ASTPKEFormer: Adaptive Spatiotemporal Prior Knowledge Embedding-Induced Transformers for Traffic Data Forecasting

Wenfeng Zhou[1] , **Xiaoyun Xia**[1,*] , **Xiangjie Kong**[2,*] , **Guojiang Shen**[2] , **Bin Chen**[1] , **Fei Wu**[1] , **Binbin Guo**[3]

[1] College of Artificial Intelligence, Jiaxing University, Jiaxing, China
[2]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China
[3]School of Information Engineering, Shaoguan University, Shaoguan, Guangdong, China
{wenfengZhou98, xiaxiaoyun}@zjxu.edu.cn, xjkong@ieee.org, gjshen1975@zjut.edu.cn, chenbin@zjxu.edu.cn, {wfmook, gbb197750}@163.com

Abstract

Traffic forecasting is fundamentally challenging due to the complex and dynamic spatiotemporal dependencies inherent in road networks. Although existing prediction models are able to achieve certain results on this task, existing Transformer-based models usually rely on simple embedding strategies and do not fully utilize the prior knowledge embedded in traffic patterns and network topology. To address these limitations, we propose ASTPKEformer, a prior knowledge-guided Transformer framework for traffic prediction. Based on the pure Transformer spatiotemporal self-attention mechanism, this model first uses a multi-scale temporal channel alignment module to generate discriminative feature embeddings. Then, it incorporates temporal prior embeddings and spatial graph structure prior embeddings to provide learning guidance for the model in both temporal and spatial dimensions. To further enhance the representation capability, an embedding cross-fusion mechanism is introduced to strengthen the interaction between the previous embeddings and the adaptive spatiotemporal embedding. Extensive experiments on six real-world traffic datasets demonstrate that ASTPKEformer overall outperforms state-of-the-art (SOTA) baselines, validating its effectiveness and strong generalization ability. The source code is available at ASTPKEformer Official Repository.

1 Introduction

Traffic data prediction is a key task in Intelligent Transportation Systems (ITS), aiming to predict future time series of road networks based on historical observation. Accurate prediction can promote proactive decision-making and optimal resource allocation, providing data support for downstream tasks such as traffic management, congestion mitigation, and travel time estimation, thereby improving the overall efficiency [Luo *et al.*, 2023b].

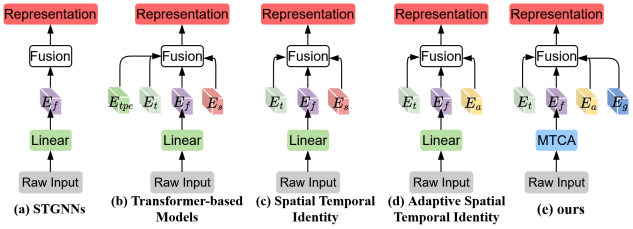


Figure 1: Input Embeddings in Different Traffic Prediction Models.

In recent years, with the rapid development of deep learning, numerous data-driven models have been proposed to capture the complex temporal dynamics and spatial correlations in traffic networks. Among them, Spatiotemporal Graph Neural Networks (STGNNs) [Yu *et al.*, 2018; Diao *et al.*, 2019; Fang *et al.*, 2019] were among the earliest models developed to jointly learn spatial and temporal dependencies, giving rise to three main architectural paradigms. First, static-graph-based approaches [Zhao *et al.*, 2019; Lv *et al.*, 2020; Wang *et al.*, 2020; Pan *et al.*, 2020] utilize predefined road network topologies to model spatial dependencies among sensors or road segments. Second, dynamic-graph-based approaches [Diao *et al.*, 2019; Zhang *et al.*, 2020; Bai *et al.*, 2020; Guo *et al.*, 2021; Zhang *et al.*, 2025] adaptively update graph structures over time to reflect the evolving spatial relationships driven by traffic dynamics. Finally, hybrid static–dynamic graph approaches [Wu *et al.*, 2019; Shao *et al.*, 2022b; Li *et al.*, 2023; Ma *et al.*, 2024] combine the advantages of both, enabling simultaneous modeling of explicit static and implicit dynamic dependencies. However, due to the inherent structural limitations of STGNNs, these models still struggle to capture long-range temporal dependencies and global contextual relationships. In contrast, Transformer-based models [Liu *et al.*, 2023; Jiang *et al.*, 2023a; Geng *et al.*, 2024; Guo *et al.*, 2025; Fang *et al.*, 2025] have attracted increasing attention for their superior capability in modeling global dependencies, evolving along two primary directions. On one hand, models based on the vanilla self-attention mechanism effectively capture long-term temporal interactions and global correlations but suffer from high computational complexity. On the other hand, models employing efficient or

*Xiaoyun Xia and Xiangjie Kong are corresponding authors.

linearized attention mechanisms aim to significantly reduce computational costs while maintaining comparable predictive performance. However, as the field progresses, it has become apparent that increasing the complexity of network architecture has not brought about significant performance improvements, which necessitates shifting our focus to the data itself.

Based on the above, this article focuses on a simple yet highly effective representation technique—*input embeddings* like [Liu *et al.*, 2023]. Fig .1 presents the principal types of embeddings commonly adopted in contemporary traffic prediction models. STGNN primarily uses feature embedding E_f [Yu *et al.*, 2018; Wu *et al.*, 2019], which projects raw inputs into a latent representation space via a linear transformation. Transformer-based models typically introduce additional time period information (hour, day, week) embedding E_t [Zheng *et al.*, 2020; Shao *et al.*, 2022a] and employ temporal position encoding E_{tpe} [Jin *et al.*, 2023] to assist the self-attention mechanism in understanding the temporal flow characteristics of the sequence. In addition, spatial embedding E_s are extensively utilized, often learned as implicit node representations to encode spatial correlations across the transportation network. More recent architectures, including STAEformer [Liu *et al.*, 2023], ASTRformer [Wang *et al.*, 2024], and DTRformer [Chen *et al.*, 2025b], all utilize adaptive spatiotemporal embedding E_a to improve the model's ability to model temporal sequence information and other traffic patterns in the input. However, most existing embedding strategies still exhibit notable limitations: (1) the spatial embedding component often **underutilizes prior structural knowledge of the road network**, limiting its expressiveness; and (2) **the interactions among different embeddings remain insufficiently modeled**, as current approaches commonly rely on simple concatenation or additive fusion (e.g., E_a and E_t), thereby lacking a deeper mechanism for cross-embedding interaction.

To address the aforementioned limitations, we incorporate spatial structural priors into the Transformer prediction framework in the form of a spatial prior embedding E_g. Our approach, referred to as *Adaptive Spatiotemporal Prior Knowledge Embedding-Induced Transformers* (ASTPKEformer), enhances the model's spatiotemporal representational capability by introducing temporal and spatial prior embeddings, while employing a cross-fusion mechanism to strengthen the deep interactions among different embeddings, particularly between the adaptive spatiotemporal embedding E_a and the prior-based embeddings (E_t and E_g). Moreover, we design a multi-scale temporal–channel attention module to further improve the modeling of feature embeddings E_f by capturing hierarchical temporal dependencies. Extensive experiments and analyses conducted on six real-world traffic datasets demonstrate that the proposed method achieves SOTA performance in traffic prediction. The main contributions of this work are summarized as follows:

- We propose a novel Transformer-based architecture that integrates spatial structure priors and temporal periodic priors into a traffic prediction framework. By embedding spatial and temporal prior knowledge into the model, the proposed ASTPKEformer effectively enhances the model's ability to capture explicit and implicit spatiotemporal dependencies in complex traffic networks.
- We construct a multi-scale spatiotemporal channel attention module, enabling the model to jointly learn multi-level temporal dynamics and cross-channel feature interactions. Furthermore, we design an embedding cross-fusion mechanism to promote deep interaction between heterogeneous embeddings, thereby improving the model's representational richness and robustness.
- Extensive experiments conducted on six real-world benchmark traffic datasets demonstrate that ASTPKEformer consistently outperforms existing SOTA models, achieving the best overall predictive performance across multiple evaluation metrics.

2 Related Work

Spatiotemporal Traffic Prediction. STGNNs and Transformers, which focus on modeling spatiotemporal dependencies, are two highly representative techniques applied to solving traffic data prediction tasks [Fang *et al.*, 2026]. For example, STGNN [Yu *et al.*, 2018] utilize graph convolutional networks based on predefined prior graph structures to extract spatiotemporal features of multi-scale traffic networks. AGCRN [Bai *et al.*, 2020] further learns adaptive graph structures to capture implicit spatial dependencies, while GWNet [Wu *et al.*, 2019] and the DGCRN [Li *et al.*, 2023] integrate static and dynamic graphs to jointly explore explicit and implicit spatial relationships. Despite their success, STGNN-based approaches still exhibit limited capability in modeling global dependencies. The self-attention mechanism in Transformers effectively captures global dependencies. Building on this, STAEFormer [Liu *et al.*, 2023] employs a vanilla Transformer architecture for sequential temporal modeling. PDFormer [Jiang *et al.*, 2023a] and STGAFormer [Geng *et al.*, 2024] perform parallel spatiotemporal modeling, where PDFormer enhances microscopic spatial representation through traffic delay-aware learning, while STGAFormer strengthens local temporal feature extraction via a gating mechanism. However, these methods still have shortcomings in the global and local modeling of spatiotemporal dependencies.

Spatiotemporal Embedding Learning. In traffic spatiotemporal prediction tasks, while complex model architectures have been extensively studied, input embedding techniques have also received considerable attention. These techniques aim to efficiently represent input data and enrich its informative content, thereby improving prediction accuracy. For example, feature embedding is widely used in various types of prediction methods, including but not limited to STGNN-based [Kong *et al.*, 2023; An *et al.*, 2025], Transformer-based [Zhao *et al.*, 2024; Jin *et al.*, 2025], and Linear-based [Zeng *et al.*, 2023] methods. Temporal and spatial embeddings are also frequently used in these methods, such as GMAN, STID, and LGSTFormer [Zhou *et al.*, 2025]. Position embedding is typically used in Transformer-based methods because attention mechanisms cannot pre-

serve sequential information between time points. Furthermore, adaptive spatiotemporal embedding was proposed in STAEformer, and the effectiveness of this simple structure has been proven; it is also widely used in methods such as ASTRformer, ImputeFormer, and STDHformer. Despite the achievements of these spatiotemporal embedding techniques, the interactions between different embeddings are often overlooked.

3 Preliminary

Traffic Spatio-Temporal Graph: Given a fixed traffic road network consisting of N nodes, the traffic spatiotemporal graph is formally defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A, X)$. Specifically, $|\mathcal{V}| = N$ denotes the set of nodes, and $\mathcal{E}$ represents the set of edges that describe the spatial connectivity among nodes. The corresponding adjacency matrix is denoted as $A \in \mathbb{R}^{N\times N}$, where each element a_{ij} equals 1 when an edge exists between node i and node j, and 0 otherwise. Furthermore, at each time step t, the dynamic graph signal of the traffic network (e.g., traffic flow or speed) is represented as a time-varying feature matrix $X_t \in \mathbb{R}^{N\times D}$, where $D = 1$ denotes the input dimension corresponding to the traffic volume.

Traffic Spatio-Temporal Graph Prediction: Given the historical traffic data $X_{t-T+1:t} \in \mathbb{R}^{T\times N\times 1}$ in the previous T time steps, the goal of traffic sptio-temporal graph prediction is to infer the traffic data $X_{t+1:t+P} \in \mathbb{R}^{P\times N\times 1}$ in the future P time steps by learning a project $\mathcal{F}(\cdot)$ based on the traffic spatiotemporal graph $\mathcal{G}$:

$$\{X_{t-T+1}, \ldots, X_t\} \xrightarrow{\mathcal{F}(\cdot)} \{X_{t+1}, \ldots, X_{t+P}\} \tag{1}$$

4 Methodology

An overview of the proposed ASTPKEformer is shown in Figure 2. Our model is composed of a multi-scale temporal–channel attention module (**MTCA Module**), an embedding layer (**Embedding Layer**), a prior knowledge embedding cross-fusion module (**PKECF Module**), vanilla Transformers are applied along the temporal and spatial dimensions respectively to construct the temporal and spatial Transformer layers , and finally, a regression layer for prediction output ((**Transformer and Regression Layer)**).

4.1 MTCA Module

As in most spatiotemporal prediction architectures, a fully connected layer is first employed to project the raw input into a d-dimension representation space, yielding the feature embedding $E_f \in \mathbb{R}^{T\times N\times d}$, where d represents the feature embedding dimnesion. Distinct from conventional designs, we further introduce a feature enhancement layer to better capture the intrinsic temporal characteristics of traffic data. Specifically, the enhancement module consists of two main components: a channel alignment block (CA) and a temporal alignment block (TA). The CA consists of two alternately executed fully connected network (FC) and layer normalization (LayerNorm), following the design principle proposed in [Luo *et al.*, 2023a]. In the first stage, a squeeze operation is applied to the feature embedding E_f along the channel dimension using an FC–LayerNorm sequence, enabling the model to recalibrate the distribution in the low-dimensional subspace. Subsequently, an excitation operation is performed through a gated threshold learning mechanism, which adaptively evaluates and emphasizes the relative importance of each channel as:

$$E_F = FC(X_{t-T+1:t}) \tag{2}$$

$$H_{CA} = ReLU(LayerNorm(FC(E_F))) \tag{3}$$

$$K_{CA} = Sigmoid(LayerNorm(FC(H_{CA}))) \tag{4}$$

$$\hat{H}_{CA} = E_F \odot K_{CA} \tag{5}$$

where H_{CA} denotes the compressed low-dimensional feature representation, K_{CA} represents the importance coefficients associated with each channel, and $\odot$ indicates the Hadamard element-wise product, with $ReLU$ and $Sigmoid$ serving as the nonlinear activation functions.

Similarly, to emphasize critical temporal points within the sequence, we design a multi-scale temporal convolution layer within the TA to evaluate the importance of different time steps from multiple receptive fields. Each temporal convolution branch is formulated as a squeeze–excitation module as:

$$H^k_{TA} = ReLU(BatchNorm(\Theta_{1,k} *_{conv} \hat{H}_{CA})) \tag{6}$$

$$K^k_{TA} = Sigmoid(BatchNorm(\Theta_{1,k} *_{conv} H^k_{TA})) \tag{7}$$

$$\hat{H}^k_{TA} = \hat{H}_{CA} \odot K^k_{TA} \tag{8}$$

where H^k_{TA} denotes the low-dimensional temporal feature representation obtained with a $1 \times k$ convolution kernel, K^k_{TA}represents the importance factors for different time steps, $\Theta_{1,k}$ and $*_{conv}$ denote the convolution kernel and the 2D convolution operation, respectively. BatchNorm indicates the Batch Normalization function. Based on this design, given k different convolution kernel sizes $\{k1, k2, \ldots, k\}$, the multi-scale temporal convolution branch produces enhanced temporal features under different receptive fields, denoted as $\{\hat{H}^{k1}_{TA}, \hat{H}^{k2}_{TA}, \ldots, \hat{H}^{k}_{TA}\}$. Finally, a FC layer is applied to get feature embedding $E_f \in \mathbb{R}^{T\times N\times d}$:

$$E_f = FC([\hat{H}^{k1}_{TA}||\hat{H}^{k2}_{TA}||\ldots||\hat{H}^{k}_{TA}]) \tag{9}$$

4.2 Embedding Layer

We improve the model's learning capacity by supplying multiple types of embedding inputs. Specifically, we design three distinct categories of embeddings, including temporal prior embedding, spatial graph-structure prior embedding, and adaptive spatiotemporal embedding.

Temporal prior embedding E_t: E_t is composed of two temporal-information embeddings: a timestamp–day embedding E_d and a day–week embedding E_w. The former encodes the sampling timestamp within a day, while the latter identifies the specific day of the week on which the observations occur. To obtain these two embeddings, we first construct two learnable parameter matrices $W_d \in \mathbb{R}^{288\times d}$ and $W_w \in \mathbb{R}^{7\times d}$ as an embedding dict. Then, the timestamp inputs $X_d \in \mathbb{R}^{T\times N}$ and day-of-week $X_w \in \mathbb{R}^{T\times N}$ inputs are used as indices to retrieve the corresponding embedding vectors, yielding $E_d \in \mathbb{R}^{T\times N\times d}$ and $E_w \in \mathbb{R}^{T\times N\times d}$. By concatenating them, we get the temporal prior embedding $E_t = [E_d||E_w] \in \mathbb{R}^{T\times N\times 2d}$.

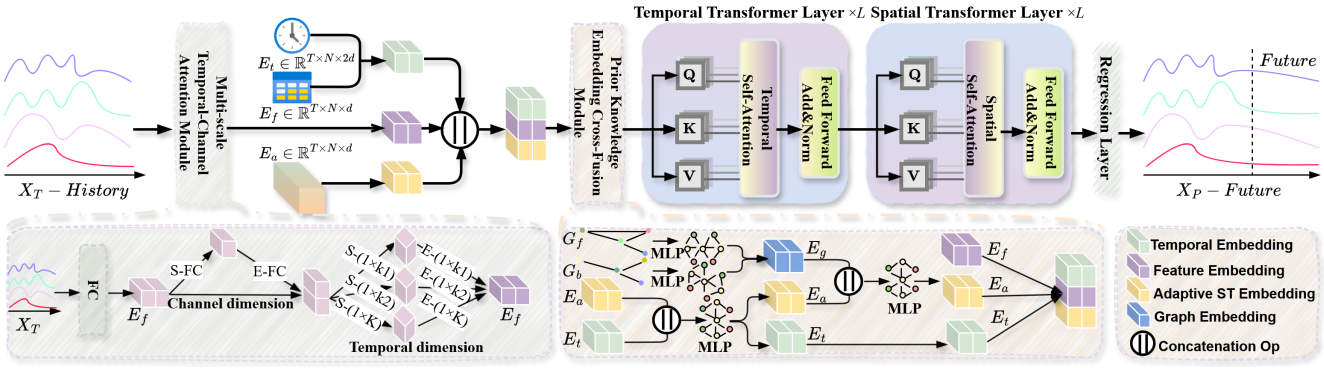


Figure 2: The Architecture of Adaptive Spatiotemporal Prior Knowledge Embedding-Induced transformer (ASTPKEformer).

Spatial graph-structure prior embedding E_g: Instead of using a learnable parameter matrix to dynamically model the spatial dependencies of the road network, we directly incorporate the adjacency matrix A of the network as the spatial embedding input. Specifically, following the procedure in [Guo *et al.*, 2025], we first compute the forward transition matrix $A_f \in \mathbb{R}^{N\times N}$ and the backward transition matrix $A_b \in \mathbb{R}^{N\times N}$ from the adjacency matrix A. These two transition matrices are then encoded through two separate multilayer perceptrons (MLPs) to obtain the graph-based representations $E_g^f \in \mathbb{R}^{N\times d_g}$ and $E_g^b \in \mathbb{R}^{N\times d_g}$, where d_g represents the graph embedding dimension. Furthermore, an *expand* operation is applied to align their dimensions with that of E_f, yielding the bidirectional spatial embeddings $E_g^f \in \mathbb{R}^{T\times N\times d_g}$ and $E_g^b \in \mathbb{R}^{T\times N\times d_g}$. By concatenating them, we get the spatial graph-structure prior embedding $E_g = [E_g^f||E_g^b] \in \mathbb{R}^{T\times N\times 2d_g}$.

Adaptive spatiotemporal embedding E_a: For a given traffic node, its traffic feature at a specific timestamp is often similar to the state observed at the same timestamp on the previous day or the previous week, and it is also influenced by the traffic conditions at adjacent time points. Moreover, the spatial dependencies among nodes in a road network are typically dynamic. Inspired by [Liu *et al.*, 2023], we introduce a learnable adaptive spatiotemporal embedding $E_a \in \mathbb{R}^{T\times N\times d_a}$, where d_a represents the adaptive spatiotemporal embedding dimension, to capture these complex spatiotemporal relationships in a unified manner.

4.3 PKECF Module

The embeddings E_t, E_g and E_a correspond to temporal, spatial, and spatiotemporal information, respectively. We design an embedding interaction module to strengthen the depth of cross-type information exchange. This module comprises two interaction processes: temporal–adaptive spatiotemporal interaction and spatial–adaptive spatiotemporal interaction.

Temporal–Adaptive spatiotemporal interaction (TASI): We first concatenate E_t and E_a to get $E_{ta} = [E_t||E_a] \in \mathbb{R}^{T\times N\times(2d+d_a)}$. Then we introduce a multilayer perceptron with residual connections (MLPRes) as the basic computational unit, where each MLPRes comprises two fully connected layers interleaved with a ReLU activation function. By sequentially stacking 2 MLPRes blocks, the interaction between E_t and E_a is achieved as:

$$E_{ta}^1 = E_{ta} + FC(ReLU(FC(E_{ta}))) \tag{10}$$

$$E_{ta}^2 = E_{ta}^1 + FC(ReLU(FC(E_{ta}^1))) \tag{11}$$

where $E_{ta}^2 \in \mathbb{R}^{T\times N\times(2d+d_a)}$. It is worth noting that we reseparate E_{ta}^2 into $\hat{E}_t$ and $\hat{E}_a$ along the channel dimension, while maintaining their original dimension size.

Spatial–Adaptive spatiotemporal interaction (SASI): Same as Temporal–Adaptive spatiotemporal interaction, we concatenate E_g and E_a to get $E_{ga} = [E_g||E_a] \in \mathbb{R}^{T\times N\times(2d_g+d_a)}$, and use 2 MLPRes blocks to achieve the interaction between E_g and E_a as:

$$E_{ga}^1 = E_{ga} + FC(ReLU(FC(E_{ga}))) \tag{12}$$

$$E_{ga}^2 = FC_{re}(E_{ga}^1 + FC(ReLU(FC(E_{ga}^1)))) \tag{13}$$

where $E_{ga}^2 \in \mathbb{R}^{T\times N\times d_a}$. Furthermore, we reduce the complexity of the E_{ga}^2 parameters by using an $FC_{re} : \mathbb{R}^{2d_g+d_a} \to \mathbb{R}^{d_a}$ layer. Finally, By concatenating the embeddings above, we can get the hidden spatio-temporal representation $H \in \mathbb{R}^{T\times N\times d_h}$ as follows:

$$H = E_f||E_{ga}^2 \tag{14}$$

where the hidden dimension d_h equals $3d + d_a$.

4.4 Transformer and Regression Layer

Transformer Layer: To learn the inherent spatiotemporal dependencies of traffic data, we use a pure self-attention mechanism along the temporal and spatial axes of the embedded input H to calculate the importance between different time points and between different nodes [Liu *et al.*, 2023]. By definition, we can obtain the query, key, and value matrix for the corresponding temporal attention based on H as:

$$Q^t = HW_Q^t, K^t = HW_K^t, V^t = HW_V^t \tag{15}$$

where $W_Q^t, W_K^t, W_V^t \in \mathbb{R}^{d_h\times d_h}$ are all learable parameters. Next, we compute the temporal self-attention score in parallel by performing matrix multiplication between Q^t and K^t as:

$$Attn^t = Softmax(\frac{Q^t {K^t}^T}{\sqrt{d_h}}) \in \mathbb{R}^{N\times T\times T} \tag{16}$$

where $Attn^t$ captures the implicit temporal dependencies. Subsequently, we output the temporal feature representation H^t as:

$$H^t = Attn^t V^t \in \mathbb{R}^{T \times N \times d_h} \tag{17}$$

Similarly, the spatial self-attention layer performs as:

$$H^s = selfAttention(H^t) \in \mathbb{R}^{T \times N \times d_h} \tag{18}$$

where H^s is the spatial feature representation. $selfAttention$ repeats the calculation process of formulas 15, 16 and 17 in the spatial axis.

Regression Layer: Finally, we used a single FC layer to output the spatio-temporal representation H^s as the future predictions as:

$$X_{t+1:t+P} = FC(H^s) \in \mathbb{R}^{P \times N \times 1} \tag{19}$$

where $X_{t+1:t+P}$ is the prediction result, P is the length of prediction.

5 Experiment

5.1 Experiment Settings

Datasets: We select six widely used traffic datasets covering two different data modalities to comprehensively evaluate the performance of ASTPKEformer. The detailed statistics of each dataset are provided in Table 1.

Datasets	Nodes	Time steps	Time range
PEMS03	358	26208	09/01/2018 - 11/30/2018
PEMS04	307	16992	01/01/2018 - 02/28/2018
PEMS07	883	28224	05/01/2017 - 08/31/2017
PEMS08	170	17856	07/01/2016 - 08/31/2016
METR-LA	207	34272	03/01/2012 - 06/30/2012
PEMS-BAY	325	52116	01/01/2017 - 05/31/2017

Table 1: Statistics of datasets.

Metrics: To evaluate the traffic prediction performance of ASTPKEformer, we adopt three commonly used evaluation metrics to quantitatively assess the experimental results: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).

Baselines: We compare ASTPKEformer with a series of representative spatio-temporal models to demonstrate the model's effectiveness. These models can be categorized into three types according to their calculation patterns: 1) Linear-based methods: STID [Shao *et al.*, 2022a], BLSTF [Chen *et al.*, 2025a]; 2) GCN-based methods: AGCRN [Bai *et al.*, 2020], GWNet [Wu *et al.*, 2019], MegaCRN [Jiang *et al.*, 2023b], HimNet [Dong *et al.*, 2024], STDN [Cao *et al.*, 2025]; 3) Attention-based methods: GMAN [Zheng *et al.*, 2020], STWave [Fang *et al.*, 2023], STAEformer [Liu *et al.*, 2023], STHDformer [Guo *et al.*, 2025].

Experimental Setup: Following the settings from previous studies [Liu *et al.*, 2023], we divide the dataset into training, validation, and test sets in a 6:2:2 ratio for traffic flow datasets and 7:1:2 ratio for traffic speed datasets based on the time order. Our experiments were conducted on a GeForce GTX 4090 GPU. All models were implemented using PyTorch. We set the model's input and output length to 1 hour, i.e., T=P=12. The optimizer was set to Adam with a decay learning rate of 0.001 and a batch size of 16. We also employed an early stopping mechanism with a $patience$ value of 10. The kernel size of MTCA is set to [1,3,5,7,9], and the graph structure embedding size is 64. Furthermore, for the six datasets in Table 1, the hidden embedding d and the adaptive spatio-temporal embedding d_a is set to $\{24, 24, 24, 32, 24, 24\}$ and $\{80, 40, 80, 80, 80, 80\}$, respectively. The graph embedding d_g maintains the same size across all datasets, 64.

5.2 Performance Analysis

We present the prediction results of all methods in Table 2 and Table 3, where $-$ indicates that no result was obtained due to excessive computational resources required. The following observations can be drawn from the prediction results: (1) The proposed ASTPKEformer achieves the best results on all six datasets, demonstrating that its enhanced embedding-interaction mechanism effectively captures spatiotemporal dependencies and yields strong generalization; (2) Despite their simplicity, MLP-based methods do not perform the worst on either flow or speed prediction tasks and often surpass models such as AGCRN, STDN, and STWave, indicating that even basic embedding strategies can provide competitive accuracy; (3) Among GCN-based approaches, MegaCRN performs best on METR-LA, while HimNet achieves superior overall results on PEMS-BAY and the flow datasets; (4) Attention-based models generally outperform GCN-based ones. STHDformer, which also integrates graph-structure embeddings, ranks second only to ASTPKEformer on flow datasets, and STAEformer delivers strong performance on speed datasets, underscoring the adaptability of attention mechanisms to spatiotemporal traffic forecasting.

5.3 Performance Evaluation of MTCA

As one of the most important components proposed in this study, we further evaluate the effect of MTCA module. We selected three comparison models that can be used in a plug-and-play manner for MTCA, including GWNet, GMAN and STWA. Table 4 shows the performance comparison results. Based on the experimental results, it can be seen that MTCA can improve the prediction accuracy of the original model to a certain extent. This shows that MTCA is applicable to different types of models and can be used as a plug-and-play module.

5.4 Ablation Experiment

To investigate the relative importance of different modules in the prediction process, we construct six model variants: (1) w/o MTCA: the multi-scale temporal–channel attention (MTCA) module is removed; (2) w/o PKECF: the embedding interaction module is removed; (3) w/o TASI: the temporal prior embedding E_t and adaptive spatiotemporal embedding E_a do not interact; (4) w/o E_a: the spatial-structure prior embedding is removed, which also eliminates the SASI process; (5) w/o SA: the spatial Transformer module is removed; (6)

Methods	Dataset	PEMS03			PEMS04			PEMS07			PEMS08		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Methods	STID	26.23	15.35	16.65	29.97	18.39	12.58	32.63	19.55	8.27	23.24	14.22	9.39
	BLSTF	**24.35**	**14.14**	**14.32**	**29.76**	18.05	11.84	32.35	18.95	7.96	23.23	13.49	8.82
	AGCRN	27.83	16.33	15.91	31.52	19.71	12.93	34.77	21.40	9.28	24.96	15.84	10.62
	GWNet	25.27	14.54	14.46	30.41	18.92	12.70	33.78	20.80	8.77	23.67	14.71	9.39
	MegaCRN	26.20	14.58	15.14	30.74	18.79	12.82	32.72	19.66	8.26	23.79	14.73	9.44
	HimNet	27.11	15.35	15.82	29.85	18.15	11.96	32.82	19.28	8.03	23.26	13.54	8.94
	STDN	28.45	15.71	16.67	30.28	18.32	12.24	33.39	20.15	9.29	24.20	14.01	9.75
	GMAN	26.68	15.69	15.90	30.84	19.16	12.97	34.20	21.15	10.37	24.11	14.61	9.86
	STWave	27.76	15.90	15.59	31.19	19.53	13.03	34.74	21.90	9.68	24.77	15.77	10.42
	STAEformer	27.00	15.15	15.21	30.11	18.21	12.10	32.60	19.22	8.02	22.98	13.55	8.90
	STHDformer	26.18	14.93	14.90	29.89	18.17	11.93	32.55	19.28	8.04	23.11	13.37	**8.73**
	Ours	26.79	14.74	14.73	29.87	**18.03**	**11.79**	**32.32**	**18.94**	**7.86**	**22.89**	**13.32**	8.74

Table 2: Performance on traffic flow datasets with best values shaded and second-best values underlined.

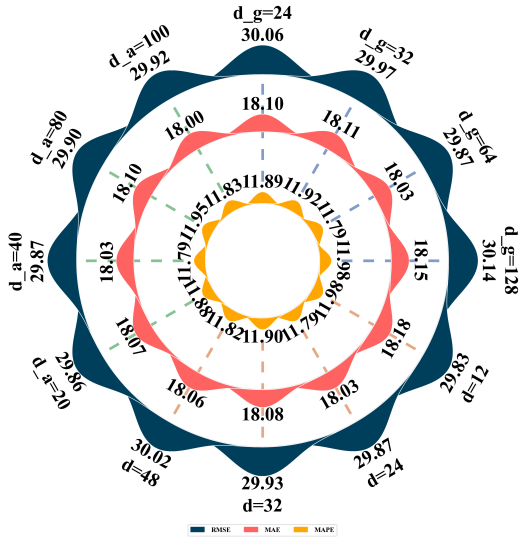
w/o TA: the temporal Transformer module is removed. Corresponding ablation experiments are conducted on the traffic flow dataset PEMS08 and the traffic speed dataset METR-LA. The results of these experiments are presented in Table 5. The results clearly show that each module contributes positively to the prediction performance. Removing the introduced E_g degrades prediction performance on both datasets. Furthermore, removing the interactive modules has a stronger negative impact, further demonstrating the importance of interactions between embeddings.

5.5 Hyperparameter Experiment

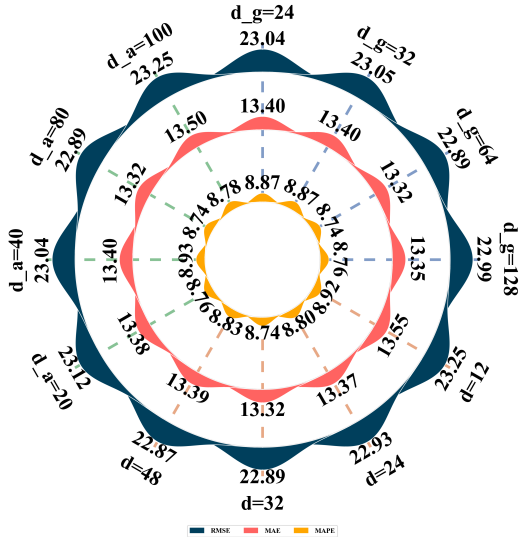
We conduct hyperparameter experiments on the PEMS04 and PEMS08 datasets to evaluate the impact of the feature embedding dimension d, the graph structure embedding dimension d_g and the adaptive spatiotemporal embedding dimension d_a on the model's predictive performance. The results are shown in Figure 3. We tune $d \in \{12, 24, 36, 48\}$, $d_g \in \{24, 32, 64, 128\}$, and $d_a \in \{20, 40, 80, 100\}$. The results show clear dataset-dependent behaviors: the optimal d and d_a vary across datasets due to differences in traffic dynamics. For PEMS04, larger d degrades performance, likely from over-parameterization, whereas for PEMS08, moderately increasing d strengthens the model's ability to capture complex temporal patterns. A similar trend holds for d_a, where datasets with higher variability or stronger spatiotemporal correlations benefit from larger adaptive embedding dimensions.

5.6 Visualization Study

We visualize the adaptive spatiotemporal embedding E_a and compare it with STAEformer using 2D t-SNE [Maaten and Hinton, 2008] for spatial analysis and time-step correlation for temporal analysis. As shown in Fig.4. ASTPKEformer exhibits clearer spatial clustering, indicating enhanced spatial representation from graph-structured priors. Temporally, unlike STAEformer's overly smooth correlations, ASTPKEformer shows a sharper decay that better captures short-term dependencies and long-term attenuation. These results demonstrate that interacting E_t with E_a effectively improves temporal modeling and validate the benefits of graph priors and strengthened embedding interactions.



(a) The results on PEMS04



(b) The results on PEMS08

Figure 3: Sensitivity analysis on PEMS04 and PEMS08 dataset.

Dataset	Methods	15min			30min			60min			Avg		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
METR-LA	STID	5.53	2.81	7.68	6.58	3.18	9.30	7.50	3.54	10.86	6.47	3.12	9.06
	BLSTF	5.51	2.80	7.46	6.58	3.17	8.98	7.47	3.53	10.59	6.44	3.11	8.79
	AGCRN	5.64	2.89	7.80	6.77	3.29	9.35	7.79	3.71	10.98	6.65	3.24	9.16
	GWNet	5.17	2.73	7.06	6.25	3.13	8.37	7.35	3.59	10.04	6.20	3.09	8.27
	MegaCRN	5.03	2.63	**6.66**	6.10	3.02	**8.06**	7.32	3.50	9.79	6.08	2.98	**7.96**
	HimNet	5.05	2.62	6.75	6.11	2.98	8.22	7.28	3.39	9.96	6.08	2.94	8.11
	STDN	5.78	2.80	7.71	6.89	3.20	9.33	7.75	3.55	10.85	6.71	3.11	9.12
	GMAN	5.69	2.84	7.78	6.53	3.13	8.98	7.53	3.51	10.59	6.52	3.12	8.97
	STWave	5.61	2.88	7.54	6.61	3.28	9.06	7.68	3.73	10.77	6.56	3.24	8.89
	STAEformer	5.12	2.65	6.93	6.04	2.97	8.20	7.05	3.35	9.80	6.00	2.94	8.12
	STHDformer	-	-	-	-	-	-	-	-	-	-	-	-
	Ours	**5.01**	**2.61**	6.76	**5.92**	**2.93**	**8.06**	**6.94**	**3.31**	**9.63**	**5.89**	**2.90**	**7.96**
PEMS-BAY	STID	2.78	1.31	2.77	3.70	1.63	3.70	4.39	1.91	4.50	3.60	1.57	3.53
	BLSTF	2.77	1.30	2.72	3.68	1.61	3.63	4.35	1.88	4.44	3.56	1.54	3.47
	AGCRN	2.93	1.39	3.04	3.92	1.72	3.95	4.65	2.00	4.72	3.80	1.65	3.78
	GWNet	2.73	1.30	2.71	3.67	1.62	3.61	4.49	1.94	4.57	3.60	1.56	3.49
	MegaCRN	2.72	1.29	2.70	3.71	1.62	3.67	4.54	1.93	4.67	3.64	1.56	3.56
	HimNet	**2.69**	**1.27**	2.71	3.64	1.62	**3.55**	4.35	**1.86**	**4.37**	3.56	1.54	3.43
	STDN	2.85	1.34	2.84	3.79	1.66	3.78	4.42	1.93	4.60	3.66	1.59	3.62
	GMAN	2.89	1.34	2.85	3.71	1.62	3.67	**4.29**	1.87	4.40	3.60	1.57	3.53
	STWave	2.90	1.38	2.97	3.88	1.72	3.95	4.60	2.03	4.77	3.76	1.66	3.77
	STAEformer	2.78	1.30	2.77	3.68	1.61	3.66	4.33	1.87	4.44	3.56	1.54	3.50
	STHDformer	-	-	-	-	-	-	-	-	-	-	-	-
	Ours	2.75	1.29	**2.69**	**3.62**	**1.59**	**3.55**	**4.29**	**1.86**	4.40	**3.52**	**1.53**	**3.42**

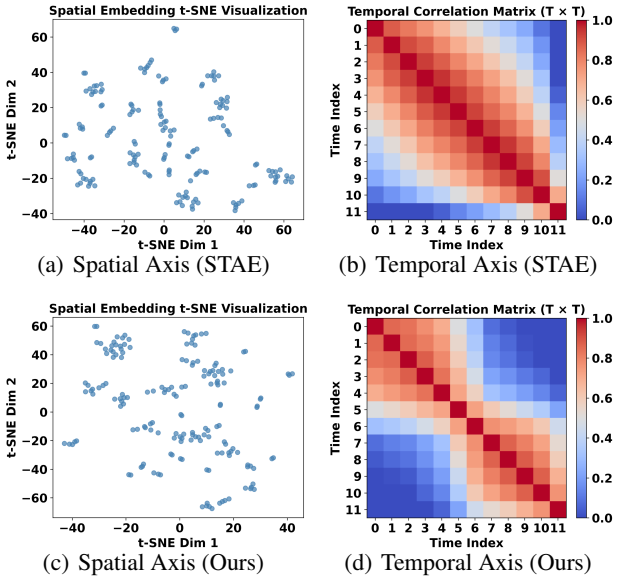
Table 3: Performance on traffic speed datasets for different horizons with best values shaded and second-best values underlined.

	PEMS04			PEMS08			METR-LA		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
GWNet	30.29	18.92	12.70	23.67	14.71	9.39	6.20	3.09	8.27
GWNet+MTCA	**30.29**	**18.75**	**12.53**	**23.54**	**14.63**	**9.34**	**6.06**	**3.00**	**8.09**
GMAN	30.84	19.16	12.97	24.11	14.61	9.86	6.52	3.12	8.97
GMAN+MTCA	**30.57**	**18.95**	**12.82**	**23.66**	**14.35**	**9.55**	**6.24**	**3.03**	**8.56**
STWA	31.19	**19.53**	13.03	**24.77**	15.77	10.42	6.56	3.24	**8.89**
STWA+MTCA	**31.16**	19.61	**12.79**	24.78	**15.70**	**10.32**	**6.41**	**3.17**	8.92

Table 4: The effect of MTCA on different models.

Method	PEMS08			METR-LA		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
ASTPKEformer	22.89	13.32	**8.74**	**5.89**	**2.90**	**7.96**
w/o MTCA	22.94	**13.30**	8.77	6.01	2.94	8.17
w/o PKECF	23.19	13.48	8.84	6.02	2.95	8.14
w/o TASA	23.08	13.42	8.79	5.93	2.92	8.06
w/o E_g	23.05	13.40	8.77	5.96	2.93	8.06
w/o SA	**22.79**	13.35	9.13	6.50	3.09	8.97
w/o TA	23.03	13.44	8.85	5.97	2.92	8.11

Table 5: Ablation study on PEMS08 and METR-LA datasets.



(a) Spatial Axis (STAE) (b) Temporal Axis (STAE)

(c) Spatial Axis (Ours) (d) Temporal Axis (Ours)

Figure 4: Visualization of E_a on PEMS08.

6 Conclusion

In this work, we propose ASTPKEformer, a Transformer-based traffic forecasting model guided by adaptive spatiotemporal prior knowledge. Building upon the standard spatiotemporal self-attention mechanism, the model first employs a multi-scale temporal–channel alignment module to generate high-quality feature embeddings. It then enriches the input representation by simultaneously integrating temporal prior embeddings and graph-structured spatial prior embeddings. Furthermore, a cross-fusion mechanism is designed to facilitate deep interactions between prior embeddings and adaptive spatiotemporal embeddings, thereby strengthening the model's capacity to capture complex spatiotemporal dependencies. Extensive experiments conducted on six real-world benchmark datasets demonstrate the effectiveness and superiority of the proposed framework. In the future, we will test the model's performance in other scenarios, such as those with missing data or few samples.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62476247 and 61703183, and in part by the National Key Research and Development Program of China under Grant 2023YFC3305900, 2023YFC3305903, and in part by Zhejiang Provincial Natural Science Foundation of China under Grant ZCLZ25F0201, D24F020013, LGG19F030010, and in part by the Qin Shen Scholar Program of Jiaxing University..

References

[An *et al.*, 2025] Yang An, Zhibin Li, Xiaoyu Li, Wei Liu, Xinghao Yang, Haoliang Sun, Meng Chen, Yu Zheng, and Yongshun Gong. Spatio-temporal multivariate probabilistic modeling for traffic prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Bai *et al.*, 2020] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in Neural Information Processing Systems*, 33:17804–17815, 2020.

[Cao *et al.*, 2025] Lingxiao Cao, Bin Wang, Guiyuan Jiang, Yanwei Yu, and Junyu Dong. Spatiotemporal-aware trend-seasonality decomposition network for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11463–11471, 2025.

[Chen *et al.*, 2025a] Jing Chen, Shixiang Pan, Weimin Peng, and Wenqiang Xu. Bilinear spatiotemporal fusion network: An efficient approach for traffic flow prediction. *Neural Networks*, 187:107382, 2025.

[Chen *et al.*, 2025b] Jing Chen, Haocheng Ye, Zhian Ying, Yuntao Sun, and Wenqiang Xu. Dynamic trend fusion module for traffic flow prediction. *Applied Soft Computing*, 174:112979, 2025.

[Diao *et al.*, 2019] Zulong Diao, Xin Wang, Dafang Zhang, Yingru Liu, Kun Xie, and Shaoyao He. Dynamic spatial-temporal graph convolutional neural networks for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 890–897, 2019.

[Dong *et al.*, 2024] Zheng Dong, Renhe Jiang, Haotian Gao, Hangchen Liu, Jinliang Deng, Qingsong Wen, and Xuan Song. Heterogeneity-informed meta-parameter learning for spatiotemporal time series forecasting. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 631–641, 2024.

[Fang *et al.*, 2019] Shen Fang, Qi Zhang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Gstnet: global spatial-temporal network for traffic flow prediction. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2286–2293, 2019.

[Fang *et al.*, 2023] Yuchen Fang, Yanjun Qin, Haiyong Luo, Fang Zhao, Bingbing Xu, Liang Zeng, and Chenxing Wang. When spatio-temporal meet wavelets: Disentangled traffic forecasting via efficient spectral graph attention networks. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 517–529. IEEE, 2023.

[Fang *et al.*, 2025] Yuchen Fang, Yuxuan Liang, Bo Hui, Zezhi Shao, Liwei Deng, Xu Liu, Xinke Jiang, and Kai Zheng. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 307–317, 2025.

[Fang *et al.*, 2026] Yuchen Fang, Hao Miao, Yuxuan Liang, Liwei Deng, Yue Cui, Ximu Zeng, Yuyang Xia, Yan Zhao, Torben Bach Pedersen, Christian S Jensen, et al. Unraveling spatio-temporal foundation models via the pipeline lens: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

[Geng *et al.*, 2024] Zili Geng, Jie Xu, Rongsen Wu, Changming Zhao, Jin Wang, Yunji Li, and Chenlin Zhang. Stgaformer: Spatial–temporal gated attention transformer based graph neural network for traffic flow forecasting. *Information Fusion*, 105:102228, 2024.

[Guo *et al.*, 2021] Shengnan Guo, Youfang Lin, Huaiyu Wan, Xiucheng Li, and Gao Cong. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5415–5428, 2021.

[Guo *et al.*, 2025] Feng Guo, Xunhuang Wang, Lyuchao Liao, Lei Zou, Tianwei Zhu, and Fumin Zou. Sthdformer: A novel dual-branch transformer for traffic flow forecasting. *Authorea Preprints*, 2025.

[Jiang *et al.*, 2023a] Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4365–4373, 2023.

[Jiang *et al.*, 2023b] Renhe Jiang, Zhaonan Wang, Jiawei Yong, Puneet Jeph, Quanjun Chen, Yasumasa Kobayashi, Xuan Song, Shintaro Fukushima, and Toyotaro Suzumura. Spatio-temporal meta-graph learning for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8078–8086, 2023.

[Jin *et al.*, 2023] Di Jin, Jiayi Shi, Rui Wang, Yawen Li, Yuxiao Huang, and Yu-Bin Yang. Trafformer: Unify time and space in traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8114–8122, 2023.

[Jin *et al.*, 2025] Di Jin, Cuiying Huo, Jiayi Shi, Dongxiao He, Jianguo Wei, and Philip S Yu. Llgformer: Learnable long-range graph transformer for traffic flow prediction. In *Proceedings of the ACM on Web Conference 2025*, pages 2860–2871, 2025.

[Kong *et al.*, 2023] Xiangjie Kong, Wenfeng Zhou, Guojiang Shen, Wenyi Zhang, Nali Liu, and Yao Yang. Dynamic graph convolutional recurrent imputation network for spatiotemporal traffic missing data. *Knowledge-Based Systems*, 261:110188, 2023.

[Li *et al.*, 2017] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural net-

work: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[Li *et al.*, 2023] Fuxian Li, Jie Feng, Huan Yan, Guangyin Jin, Fan Yang, Funing Sun, Depeng Jin, and Yong Li. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. *ACM Transactions on Knowledge Discovery from Data*, 17(1):1–21, 2023.

[Liu *et al.*, 2023] Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4125–4129, 2023.

[Luo *et al.*, 2023a] Xunlian Luo, Chunjiang Zhu, Detian Zhang, and Qing Li. Dynamic graph convolutional network with attention fusion for traffic flow prediction. *arXiv preprint arXiv:2302.12598*, 2023.

[Luo *et al.*, 2023b] Xunlian Luo, Chunjiang Zhu, Detian Zhang, and Qing Li. Stg4traffic: A survey and benchmark of spatial-temporal graph neural networks for traffic prediction. *CoRR*, 2023.

[Lv *et al.*, 2020] Mingqi Lv, Zhaoxiong Hong, Ling Chen, Tieming Chen, Tiantian Zhu, and Shouling Ji. Temporal multi-graph convolutional network for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 22(6):3337–3348, 2020.

[Ma *et al.*, 2024] Ying Ma, Haijie Lou, Ming Yan, Fanghui Sun, and Guoqi Li. Spatio-temporal fusion graph convolutional network for traffic flow forecasting. *Information Fusion*, 104:102196, 2024.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[Pan *et al.*, 2020] Zheyi Pan, Wentao Zhang, Yuxuan Liang, Weinan Zhang, Yong Yu, Junbo Zhang, and Yu Zheng. Spatio-temporal meta learning for urban traffic prediction. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1462–1476, 2020.

[Shao *et al.*, 2022a] Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4454–4458, 2022.

[Shao *et al.*, 2022b] Zezhi Shao, Zhao Zhang, Wei Wei, Fei Wang, Yongjun Xu, Xin Cao, and Christian S Jensen. Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. *Proceedings of the VLDB Endowment*, 15(11):2733–2746, 2022.

[Song *et al.*, 2020] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 914–921, 2020.

[Wang *et al.*, 2020] Xiaoyang Wang, Yao Ma, Yiqi Wang, Wei Jin, Xin Wang, Jiliang Tang, Caiyan Jia, and Jian Yu. Traffic flow prediction via spatial temporal graph neural network. In *Proceedings of the Web Conference 2020*, pages 1082–1092, 2020.

[Wang *et al.*, 2024] Ruidong Wang, Liang Xi, Jinlin Ye, Fengbin Zhang, Xu Yu, and Lingwei Xu. Adaptive spatio-temporal relation based transformer for traffic flow prediction. *IEEE Transactions on Vehicular Technology*, 74(2):2220–2230, 2024.

[Wu *et al.*, 2019] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1907–1913, 2019.

[Yu *et al.*, 2018] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3634–3640. International Joint Conferences on Artificial Intelligence Organization, 2018.

[Zeng *et al.*, 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11121–11128, 2023.

[Zhang *et al.*, 2020] Qi Zhang, Jianlong Chang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Spatio-temporal graph structure learning for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1177–1185, 2020.

[Zhang *et al.*, 2025] Xiaomei Zhang, Ziqin Jiang, and Ping Lou. Triple dynamic graph convolutional recurrent network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[Zhao *et al.*, 2019] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21(9):3848–3858, 2019.

[Zhao *et al.*, 2024] Zhenzhen Zhao, Guojiang Shen, Lei Wang, and Xiangjie Kong. Graph spatial-temporal transformer network for traffic prediction. *Big Data Research*, 36:100427, 2024.

[Zheng *et al.*, 2020] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1234–1241, 2020.

[Zhou *et al.*, 2025] Wenfeng Zhou, Guojiang Shen, Zhenzhen Zhao, Zhaolin Deng, Tao Tang, Xiangjie Kong, Amr Tolba, and Osama Alfarraj. A transformer-based approach for traffic prediction with fusion spatiotemporal attention. *Knowledge-Based Systems*, page 114466, 2025.