
STG-Mamba: Spatial-Temporal Graph Learning via Selective State Space Model

Lincan Li, Hanchen Wang, Wenjie Zhang*, Adelle Coster

University of New South Wales
Sydney, Australia

lincan.li@unsw.edu.au, hanchen.wang@unsw.edu.au,
wenjie.zhang@unsw.edu.au, a.coster@unsw.edu.au

Abstract

Spatial-Temporal Graph (STG) data is characterized as dynamic, heterogenous, and non-stationary, leading to the continuous challenge of spatial-temporal graph learning. In the past few years, various GNN-based methods have been proposed to solely focus on mimicking the relationships among node individuals of the STG network, ignoring the significance of modeling the intrinsic features that exist in STG system over time. In contrast, modern Selective State Space Models (SSSMs) present a new approach which treat STG Network as a system, and meticulously explore the STG system's dynamic state evolution across temporal dimension. In this work, we introduce Spatial-Temporal Graph Mamba (STG-Mamba) as the first exploration of leveraging the powerful selective state space models for STG learning by treating STG Network as a system, and employing the Spatial-Temporal Selective State Space Module (ST-S3M) to precisely focus on the selected STG latent features. Furthermore, to strengthen GNN's ability of modeling STG data under the setting of selective state space models, we propose Kalman Filtering Graph Neural Networks (KFGN) for dynamically integrate and upgrade the STG embeddings from different temporal granularities through a learnable Kalman Filtering statistical theory-based approach. Extensive empirical studies are conducted on three benchmark STG forecasting datasets, demonstrating the performance superiority and computational efficiency of STG-Mamba. It not only surpasses existing state-of-the-art methods in terms of STG forecasting performance, but also effectively alleviate the computational bottleneck of large-scale graph networks in reducing the computational cost of FLOPs and test inference time. The implementation code is available at: <https://github.com/LincanLi98/STG-Mamba>.

1 Introduction

Spatial-temporal graph data is a kind of Non-Euclidean data which widely exists in our daily life, such as urban traffic network, metro system InFlow/OutFlow, social networks, regional energy load, weather observations, etc. Owning to the strong ability of capturing dynamic features across both temporal and spatial dimension, attention-based methods, such as Spatial-Temporal Transformer and its variants have recently been broadly implemented in spatial-temporal graph (STG) learning tasks. In the domain of STG traffic forecasting, seminal works have emerged that leverage these methods to enhance prediction accuracy. ASTGCN Guo et al. [2019] introduces an innovative approach combining spatial-attention & temporal attention mechanisms with graph neural networks. ADCT-Net Kong et al. [2024] presents to incorporate a dual-graphic cross-fusion Transformer for

*Corresponding author.

dynamic traffic forecasting. Huo et al. [2023] integrates hierarchical graph convolutional networks with Transformer networks to capture sophisticated STG relationships among traffic data. For weather forecasting, Liang et al. [2023] exemplifies how Transformer-based models can be utilized for accurate air quality forecasting across vast geographical areas. Chen et al. [2023] further demonstrates the power of GNNs in predicting air quality, emphasizing the importance of group-aware mechanisms in enhancing STG forecasting performance. In the realm of Social Recommendation, Zhang et al. [2023] introduces an Adaptive Disentangled Transformer (ADT) framework that optimizes the disentanglement of attention heads in different layers of Transformer architecture. Wei et al. [2023] explores the improved efficiency of multimedia recommendation by introducing modal-specific embeddings and a lightweight self-attention.

Although Transformer-based methods demonstrate notable improvements in STG relationship learning Chen et al. [2022], Cong et al. [2021], their modeling ability on large-scale STG networks and Long-Term Sequential Forecasting (LTSF) tasks is greatly hindered by the quadratic computational complexity $O(n^2)$ provided by its core component—attention mechanism Zhou et al. [2021, 2022], Rampášek et al. [2022]. Moreover, attention mechanisms usually approximate the full-length historical data, or encode all the context features, which may not be necessary for modeling the long-term dependencies of STG learning tasks. As such, the research world has been actively looking for counterparts of Attention-based models.

Lately, with the popularization of Mamba Gu and Dao [2023], Wang et al. [2024], Liu et al. [2024], Selective State Space Models (SSSM) have aroused considerable interest among researchers from computer vision and Natural Language Processing. SSSM is a variant of the State Space Models, which originated from Control Science and Engineering fields Lee et al. [1994], Friedland [2012]. State Space Models provide a professional framework to describe a physical system's dynamics state evolution through sets of input, output, and state variables related by first-order differential or difference equations. The approach allows for a compact way to model and analyze systems with multiple inputs and outputs (MIMO) Aoki [2013].

STG learning can be treated as a complex process of understanding and forecasting the evolution of STG networks over time, which highly resembles the state space transition process where each state encapsulates the current configuration of a system and transitions represent changes over time. Deep learning-based SSSMs bring new horizons to STG learning tasks. However, great challenges pose in accurately and effectively adopting Mamba architectures for STG modeling, which is characterized as heterogeneous, dynamic, and non-stationary.

Motivated by the excellent long-term contextual modeling ability and low computational overhead of SSSMs, we propose Spatial-Temporal Graph Mamba (STG-Mamba) as the first deep learning-based SSSM for effective data-centric STG learning. The main contributions of this work are summarized in the following:

- We make the first attempt to adapting SSSMs for STG learning tasks. A simple yet elegant way is employed to extend selective state space models to handle STG data. Specifically, we formulate the framework following the Stacked Residual Encoder Fashion, with N stacked Graph Selective State Space Block (GS3B) as basic modules and the proposed Kalman Filtering Graph Neural Network (KFGN) within it, to facilitate efficient Seq2Seq modeling.
- A novel Spatial-Temporal Selective State Space Module (ST-S3M) is proposed for STG network pioneering integration with selective state space models, which performs input-dependent adaptive spatial-temporal graph feature selection. ST-S3M is capable of capturing the sophisticated heterogenous dependencies among STG network, and accurately modeling the dynamic evolution of STG system over time, making it highly fitted for STG learning.
- We introduce KFGN as the specialized adaptive spatiotemporal graph generation and upgrading method, which smoothly fits in within the SSSM-based context. In KFGN, the Dynamic Filter Linear serves as dynamic GNN module which generating input-specific graph structures. The KF-Upgrading mechanism models STG inputs from different temporal granularities as parallel streams, and the output embeddings are integrated through the statistical theory learning of Kalman Filtering for optimization.
- Extensive evaluations are carried on three open-sourced benchmark STG dataset. Results demonstrate that our STG-Mamba not only exceed other benchmark methods in terms of

STG prediction performance, but also achieve $O(n)$ computational complexity, remarkably reducing the computational overhead compared with Transformers-based methods.

2 Preliminaries

Spatial-Temporal Graph (STG). STG is a kind of non-Euclidean structured data, which extensively existed in daily-life, business, and society. For a given graph network, if the node individuals have certain spatial connectivity at any given time step τ , and the node features and/or graph structures are dynamically evolving over temporal axis, it could be defined as an STG network. Formally, we use $\mathbb{G}^{ST} = (V^{ST}, E^{ST}, A^{ST})$ to represent a given STG network. Here, V^{ST} denotes the set of graph vertices (nodes) within the network, E^{ST} denotes the edges of the graph network. A^{ST} is the graph adjacency matrix, which indicates the calculated proximity between node individuals.

Spatial-Temporal Graph System (STG system). We for the first time define a network which consists of STG data as a Spatial-Temporal Graph System (STG system). In the theory foundation, a system is defined as the representation of a physical process, comprising state variables that describe the system’s current condition, input variables influencing the system’s state, and output variables reflecting the system’s response. For STG System, this framework is adapted to accommodate spatial dependencies and temporal evolution, structuring the system with nodes representing spatial entities, edges denoting spatial interactions, and state variables evolving over time to capture the dynamics of STG data.

State Space Models (SSMs). Deep Learning-based SSMs are a lately invented class of sequential models that are closely related to RNN architectures and classical state space models. They are characterized by a particular continuous system model that maps a multi-dimensional input sequence to a corresponding output sequence through an implicit latent state representation. SSMs are defined by four parameters ($\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$) that specify how the input (control signal) and the current state determine the next state and output. This framework allows for efficient sequential modeling by enabling both linear and non-linear computations. As a variant of SSMs, SSSMs emphasize how to build the selection mechanism upon SSMs, which is highly resemble to the core idea of attention mechanism, making it a competitive counterpart of Transformer architecture.

STG Forecasting based on SSSMs. Employing SSSMs for spatial-temporal graph forecasting, the problem can be formulated as dynamically identifying and utilizing relevant portions of historical spatiotemporal data and graph structures to predict the future states of a STG system.

Given spatiotemporal graph $\mathbb{G}^{ST} = (V^{ST}, E^{ST}, A^{ST})$ and historical spatial-temporal sequential data $X_{t-p+1:t} = \{X_{t-p+1}, X_{t-p+2}, \dots, X_t\}$, the objective is to utilize SSSMs to forecast future STG system’s states $\hat{X}_{t+1:t+k} = \{\hat{X}_{t+1}, \dots, \hat{X}_{t+k}\}$. This process is achieved by learning a mapping function $F_{SSSM}(\cdot)$ that dynamically selects relevant state transitions and interactions for prediction:

$$F_{SSSM}(X_{t-p+1:t}; \mathbb{G}^{ST}) = \hat{X}_{t+1:t+k} \quad (1)$$

3 Methodology

In this section, we start with introducing the model architecture of STG-Mamba and the formulation of STG Selective State Space Models. Following that, we elaborate the key contributions of this work, including the Kalman Filtering Graph Neural Networks (Section 3.2) which specifically designed for SSSMs to enable adaptive statistical-based graph learning under contexts with noisy data and uncertainty; and the Spatial-Temporal Selective State Space Module (ST-S3M) which serves as the state space selection mechanism for spatial-temporal graph learning. Finally, we describe the computational efficiency of STG-Mamba in Section 3.4.

3.1 Formulation of STG Selective State Space Models

SSSMs like structured state space sequential models (S4) and Mamba, initially designed for sequential data, are adapted for Spatial-Temporal Graph (STG) data modeling by leveraging their structure to handle spatiotemporal dependencies dynamically. The adaptation involves discretizing the continuous system described by linear Ordinary Differential Equations (ODEs) into a format amenable to deep

neural networks modeling and optimization, enhancing computational efficiency while capturing the evolving dynamics of the STG networks.

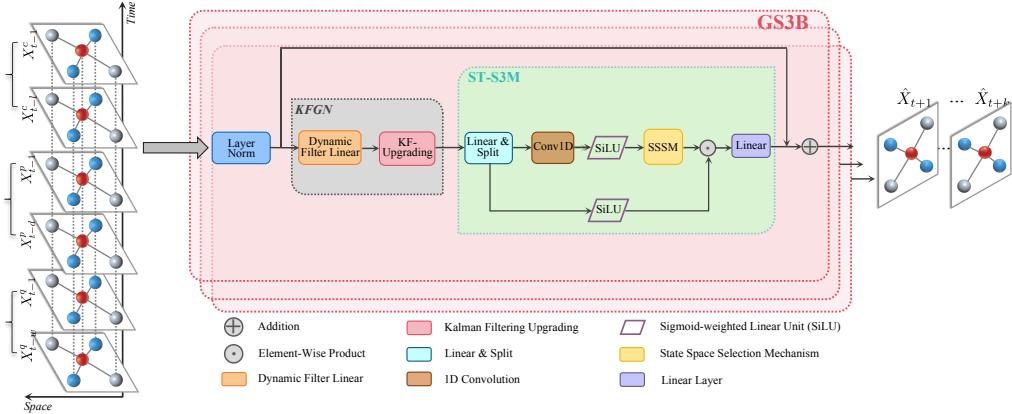


Figure 1: The comprehensive illustration of STG-Mamba’s Model Architecture.

Modern selective state space models, such as Mamba, build upon classical continuous systems by incorporating deep learning techniques and discretization methods suitable for complex time-series data. Mamba specifically maps a one-dimensional input sequence $x(t) \in \mathbb{R}$ through intermediate implicit states $h(t) \in \mathbb{R}^N$, culminating in an output $y(t) \in \mathbb{R}$. The fundamental process is described by a modified set of state space equations that include discretization adjustments to accommodate deep learning frameworks:

$$\begin{aligned} h'(t) &= \bar{\mathbf{A}}h(t) + \bar{\mathbf{B}}x(t) \\ y(t) &= \mathbf{C}h(t) \end{aligned} \quad (2)$$

where $\bar{\mathbf{A}} \in \mathbb{R}^{N \times N}$ and $\bar{\mathbf{B}} \in \mathbb{R}^{N \times 1}$ are discretized representations of the continuous state transition and input matrices, respectively. $\mathbf{C} \in \mathbb{R}^{1 \times N}$ is the output matrices. The discretization process involves transformation techniques that convert \mathbf{A} and \mathbf{B} from their continuous forms to computational efficient formats. These transformations are determined by a selective scan algorithm that utilizes a convolutional approach to model interactions across temporal dimension:

$$\begin{aligned} \bar{\mathbf{A}} &= \exp(\Delta \mathbf{A}) \\ \bar{\mathbf{B}} &= \Delta \mathbf{B} \cdot (\exp(\Delta \mathbf{A}) - \mathbf{I}) \end{aligned} \quad (3)$$

Mamba extends classic state space model functionalities by incorporating a selective mechanism that dynamically adapts to the input sequence characteristics. This adaptability is achieved through a novel architecture that selectively applies convolution kernels across the sequence, optimizing the model’s response to spatiotemporal variations:

$$\bar{\mathbf{K}} = (\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \dots, \mathbf{C}\bar{\mathbf{A}}^{L-1}\bar{\mathbf{B}}) \quad (4)$$

where $\bar{\mathbf{K}}$ denotes a structured convolutional kernel, capturing dependencies across spatial and temporal dimensions, and L represents the length of the input sequence x . This selective convolutional approach significantly enhances the model’s capability to forecast spatiotemporal data by focusing computation on critical features of the data sequence, thus improving both accuracy and efficiency in predictions.

Figure 1 illustrates the comprehensive architecture of the proposed STG-Mamba. Specifically, we formulate the overall architecture following the Residual Encoder fashion for efficient sequential data modeling and prediction. In STG-Mamba, we leverage the Graph Selective State Space Block (GS3B) as the basic Encoder module, and repeat it N times. GS3B consists of several networks and operations, including Layer Norm, Kalman Filter Graph Neural Networks (KFGN) which consists of DynamicFilter-GNN followed by Kalman Filter Upgrading (KF-Upgrading), Spatial-Temporal

Selective State Space Module (ST-S3M) which consists of Linear & Split, Conv1D, SiLU, State Space Selection Mechanism (SSSM), Element-Wise Concatenation, as well as th Residual Connections for every GS3B Encoder layer.

3.2 Kalman Filtering Graph Neural Networks

The motivation for employing Kalman Filtering-based methods in constructing GNNs stems from the need to enhance the reliability and accuracy of spatial-temporal forecasting. STG big data (such as traffic sensor records/weather station records/urban metro system In/Out records) often contains inherent biases and noises that existing methods typically overlook. By integrating Kalman Filtering-based optimization and upgrading, our KFGN effectively addresses these inaccuracies by dynamically weighing the reliability of data streams from different temporal granularity, optimizing the fusion of these data streams based on the estimated variances. This approach not only corrects dataset inherent errors but also significantly improves the model’s capacity to capture the complex inter-dependencies within STG patterns.

As depicted in Figure 2, the KFGN pipeline consists of two key steps. In the first step, embeddings generated from the model inputs of different temporal granularities (i.e., recent steps/periodic steps/trend steps) are sent into a DynamicFilter-GNN module for processing. Then, in the second step, the outputs of DynamicFilter-GNN module are integrated and optimized through the Kalman Filtering Upgrading module. In the initial stage of DynamicFilter-GNN, a learnable Base Filter parameter matrix of size $\mathbb{R}^{\text{in_fea} \times \text{in_fea}}$ is defined. It’s employed to transform the graph adjacency matrix, thereby dynamically adjusting the connection degrees between nodes. Following that, we uniformly initialize the weights and biases, with an initialization standard deviation stdv calculated as the reciprocal of the square root of the number of input features: $\text{stdv} = \frac{1}{\sqrt{\text{in_fea}}}$. This uniform initialization is critical for ensuring the module starts from a neutral point.

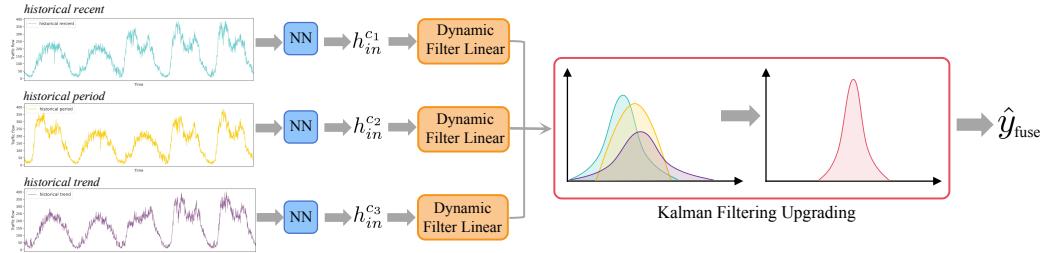


Figure 2: Kalman Filtering Graph Neural Networks.

Subsequently, we transformer the Base Filter via a linear transformation layer, and combine it with the original adjacency matrix $A_{\text{ini}}^{c_i}$, resulting in a dynamically adjusted adjacency matrix $A_{\text{dyn}}^{c_i}$. Let $c_{i,i=\{1,2,3\}} \in \{r,p,q\}$ represents the input data from different temporal granularities, (i.e., $c_1 = r$ denotes historical recent, $c_2 = p$ denotes historical period, $c_3 = q$ denotes historical trend). Leveraging $A_{\text{dyn}}^{c_i}$, input embedding $h_{in}^{c_i}$, weights $W_{DF}^{c_i}$ and bias $b_{DF}^{c_i}$, the input embedding undergoes a graph convolution:

$$h_{DF}^{c_i} = h_{in}^{c_i} \cdot (A_{\text{dyn}}^{c_i} \cdot W_{DF}^{c_i}) + b_{DF}^{c_i} \quad (5)$$

This design enables the model to dynamically adjust the strengths of connections between nodes within the graph based on the learned adjustments.

Following DynamicFilter-GNN, the next step is Kalman Filtering Upgrading. Since the input data are subjected to the same dataset but within different temporal granularities, they are assumed to follow Gaussian Distributions:

$$y_q(x; \mu_q, \sigma_q) \triangleq \frac{\exp^{-\frac{(x-\mu_q)^2}{2\sigma_q^2}}}{\sqrt{2\pi\sigma_q^2}}; y_p(x; \mu_p, \sigma_p) \triangleq \frac{\exp^{-\frac{(x-\mu_p)^2}{2\sigma_p^2}}}{\sqrt{2\pi\sigma_p^2}}; y_r(x; \mu_r, \sigma_r) \triangleq \frac{\exp^{-\frac{(x-\mu_r)^2}{2\sigma_r^2}}}{\sqrt{2\pi\sigma_r^2}}; \quad (6)$$

where subscript $q = c_3, p = c_2, r = c_1$ denote historical trend/period/recent dataset, respectively.

We employ Kalman Filtering method to derive accurate information from the multi-granularity observation sets. Specifically, their individual probability distribution functions can be integrated together by multiplication:

$$\begin{aligned} y_{fuse}(x; \mu_q, \sigma_q, \mu_p, \sigma_p, \mu_r, \sigma_r) &= \frac{\exp^{-\frac{(x-\mu_q)^2}{2\sigma_q^2}}}{\sqrt{2\pi\sigma_q^2}} \times \frac{\exp^{-\frac{(x-\mu_p)^2}{2\sigma_p^2}}}{\sqrt{2\pi\sigma_p^2}} \times \frac{\exp^{-\frac{(x-\mu_r)^2}{2\sigma_r^2}}}{\sqrt{2\pi\sigma_r^2}} \\ &= \frac{1}{(2\pi)^{3/2}\sqrt{\sigma_q^2\sigma_p^2\sigma_r^2}} \exp^{-\left(\frac{(x-\mu_q)^2}{2\sigma_q^2} + \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_r)^2}{2\sigma_r^2}\right)} \quad (7) \end{aligned}$$

By reorganizing Eq. 7 into a simplified version, we have:

$$y_{fuse}(x; \mu_{fuse}, \sigma_{fuse}) = \frac{1}{\sqrt{2\pi\sigma_{fuse}^2}} \exp^{-\frac{(x-\mu_{fuse})^2}{2\sigma_{fuse}^2}} \quad (8)$$

where $\mu_{fuse} = \frac{\mu_q/\sigma_q^2 + \mu_p/\sigma_p^2 + \mu_r/\sigma_r^2}{1/\sigma_q^2 + 1/\sigma_p^2 + 1/\sigma_r^2}$ and $\sigma_{fuse}^2 = \frac{1}{1/\sigma_q^2 + 1/\sigma_p^2 + 1/\sigma_r^2}$.

To simplify the representation of μ_{fuse} and σ_{fuse}^2 , we hereby introduce parameters $\omega_q = \frac{1}{\sigma_q^2}$, $\omega_p = \frac{1}{\sigma_p^2}$, $\omega_r = \frac{1}{\sigma_r^2}$. Then, μ_{fuse} and σ_{fuse}^2 can be re-written as:

$$\mu_{fuse} = \frac{\mu_q\omega_q + \mu_p\omega_p + \mu_r\omega_r}{\omega_q + \omega_p + \omega_r}; \sigma_{fuse}^2 = \frac{1}{\omega_q + \omega_p + \omega_r} \quad (9)$$

This means that observations from different branches can be effectively integrated using weighted sum, and the weight is a combination of variances:

$$\begin{aligned} \mu_{fuse} &= \mu_q\left(\frac{\omega_q}{\omega_q + \omega_p + \omega_r}\right) + \mu_p\left(\frac{\omega_p}{\omega_q + \omega_p + \omega_r}\right) + \mu_r\left(\frac{\omega_r}{\omega_q + \omega_p + \omega_r}\right) \\ &\Downarrow \\ y_{fuse} &= y_q\left(\frac{\omega_q}{\omega_q + \omega_p + \omega_r}\right) + y_p\left(\frac{\omega_p}{\omega_q + \omega_p + \omega_r}\right) + y_r\left(\frac{\omega_r}{\omega_q + \omega_p + \omega_r}\right) \end{aligned} \quad (10)$$

Directly calculating the variance of an observation set requires the complete data, which is computationally expensive. Therefore, we estimate the variance by computing the variance distribution of each training samples:

$$\mathbb{E}[\sigma_{\{q,p,r\}}^2] = \frac{1}{N_m} \sum_i \frac{(S_i - \bar{S})^2}{L} \quad (11)$$

where L is the length of each sample sequence, and N_m is the number of data samples. S_i denotes the i -th observation, and \bar{S} denotes the average value of all observed samples. To further improve the integration and upgrading, we add two learnable weight params ϵ, φ to balance different observation branches. To this end, based on the formulation of y_{fuse} in Eq. 10, the output of Kalman Filtering Upgrading module is:

$$\tilde{y}_{fuse} = \frac{\epsilon \cdot (\hat{y}_q\omega_q) + \varphi \cdot (\hat{y}_p\omega_p) + \hat{y}_r\omega_r}{\omega_q + \omega_p + \omega_r} \quad (12)$$

where $\omega_q = \frac{1}{\sigma_q^2}$, $\omega_p = \frac{1}{\sigma_p^2}$, $\omega_r = \frac{1}{\sigma_r^2}$, ϵ and φ are trainable weights. Finally, for the convenience of neural network training and ensure the scalability, we remove the constant denominator $(\omega_q + \omega_p + \omega_r)$ in Eq. 12. As such, the final output of Kalman Filtering Upgrading is defined as:

$$\hat{y}_{fuse} = \epsilon \cdot (\hat{y}_q\omega_q) + \varphi \cdot (\hat{y}_p\omega_p) + \hat{y}_r\omega_r \quad (13)$$

Here, \hat{y}_{fuse} is the final output of KF-Upgrading module. Note that our approach is an optimized version of the classic Kalman Filtering theory, which is specifically designed for deep learning-based methodology, ensuring both computationally efficiency, dynamic hierarchical STG feature fusion, and improved accuracy.

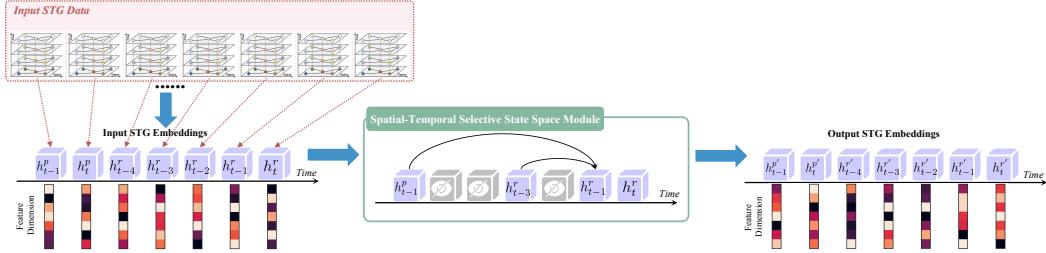


Figure 3: Illustration of the spatial-temporal state space selection mechanism that facilitates input-dependent feature selection and focusing.

3.3 ST-S3M: Spatial-Temporal Selective State Space Module

ST-S3M plays the significant role of adaptive STG feature selection, serving as the counterpart of Attention mechanism. The architecture of ST-S3M is shown in Figure 1. We also give an vivid illustration of the STG feature selection procedure in Figure 4.

After being processed by KFGN module for dynamic spatiotemporal dependency modeling and statistical-based integration & upgrading. The generated embedding \hat{h}_{fuse} is sent to ST-S3M for input-specific dynamic feature selection. In ST-S3M, the input STG embedding passes through a Linear Layer, followed by a splitting operation. Let b denotes batch size, l denotes sequence length, d_{model} be the feature dimension, d_{inner} be the model inner feature dimension, we have:

$$\begin{aligned} h_{\text{main-res}} &= W_{in}\hat{h}_{\text{fuse}} + b_{in} \\ (h_{\text{main}}, res) &= \text{split}(h_{\text{main-res}}) \end{aligned} \quad (14)$$

where $\hat{h}_{\text{fuse}} \in \mathbb{R}^{b \times l \times d_{model}}$, $W_{in} \in \mathbb{R}^{2d_{inner} \times d_{model}}$, $b_{in} \in \mathbb{R}^{2d_{inner}}$, $h_{\text{main-res}} \in \mathbb{R}^{b \times l \times 2d_{inner}}$, $h_{\text{main}}, res \in \mathbb{R}^{b \times l \times d_{inner}}$

Then, h_{main} flows into a 1D convolution layer, followed by SiLU activation function:

$$\begin{aligned} h'_{\text{main}} &= \text{Conv1D}(h_{\text{main}}) \\ h''_{\text{main}} &= \text{SiLU}(h'_{\text{main}}) \end{aligned} \quad (15)$$

where $h''_{\text{main}} \in \mathbb{R}^{b \times l \times d_{inner}}$. The output of SiLU is sent to state space selection mechanism for processing:

$$h_{\text{sssm}} = \text{SSSM}(h''_{\text{main}}) \quad (16)$$

where $h_{\text{sssm}} \in \mathbb{R}^{b \times l \times d_{inner}}$. On the same time, the residual part res is also processed by SiLU activation, and finally we employ element-wise dot product to fuse the main STG embedding and its residual: $h_{\text{sssm}} \odot \text{SiLU}(res)$. The fused result is transformed through a Linear projection:

$$h_{\text{out}} = W_{out}(h_{\text{sssm}} \odot \text{SiLU}(res)) + b_{out} \quad (17)$$

where $h_{\text{out}} \in \mathbb{R}^{b \times l \times d_{model}}$ is the final output of ST-S3M. $W_{out} \in \mathbb{R}^{d_{model} \times d_{inner}}$, $b_{out} \in \mathbb{R}^{d_{model}}$.

The State Space Selection Mechanism in ST-S3M plays a pivotal role in adaptive spatiotemporal feature selection. We detail the parameter computation and updating process of the State Space Selection Mechanism in Algorithm 1. Further, the pseudo-code for the employed Selective Scan method is presented in Algorithm 2.

3.4 Computational Efficiency of STG-Mamba

Leveraging a data-specific selection process, the increasing of sequence length L or graph network node scale N by k -fold can results in huge expansion of model parameters for Transformer-based methods. STG-Mamba introduces a sophisticated algorithm that is conscious of hardware capabilities, exploiting the layered structure of GPU memory to minimize the ensuing cost. In detail, for a batch size of B , STG feature dimension of d' , and network node number of N , STG-Mamba reads the

Algorithm 1 State Space Selection Mechanism Parameter Calculation and Upgrading.**Input:** $x \in \mathbb{R}^{b \times l \times d_{\text{model}}}$, learnable model params $A_{log} \in \mathbb{R}^{d_{\text{in}} \times n}$ and $D \in \mathbb{R}^{d_{\text{in}}}$;**Output:** $y \in \mathbb{R}^{b \times l \times d_{\text{inner}}}$;

- 1: Compute input-independent params $A \in \mathbb{R}^{d_{\text{in}} \times n}$ and $D \in \mathbb{R}^{d_{\text{in}}}$:
- 2: $A = -\exp(A_{log})$, $D = D$;
- 3: Process input x with Linear Layer *Input_proj*:
- 4: $x_{\text{dbl}} = \text{Input_proj}(x)$, $x_{\text{dbl}} \in \mathbb{R}^{b \times l \times (dt_{\text{rank}} + 2n)}$;
- 5: Dividing x_{dbl} in order to get the input-dependent params Δ, B, C , where $\Delta \in \mathbb{R}^{b \times l \times dt_{\text{rank}}}$, $B, C \in \mathbb{R}^{b \times l \times n}$;
- 6: Use Linear Layer *dt_proj* and *SoftPlus* activation to adjust Δ :
- 7: $\Delta^* = \text{SoftPlus}(\text{dt_proj}(\Delta))$, where $\Delta^* \in \mathbb{R}^{b \times l \times d_{\text{in}}}$;
- 8: Employ the *SelectiveScan* function to proceed with x, Δ^*, A, B, C, D :
- 9: $y = \text{SelectiveScan}(x, \Delta^*, A, B, C, D)$;
- 10: Return the Output y .

Algorithm 2 Selective Scan Algorithm**Input:** u, Δ^*, A, B, C, D ;**Output:** z ;

- 1: Obtain the dim of u as (b, l, d_{in}) and the second dim of A as n ;
- 2: Discretize continuous parameters A and B :
- 3: $\Delta A \leftarrow \exp(\text{einsum}(\Delta^*, A))$, where $\Delta A \in \mathbb{R}^{b \times l \times d_{\text{in}} \times n}$;
- 4: $\Delta B_u \leftarrow \text{einsum}(\Delta^*, B, u)$, where $\Delta B_u \in \mathbb{R}^{b \times l \times d_{\text{in}} \times n}$;
- 5: Initialize state x as zeroes with dimension $\mathbb{R}^{b \times d_{\text{in}} \times n}$;
- 6: **for** i from 1 to l **do**
- 7: $x \leftarrow \Delta A[:, i] \times x + \Delta B_u[:, i]$;
- 8: $z \leftarrow \text{einsum}(x, C[:, i, :])$;
- 9: Append z to the output list z_s ;
- 10: **end for**
- 11: $z \leftarrow \text{stack}(z_s)$, where $z \in \mathbb{R}^{b \times l \times d_{\text{in}}}$;
- 12: Adding the direct gain D :
- 13: $z \leftarrow z + u \times D$
- 14: Return z .

$O(BLd' + Nd')$ input data of $\mathbf{A}, \mathbf{B}, \mathbf{C}, \Delta$ from High Bandwidth Memory (HBM), processes the intermediate stages sized $O(BLd'N)$ in Static Random-Access Memory (SRAM), and dispatches the ultimate output, sized $O(BLd')$, back to HBM. In contrast, if a Transformer-based model is employed, the HBM will read $O(BL^2d')$ input data, and the computational size remains $O(BL^2d')$ in SRAM. In common practices we often have $BL^2d \gg BLd' + Nd'$, which means SSSM-based approach can effectively decrease the computational overhead in large-scale parallel processing, such as neural network training. Additionally, by not retaining intermediate states, it also diminishes memory usage, since these states are recalculated during the gradient assessment in reverse phase. SSSM's GPU-optimized design ensures STG-Mamba operates with a linear time complexity $O(L)$ relative to the length of the input sequence, marking a substantial speed advantage over traditional transformers' dense attention computations, which exhibit quadratic time complexity $O(L^2)$.

4 Experiments

4.1 Dataset Statistics and Baseline Methods

Datasets. For experimental evaluation, we meticulously select three real-world STG datasets from California road network speed records (PeMS04) Song et al. [2020], Hangzhou urban metro system entry/exit records (HZMetro) Liu et al. [2022], and weather station records across mainland China (KnowAir) Wang et al. [2020]. Acquired from the Caltrans Performance Measurement System (PeMS), PeMS04 records the average traffic speed from a total number of 307 sensors distributed on California Highway system. The time range of PeMS04 is 1st/Jan/2018 - 28th/Feb/2018, with 5-minute time interval. HZMetro was created with the In/Out transaction records of 80 stations of

Hangzhou Metro system from 1st Jan 2019 to 25th Jan 2019, which takes 15-minute as data collection time interval. KnowAir documents weather observations in every 3-hour, encompassing a dataset that spans from 2015 to 2018. Within the KnowAir dataset, there are three subsets. In this study, we exclusively use subset No. 3, covering the period from 1/Sep/2016 to 31/Jan/2017. The detailed dataset statistics and descriptions are summarized in Table 1.

Table 1: Statistics and descriptions of the three datasets used for experimental evaluation.

Dataset	PeMS04	HZMetro	KnowAir
City	California, USA	Hangzhou, China	184 main cities across China
Data Type	Network traffic speed	Station InFlow/OutFlow	Weather records
Total Nodes	307	80	184
Time Interval	5-minute	15-minute	3-hour
Time Range	1/Jan/2018 - 28/Feb/2018	1/Jan/2019 - 25/Jan/2019	1/Sep/2016 - 31/Jan/2017
Dataset Length	16,992	2,400	1,224

Baseline Methods. For fair comparison, we consider both benchmark spatiotemporal graph neural network (STGNN)-based methods, and Transformer-based methods for STG learning.

STGNN-based Methods: (I) STGCN Yu et al. [2018], (II) STSGCN Song et al. [2020], (III) STG-NCDE Choi et al. [2022], (IV) DDGCRN Weng et al. [2023].

Attention (Transformer)-based Methods: (V) ASTGCN Guo et al. [2019], (VI) ASTGNN Guo et al. [2021], (VII) PDFFormer Jiang et al. [2023], (VIII) STAEformer Liu et al. [2023], (IX) Multi-SPANS Zou et al. [2024].

4.2 Implementation Settings

We split all the datasets with a ratio of 6:2:2 for Training/Validation/Testing sets along the time axis, respectively. Before starting model training session, all the data samples are normalized into range $[0, 1]$ with MinMax normalization method. We set the batch size as 48 and employ AdamW as the Optimizer. The learning rate is initialized as $1e^{-4}$, with a weight decay of $1e^{-2}$. CosineAnnealingLR Loshchilov and Hutter [2016] is adopted as the learning rate scheduler, with a maximum of 50 iterations, and a minimum learning rate $1e^{-5}$. Model training epochs is set as 100. The number of GS3B Encoder Layer is set as $N = 4$. We use the historical ground-truth data from the past 12 steps (for historical recent/period/trend scales) to forecast the future 12 time steps. Mean Squared Error (MSE) is employed as the Loss Function for optimization. And finally, we employ RMSE, MAE, MAPE as the evaluation metrics. All experiments are conducted under one NVIDIA A100 80GB GPU for neural computing acceleration.

4.3 Results Evaluation and Comparison

A comprehensive evaluation between STG-Mamba and the proposed baseline methods is conducted, and we report the RMSE/MAE/MAPE performance results in Table 2. Our STG-Mamba consistently outperforms other benchmark methods on the three tasks, except for the MAPE criteria of PeMS04 (Flow). In PeMS04 (Flow) forecasting task, STAEformer shows slightly better MAPE results than STG-Mamba. It can also be easily observed that Transformer-based methods which integrate GNNs inside the architecture always gain superior performance than sole GNN-based methods. For instance, PDFFormer, STAEformer, and MultiSPANS always show better performance than STGCN, STSGCN, and STG-NCDE. The reason is that although GNN is an effective approach of modeling STG tasks, Transformer-based methods have stronger ability to capture local and long-term dynamic dependencies. Within the GNN-based methods, STG-NCDE and DDGCRN stand out for their competitive performance. STG-NCDE models continuous-time dynamics within STG networks through neural controlled differential equations (NCDE), offering a more precise representation of spatiotemporal feature changes over time compared to other discrete GNN-based methods. Figure 4 presents a direct performance visualization comparison among STG-Mamba and other superior Transformer-based baselines (ASTGNN/PDFFormer/STAEformer) on PeMS04, HZMetro, and KnowAir dataset, respectively. We visualize the predictions over 72 consecutive time steps using the average results of multiple 12-step forecast settings. For each dataset, four node individuals are randomly selected (i.e., 4 random traffic sensors of PeMS04, 4 random metro stations of HZMetro, and 4 random weather stations of KnowAir). The results indicate that STG-Mamba

consistently yields the prediction curve closest to the ground truth waveform, featuring more accurate STG sequence peak values and fewer fluctuations.

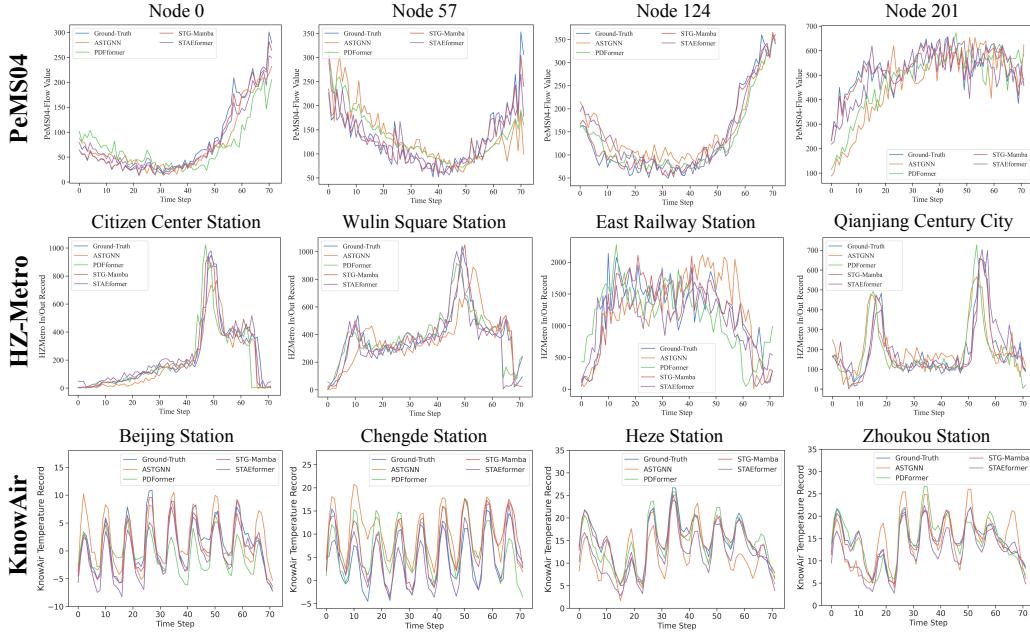


Figure 4: STG forecasting results visualization and comparison among STG-Mamba and other superior Transformer-based baselines.

As a new contender poised to challenge Transformer architecture, Mamba introduces an innovative approach by replacing the foundational Attention mechanism with modern selective state space model (SSSM). SSSM diverges from Attention by utilizing a continuous-time model that captures the dynamic evolution of spatial-temporal dependencies more naturally and efficiently. Unlike Attention, which computes weights across all pairs of input and output positions and can be computationally intensive, Mamba’s selective state space model directly captures the temporal evolution of features, enabling it to scale more gracefully with the sequence length and complexity of spatial-temporal data. Furthermore, we facilitate STG-Mamba with Deep Kalman Filtering-based adaptive spatiotemporal graph evolution, and the elaborately designed Graph Selective State Space Block (GS3B), making it inherit the advantages of modern SSSM, and highly suitable for STG learning tasks.

Table 2: Performance Evaluation and Comparison with Benchmark Baselines on PeMS04, HZMetro, KnowAir Datasets.

Model	PeMS04 (Flow)			HZMetro			KnowAir		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
STGCN	35.55	22.70	14.59	34.85	21.33	13.47	20.64	15.27	12.92
STSGCN	33.65	21.19	13.90	33.24	20.79	13.06	20.25	15.22	12.69
STG-NCDE	31.09	19.21	12.76	32.91	20.75	12.88	20.08	14.93	11.87
DDGCRN	30.51	18.45	12.19	31.69	19.52	12.45	19.63	12.45	11.84
ASTGCN	35.22	22.93	16.56	34.36	21.12	13.50	20.37	15.07	12.11
ASTGNN	31.16	19.26	12.65	32.75	20.63	12.47	19.96	12.38	11.23
PDFFormer	29.97	18.32	12.10	30.18	19.13	11.92	18.65	12.21	11.06
STAEformer	30.18	18.22	11.98	29.94	18.85	12.03	18.49	11.93	10.75
MultiSPANS	30.46	19.07	13.29	30.31	18.97	11.85	18.57	12.24	10.88
STG-Mamba	29.53	18.09	12.12	29.26	18.38	11.67	18.05	11.73	10.66

4.4 Robustness Analysis

STG data exhibit notable periodicity and diversity. For urban mobility/traffic data, there are distinct differences in traffic characteristics between the peak travel periods in the morning and evening and

Table 3: Robustness analysis on PeMS04. We compare the forecasting performance among STG-Mamba and 4 superior baseline methods during rush/non rush hours, weekend/non weekend settings.

Models	Rush Hours			Non Rush Hours			Weekend			Non Weekend		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
ASTGNN	33.57	21.05	13.81	29.76	18.22	12.49	30.66	19.08	13.35	31.82	19.53	12.28
PDFormer	31.69	19.54	12.38	28.95	18.16	11.97	29.87	18.26	12.04	30.39	18.68	12.23
STAEformer	31.36	19.10	12.22	29.47	17.98	11.93	29.83	18.19	11.96	30.26	18.42	12.27
MultiSPAN	30.95	19.28	13.31	30.04	18.37	12.54	30.13	18.42	12.73	30.52	19.31	13.34
STG-Mamba	29.58	18.13	12.15	29.23	18.04	11.91	29.31	18.05	11.92	29.57	18.15	12.16

the none-peak periods. Moreover, the traffic patterns over the 24 hours of weekdays and those of weekends also differ substantially. In light of the variations in STG characteristics due to external environmental changes, it is imperative to assess whether deep learning models can accurately and effectively model the spatial-temporal dependencies under various external conditions.

As such, we establish four distinct external scenarios: (a) Traffic rush hours: Monday to Friday, 8:00-11:00 AM and 16:00-19:00 PM; (b) Non-rush hours: encompassing the remaining time periods from Monday to Friday; (c) Weekend: including all hours on Saturday and Sunday; and (d) Non-weekend: including all time periods from Monday to Friday. Extensive experiments were conducted on the PeMS04 dataset. Table 3 presents the STG forecasting results corresponding to the four traffic scenarios. We compare the prediction performance of STG-Mamba with the other four superior baselines, and report the RMSE/MAE/MAPE metrics.

An ideal deep learning-based STG system should maintain robustness to external disturbances (changes of external environment or events). This is specifically reflected in the fact that when facing different scenarios, the model performance should maintain as little difference as possible, rather than being good at some scenarios but being bad at other occasions.

Among the compared methods, ASTGNN, PDFormer, and STAEformer exhibit relatively large differences in the prediction results between the Rush hour and Non-rush Hour scenarios. In the Weekend/Non Weekend scenario, although the performance differences of each model are not as significant as in the Rush hour/Non-Rush hour scenario, we observe that STG-Mamba still shows the smallest performance difference (the maximum robustness) and achieves the best RMSE/MAE/MAPE metrics. In summary, STG-Mamba proves the best robustness and outperforms existing baseline methods even in face of complex and diversity traffic scenes.

4.5 Ablation Study

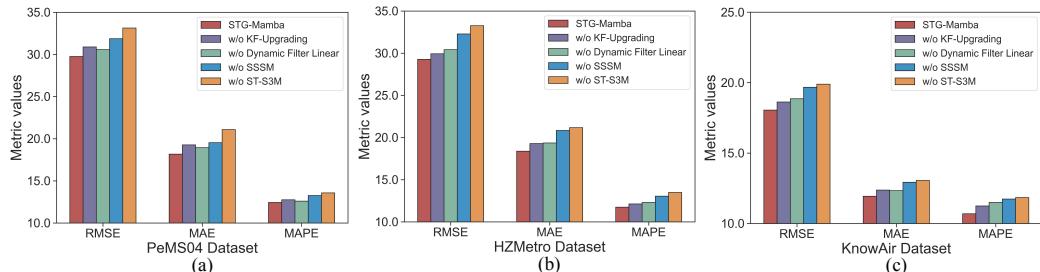


Figure 5: Model component analysis of STG-Mamba.

To investigate the effectiveness of each model component within STG-Mamba, we further design three kinds of model variants, and evaluate their forecasting performance on PeMS04, HZMetro, and KnowAir dataset. The detailed description of these model variants are introduced in the following:

1. STG-Mamba: The full STG-Mamba model without any modification.
2. STG-Mamba w/o KF-Upgrading: We replace the KF-Upgrading module in KFGN with the simple sum average operation to fuse the three temporal branches.
3. STG-Mamba w/o Dynamic Filter Linear: We replace the proposed Dynamic Filter Linear in KFGN with the basic static graph convolution.

4. STG-Mamba w/o SSSM: The State Space Selection Mechanism (SSSM) is removed from the ST-S3M module.
5. STG-Mamba w/o ST-S3M: The whole ST-S3M module is removed from GS3B Encoder.

As illustrated in Figure 5, STG-Mamba w/o KF-Upgrading shows worse performance than the full model, demonstrating the necessity and suitability of employing statistical learning-based GNN state space upgrading and optimization for SSSM-based methods. The decrease of performance due to the absence of Dynamic Filter Linear demonstrate its effectiveness, and the necessity of designing a suitable adaptive GNN for STG feature learning. Furthermore, removing the State Space Selection Mechanism in ST-S3M results in a substantial degradation of model capability, proving the feasibility of using SSSM as the substitute for Attention mechanism. Finally, the removal of ST-S3M makes the STG-Mamba degrade to a plain GNN-based model, resulting in the lowest performance.

5 Conclusions

In this work, we for the first time introduce deep learning-based selective state space models (SSSM) for spatial-temporal graph learning tasks. We propose STG-Mamba that leverages modern SSSM for accurate and efficient STG forecasting. STG-Mamba is formulated following the stacked Residual Encoder architecture, with the ST-S3M module facilitating input-dependent graph evolution and node feature selection, successfully integrates STG network with SSSMs. Through the Kalman Filtering Graph Neural Networks (KFGN), the learned STG embeddings achieve a smooth optimization & upgrading based on statistical learning, aligning with the whole STG selective state space modeling process. Compared with Attention-based methods, STG-Mamba is able to achieve linear-time complexity and substantial reduction in FLOPs and test inference time. We conduct extensive empirical studies on three open-sourced STG datasets from urban traffic, metro In/Out flow, and weather records. Experimental results demonstrate the consistent superiority of STG-Mamba over other benchmark methods. We believe that STG-Mamba presents a brand new promising approach to general STG learning fields, offering competitive model performance and input-dependent context-aware feature selection under an affordable computational cost.

References

- Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 922–929, 2019.
- Jianlei Kong, Xiaomeng Fan, Min Zuo, Muhammet Deveci, Xuebo Jin, and Kaiyang Zhong. Adct-net: Adaptive traffic forecasting neural network via dual-graphic cross-fused transformer. *Information Fusion*, 103:102122, 2024.
- Guangyu Huo, Yong Zhang, Boyue Wang, Junbin Gao, Yongli Hu, and Baocai Yin. Hierarchical spatio-temporal graph convolutional networks and transformer network for traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 24(4):3855–3867, 2023.
- Yuxuan Liang, Yutong Xia, Songyu Ke, Yiwei Wang, Qingsong Wen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. Airformer: Predicting nationwide air quality in china with transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14329–14337, 2023.
- Ling Chen, Jiahui Xu, Binqing Wu, and Jianlong Huang. Group-aware graph neural network for nationwide city air quality forecasting. *ACM Transactions on Knowledge Discovery from Data*, 18(3):1–20, 2023.
- Yipeng Zhang, Xin Wang, Hong Chen, and Wenwu Zhu. Adaptive disentangled transformer for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 3434–3445, 2023.
- Yinwei Wei, Wenqi Liu, Fan Liu, Xiang Wang, Liqiang Nie, and Tat-Seng Chua. Lightgt: A light graph transformer for multimedia recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1508–1517, 2023.

Changlu Chen, Yanbin Liu, Ling Chen, and Chengqi Zhang. Bidirectional spatial-temporal adaptive transformer for urban traffic flow forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16372–16382, 2021.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pages 27268–27286, 2022.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Chloe Wang, Oleksii Tsepa, Jun Ma, and Bo Wang. Graph-mamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv preprint arXiv:2402.00789*, 2024.

Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. *arXiv preprint arXiv:2402.03302*, 2024.

Jay H Lee, Manfred Morari, and Carlos E Garcia. State-space interpretation of model predictive control. *Automatica*, 30(4):707–717, 1994.

Bernard Friedland. *Control system design: an introduction to state-space methods*. 2012.

Masanao Aoki. *State space modeling of time series*. 2013.

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.

Lingbo Liu, Jingwen Chen, Hefeng Wu, Jiajie Zhen, Guanbin Li, and Liang Lin. Physical-virtual collaboration modeling for intra- and inter-station metro ridership prediction. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3377–3391, 2022.

Shuo Wang, Yanran Li, Jiang Zhang, Qingye Meng, Lingwei Meng, and Fei Gao. Pm2.5-gnn: A domain knowledge enhanced graph neural network for pm2.5 forecasting. In *Proceedings of the 28th international conference on advances in geographic information systems*, pages 163–166, 2020.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3634–3640, 2018.

Jeongwhan Choi, Hwangyong Choi, Jeehyun Hwang, and Noseong Park. Graph neural controlled differential equations for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6367–6374, 2022.

Wenchao Weng, Jin Fan, Huifeng Wu, Yujie Hu, Hao Tian, Fu Zhu, and Jia Wu. A decomposition dynamic graph convolutional recurrent network for traffic forecasting. *Pattern Recognition*, 142: 109670, 2023.

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.

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.

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.

Dongcheng Zou, Senzhang Wang, Xuefeng Li, Hao Peng, Yuandong Wang, Chunyang Liu, Kehua Sheng, and Bo Zhang. Multispans: A multi-range spatial-temporal transformer network for traffic forecast via structural entropy optimization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1032–1041, 2024.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.