245	**19.643%**	7.918	11.924	19.643%
ARIMA	7.339	11.569	20.076%	7.364	11.220	21.057%	7.584	11.517	22.359%	7.404	11.379	20.309%
SVM	7.257	11.221	**17.906%**	7.340	11.333	21.218%	7.485	11.383	20.962%	7.393	11.402	19.813%
LSTM NN	7.120	11.266	20.000%	7.343	11.282	21.355%	7.637	11.683	23.283%	7.376	11.405	20.747%
Bi-LSTM NN	7.105	11.266	21.998%	7.261	11.220	20.989%	7.538	11.616	23.969%	7.302	11.342	21.145%
FI-RNNs	7.122	11.288	19.045%	7.319	11.262	21.030%	7.568	11.639	22.704%	7.345	11.382	20.123%
PSPNN	7.053	11.129	19.473%	7.083	10.922	20.651%	7.324	11.290	22.550%	7.125	11.066	20.143%
MDL	7.116	11.169	19.181%	7.222	11.012	20.194%	7.557	11.413	21.811%	7.300	11.180	19.824%
T-GCN	7.293	11.340	19.257%	7.260	11.030	20.114%	7.413	11.251	21.609%	7.317	11.195	19.634%
DCRNN	7.254	11.237	22.607%	7.390	11.361	21.281%	7.468	11.478	21.599%	7.381	11.374	21.614%
GMAN	7.000	11.003	20.404%	6.978	10.726	20.621%	7.030	10.907	21.853%	6.967	10.830	19.950%
AST-GAT	6.892	10.923	20.166%	6.962	10.729	21.109%	7.007	10.842	22.129%	6.913	**10.746**	20.226%
ST-GRAT	**6.771**	**10.777**	18.527%	6.995	11.055	19.349%	7.274	11.466	20.177%	7.039	11.134	19.463%
MT-STGIN	6.884	10.983	18.313%	**6.750**	**10.592**	**18.853%**	**6.883**	**10.829**	20.219%	**6.830**	10.771	**18.302%**

Graph attention networks (GATs) as a critical tool instead of GCNs to model the dynamic spatial correlation of traffic speed in non-Euclidean space; in addition, the temporal attention methods based on Transformer are incorporated into spatiotemporal models to extract dynamic temporal correlation, such as GMAN, AST-GAT, and ST-GRAT. Tables 2, 3, and 4 show that these three models outperform other baseline models regarding the three metrics. For example, for sixth-time step traffic speed forecasting compared with spatiotemporal methods PSPNN, MDL, T-GCN, and DCRNN, GMAN respectively improved MAE by 7.810%,11.786%, 13.648%, and 12.098% in ETTG; by 13.738%, 10.809%, 6.209%, and 7.584% in GTG; by 4.014%, 6.974%, 5.167%, and 6.230% in GTET. These results demonstrate that spatiotemporal methods based on GATs and temporal attention networks are more adaptive for long-term highway traffic speed prediction in non-Euclidean space. However, for ST-GRAT in Fig. 9, the speed prediction performance decreases quickly with time steps because dynamic decoding leads to prediction error propagation in the spatial and temporal dimensions.

In this paper, we absorb the advantages of graph attention networks in modeling dynamic spatial correlation and temporal attention in extracting time series dependency. The proposed model first designed the ST-Block to extract the dynamic spatiotemporal correlation of the highway network, a generative inference architecture is then suggested to avoid error propagation in spatial and time dimensions, a multi-task learning method is finally used to predict traffic speed on different types of roads because of heterogeneity and shares the underlying network parameters. According to Fig. 9 and Tables 2, 3, and 4, the

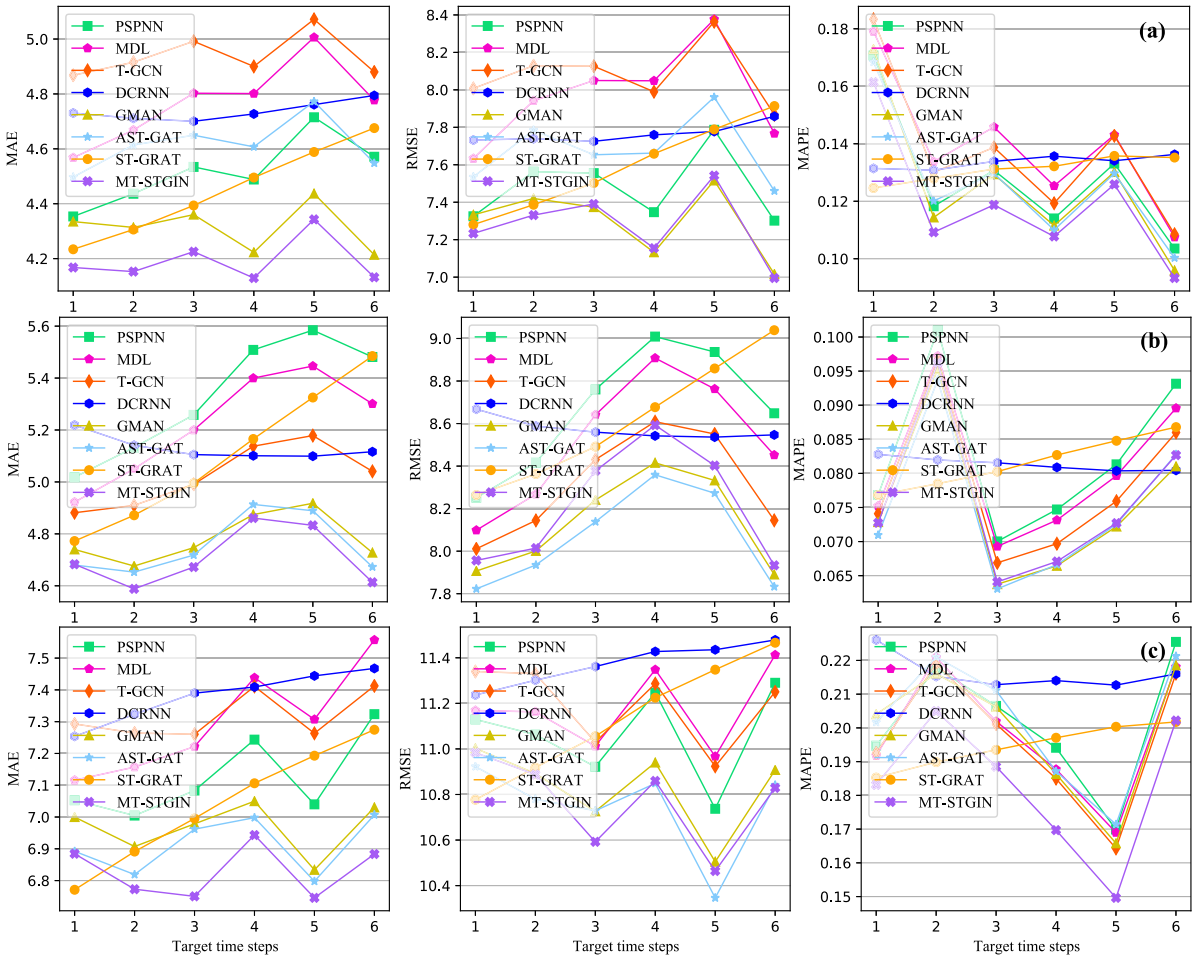


Fig. 9. Long-term highway traffic speed prediction ability: row **(a)** presents the prediction error of each step in task ETTG; row **(b)** indicates the performance of each step in task GTG; row **(c)** reflects the prediction accuracy of each step in task GTET.

proposed method outperforms all baselines on three tasks regarding the three metrics. Compared with spatiotemporal methods GMAN, AST-GAT, and ST-GRAT for sixth-time step traffic speed forecasting, MT-STGIN improved MAE by 1.946%, 9.147%, and 11.634% in ETTG; by 2.411%, 1.263%, and 15.880% in GTG; by 2.091%, 1.770%, and 5.375% in GTET. In addition, for the next six-time steps prediction, MT-STGIN outperforms absolute advantages in different tasks. For instance, MT-STGIN lowered MAE by 2.828%, 9.166%, and 5.777% in ETTG compared with GMAN, AST-GAT, and ST-GRAT; by 1.506%, 0.968%, and 7.722% in GTG; by 1.966%, 1.201%, and 2.969% in GTET. The comparison results validate that MT-STGIN is more adaptable than baselines for long-term highway traffic speed prediction, and the accuracy is not affected by the length of predicting time steps in different tasks. We argue that long-term traffic speed forecasting is more beneficial to practical applications, e.g., it allows transportation agencies to have more time to take actions to optimize the traffic according to the prediction. Therefore, when MT-STGIN obtains a high level of prediction performance, the benefits become more apparent in long-term highway traffic speed prediction.

Note, for GTG, the performance of MT-STGIN low than GMAN and AST-GAT in terms of RMSE and MAPE; inferior to GMAN regarding MAPE. These experimental results indicate that: (1) first, MT-STGIN balances the prediction accuracy on each sub-task, leading to performance lower than GMAN in terms of MAPE but superior to other baselines; and (2) second, MT-STGIN needs to fit speed fluctuation, which leads to prediction shift in rare time steps, causing the performance to be inferior to GMAN and AST-GAT in terms of RMSE and MAPE. Fortunately, multi-task learning ensures the advantage of sub-tasks, especially regarding MAE, as shown in 2, 3, and 4. Additionally, the above two aspects will be further demonstrated in Sections 5.4.2, 5.4.4, and 5.4.5.

5.4.2. *Influence of each essential component*

To verify the effectiveness of each essential component of the proposed MT-STGIN model, five variants are compared in this part, including MT-STGIN-1, −2, −3, −4, and −5 models. MT-STGIN-1 does not consider semantic enhancement and uses a fully connected layer instead. MT-STGIN-2 does not consider the physical relationship in the highway network. MT-STGIN-3 does not consider the fusion gate mechanism on MT-STGIN and uses addition operation instead. MT-STGIN-4 does not consider error propagation in the long-term prediction stage and uses dynamic decoding. MT-STGIN-5 removes the multi-task learning component and uses two fully connected layers instead. As Table 5 shows, each component's contribution to MT-STGIN is heightened in the following comparisons,

MT-STGIN-1 Compared with MT-STGIN, the performance is decreased, and MT-STGIN-1 increased MAE and RMSE by 0.541% and 0.124% in ETTG; 0.170% and −0.085% in GTG; 0.029% and 0.028% in GTET. The experimental results prove that the semantic enhancement module is favorable for long-term traffic speed prediction even faces traffic fluctuation. In addition, MT-STGIN-1 lowered the MAPE compared to MT-STGIN in three tasks. This phenomenon is from the MT-STGIN to overcome semantic information sparsity via 1-D CNNs with n-gram, which helps the model to percept the local traffic conditions in time dimension for contextual semantics reinforcement; however, the

Table 5
Average performance of the next six-time steps prediction for different variants.

Task	Metric	MT-STGIN-1	MT-STGIN-2	MT-STGIN-3	MT-STGIN-4	MT-STGIN-5	MT-STGIN
ETTG	MAE	4.215	4.231	4.207	4.296	4.200	**4.192**
	RMSE	7.286	7.302	7.270	7.349	**7.262**	7.277
	MAPE	**11.884%**	11.992%	12.079%	12.145%	12.059%	11.938%
GTG	MAE	4.716	4.734	4.728	4.735	4.715	**4.708**
	RMSE	8.210	8.252	8.206	8.244	**8.209**	8.217
	MAPE	**7.592%**	7.645%	7.612%	7.644%	7.608%	**7.592%**
GTET	MAE	6.832	6.839	6.840	6.977	6.855	**6.830**
	RMSE	10.774	10.789	10.778	10.929	10.798	**10.771**
	MAPE	**18.230%**	18.322%	18.262%	19.004%	18.658%	18.302%

Table 6
Computation cost during the training and inference phases, an example of GTG.

Model	Parameters	Training/(100 iterations) (batch size = 64)		Inference (batch size = 1)	
		Time cost	GPU memory usage	Time cost	GPU memory usage
HA[a]	–	0.001 (min)	–	0.001 (min)	–
ARIMA[a]	–	56.482 (min)	–	2.036 (min)	–
SVM[a]	–	17.186 (min)	–	5.902 (min)	–
LSTM NN	61,249	0.249 (min)	1011 MiB	0.238 (min)	503 MiB
Bi-LSTM NN	284,033	0.414 (min)	2547 MiB	0.720 (min)	507 MiB
FI-RNNs	53,575	0.278 (min)	3439 MiB	0.215 (min)	1963 MiB
PSPNN	141,542	0.276 (min)	4543 MiB	0.280 (min)	2473 MiB
MDL	279,553	0.294 (min)	5095 MiB	0.307 (min)	5035 MiB
T-GCN	17,222	0.086 (min)	755 MiB	0.108 (min)	501 MiB
DCRNN	372,352	0.401 (min)	2547 MiB	0.415 (min)	507 MiB
GMAN	222,721	0.365 (min)	9583 MiB	0.199 (min)	4011 MiB
AST-GAT	219,910	0.803 (min)	9071 MiB	0.591 (min)	9319 MiB
ST-GRAT	293,697	0.203 (min)	5425 MiB	1.171 (min)	1643 MiB
MT-STGIN	344,387	0.400 (min)	12 415 MiB	0.237 (min)	6059 MiB

[a] Means the model train one time on the whole training set.

MT-STGIN-1 lacks consideration of traffic speed fluctuation, making it challenging to adapt to the complex real-world traffic environment.

MT-STGIN-2 The performance of MT-STGIN-2 is low than MT-STGIN regarding the three metrics on all tasks. For example, compared with MT-STGIN for the next six-time steps forecasting, MT-STGIN-2 increased MAE, RMSE, and MAPE by 0.922%, 1.027%, and 0.453% in ETTG; by 0.549%, 0.424%, and 0.693% in GTG; by 0.132%, 0.167%, and 0.109% in GTET. The experimental results prove that the physical relationship has a non-negligible influence on long-term speed prediction. Moreover, MT-STGIN-2 is also lower than MT-STGIN-1, reflecting the importance of physical relationship priors to semantic enhancement.

MT-STGIN-3 The results of MT-STGIN-3 are inferior to MT-STGIN regarding the three metrics on all tasks. For instance, MT-STGIN-3 increased MAE and MAPE by 0.357% and 1.167% in ETTG, respectively, compared with MT-STGIN; by 0.423% and 0.263% in GTG; increased MAE and RMSE by 0.146% and 0.065% in GTET. The experiments reflect that studying how to embed the physical relationship into dynamic spatial correlation and how to combine spatial and temporal correlations adaptively is meaningful. There is no doubt that the fusion gate network is an effective way to solve the feature fusion problem we encountered.

MT-STGIN-4 MT-STGIN-4 is a typical dynamic decoding case. Compared with MT-STGIN, MT-STGIN-4 is inferior regarding the three metrics on all tasks. For example, MT-STGIN-4 increased MAE, RMSE, and MAPE by 2.421%, 0.980%, and 1.704% in ETTG; by 0.570%, 0.328%, and 0.680% in GTG; by 2.107%, 1.446%, and 3.694% in GTET. In addition, the results also present that MT-STGIN-4 is the worst in the variants. These comparisons verify that dynamic decoding is a critical flaw in prediction processing, and generative inference as an alternative technique can avoid prediction error propagation in spatial and temporal dimensions.

MT-STGIN-5 Traffic speed varies across types of road segments making speed forecasts deviate significantly from the observed value. Multi-task learning separates speed prediction on the whole network into three relative tasks, and this inspiration is proven in Table 5. Compared MT-STGIN-5 with MT-STGIN, the advantage of multi-task learning in highway traffic speed perdition becomes obvious. For the next six-time steps prediction, MT-STGIN-5 increased MAE and MAPE by 0.190% and 1.003% in ETTG; by 0.148% and 0.210% in GTG; increased MAE, RMSE, and MAPE by 0.365%, 0.250%, and 1.908% in GTET. The compassion results verify the necessity of multi-task learning on long-term highway traffic speed prediction.

Through comparisons, we have combined the advantages of all the variants and highlighted the contributions of each essential component of the proposed MT-STGIN.

5.4.3. Computation cost

Tables 2, 3, and 4 present the prediction performance comparisons, and the computation costs of the baselines and the proposed model on six target time steps are shown in Table 6,

Table 6 presents the computation cost of MT-STGIN and baseline methods for predicting highway traffic speed, including total parameters, time cost, and GPU memory usage. According to Tables 2, 3, and 4, the ARIMA and SVR possess high time costs and HA vice versa, but weak prediction performance in both the training and inference stages. T-GCN consumes less time and GPU memory than other deep learning baselines during the training and inference phases, but its performance is inferior. For the two optimal baselines shown in Tables 2, 3, and 4, GMAN and AST-GAT, achieving high prediction performance require more time cost and GPU memory in both training and stages, especially AST-GAT, as shown in Table 6. Additionally, due to the difference in data loading, such as the proposed method uploading the whole training set to GPU, the memory consumption of GMAN and AST-GAT in the training and inference phases is more than that of other baselines, but the forecasting accuracy is high. This paper seeks to devise a faster, more efficient, and low-complexity model to achieve high prediction performance. Therefore, MT-STGIN is proposed to have superior performance, even though its model complexity is higher than that of GMAN and AST-GAT. In the training phase, the time cost is

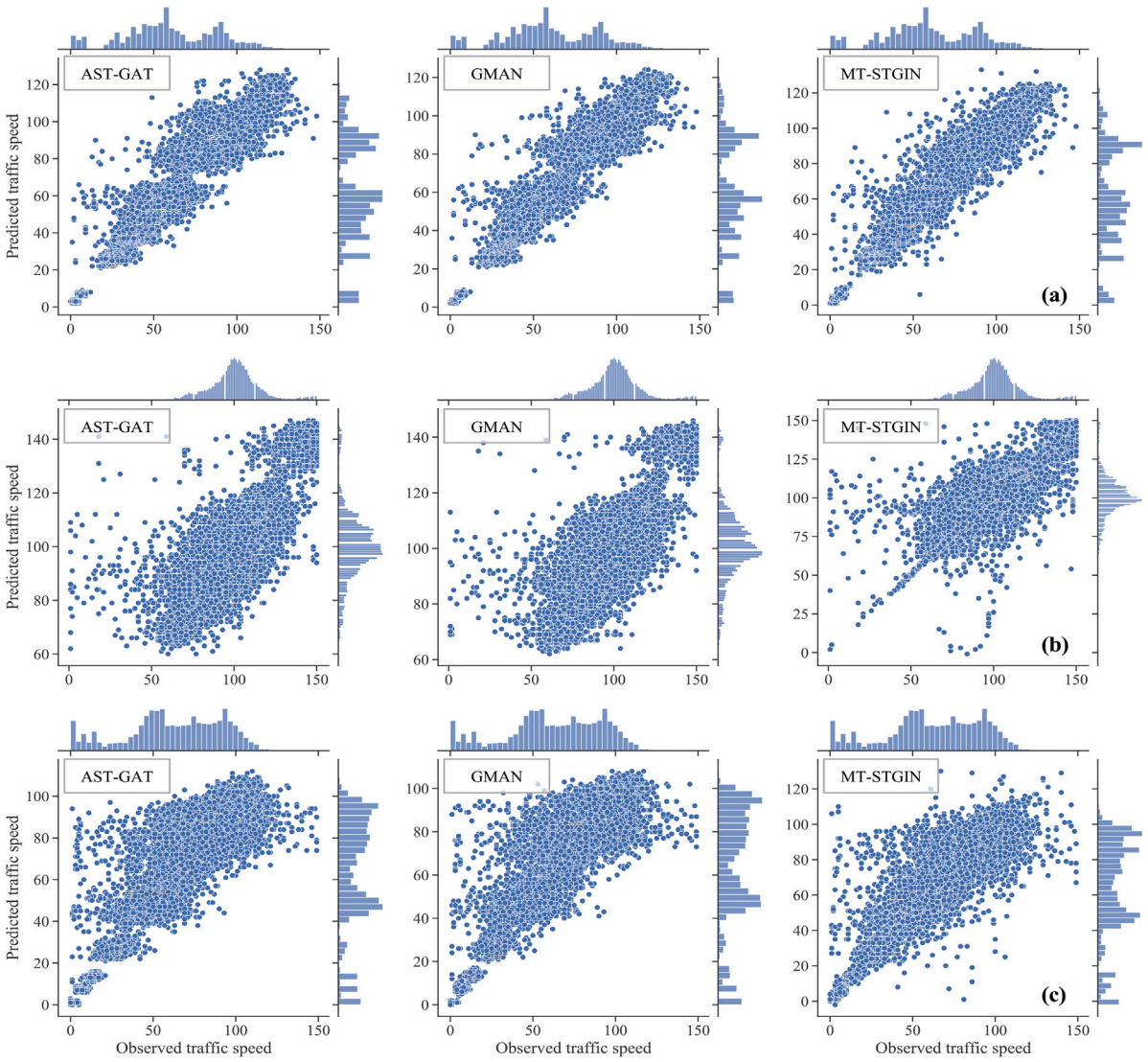


Fig. 10. Degree of fit between the observed and predicted traffic speed values. The blue dots indicate the degree of deviation between the observed and predicted values, and the blue histograms represent the distribution of observed and predicted. **(a)** relevant results in task ETTG; **(b)** in task GTG; **(c)** in task GTET.

approximately equal to GMAN and less than AST-GAT, and the GPU memory usage of MT-STGIN is higher than those of GMAN and AST-GAT. In contrast, in the inference phase, the time cost and GPU memory usage are minimal compared with AST-GAT.

We prefer a faster, more efficient, low-complexity model that even if using more GPU memory while maintaining high prediction precision. MT-STGIN provides long-term forecasts in a single pass, reducing the time required for inference compared to baselines such as Bi-LSTM NN, DCRNN, and AST-GAT. The computation cost further validates the superiority of MT-STGIN in long-term highway traffic speed prediction.

5.4.4. Fitting performance

To better demonstrate the performance of MT-STGIN, we compare it with the other two optimal spatiotemporal baseline models and visualize the fitting results. Fig. 10 shows the visualization results of the predictive fit ability over six target time steps, and we note the following three findings:

1. Traffic condition is easily affected by traffic incidents, such as agglomerate fog, which may cause speed to fall low value, called speed fluctuation. Especially the proposed method has an absolute advantage when traffic speed is below 30 km/h compared with GMAN and AST-GAT. For instance, GMAN and AST-GAT failed to predict the traffic velocity under 60 km/h in task GTG entirely. The performance of MT-STGIN on low traffic velocity may benefit from traffic pattern (e.g., the periodic pattern of traffic speeds shown in Fig. 11) learning and semantic enhancement.
2. MT-STGIN presents a significant fitting performance on ETTG, GTG, and GTET when the speed exceeds 30 km/h; the discrete degree of blue dots contained in MT-STGIN is lower than that in the other two models, except for rares. In addition, when the traffic speeds are between 50 km/h and 120 km/h, the distributions of observed and predicted keep consistent, as indicated by the blue histograms. The comparison results demonstrate that MT-STGIN has excellent fitting performance at various scope traffic speeds and may express good application prospects.
3. However, a few blue dots deviate from the diagonal for the proposed model. For example, in the tasks GTG and GTET, when the traffic speeds exceed 50 km/h, the discrete degree of a few blue dots contained in MT-STGIN is higher than in the other two models. This phenomenon is from the MT-STGIN fit the

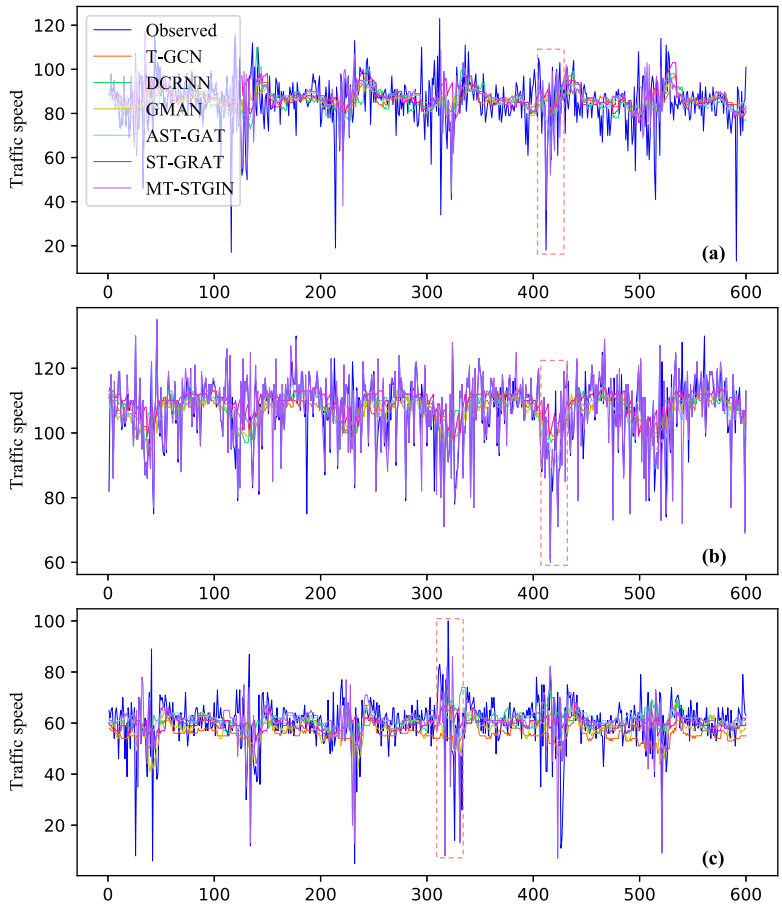


Fig. 11. The visualization results for prediction over six-time steps, and the example length of the test set is six hundred time steps. **(a)** sample from task ETTG, **(b)** from task GTG, and **(c)** from task GTET.

dynamic traffic fluctuation, which causes model prediction shifts in rare cases; however, baselines keep prediction in a range with traffic speeds (proved in the following Fig. 11), making it challenging to adapt to the complex real-world traffic condition. The practical applicability value of MT-STGIN is confirmed once more.

5.4.5. Case study

Three road segments are exampled from these three tasks, respectively, and visualized the prediction results for the six-time steps horizon. In the experiment, one hundred continuous samples are randomly sliced from the test set, and the samples' time interval is 2021.8.13 18:45 to 2021.8.20 00:45. Fig. 11 shows that MT-STGIN can accurately fit the changing trend of traffic speed and adapt to complex speed fluctuations, compared with optimal baseline models based on graph neural networks, T-GCN, DCRNN, GMAN, AST-GAT, and ST-GRAT.

For example, the traffic speeds present huge differences between different types of road segments; however, the performance of MT-STGIN keeps steady compared with baseline models. Combining this result with comparisons in Table 5 validates that traffic speed heterogeneity is a non-negligible factor in the highway system, which may cause the prediction model to discriminate the traffic condition limited. In addition, MT-STGIN still consistently achieves better results than other baselines, makes predictions close to actual observations, and conquers speed fluctuations, e.g., the red dashed boxes in Fig. 11(a), (b), and (c). These properties play a vital role in future travel services and traffic control. Moreover, the conclusion conjectures clarified in Section 5.4.1 can be verified in these cases; that is, the proposed method is more adaptive to traffic fluctuation, leading to model prediction shift in rare time steps, which causes performance to be inferior to GMAN and AST-GAT regarding RMSE and MAPE, as shown in Tables 3 and 4, but high accuracy in terms of MAE.

6. Conclusion

This paper proposes a multi-task-based spatiotemporal generative inference network (MT-STGIN) that can predict highway traffic speed accurately. The ST-block is first designed to be used on the encoder and decoder: semantic layer enhances the contextual semantic; F-Gate incorporates the physical relationship extracted by multi-head GCN into dynamic spatial correlation modeled by multi-head GAT; temporal attention network models the dynamic temporal dependency, which is later combined with spatial correlation through F-Gate. A generative inference is proposed to generate the target hidden outputs rather than dynamic step-by-step decoding to avoid error propagation in long-term prediction. Finally, a multi-task learning method divides long-term highway traffic speed prediction into three types of tasks, sharing the underlying network parameters and overcoming speed heterogeneity on the highway network.

Experiments on real-world highway traffic dataset show that MT-STGIN achieves state-of-the-art results compared with the baseline models, with a more evident advantage in long-term highway traffic speed prediction. Furthermore, comparing the prediction performance of all variants with MT-STGIN, the contributions of each part of the proposed method are highlighted. However, in this paper, we do not consider the impact of meteorological phenomena on traffic speed on the highway; we will embed this property into the model as additional information in future work. In addition, we consider incorporating existing tricks, such as multiscale speeds used in AST-GAT, into the proposed approach, achieving a more accurate prediction performance.

CRediT authorship contribution statement

Guojian Zou: Data curation, Writing – original draft, Visualization, Investigation, Writing – review & editing. **Ziliang Lai:** Conceptualization, Methodology. **Ting Wang:** Conceptualization. **Zongshi Liu:** Conceptualization. **Jingjue Bao:** Conceptualization. **Changxi Ma:** Conceptualization. **Ye Li:** Supervision, Writing – review & editing. **Jing Fan:** Software, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: We have no known competing financial interests or personal relationships.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported by the project of the China Scholarship Council (No. 202306260111), the National Key Research and Development Program of China (No. 2018YFB1601301), the National Natural Science Foundation of China (No. 71961137006, 52062027).

References

Afrin, T., & Yodo, N. (2022). A long short-term memory-based correlated traffic data prediction framework. *Knowledge-Based Systems*, *237*, Article 107755.

Ahmed, M. S., & Cook, A. R. (1979). *Analysis of freeway traffic time-series data by using Box-Jenkins techniques*. (722).

Cao, S., Lu, W., Zhou, J., & Li, X. (2018). Cw2vec: Learning chinese word embeddings with stroke n-gram information. *vol. 32*, In *Proceedings of the AAAI conference on artificial intelligence*. (1).

Chen, X., Wang, Z., Hua, Q., Shang, W.-L., Luo, Q., & Yu, K. (2022). AI-empowered speed extraction via port-like videos for vehicular trajectory analysis. *IEEE Transactions on Intelligent Transportation Systems*, *24*(4), 4541–4552.

Chen, X., Wu, S., Shi, C., Huang, Y., Yang, Y., Ke, R., et al. (2020). Sensing data supported traffic flow prediction via denoising schemes and ANN: A comparison. *IEEE Sensors Journal*, *20*(23), 14317–14328.

Cheng, R., Lyu, H., Zheng, Y., & Ge, H. (2022). Modeling and stability analysis of cyberattack effects on heterogeneous intelligent traffic flow. *Physica A. Statistical Mechanics and its Applications*, *604*, Article 127941.

Duan, P., Mao, G., Zhang, C., & Wang, S. (2016). STARIMA-based traffic prediction with time-varying lags. In *2016 IEEE 19th international conference on intelligent transportation systems (ITSC)* (pp. 1610–1615). IEEE.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2303.10130.

Fang, Y., Zhao, F., Qin, Y., Luo, H., & Wang, C. (2022). Learning all dynamics: Traffic forecasting via locality-aware spatio-temporal joint transformer. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 23433–23446.

Feng, A., & Tassiulas, L. (2022). Adaptive graph spatial-temporal transformer network for traffic forecasting. In *Proceedings of the 31st ACM international conference on information & knowledge management* (pp. 3933–3937).

Gao, M., Li, J.-Y., Chen, C.-H., Li, Y., Zhang, J., & Zhan, Z.-H. (2023). Enhanced multi-task learning and knowledge graph-based recommender system. *IEEE Transactions on Knowledge and Data Engineering*.

Gu, Y., Lu, W., Qin, L., Li, M., & Shao, Z. (2019). Short-term prediction of lane-level traffic speeds: A fusion deep learning model. *Transportation Research Part C: Emerging Technologies*, *106*, 1–16.

Hong, W.-C. (2011). Traffic flow forecasting by seasonal SVR with chaotic simulated annealing algorithm. *Neurocomputing*, *74*(12–13), 2096–2107.

Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L., & Shao, L. (2023). Normalization techniques in training dnns: Methodology, analysis and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Huang, X., Ye, Y., Ding, W., Yang, X., & Xiong, L. (2022). Multi-mode dynamic residual graph convolution network for traffic flow prediction. *Information Sciences*, *609*, 548–564.

James, J., Markos, C., & Zhang, S. (2021). Long-term urban traffic speed prediction with deep learning on graphs. *IEEE Transactions on Intelligent Transportation Systems*.

Jia, D., Chen, H., Zheng, Z., Watling, D., Connors, R., Gao, J., et al. (2021). An enhanced predictive cruise control system design with data-driven traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*.

Jia, Y., Wu, J., & Du, Y. (2016). Traffic speed prediction using deep learning method. In *2016 IEEE 19th international conference on intelligent transportation systems (ITSC)* (pp. 1217–1222). IEEE.

Jiang, B., & Fei, Y. (2016). Vehicle speed prediction by two-level data driven models in vehicular networks. *IEEE Transactions on Intelligent Transportation Systems*, *18*(7), 1793–1801.

Jiang, W., & Luo, J. (2022). Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, *207*, Article 117921.

Jin, G., Liang, Y., Fang, Y., Huang, J., Zhang, J., & Zheng, Y. (2023). Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. arXiv preprint arXiv:2303.14483.

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186).

Lee, Y., Jeon, H., & Sohn, K. (2020). Predicting short-term traffic speed using a deep neural network to accommodate citywide spatio-temporal correlations. *IEEE Transactions on Intelligent Transportation Systems*, *22*(3), 1435–1448.

Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., & Tian, Q. (2021). Multiscale spatio-temporal graph neural networks for 3D skeleton-based motion prediction. *IEEE Transactions on Image Processing*, *30*, 7760–7775.

Li, D., & Lasenby, J. (2021). Spatiotemporal attention-based graph convolution network for segment-level traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*.

Li, Z., Lu, C., Yi, Y., & Gong, J. (2021). A hierarchical framework for interactive behaviour prediction of heterogeneous traffic participants based on graph neural network. *IEEE Transactions on Intelligent Transportation Systems*.

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International conference on learning representations*.

Lin, L., Li, J., Chen, F., Ye, J., & Huai, J. (2017). Road traffic speed prediction: a probabilistic model fusing multi-source data. *IEEE Transactions on Knowledge and Data Engineering*, *30*(7), 1310–1323.

Liu, Z., Li, H., Wang, H., Liao, Y., Liu, X., & Wu, G. (2023). A novel pipelined end-to-end relation extraction framework with entity mentions and contextual semantic representation. *Expert Systems with Applications*, *228*, Article 120435.

Liu, L., Qiu, Z., Li, G., Wang, Q., Ouyang, W., & Lin, L. (2019). Contextualized spatial–temporal network for taxi origin-destination demand prediction. *IEEE Transactions on Intelligent Transportation Systems*, *20*(10), 3875–3887.

Lu, W., Rui, Y., & Ran, B. (2020). Lane-level traffic speed forecasting: A novel mixed deep learning model. *IEEE Transactions on Intelligent Transportation Systems*.

Lu, Y., Wang, Q., Ma, S., Geng, T., Chen, Y. V., Chen, H., et al. (2023). Transflow: Transformer as flow learner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18063–18073).

Lv, Z., Xu, J., Zheng, K., Yin, H., Zhao, P., & Zhou, X. (2018). Lc-rnn: A deep learning model for traffic speed prediction.. In *IJCAI* (pp. 3470–3476).

Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C (Emerging Technologies)*, *54*, 187–197.

Ma, C., Zhao, Y., Dai, G., Xu, X., & Wong, S.-C. (2022). A novel STFSA-CNN-gru hybrid model for short-term traffic speed prediction. *IEEE Transactions on Intelligent Transportation Systems*, *24*(4), 3728–3737.

Magazzino, C., & Mele, M. (2021). On the relationship between transportation infrastructure and economic development in China. *Research in Transportation Economics*, *88*, Article 100947.

Meng, X., Fu, H., Peng, L., Liu, G., Yu, Y., Wang, Z., et al. (2020). D-LSTM: Short-term road traffic speed prediction model based on gps positioning data. *IEEE Transactions on Intelligent Transportation Systems*.

Park, C., Lee, C., Bahng, H., Tae, Y., Jin, S., Kim, K., et al. (2020). ST-GRAT: A novel spatio-temporal graph attention networks for accurately forecasting dynamically changing road speed. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 1215–1224).

Qiu, Z., Zhu, T., Jin, Y., Sun, L., & Du, B. (2023). A graph attention fusion network for event-driven traffic speed prediction. *Information Sciences*, *622*, 405–423.

Qu, L., Lyu, J., Li, W., Ma, D., & Fan, H. (2021). Features injected recurrent neural networks for short-term traffic speed prediction. *Neurocomputing*, *451*, 290–304.

Rempe, F., Franeck, P., & Bogenberger, K. (2022). On the estimation of traffic speeds with deep convolutional neural networks given probe data. *Transportation Research Part C: Emerging Technologies*, *134*, Article 103448.

Shin, J., & Sunwoo, M. (2018). Vehicle speed prediction using a Markov chain with speed constraints. *IEEE Transactions on Intelligent Transportation Systems*, *20*(9), 3201–3211.

Song, C., Lee, H., Kang, C., Lee, W., Kim, Y. B., & Cha, S. W. (2017). Traffic speed prediction under weekday using convolutional neural networks concepts. In *2017 IEEE intelligent vehicles symposium (IV)* (pp. 1293–1298). IEEE.

Tang, J., Liu, F., Zou, Y., Zhang, W., & Wang, Y. (2017). An improved fuzzy neural network for traffic speed prediction considering periodic characteristic. *IEEE Transactions on Intelligent Transportation Systems*, *18*(9), 2340–2350.

Tang, W., Yiu, K., Chan, K., & Zhang, K. (2023). Conjoining congestion speed-cycle patterns and deep learning neural network for short-term traffic speed forecasting. *Applied Soft Computing*, *138*, Article 110154.

Vanajakshi, L., & Rilett, L. R. (2004). A comparison of the performance of artificial neural networks and support vector machines for the prediction of traffic speed. In *IEEE intelligent vehicles symposium, 2004* (pp. 194–199). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wang, J., Chen, R., & He, Z. (2019). Traffic speed prediction for urban transportation network: A path based deep learning approach. *Transportation Research Part C (Emerging Technologies)*, *100*, 372–385.

Wang, T., Cheng, R., & Wu, Y. (2022). Stability analysis of heterogeneous traffic flow influenced by memory feedback control signal. *Applied Mathematical Modelling*, *109*, 693–708.

Wang, H., Liu, L., Dong, S., Qian, Z., & Wei, H. (2016). A novel work zone short-term vehicle-type specific traffic speed prediction model through the hybrid EMD–ARIMA framework. *Transportmetrica B: Transport Dynamics*, *4*(3), 159–186.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4–24.

Yang, H., Liu, C., Zhu, M., Ban, X., & Wang, Y. (2021). How fast you will drive? Predicting speed of customized paths by deep neural network. *IEEE Transactions on Intelligent Transportation Systems*.

Yi, H., & Bui, K.-H. N. (2020). An automated hyperparameter search-based deep learning model for highway traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*.

Yu, B., Lee, Y., & Sohn, K. (2020). Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network (GCN). *Transportation Research Part C: Emerging Technologies*, *114*, 189–204.

Zafeiriou, S., Bronstein, M., Cohen, T., Vinyals, O., Song, L., Leskovec, J., et al. (2022). Guest editorial: Non-euclidean machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(02), 723–726.

Zang, D., Ling, J., Wei, Z., Tang, K., & Cheng, J. (2018). Long-term traffic speed prediction based on multiscale spatio-temporal feature learning network. *IEEE Transactions on Intelligent Transportation Systems*, *20*(10), 3700–3709.

Zhang, W., Feng, Y., Lu, K., Song, Y., & Wang, Y. (2020). Speed prediction based on a traffic factor state network model. *IEEE Transactions on Intelligent Transportation Systems*, *22*(5), 3112–3122.

Zhang, Z., Li, Y., Song, H., & Dong, H. (2021). Multiple dynamic graph based traffic speed prediction method. *Neurocomputing*, *461*, 109–117.

Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., et al. (2019). T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, *21*(9), 3848–3858.

Zheng, C., Fan, X., Wang, C., & Qi, J. (2020). Gman: A graph multi-attention network for traffic prediction. *vol. 34*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1234–1241). (01).

Zhou, Y., Li, J., Chi, J., Tang, W., & Zheng, Y. (2022). Set-CNN: A text convolutional neural network based on semantic extension for short text classification. *Knowledge-Based Systems*, *257*, Article 109948.

Zhou, L., Zhang, S., Yu, J., & Chen, X. (2019). Spatial–temporal deep tensor neural networks for large-scale urban network speed prediction. *IEEE Transactions on Intelligent Transportation Systems*, *21*(9), 3718–3729.

Zou, G., Lai, Z., Ma, C., Li, Y., & Wang, T. (2023). A novel spatio-temporal generative inference network for predicting the long-term highway traffic speed. *Transportation Research Part C (Emerging Technologies)*, *154*, Article 104263.

Zou, G., Lai, Z., Ma, C., Tu, M., Fan, J., & Li, Y. (2023). When will we arrive? A novel multi-task spatio-temporal attention network based on individual preference for estimating travel time. *IEEE Transactions on Intelligent Transportation Systems*.