# STPformer: Mutation-Aware Spatial-Temporal Pivotal Attention Networks for Transformer-based Traffic Forecasting

Hongyang Su[1], Chenyun Yu[2(✉)], Qingcai Chen[1(✉)], Beibei Kong[3], Lei Cheng[3], Chengxiang Zhuo[3], Zang Li[3], and Xiaolong Wang[1]

[1] Harbin Institute of Technology, Shenzhen, China
suhongyang@stu.hit.edu.cn, {qingcai.chen, xlwangsz}@hit.edu.cn
[2] Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
yuchy35@mail.sysu.edu.cn
[3] Platform and Content Group, Tencent, Shenzhen, Guangdong, China
{echokong, raycheng, felixzhuo, gavinzli}@tencent.com

**Abstract.** The primary challenge in traffic forecasting lies in effectively capturing the spatio-temporal patterns in traffic data. Recent studies have highlighted the importance of pivotal nodes in road networks, which exhibit dominant impacts due to their prominent role in flow distribution. However, existing methods focus solely on the pivotal properties of the spatial dimension, inevitably diminishing the synchronisation of spatio-temporal patterns. Additionally, nodes with critical spatial semantic attributes are often overlooked. Despite their limited capacity for traffic distribution, these nodes are equally influential due to their strategic geographic positioning or intricate location characteristics. To overcome those limitations, we introduce a novel Spatial-Temporal Pivotal Attention Networks (STPformer) for traffic forecasting. Specifically, our model incorporates a mutation-aware pivotal temporal attention mechanism, which is integrated with Hawkes process, ensuring precise attention to the transition patterns from historical to future sequences. Moreover, the pivotal spatial attention integrated with a probabilistic sparsification mechanism is proposed to adaptively capture the spatial heterogeneity of nodes with significant spatial semantic attributes. By integrating these two innovative components into a Transformer-based architecture, STPformer efficiently learns fine-grained and synchronised spatial-temporal dependencies through stacked layers. Comprehensive experiments have demonstrated the superiority of STPformer in precision, efficiency, scalability, and interpretability.

**Keywords:** Traffic forecasting · Spatio-temporal patterns · Pivotal temporal attention · Pivotal spatial attention · Transformer

## 1 Introduction

Traffic forecasting, a typical task in spatio-temporal data mining, has attracted significant interest across various intelligent transportation applications. It aims

to accurately predict the future traffic state (in terms of flow or speed) at specific locations or sensors, offering valuable insights for dynamic traffic management, travel duration optimization and intelligent route planning, amongst others [13, 2, 16]. It is noteworthy that traffic flow on any given road is not only determined by its historical traffic state, but also affected by the traffic conditions on adjacent roads, including both upstream and downstream. This influence stems from the intrinsic interconnectivity within road networks, which is known as spatio-temporal patterns [25, 19]. Consequently, how to accurately learn the spatio-temporal patterns from observed traffic data poses a critical challenge in the realm of traffic forecasting.

Early studies utilized convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for traffic prediction, but these models are limited by their neglect of spatial correlations among sensors [21, 15]. In response, subsequent studies have turned to spatial-temporal graph models, which integrate temporal convolution and graph convolution, and have yielded promising results [13, 25]. This success is primarily attributed to the effective alignment of sensors and roads within the road network to the corresponding graph nodes and edges. Nonetheless, the intricate spatio-temporal dynamics inherent in traffic data are not fully captured by standard convolutional components. Consequently, a multitude of studies have dedicated to the enhancement of spatio-temporal pattern recognition, each focusing on the diverse attributes of traffic data that have a profound impact on the prediction performance, including dynamic or static, local or global, synchronous or hierarchical spatio-temporal patterns [24, 22, 12, 10]. Additionally, transformer-based models have demonstrated remarkable performance in traffic prediction [14, 8, 7]. The seamless integration of attention modules within these spatio-temporal frameworks significantly enhances the model's perceptual capability, while the spatial-temporal embedding facilitates the efficient aggregation of data patterns across multiple dimensions, thereby capturing the complex spatio-temporal dynamics of traffic flow more effectively. To better leverage the spatial information embedded in traffic data, recent research has concentrated on identifying key nodes within the road network, known as pivotal nodes, which exhibit more complex spatio-temporal patterns due to their significant role in traffic flow distribution [9]. Although this focus on pivotal nodes brings a new perspective to traffic forecasting, there are still several challenges that remain to be addressed:

**CH1. The pivotal characteristics in the temporal dimension are overlooked.** Existing studies focus only on pivotal nodes in the spatial dimension, but none of them pay enough attention to the temporal pivotal attributes. We argue that the historical time steps of traffic data possess significant pivotal properties, which we define as pivotal temporal attributes that are crucial for understanding traffic dynamics. Different from traditional periodic temporal attributes (e.g., holidays or peak hours), the pivotal temporal attributes are more concerned with the transition patterns between historical and future sequences. Therefore, they exhibit heightened sensitivity to mutations in traffic

flow, presenting a more significant challenge for accurate detection in complex traffic environments.

**CH2. The global pivotal nodes with crucial semantic attributes in the spatial dimension are not adequately recognized.** It is intuitive to regard nodes with strong traffic distribution capabilities as pivotal nodes. For instance, STPGNN [9] differentiates between pivotal and non-pivotal nodes using a pivotal node identification module and employs graph convolutional networks (GCNs) to capture the spatial dependencies for these two groups separately. However, existing approaches for identifying pivotal nodes often lack a global perspective. In practical scenarios, nodes that might not possess strong traffic distribution capabilities but are situated in critical locations with intricate spatial patterns also warrant special attention. Furthermore, to the best of our knowledge, none of the current methods for exploring spatial pivotal nodes guarantee the synchronization properties of spatio-temporal patterns, which are also vital attributes that impact prediction performance [18, 12].

In this paper, we propose a Spatial-Temporal Pivotal Attention Network for Transformer-based traffic forecasting (STPformer), which concentrates on learning precise and fine-grained spatio-temporal patterns. **As for the first challenge**, we propose the mutation-aware Pivotal Temporal Transformer (PT-Trans) to capture the pivotal attributes from the temporal perspective. Specifically, we design a pivotal temporal attention module with the Hawkes process [11]. By decoupling the temporal patterns of the traffic data into a data-driven and continuous time point process, PT-Trans attaches more attention to the transition patterns at the tail of historical sequences and facilitates the perception for sudden changes in traffic flow. **To address the second challenge,** we propose the Pivotal Spatial Transformer (PS-Trans) to capture pivotal spatial nodes without losing critical semantic attributes. In particular, we elaborately design a pivotal spatial attention module with a probabilistic sparsification mechanism, which allows for the efficient and adaptive identification of pivotal nodes. By focusing on these pivotal nodes and amplifying their influence, our model can learn more precise spatio-temporal patterns in intricate traffic scenarios. In summary, the core contributions of this work are as follows:

**(1)** We introduce a novel end-to-end spatial-temporal framework for traffic forecasting, which enhances the prediction accuracy by incorporating pivotal attributes across both temporal and spatial dimensions. Our approach marks a pioneering effort in proposing novel methods that adaptively identify pivotal temporal attributes within historical time steps and pivotal spatial attributes of the road network, synchronously. **(2)** We propose a novel mutation-aware pivotal temporal attention mechanism with Hawkes process to capture temporal pivotal attributes of historical time steps. Furthermore, we design a pivotal spatial attention with the ProbSparse mechanism to adaptively identify spatial pivotal nodes with global and semantic properties. These attention mechanisms can be seamlessly integrated into various Transformer-based traffic prediction frameworks. **(3)** Extensive experiments in different traffic prediction scenarios have demonstrated the effectiveness of STPformer. Furthermore, visualized analyses

have also confirmed our method's ability to identify semantic pivotal nodes and handle traffic trends that changes frequently.

## 2 Related Work

Early studies treated traffic forecasting as a time-series analysis problem, initially applying conventional time series and machine learning models [1, 28], and later progressing to the use of RNNs and CNNs [21, 15]. As deep learning advances, spatial-temporal graph models and Transformer-based approaches have achieved considerable improvements. Next, we will introduce key milestones in traffic prediction from these two perspectives.

**Spatial-temporal graph models** utilize GCNs to extract the spatial dependencies of traffic data and perform traffic prediction task in conjunction with sequential forecasting interfaces. For example, DCRNN [13] treats the spatial dependencies between locations as a diffusion process on a directed graph and captures the corresponding spatial patterns by random walk. STGCN [25] simulates the spatio-temporal patterns of traffic data by combining the graph convolution module and the temporal gating unit. However, conventional spatial-temporal graph models often depend on static spatial information and fail to handle dynamic traffic patterns. To overcome this limitation, GraphWavent, AGCRN and LSGCN [24, 2, 6] propose graph representation learning techniques to adaptively learn dynamic spatial patterns, releasing traffic prediction models from the constraints of predefined graph structures. In addition, some studies have delved into the learning of more complicated spatio-temporal patterns. STSGCN and STFGNN [18, 12] confirm the importance of spatio-temporal patterns in terms of synchronisation, localisation and globalisation. Z-Gcnets [3] enhances the accuracy and robustness of predictions by exploring spatial topological details of traffic data. More recently, STPGNN [9] offers a new insight for capturing spatio-temporal patterns of critical nodes, which firstly identifies those spatial nodes with substantial traffic distribution capabilities and then focuses on exploring more complicated spatial dependencies.

**Attention and Transformer-based methods** have achieved remarkable breakthroughs in traffic prediction. For example, GMAN and EASTAN [26, 19] propose to characterise traffic data through a fused embedding approach and extract spatio-temporal patterns using spatial-temporal attention, thereby enhancing the representation of complex spatio-temporal patterns. Notably, Trafformer [8] incorporates spatio-temporal graph information into an encoder-decoder framework and employs a sparse mechanism to improve computational efficiency. STAEformer [14] introduces an innovative technique for embedding dynamic spatio-temporal patterns, while PDFormer [7] enhances the spatial attention module to identify long-term spatial patterns using graph masking strategies.

Despite the impressive results of the aforementioned studies, it's regrettable that none of them can comprehensively accommodate heterogeneous spatio-temporal patterns from various perspectives. For instance, STPGNN ignores the pivotal properties of traffic data in the temporal dimension and those piv-

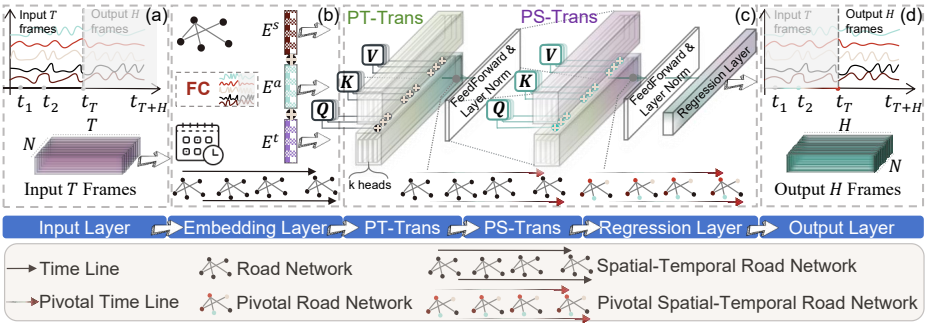


Fig. 1: The overall framework of STPformer.

otal spatial nodes unrelated to traffic distribution capabilities. Furthermore, the inherent structure of existing Transformer-based models poses challenges in exploring complex patterns, such as synchronous spatio-temporal patterns and pivotal spatial dependencies. Motivated by these studies, our work addresses these limitations by **1)** concentrating on the complex patterns surrounding pivotal nodes in the spatio-temporal domain and **2)** enhancing the Transformer-based traffic forecasting model with more flexible and efficient attention mechanisms.

## 3 Preliminary

Traffic forecasting aims at predicting traffic state over a future period of time from historical traffic series (e.g., flow or speed). Formally, given a historical time series $\mathbf{X}_{t-T+1:t}$ with $T$ frames and a future time series $\mathbf{X}_{t+1:t+H}$ with $H$ frames, our goal is to learn a mapping function $f(\cdot)$ that satisfies:

$$\mathbf{X}_{t+1:t+H} = f(\mathbf{X}_{t-T+1:t}|\Theta), \tag{1}$$

where $\mathbf{X}_t \in \mathbb{R}^{N\times d}$, $N$ is the number of spatial nodes, $d$ is the dimension of the traffic feature to be predicted, such as speed or flow. $\Theta$ is the model parameter.

## 4 The Model

### 4.1 Overall Architecture

As depicted in Fig. 1, the input frames are first processed by the spatial-temporal fusion embedding layer to yield the spatial embedding $\mathbf{E}^s$, the temporal embedding $\mathbf{E}^t$, and the hidden spatio-temporal representation $\mathbf{E}^a$. Subsequently, these three embeddings are concatenated and fed into our meticulously crafted Pivotal Temporal Transformer (PT-Trans) and Pivotal Spatial Transformer (PS-Trans) layers. Notably, PT-Trans and PS-Trans are constructed by stacking the proposed pivotal temporal and spatial attention modules, with feedforward networks and layer normalization applied at the transition of each layer. Ultimately, the prediction results are derived from the regression layer.

### 4.2 Heterogeneous Spatio-Temporal Pattern Embedding Layer

Spatial-temporal fusion embedding is a higher-order representation of spatio-temporal data that absorbs the heterogeneity of traffic patterns. More recently, numerous efforts have confirmed the effectiveness of the spatial-temporal embedding of Transformer-based traffic forecasting models, which has also derived corresponding research branches [26, 19, 17, 14]. Thus, we follow the Transformer-based framework [14], and obtain the spatial-temporal embedding $\mathbf{Z}$ by $\mathbf{Z} = \mathbf{E}^s \oplus \mathbf{E}^t \oplus \mathbf{E}^a$, where $\oplus$ is the concatenation operation, $\mathbf{E}^s \in \mathbb{R}^{T\times N\times d_s}$ is the learnable spatial embedding to reflect dynamic spatial heterogeneity, $\mathbf{E}^t \in \mathbb{R}^{T\times N\times d_t}$ is the learnable temporal embedding to reflect periodic temporal heterogeneity, and $\mathbf{E}^a \in \mathbb{R}^{T\times N\times d_a}$ is the input frame processed by the fully-connected layer. The aforementioned dimensions ($d_s$, $d_a$ and $d_t$) are set manually and subsequently integrated into the model dimension.

Next, given the spatial-temporal embedding $\mathbf{Z} \in \mathbb{R}^{T\times N\times d}$ with $T$ frames, $N$ nodes and $d$ feature dimensions, the model will capture temporal and spatial patterns through the application of standard attention mechanisms as:

$$\mathbf{A}^{(T)} = Attn(\mathbf{Q}^{(T)}, \mathbf{K}^{(T)}) = softmax(\mathbf{Q}^{(T)}{\mathbf{K}^{(T)}}^{\mathrm{T}}/\sqrt{D_k}), \tag{2}$$

$$\mathbf{A}^{(S)} = Attn(\mathbf{Q}^{(S)}, \mathbf{K}^{(S)}) = softmax(\mathbf{Q}^{(S)}{\mathbf{K}^{(S)}}^{\mathrm{T}}/\sqrt{D_k}), \tag{3}$$

where $\mathbf{A}^{(T)} \in \mathbb{R}^{N\times T\times T}$ and $\mathbf{A}^{(S)} \in \mathbb{R}^{T\times N\times N}$ are used to jointly capture temporal and spatial correlations of traffic signals. $\{\mathbf{Q}^{(T)}, \mathbf{K}^{(T)}, \mathbf{V}^{(T)}\} = \mathbf{Z}\{\mathbf{W}_Q^T, \mathbf{W}_K^T, \mathbf{W}_V^T\}$ and $\{\mathbf{Q}^{(S)}, \mathbf{K}^{(S)}, \mathbf{V}^{(S)}\} = \mathbf{Z}'\{\mathbf{W}_Q^S, \mathbf{W}_K^S, \mathbf{W}_V^S\}$ are query, key and value matrices in temporal and spatial dimensions, respectively. $\mathbf{Z}' \in \mathbb{R}^{T\times N\times d}$ is the high-order representation of features obtained from temporal attention. $\mathbf{W}_Q^T, \mathbf{W}_K^T, \mathbf{W}_V^T, \mathbf{W}_Q^S, \mathbf{W}_K^S, \mathbf{W}_V^S \in \mathbb{R}^{d\times d}$ are learnable parameters, $D_k$ is the standardization factor.

### 4.3 Mutation-Aware Pivotal Temporal Attention Module

Advanced traffic prediction methods have made significant strides in capturing global, periodic, and dynamic temporal patterns [26, 18, 14]. However, existing studies often overlook the trend signals of traffic data in fluctuating states, such

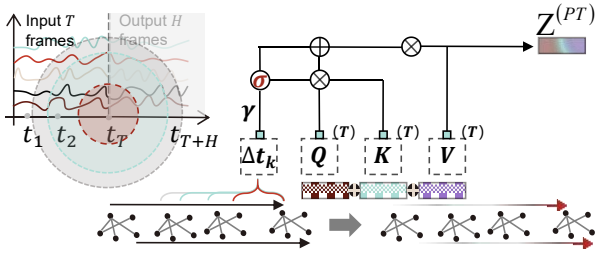


Fig. 2: A sketch of Pivotal Temporal Attention (PTA) module.

as the reversal of flow direction in adjacent time intervals, which typically convey complex patterns and critical information. Considering that the continuity of traffic status changes allows future sequences to inherit underlying temporal patterns from historical data, we are motivated to place greater emphasis on the tail part of historical sequences, particularly the time points with abrupt flow changes, which we define as pivotal temporal attributes. With this perspective, we propose a novel pivotal temporal attention (**PTA**) module to learn the transition process from historical to future sequences, as shown in Fig. 2.

Specifically, we incorporate the Hawkes process [11] into the computational paradigm of temporal attention and design a mutation-aware decay coefficient $\gamma$ to enhance the capability of handling traffic fluctuations. In other words, we aim to assign greater temporal attention scores to the tail of sequences which occurs flow mutation. Formally, our PTA module can be represented as:

$$\mathbf{A}^{(PT)} = softmax(\frac{\mathbf{Q}^{(T)}{\mathbf{K}^{(T)}}^{\mathrm{T}} \times (1 + \sigma exp(-\gamma \Delta t_k))}{\sqrt{d_k}}), \tag{4}$$

$$\gamma = \begin{cases} \min\{0.5, 1 - \frac{|\Delta f_k| + |\Delta f_{k-1}|}{\sum_{k=1}^{k=T} |\Delta f_k|}\}, & \Delta f_k \times \Delta f_{k-1} < 0 \\ 1, & others \end{cases}, \tag{5}$$

where $\mathbf{A}^{(PT)} \in \mathbb{R}^{N \times T \times T}$ denotes the pivotal temporal coefficient matrix, $\sigma$ is the excitation coefficient with batch normalization, $\Delta t_k \in [0, T-1]$ represents the temporal gap between the historical time step and the starting position of the predicted time step, $\gamma$ is a decay rate associated with flow mutation, and $\Delta f_k = f_k - f_{k-1}$ refers to the fluctuation in traffic flow at time $k$. Note that we set $\Delta f_0 = 0$, and the condition $\Delta f_k \times \Delta f_{k-1} < 0$ in Eq. 5 indicates that there is a reversal in the flow direction during this time interval. Since a smaller value of $\gamma$ is assigned to such situation according to the magnitude of flow change, our model attaches more attention to flow mutation, with the attention score being higher at a more recent time $k$. Upon $\mathbf{A}^{(PT)}$, we then derive the output of the pivotal temporal attention module $\mathbf{Z}^{(PT)} \in \mathbb{R}^{T \times N \times d}$ as follows:

$$\mathbf{Z}^{(PT)} = \mathbf{A}^{(PT)} \mathbf{V}^{(T)}. \tag{6}$$

### 4.4 Pivotal Spatial Attention Module

Recent research [9] has confirmed the critical role of pivotal nodes due to their outstanding ability to distribute traffic flow. However, how to effectively identify these pivotal nodes from a comprehensive perspective and capture spatio-temporal patterns in complex traffic environments warrant further exploration. **Firstly**, the current approach for identifying pivotal nodes often overlooks those nodes with less prominent traffic distribution capabilities but substantial semantic spatial attributes. In practical applications, these nodes are equally noteworthy due to their critical locations and intricate spatio-temporal patterns. **Secondly**, the existing *identification-then-computation* scheme lacks a holistic

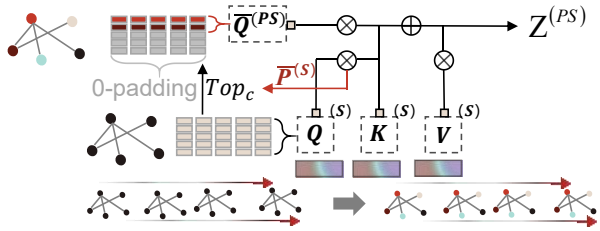


Fig. 3: A sketch of Pivotal Spatial Attention (PSA) module.

perspective, as it processes spatial patterns for pivotal and non-pivotal nodes separately, thereby diminishing the spatial dependencies between different types of nodes. Furthermore, the synchronisation of spatio-temporal patterns is not guaranteed due to the internal structure of traditional spatial-temporal graph models. Consequently, we propose a pivotal spatial attention (**PSA**) module, which adopts a novel *computation-then-identification* paradigm to adaptively identify pivotal nodes. This process is data-driven, and we compute spatial dependencies for all nodes to preserve semantic spatial properties. More importantly, the PSA module can be flexibly integrated into various Transformer-based architectures to ensure the synchronisation of spatio-temporal patterns.

As shown in Fig. 3, we initially calculate the spatial associations among sensor nodes utilizing query-key pairs and identify the Top-$c$ results. Following this, we construct the pivotal query matrix $\overline{\mathbf{Q}}^{(PS)}$, where each row represents the spatial relationships between a pivotal node and the remaining nodes. Finally, $\overline{\mathbf{Q}}^{(PS)}$ is integrated into the calculation process of spatial attention. Note that attention-based computational paradigm involves repetitive dot product operations, and computing the spatial associations of all query-key pairs incurs extra costs, which will inevitably affect model efficiency. Inspired by the probabilistic sparsification (ProbSparse) mechanism [27] in simplifying the operations in long sequence problems, we elaborately introduce the ProbSparse mechanism into the construction process of $\overline{\mathbf{Q}}^{(PS)}$ to improve efficiency. Formally, the proposed PSA module can be represented as:

$$\mathbf{A}^{(PS)} = \mathbf{A}^{(S)} + softmax(\overline{\mathbf{Q}}^{(PS)}{\mathbf{K}^{(S)}}^{\mathrm{T}}/\sqrt{d_k}), \tag{7}$$

$$\overline{\mathbf{Q}}^{(PS)} = Top_c(c, \mathbf{Q}^{(S)}|\overline{P}^{(S)}), \tag{8}$$

$$\overline{P}^{(S)} = M(\mathbf{q}_i^{(s)}, \mathbf{K}^{(S)}) = max_j\{\frac{\mathbf{q}_i^{(s)}{\mathbf{k}_j^{(s)}}^{\mathrm{T}}}{\sqrt{d_k}}\} - \frac{1}{N}\sum_{j=1}^{N}\frac{\mathbf{q}_i^{(s)}{\mathbf{k}_j^{(s)}}^{\mathrm{T}}}{\sqrt{d_k}}, \tag{9}$$

where $\mathbf{A}^{(PS)} \in \mathbb{R}^{T\times N\times N}$ is the pivotal spatial attention matrix. Form Eq. 7, we can observe that $\mathbf{A}^{(PS)}$ superimposes the influence of semantic pivotal nodes on the basis of original attention matrix $\mathbf{A}^{(S)}$. In Eq. 8, $c$ is the sampling factor, which determines the number of pivotal nodes as $cln(N)$, and the pivotal spatial

**Algorithm 1** Calculation process of STPformer

**Require**: Input frame $\mathbf{X}_{t-T+1:t} \in \mathbb{R}^{T\times N\times d}$, sampling factor $c$, dimensions of adaptive embeddings $d_s$, $d_t$ and $d_a$, model hyperparameters.

Set parametric embeddings by $\mathbf{E}^s \in \mathbb{R}^{T\times N\times d_s}$, $\mathbf{E}^t \in \mathbb{R}^{T\times N\times d_t}$, $\mathbf{E}^a \in \mathbb{R}^{T\times N\times d_a}$;
Get spatial-temporal fusion embedding by $\mathbf{Z} = \mathbf{E}^s \oplus \mathbf{E}^t \oplus \mathbf{E}^a$;
Obtain $\mathbf{Q}^{(T)}, \mathbf{K}^{(T)}, \mathbf{V}^{(T)}$ by the linear transformation of $\mathbf{Z}$;
Compute pivotal temporal attention matrix $\mathbf{A}^{(PT)}$ by Eq. 4;
Get the output of the PTA module $\mathbf{Z}^{(PT)}$ by Eq. 6;
Obtain $\mathbf{Q}^{(S)}, \mathbf{K}^{(S)}, \mathbf{V}^{(S)}$ by the linear transformation of $\mathbf{Z}^{(PT)}$;
Get the pivotal spatial query matrix $\overline{\mathbf{Q}}^{(PS)}$ by Eq. 8, which uses Eq. 9 for the efficient calculation of query-key pairs;
Compute pivotal spatial attention matrix $\mathbf{A}^{(PS)}$ by Eq. 7;
Get the output of the PSA module $\mathbf{Z}^{(PS)}$ by Eq. 10;
**Return** prediction results through the regression layer by $\hat{\mathbf{Y}} = FCs(\mathbf{Z}^{(PS)})$.

query matrix $\overline{\mathbf{Q}}^{(PS)}$ is constructed from the original spatial query matrix $\mathbf{Q}^{(S)}$ using the ProbSparse method $\overline{P}^{(S)}$. Therefore, the complexity of the pivotal spatial attention is reduced from $O(N^2)$ to $O(Nln(N))$. Then, the output of the PSA module, i.e. $\mathbf{Z}^{(PS)} \in \mathbb{R}^{T\times N\times d}$, can be obtained as:

$$\mathbf{Z}^{(PS)} = \mathbf{A}^{(PS)}\mathbf{V}^{(S)}. \tag{10}$$

Finally, we utilize a fully connected layer $FCs(\cdot)$ as the regression layer to transform the spatio-temporal representation $\mathbf{Z}^{(PS)}$ into prediction results, denoted by:

$$\hat{\mathbf{Y}} = FCs(\mathbf{Z}^{(PS)}), \tag{11}$$

where $\hat{\mathbf{Y}} \in \mathbb{R}^{H\times N\times d_o}$, $H$ is the prediction horizon and $d_o = 1$ indicates the dimension of output features. We give the calculation pseudo-code of the proposed STPformer in Algorithm 1.

### 4.5 Loss Function

We choose the mean absolute error (MAE) as the loss function. The objective function can be represented as:

$$\mathcal{L}(\hat{\mathbf{Y}}; \Theta) = \frac{1}{NH}\sum_{i=1}^{i=H}\sum_{j=1}^{j=N}|\hat{\mathbf{Y}}_{ij} - \mathbf{Y}_{ij}|, \tag{12}$$

where $\mathbf{Y}$ is the ground truth, and $\Theta$ denotes all learnable model parameters.

## 5 Experiments

### 5.1 Experimental Setup

**Datasets.** Detailed statistics of datasets are summarized in Tab. 1. There are two types of public traffic benchmark datasets are selected, including 2 speed

Table 1: Dataset analysis and description.

| Datasets | #Nodes | #Edges | #Time Steps (5 mins) | #Time Range | #Signals |
|---|---|---|---|---|---|
| METR-LA | 207 | 1515 | 34272 | 03/01/2012-06/30/2012 | speed |
| PEMS-BAY | 325 | 2369 | 52116 | 01/01/2017-05/31/2017 | speed |
| PEMS04 | 307 | 340 | 16992 | 01/01/2018-02/28/2018 | flow |
| PEMS07 | 883 | 866 | 28224 | 05/01/2017-08/31/2017 | flow |
| PEMS08 | 170 | 295 | 17856 | 07/01/2016-08/31/2016 | flow |

datasets [13] and 4 flow datasets [18]. In the data processing stage, we perform Z-score normalization on the raw inputs.

**Baselines.** We classify diverse benchmarks into 5 categories. **1)** Statistical-based methods: VAR [28]; **2)** machine learning-based methods: LSTM [21]; **3)** spatial-temporal graph based methods: STGCN [25], DCRNN [13], GWNet [24], MTGNN [23], AGCRN [2], and DSTAGNN [10] and STPGNN [9]; **4)** attention and Transformer-based models: ASTGCN [5], GMAN [26], PDFormer [7] and Trafformer [8]; **5)** other competitive methods: ST-Norm [4], GTS [16], STID [17] and ModWaveMLP [20].

**Settings.** Referring to existing studies [7, 9], the flow and speed datasets are divided into training, validation, and test sets with ratios of 6:2:2 and 7:1:2, respectively. We set the input and output horizons to 12 (60 minutes), which are common settings in real-time traffic prediction tasks. For STPformer, we adopt the Adam optimizer and select the optimal learning rate from {0.01, 0.001, 0.0001} using a grid search strategy. We set the maximum training epoch to 300 and use an early stopping strategy to prevent overfitting, and the batch size is set to 16. Referring to [14], we set the dimension of the spatial-temporal fusion embedding to 152. All attention modules use multi-head mechanism with 4 heads. The dimension of the hidden layer in the feed-forward neural network is set to 256. The sampling factor is chosen from {5, 10, 15, 20, 25}. In addition, we run the model independently at least 5 times on all datasets separately on one single Tesla-V100 32GB GPU and take the average values as final results[1].

### 5.2 Results and Analysis

Tabs. 2 and 3 summarize the forecasting evaluation results of all the baselines on the speed and flow datasets. The best results are highlighted in boldface, and the second-best results are underlined.

**Traffic Speed Forecasting.** From Tab. 2, we can observe that: **(1)** Our STPformer significantly improves the inference performance across all datasets, demonstrating its success in enhancing prediction capacity in traffic forecasting problems. **(2)** The proposed method outperforms the most relevant works, i.e., PDformer and Trafformer. We observe that the performance improvement of

[1] Codes are available at: `https://github.com/dasfaa2025authors/STPformer`

Table 2: Performance comparison of different baselines for traffic speed forecasting on METR-LA and PEMS-BAY. Results marked with * indicate that the improvement is **statistically significant** compared with the best baselines.

| Datasets | Metrics | VAR | LSTM | STGCN | DCRNN | ASTGCN | GWNet | MTGNN | GMAN | AGCRN | ST-Norm | GTS | STID | PDFormer | Trafformer | ModWaveMLP | **STPformer** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| METR-LA | MAE | 4.42 | 3.44 | 2.75 | 2.67 | 4.86 | 2.70 | 2.69 | 2.80 | 2.85 | 2.81 | 2.66 | 2.82 | 2.83 | 2.78 | 2.68 | **2.65** |
| 15 mins | RMSE | 7.80 | 6.30 | 5.29 | 5.16 | 9.27 | 5.19 | 5.18 | 5.55 | 5.53 | 5.57 | 5.22 | 5.53 | 5.45 | 5.35 | 5.27 | **5.08*** |
| | MAPE | 13.00% | 9.60% | 7.10% | 6.86% | 9.21% | 6.97% | 6.88% | 7.41% | 7.63% | 7.40% | 7.03% | 7.75% | 7.77% | 7.32% | 7.06% | **6.79%*** |
| | MAE | 5.41 | 3.77 | 3.15 | 3.12 | 5.43 | 3.12 | 3.05 | 3.12 | 3.20 | 3.18 | 3.04 | 3.19 | 3.20 | 3.05 | 2.99 | **2.97** |
| 30 mins | RMSE | 9.13 | 7.23 | 6.35 | 6.27 | 10.61 | 6.32 | 6.17 | 6.49 | 6.52 | 6.59 | 6.25 | 6.57 | 6.46 | 6.18 | 6.22 | **6.00*** |
| | MAPE | 12.70% | 10.09% | 8.62% | 8.42% | 10.13% | 8.47% | 8.19% | 8.73% | 9.02% | 8.47% | 8.48% | 9.39% | 9.19% | 8.67% | 8.41% | **8.07%*** |
| | MAE | 6.52 | 4.37 | 3.60 | 3.54 | 6.51 | 3.55 | 3.49 | 3.44 | 3.59 | 3.57 | 3.44 | 3.55 | 3.62 | 3.41 | 3.38 | **3.33** |
| 60 mins | RMSE | 10.11 | 8.69 | 7.43 | 7.47 | 12.52 | 7.39 | 7.23 | 7.35 | 7.45 | 7.51 | 7.28 | 7.55 | 7.47 | 7.17 | 7.13 | **6.99*** |
| | MAPE | 15.80% | 14.00% | 10.35% | 10.32% | 11.64% | 10.02% | 9.87% | 10.07% | 10.47% | 10.24% | 10.01% | 10.95% | 10.91% | 9.96% | 9.82% | **9.52%*** |
| PEMS-BAY | MAE | 1.74 | 2.05 | 1.35 | 1.31 | 1.52 | 1.33 | 1.32 | 1.34 | 1.37 | 1.34 | 1.34 | 1.30 | 1.32 | 1.31 | 1.30 | **1.29** |
| 15 mins | RMSE | 3.16 | 4.19 | 2.88 | 2.80 | 3.13 | 2.82 | 2.79 | 2.92 | 2.93 | 2.88 | 2.83 | 2.81 | 2.83 | 2.83 | 2.79 | **2.75** |
| | MAPE | 3.60% | 4.80% | 2.88% | 2.73% | 3.22% | 2.77% | 2.77% | 2.88% | 2.95% | 2.82% | 2.83% | 2.73% | 2.78% | 2.92% | 2.82% | **2.71%*** |
| | MAE | 2.32 | 2.20 | 1.69 | 1.67 | 2.01 | 1.64 | 1.65 | 1.65 | 1.70 | 1.67 | 1.66 | 1.62 | 1.64 | 1.61 | 1.59 | **1.59** |
| 30 mins | RMSE | 4.25 | 4.55 | 3.83 | 3.81 | 4.27 | 3.70 | 3.74 | 3.81 | 3.89 | 3.83 | 3.78 | 3.72 | 3.79 | 3.74 | 3.72 | **3.63*** |
| | MAPE | 5.00% | 5.20% | 3.85 | 3.75% | 4.48% | 3.70% | 3.69% | 3.71% | 3.88% | 3.75% | 3.79% | 3.68% | 3.71% | 3.82 | 3.68% | **3.57%*** |
| | MAE | 2.93 | 2.37 | 2.01% | 1.99 | 2.61 | 2.01 | 1.94 | 1.89 | 1.99 | 1.96 | 1.95 | 1.89 | 1.91 | 1.88 | 1.89 | **1.86** |
| 60 mins | RMSE | 5.44 | 4.96 | 4.56 | 4.66 | 5.42 | 4.56 | 4.49 | 4.38 | 4.64 | 4.52 | 4.46 | 4.40 | 4.43 | 4.38 | 4.38 | **4.30*** |
| | MAPE | 6.50% | 5.70% | 4.74% | 4.73% | 6.00% | 4.67% | 4.53% | 4.51% | 4.72% | 4.62% | 4.62% | 4.47% | 4.51% | 4.59% | 4.47% | **4.34%*** |

Table 3: Performance comparison for traffic flow forecasting.

| Datasets | PEMS04 | | | PEMS07 | | | PEMS08 | | |
|---|---|---|---|---|---|---|---|---|---|
| Metrics | MAE | RMSE | MAPE | MAE | RMSE | MAPE | MAE | RMSE | MAPE |
| VAR | 29.03 | 44.13 | 19.96% | 33.27 | 49.89 | 16.82% | 25.40 | 40.37 | 18.85% |
| LSTM | 27.09 | 40.97 | 18.07% | 30.51 | 44.96 | 23.62% | 22.18 | 33.05 | 15.58% |
| STGCN | 25.34 | 38.76 | 18.82% | 29.26 | 44.96 | 14.04% | 21.69 | 32.47 | 16.17% |
| DCRNN | 21.30 | 33.58 | 14.05% | 22.70 | 35.46 | 9.90% | 16.64 | 26.14 | 10.70% |
| ASTGCN | 22.48 | 34.93 | 16.04% | 27.32 | 41.28 | 12.74% | 18.97 | 28.71 | 12.32% |
| GWNet | 18.97 | 30.32 | 14.26% | 20.25 | 33.32 | 8.63% | 14.67 | 23.49 | 9.52% |
| MTGNN | 19.50 | 32.00 | 14.04% | 20.94 | 34.03 | 9.10% | 15.31 | 24.42 | 10.70% |
| GMAN | 19.97 | 31.69 | 13.62% | 22.37 | 35.44 | 11.31% | 15.18 | 24.45 | 9.85% |
| AGCRN | 19.76 | 31.82 | 12.92% | 21.20 | 35.01 | 8.89% | 15.95 | 25.22 | 10.09% |
| ST-Norm | 18.95 | 30.96 | 12.67% | 20.52 | 34.66 | 8.78% | 15.48 | 24.82 | 9.77% |
| GTS | 20.85 | 32.81 | 14.34% | 22.15 | 35.13 | 9.40% | 16.33 | 26.02 | 10.01% |
| DSTAGNN | 19.30 | 31.46 | 12.70% | 21.42 | 34.51 | 9.01% | 15.67 | 24.77 | 9.94% |
| STID | 18.29 | 29.95 | 12.49% | 19.54 | 32.85 | 8.25% | 14.20 | 23.49 | 9.28% |
| PDFormer | 18.32 | 29.97 | 12.10% | 19.83 | 32.87 | 8.53% | 13.58 | 23.51 | 9.05% |
| STPGNN | 18.34 | **29.64** | 12.49% | 20.52 | 33.38 | 8.75% | 13.90 | 23.05 | 9.01% |
| **STPformer** | **18.15** | 29.94 | **11.95%** | **19.14** | **32.51** | **8.02%** | **13.37** | **23.03** | **8.79%** |

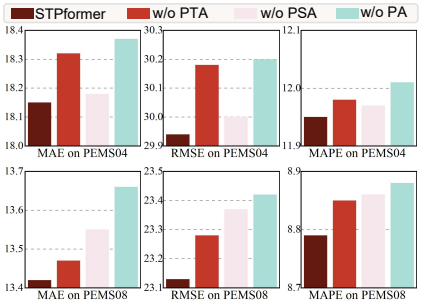


Fig. 4: Ablation study on PEMS04 and PEMS08 datasets.

Transformer-based models is progressively weakening, which can be attributed to their inability to capture more fine-grained traffic patterns, such as pivotal spatio-temporal attributes. **(3)** The proposed method outperforms the state-of-the-art benchmark ModWaveMLPs[2] on MAPE by decreasing 3.64% (on METR-LA), and 3.27% (on PEMS-BAY) in average. **(4)** The proposed STPformer presents significantly better results than spatial-temporal graph models in both short-term (15 minutes) and long-term (60 minutes) predictions. This reveals that the spatial-temporal fusion embedding and the attention mechanism acquires better prediction capacity than the GCN-based models[3].

**Traffic Flow Forecasting.** Within this setting, STPGNN and STID are competitive state-of-the-art benchmarks. It is important to note that not all models perform well in both speed and flow prediction. For example, Trafformer and ModWaveMLP are reported to be optimal for speed prediction. In contrast, our

[2] Updated results are available at `https://github.com/Kqingzheng/ModWaveMLP`

[3] It should be noted that the results reported in the STPGNN paper are not available for the speed datasets in our task settings.

STPformer easily adapts from speed prediction to flow prediction while retaining its superior performance. From Tab. 3, we observe that: **(1)** Our STPformer greatly outperforms other models and the findings (1) & (2) in the speed forecasting also hold for the flow forecasting. **(2)** The STPformer model shows better results than STPGNN and STID. Specifically, the MAE decreases by 2.05% on PEMS07 and 1.55% on PEMS08, benefiting from our additional considerations for pivotal attributes in both temporal and spatial dimensions. On the PEMS04 dataset, STPGNN performs better on RMSE, while our method surpasses on other metrics. We attribute this to a specific case where the effectiveness of forecasting capacity is influenced by the problem's scalability.

### 5.3 Ablation Study

To evaluate the effectiveness of the proposed pivotal spatial-temporal components, we design three different variants:

- **w/o PTA**: Replacing pivotal temporal attention module with standard temporal attention module.
- **w/o PSA**: Replacing pivotal spatial attention module with standard spatial attention module.
- **w/o PA**: Replacing all pivotal attention components with standard attention components.

The experimental results for the different variants on PEMS04 and PEMS08 are presented in Fig. 4. It can be observed that: **(1)** STPformer significantly outperforms all other variants, and the performance degradation of the **w/o PA** model is particularly noticeable, thereby demonstrating the effectiveness of our proposed approaches. **(2)** It is also important to note that the **w/o PA** variant of our STPformer essentially degenerates into the backbone model STAEformer [14]. More importantly, despite the performance degradation of both the **w/o PTA** and **w/o PSA** models, they still outperform all other benchmarks, further demonstrating the superiority of our methods.

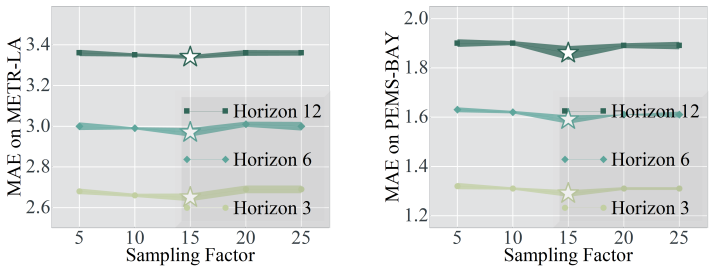


Fig. 5: Parameter sensitivity on METR-LA and PEMS-BAY datasets. Best results are marked with stars. Horizons 3, 6, and 12 correspond to forecast steps of 15 minutes, 30 minutes, and 60 minutes, respectively.

Table 4: Efficiency comparison between STPformer and Transformer-based traffic forecasting backbone (w/o PSA).

| Training Time (s/epoch) | PEMS04 | PEMS08 |
|---|---|---|
| w/o PSA (3-layer spatial attention) | 115.25 | 69.25 |
| STPformer (2-layer spatial attention and 1-layer PSA) | **109.62** | **67.58** |
| Parameter Numbers (E+06) | PEMS04 | PEMS08 |
| w/o PSA (3-layer spatial attention) | 1.35 | 1.22 |
| STPformer (2-layer spatial attention and 1-layer PSA) | **1.33** | **1.20** |

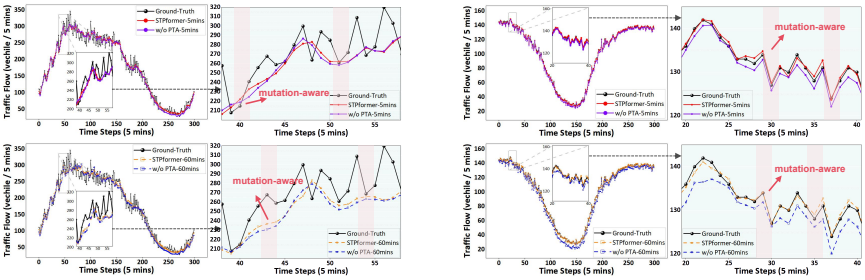


Fig. 6: Traffic forecasting comparison between **w/o PTA** and STPformer on Node #307 (PEMS04) and Node #150 (PEMS08). Please zoom in the plot area for a better view.

### 5.4 Model Analysis

**Sensitivity to *c*.** We analyze the effect of the sampling factor $c$ on the STPformer using both the METR-LA and PEMS-BAY datasets. Considering the variability of road environments, we gradually increase $c$ from a low value ($c = 5$) to a high value ($c = 25$) to cover as many global semantic pivotal nodes as possible. As shown in Fig. 5, we observe that the prediction error initially decreases, then increases, and finally stabilizes. This is because a small sampling factor fails to cover all pivotal nodes, while a large sampling factor blurs the boundary between pivotal and non-pivotal nodes, effectively degrading our PSA module into an ordinary spatial attention module. Additionally, our approach is flexible enough to adapt to datasets of different scales. The higher the number of nodes in the dataset, the more pivotal nodes are selected.

**PSA module can be seamlessly integrated into an attention-based model.** Tab. 4 compares the model efficiency of STPformer and Transformer-based backbone (w/o PSA), clearly showing the advantages of STPformer in terms of training time and parameter size. Note that the Transformer-based backbone requires 3 layers of attention to achieve the desired performance, whereas our STPformer uses only a 2-layer spatial attention module, thereby outperforming the traditional spatial attention mechanism in terms of efficiency.

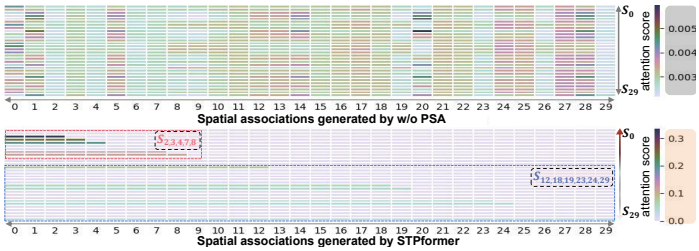


Fig. 7: Heatmaps of spatial associations between sensors from the PEMS-BAY dataset. The top 5/top 6-10 results are marked with red/ blue boxes, respectively.

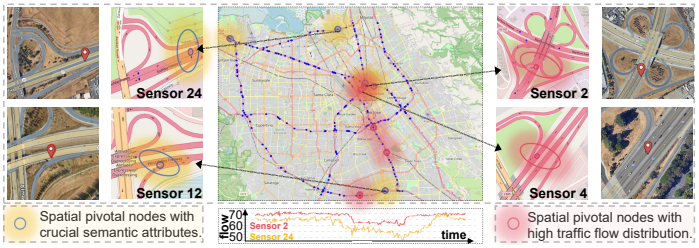


Fig. 8: Pivotal nodes discovery on the PEMS-BAY dataset. Pivotal nodes identified by STPformer with red circles for top 5 and blue circles for the top 6-10.

### 5.5 Case Study: How well STPformer works?

**Capability of Pivotal Temporal Attention in Identifying Flow Mutations.** Fig. 6 shows the prediction curves of STPformer and **w/o PTA**, as well as the ground truth. We evaluate the different models in both short-term (5 minutes) and long-term (60 minutes) predictions. We observe that: **(1)** STPformer can identify the moment of flow mutation better than **w/o PTA** and maintains an overall advantage in precision, which proves the effectiveness of our PTA module; and **(2)** as the prediction length increases, STPformer demonstrates a significant advantage in both fitting ability and mutation detection.

**Capability of Pivotal Spatial Attention in Identifying Pivotal Nodes.** As shown in Fig. 7, we first randomly selected 30 sensors from the PEMS-BAY dataset to assess the effectiveness of the PSA module. Then, we created heat maps to visualize the spatio-temporal correlations captured by **w/o PSA** and STPformer. We can observe that: **(1)** the proposed PSA module demonstrates an excellent ability to capture clear and distinguishable traffic patterns. **(2)** Compared to conventional spatial attention, our PSA module supports direct backtracking to obvious potential associations at the node level, thereby enhancing its interpretability. Furthermore, the pivotal nodes identified by STPformer are illustrated in Fig. 8, accompanied by detailed traffic conditions for in-depth analysis. Our STPformer excels at pinpointing nodes with significant

traffic concentration, such as Sensor 2 at a roundabout location and Sensor 4 at a highway intersection. More importantly, STPformer is able to identify the semantic pivotal nodes, which are marked in blue. For example, Sensors 24 and 12 are located upstream of multiple roads and possess more important spatial attributes. The above findings further confirm the effectiveness of our methods.

## 6 Conclusion and Future Work

We propose a novel transformer-based traffic forecasting framework, STPformer, which efficiently identifies pivotal attributes in both temporal and spatial dimensions. To the best of our knowledge, we are the first to address the traffic forecasting problem by considering both pivotal temporal attributes and semantic pivotal spatial nodes. More importantly, the proposed methods can be flexibly integrated into any attention-based traffic prediction backbone, thereby providing more transparent and interpretable prediction results for the inference process. Experimental results have demonstrated the superiority of our method over state-of-the-art baselines. Extensive analysis has also confirmed the ability of our method to detect changes in traffic trends and to identify global or local, flow-aware or semantically-attributed, pivotal or non-pivotal nodes. For future work, we intend to integrate our methods into more sophisticated traffic prediction backbones and spatio-temporal prediction tasks to broaden the model's applicability and robustness.

**Acknowledgements** This work is supported by the National Natural Science Foundation of China [No. 62276075] and the Shenzhen Science and Technology Program [No. KCXFZ20230731093001002, KJZD20230923115113026].

## References

1. Ahmed, M.S., Cook, A.R.: Analysis of Freeway Traffic Time-series Data by Using Box-Jenkins Techniques. No. 722 (1979)
2. Bai, L., Yao, L., Li, C., Wang, X., Wang, C.: Adaptive graph convolutional recurrent network for traffic forecasting. In: Neurips (2020)
3. Chen, Y., Segovia-Dominguez, I., Gel, Y.R.: Z-gcnets: Time zigzags at graph convolutional networks for time series forecasting. In: ICML. pp. 1684–1694 (2021)
4. Deng, J., Chen, X., Jiang, R., Song, X., Tsang, I.W.: St-norm: Spatial and temporal normalization for multi-variate time series forecasting. In: SIGKDD. pp. 269–278 (2021)
5. Guo, S., Lin, Y., Feng, N., Song, C., Wan, H.: Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In: AAAI. pp. 922–929 (2019)
6. Huang, R., Huang, C., Liu, Y., Dai, G., Kong, W.: Lsgcn: Long short-term traffic prediction with graph convolutional networks. In: IJCAI. pp. 2355–2361 (2020)
7. Jiang, J., Han, C., Zhao, W.X., Wang, J.: Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In: AAAI. pp. 4365–4373 (2023)

8. Jin, D., Shi, J., Wang, R., Li, Y., Huang, Y., Yang, Y.B.: Trafformer: Unify time and space in traffic prediction. In: AAAI. pp. 8114–8122 (2023)
9. Kong, W., Guo, Z., Liu, Y.: Spatio-temporal pivotal graph neural networks for traffic flow forecasting. In: AAAI. pp. 8627–8635 (2024)
10. Lan, S., Ma, Y., Huang, W., Wang, W., Yang, H., Li, P.: Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In: ICML. pp. 11906–11917 (2022)
11. Laub, P.J., Taimre, T., Pollett, P.K.: Hawkes Processes. arXiv e-prints p. arXiv:1507.02822 (2015). https://doi.org/10.48550/arXiv.1507.02822
12. Li, M., Zhu, Z.: Spatial-temporal fusion graph neural networks for traffic flow forecasting. In: AAAI. pp. 4189–4196 (2021)
13. Li, Y., Yu, R., Shahabi, C., Liu, Y.: Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In: ICLR (2018)
14. Liu, H., Dong, Z., Jiang, R., Deng, J., Deng, J., Chen, Q., Song, X.: Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In: CIKM. pp. 4125–4129 (2023)
15. Seo, Y., Defferrard, M., Vandergheynst, P., Bresson, X.: Structured sequence modeling with graph convolutional recurrent networks. In: Neurips. pp. 362–373 (2018)
16. Shang, C., Chen, J., Bi, J.: Discrete graph structure learning for forecasting multiple time series. In: ICLR (2021)
17. Shao, Z., Zhang, Z., Wang, F., Wei, W., Xu, Y.: Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In: CIKM. pp. 4454–4458 (2022)
18. Song, C., Lin, Y., Guo, S., Wan, H.: Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In: AAAI. pp. 914–921 (2020)
19. Su, H., Wang, X., Chen, Q., Qin, Y.: Efficient adaptive spatial-temporal attention network for traffic flow forecasting. In: ECML-PKDD. pp. 205–220 (2023)
20. Sun, K., Liu, P., Li, P., Liao, Z.: Modwavemlp: Mlp-based mode decomposition and wavelet denoising model to defeat complex structures in traffic forecasting. In: AAAI. pp. 9035–9043 (2024)
21. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Neurips. pp. 3104–3112 (2014)
22. Wang, X., Ma, Y., Wang, Y., Jin, W., Wang, X., Tang, J., Jia, C., Yu, J.: Traffic flow prediction via spatial temporal graph neural network. In: WWW. pp. 1082–1092 (2020)
23. Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., Zhang, C.: Connecting the dots: Multivariate time series forecasting with graph neural networks. In: SIGKDD. pp. 753–763 (2020)
24. Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C.: Graph wavenet for deep spatial-temporal graph modeling. In: IJCAI. pp. 1907–1913 (2019)
25. Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In: IJCAI. pp. 3634–3640 (2018)
26. Zheng, C., Fan, X., Wang, C., Qi, J.: Gman: A graph multi-attention network for traffic prediction. In: AAAI. pp. 1234–1241 (2020)
27. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In: AAAI. pp. 11106–11115 (2021)
28. Zivot, E., Wang, J.: Vector autoregressive models for multivariate time series. Modeling Financial Time Series with S-Plus® pp. 385–429 (2006)