Riding the Wave: Multi-Scale Spatial-Temporal Graph Learning for Highway Traffic Flow Prediction Under Overload Scenarios

Xigang Sun[1], **Jiahui Jin**[1*], **Hancheng Wang**[1], **Xiangguo Sun**[2], **Xiaoliang Wang**[3] and **Jun Zhu**[3]
[1]School of Computer Science and Engineering, Southeast University, China
[2]SEEM, The Chinese University of Hong Kong, China
[3]Zhejiang Mobile Digital Intelligence Technology Co., Ltd., China
{xigangsun, jjin, hanchengwang13}@seu.edu.cn, xiangguosun@cuhk.edu.hk,
{wangxiaoliang2,zhujun}@zj.chinamobile.com

Abstract

Highway traffic flow prediction under overload scenarios (HIPO) is a critical problem in intelligent transportation systems, which aims to forecast future traffic patterns on highway segments during periods of exceptionally high demand. Despite its importance, this problem has rarely been explored in recent research due to the unique challenges posed by irregular flow patterns, complex traffic behaviors, and sparse contextual data. In this paper, we propose a Heterogeneous Spatial-Temporal graph network With Adaptive contrastiVE learning (HST-WAVE) to address the HIPO problem. Specifically, we first construct a heterogeneous traffic graph according to the physical highway structure. Then, we develop a multi-scale temporal weaving Transformer and a coupled heterogeneous graph attention network to capture the irregular traffic flow patterns and complex transition behaviors. Furthermore, we introduce an adaptive temporal enhancement contrastive learning strategy to bridge the gap between divergent temporal patterns and mitigate data sparsity. We conduct extensive experiments on two real-world highway network datasets (No. G56 and G60 in Hangzhou, China), showing that our model can effectively handle the HIPO problem and achieve state-of-the-art performance. The source code is available at https://github.com/luck-seu/HST-WAVE.

1 Introduction

Highway traffic overload is a condition where the volume of vehicles on a highway exceeds its designed or operational capacity, leading to congestion, reduced speeds, and inefficiencies in traffic flow [Jin *et al.*, 2023b; Cui *et al.*, 2020]. According to TomTom's 2024 statistics, the time spent by users traveling through Indianapolis highway sections increased by 27% due to traffic overload [Martichoux, 2025]. Predicting highway traffic flow under such situations can help alleviate bottlenecks by enabling traffic diversion or recommending optimal entry and exit times, reducing negative impacts on society and the economy.

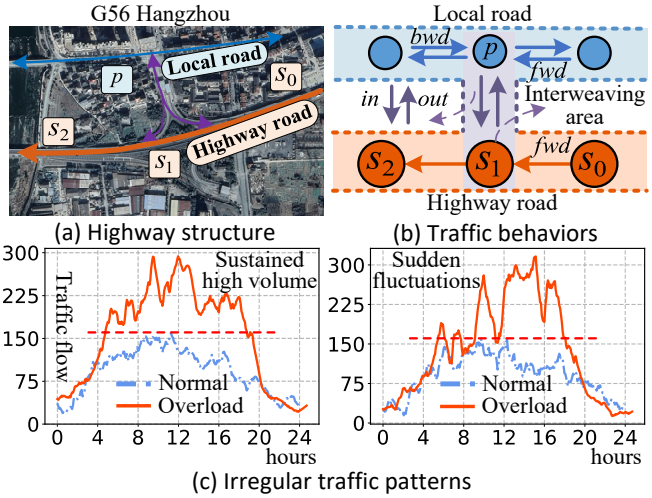

Figure 1: An example of a road patch in the G56 highway. In (a), s_* and p denote the highway and local segments, and s_1 is an interweaving area. In (b), the terms fwd and bwd represent the forward and backward flows on the highway or local segments, while in and out denote the entry and exit flows within the interweaving areas. (c) shows the irregular traffic patterns under overload scenarios.

Highway overload scenarios differ from normal scenarios in both traffic dynamics and spatial interactions. We illustrate these differences using a case study of Highway G56 in Hangzhou, China, as depicted in Figure 1. In overload scenarios, the highway traffic flow consistently surpasses the peak threshold observed under normal conditions (Figure 1(c)), often causing severe traffic congestion. Such high traffic volume usually compels drivers to divert to parallel local roads to bypass congestion (Figure 1(b)), leading to spatial interactions between highways and local roads. However, most existing methods [Jin *et al.*, 2023a; Wu *et al.*, 2019; Han *et al.*, 2024] primarily focus on normal scenarios and overlook the distinct characteristics of overload scenarios.

We study the problem of HIghway traffic flow Prediction under Overload scenarios, abbreviated as HIPO. This problem is more challenging than ordinary traffic prediction tasks, which is reflected in the following crucial difficulties: *1) Irregular traffic patterns.* Traffic patterns during overload pe-

*Corresponding author

riods exhibit significant changes compared to normal periods. As shown in Figure 1 (c), traffic volume often exceeds highway capacity, causing prolonged congestion. Incidents and interweaving behaviors add sudden fluctuations, further destabilizing traffic flow. These unpredictable disruptions break regular traffic patterns, making traditional methods designed for consistent dynamics ineffective in capturing and predicting irregular patterns in overload scenarios. *2) Complex traffic behaviors.* In Figure 1 (a) and (b), the traffic flow on highway segment s_0 can enter local segment p through the interweaving area s_1 when vehicles either reach their destinations or seek to avoid congestion. Conversely, segment p can also serve as the entry point for the highway, merging into segment s_2 through area s_1. These complex traffic behaviors obviously affect highway segments' flows. However, previous data-driven prediction approaches assume that vehicles remain solely on highway segments without accounting for diversion to local roads. *3) Insufficient contextual data.* Highway overload scenarios, such as holidays, are concentrated in a few days, inherently leading to data sparsity. Other approaches that leverage point-of-interests or similar urban information for traffic prediction become ineffective on highway networks. The limited data is insufficient to support training a model with strong generalization ability.

To overcome these challenges, we propose a Heterogeneous Spatial-Temporal graph network With Adaptive contrastiVE learning (HST-WAVE) to address the HIPO problem. Specifically, we first construct a heterogeneous traffic graph (HTG) to depict various transfer interactions on highway networks. We then develop a multi-scale weaving Transformer network to adapt to irregular traffic patterns. A coupled heterogeneous graph attention network performed on HTG is delivered to learn the complex traffic behaviors. Both types of networks learn alternately to form a heterogeneous spatial-temporal module (HSTM) as the primary learner. To further improve the model's generality when facing irregular traffic patterns while alleviating the data sparsity, we develop four temporal augmentation strategies (flip, mask, replace, and noise) in a mini-batch data to implement contrastive learning. Moreover, we incorporate an adaptive learnable temporal pattern with a fixed length, concatenated with the original pattern, to retain common temporal knowledge during training. Our contributions are:

- We investigate a significant yet often neglected problem: highway traffic flow prediction under overload scenarios (HIPO). This problem is notably different from existing traffic prediction tasks due to the unique challenges caused by irregular flow patterns, complex traffic behaviors, and limited contextual data.
- We propose the HST-WAVE framework to address the HIPO problem. Our approach introduces a multi-scale temporal weaving module and a heterogeneous spatial interaction module, which effectively captures irregular temporal patterns and complex traffic behavior dynamics. Additionally, we propose an adaptive temporal enhancement contrastive learning strategy to enhance the model's generalization capability and mitigate data sparsity in overloaded scenarios.
- Extensive experiments on two real-world highway datasets demonstrate that our model achieves state-of-the-art performance, confirming its effectiveness in addressing the HIPO problem.

2 Problem Formulation for HIPO

We study the HIPO problem on a highway road that runs parallel to a local road. Since the highway comprises two separate, non-interfering bidirectional lanes, this study focuses solely on the single-directional highway with the highest traffic flow for simplicity.

Highway Network. A highway network consists of a highway road $\mathcal{S}$ and a parallel local road $\mathcal{P}$, denoted by $\mathcal{V} = \{\mathcal{S}, \mathcal{P}\}$. The highway road $\mathcal{S} = \{s_1, s_2, ..., s_n\}$ is composed of directed highway segments. The interweaving area $\mathcal{S}^{\mathcal{P}} \subset \mathcal{S}$ is a set of highway segments connecting the local road $\mathcal{P}$. The local road $\mathcal{P} = \{p_1, p_2, ..., p_m\}$ is made up of bidirectional local segments, where each segment $p_i \leftrightarrow p_{i+1}$. Within the network, highway segments are sequentially connected in one direction ($s_i \rightarrow s_{i+1}$) while the interweaving area $s \in \mathcal{S}^{\mathcal{P}}$ serves as an interaction point linked with the corresponding local segment $p \in \mathcal{P}$.

Overload Scenario. An overload scenario is a condition of the highway network where the traffic demand significantly exceeds the highway's capacity. Overload scenarios typically arise during holidays, road maintenance, or special events, performing irregular temporal patterns and increased interactions between $\mathcal{S}^{\mathcal{P}}$ and $\mathcal{P}$.

Problem Formulation. Given a highway network $\{\mathcal{S}, \mathcal{P}\}$, we assign a traffic flow placeholder matrix into each type road, i.e., $X = \{X^{\mathcal{S}}, X^{\mathcal{P}}\}$. We intercept a time interval with $T + H$ steps during overload periods. Then we record the traffic flow in historical T-steps on all road segments at time step t, i.e., $X^{\mathcal{S}}_{t-T:t} = \{x^{s_1}_{t-T:t}, x^{s_2}_{t-T:t}, ..., x^{s_n}_{t-T:t}\}$, $X^{\mathcal{P}}_{t-T:t} = \{x^{p_1}_{t-T:t}, x^{p_2}_{t-T:t}, ..., x^{p_m}_{t-T:t}\}$, and $X_{t-T:t} = \{X^{\mathcal{S}}_{t-T:t}, X^{\mathcal{P}}_{t-T:t}\}$. The HIPO problem is viewed as a time series auto-regressive task, given the historically observed traffic flow matrix $X_{t-T:t}$, we aim to predict the traffic flow $X^{\mathcal{S}}_{t:t+H}$ within future H steps on highway segment set $\mathcal{S}$. The HIPO problem can be formulated as:

$$X^{\mathcal{S}}_{t:t+H} = f(X_{t-T:t}, \mathcal{S}, \mathcal{P}), \tag{1}$$

where $f(\cdot)$ is the traffic prediction function.

3 HST-WAVE Framework

This section presents the HST-WAVE framework for solving the HIPO problem, as shown in Figure 2. The framework consists of two components: 1) Heterogeneous Spatial-Temporal Module (HSTM) blocks integrating a Multi-Scale Weaving Transformer (MSWT) and a Coupled Heterogeneous Graph Attention Network (CHGAN); 2) an Adaptive Temporal Enhancement Contrastive Learning (ATECL) strategy integrating temporal enhanced contrastive learning and learnable temporal patterns.

3.1 Traffic Modeling with HSTM

Most current traffic flow prediction methods [Yu *et al.*, 2018; Kong *et al.*, 2024; Zhou *et al.*, 2024; Jin *et al.*, 2023a] de-

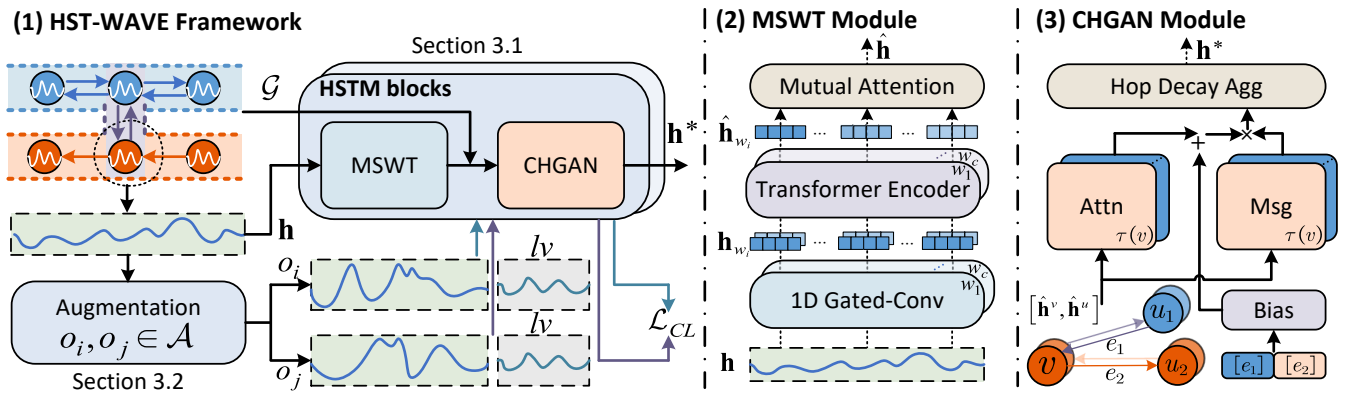

Figure 2: The overview of HST-WAVE's architecture. (1) The main work process for HST-WAVE, the lower part is a temporal enhancement contrastive learning module. (2) A multi-scale weaving Transformer comprises multi-scale 1D-gated CNNs, a Transformer encoder, and inter-scale mutual attention. (3) A coupled heterogeneous graph attention network includes a bidirectional heterogeneous attention mechanism and a hop decay aggregation.

pend on homogeneous graph modeling and simplistic temporal learning mechanisms, which are inadequate for highway traffic modeling under overload conditions. To address this limitation, we design a novel heterogeneous spatial-temporal block, termed HSTM, to effectively learn irregular traffic patterns and complex traffic behaviors.

The HSTM block is built upon a heterogeneous traffic graph (HTG) that captures the diverse types of road segments and their transfer interactions, as illustrated in Figure 1 (b). In this representation, each road segment—whether on a highway or a local road—is modeled as a node, while the transfer interactions between segments are represented as edges. An HTG is formally defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{T}, \mathcal{R}\}$, where $\mathcal{V} = \{\mathcal{S}, \mathcal{P}\}$ is the set of nodes, $\mathcal{E}$ is the set of edges, $\mathcal{T} = \{\text{highway}, \text{local}\}$ is the set of node types, $\mathcal{R} = \{\text{in}, \text{out}, \text{fwd}, \text{bwd}\}$ is the set of edge types, and $N = |\mathcal{V}|$ is the number of nodes. We define $\tau(v) \in \mathcal{T}$ and $\phi(e) \in \mathcal{R}$ as type mapping functions for nodes and edges, respectively, where $v \in \mathcal{V}$ and $e \in \mathcal{E}$. Specifically:

$$\tau(v) = \begin{cases} \text{highway}, & \forall v \in \mathcal{S}, \\ \text{local}, & \forall v \in \mathcal{P}, \end{cases}$$

$$\phi(e) = \begin{cases} \text{fwd}, & e = (s_i, s_j), \\ \text{fwd } \textbf{or } \text{bwd}, & e = (p_i, p_j) \textbf{ or } (p_j, p_i), \\ \text{in } \textbf{or } \text{out}, & e = (p_i, s'_i) \textbf{ or } (s'_i, p_i), \end{cases}$$

where $s_* \in \mathcal{S}$, $s'_* \in \mathcal{S}^{\mathcal{P}}$, and $p_* \in \mathcal{P}$ are nodes in $\mathcal{V}$. The direction of an edge is specified by $i < j$.

Each node $v \in \mathcal{V}$ is associated with temporal features $\mathbf{h}^v \in \mathbb{R}^{T \times d}$, representing the traffic flow on the corresponding highway or local road segment over the past T time steps, where d is the dimension of the temporal feature, and $\mathbf{H} \in \mathbb{R}^{N \times T \times d}$ denotes the temporal features of all nodes. Given the traffic flow matrix $X_{t-T:T}$, we construct the initial feature $\mathbf{H}_0$ by incorporating the traffic flow value, time of day, and day of the week for each time step.

We then input $\mathbf{H}_0$ into the HSTM blocks, which are formed by alternately stacking MSWT and CHGAN:

$$\hat{\mathbf{H}}_{l-1} = \text{MSWT}(\mathbf{H}_{l-1}), \ \mathbf{H}_l = \text{CHGAN}(\hat{\mathbf{H}}_{l-1}, \mathcal{G}), \quad (2)$$

where $\mathbf{H}_l \in \mathbb{R}^{N \times T \times d}$ is the output of all nodes, and l denotes the number of layers. MSWT and CHGAN learn the complex temporal and spatial dynamic features of highway traffic data.

Learning Irregular Traffic Patterns with MSWT

Traffic flow in overload scenarios exhibits sustained high volumes and frequent fluctuations. Existing methods in temporal learning typically rely on time-step-based models [Yu *et al.*, 2018; Jin *et al.*, 2023a], which struggle to capture the dependencies with mixed irregular temporal patterns. We propose a multi-scale weaving Transformer (MSWT), incorporating single- and multi-window weaving attention to handle such irregular traffic patterns. MSWT module inputs $\mathbf{H}_{l-1}$ that contains all nodes' temporal features and transforms $\mathbf{h}^v_{l-1}$ of each node v into the output $\hat{\mathbf{h}}^v_{l-1}$. We omit the subscripts v and $l-1$ for convenience.

We first improve the gated CNNs [Yu *et al.*, 2018] by using multiple temporal scales to decompose the irregular temporal patterns into multi-scale window pattern representations. Given the temporal feature $\mathbf{h}$ for node v, we employ 1-D CNNs with c scale kernels combined with the gated linear unit (GLU) along the temporal dimension T. For the kernel size w_i, the window pattern is sliding on $\mathbf{h}$ with padding $w_i - 1$ zeros, denoted as $\text{Pad}_{w_i}(\mathbf{h})$. The convolution kernel $F_{w_i} \in \mathbb{R}^{w_i \times d \times 2d}$ implements on it and following the GLU operation:

$$\mathbf{h}_{w_i} = P \odot \sigma(Q), \quad (3)$$

where $P, Q \in \mathbb{R}^{T \times d}$ are two parts with the d dimension in the output, which are computed by $P||Q = F_{w_i} * \text{Pad}_{w_i}(\mathbf{h})$. The function $\sigma(Q)$ is a sigmoid gate determining the relevance of input P from the current states for uncovering compositional structure and dynamic variations in each window.

We then input the window pattern embedding $\mathbf{h}_{w_i} \in \mathbb{R}^{T \times d}$ of each granular into a base Transformer encoder to achieve the interaction modeling between single-scale windows:

$$\hat{\mathbf{h}}_{w_i} = \text{TransformerEncoder}_{w_i}(\mathbf{h}_{w_i}). \quad (4)$$

In the multi-scale window design, small-scale windows capture fine-grained, localized traffic variations, while large-scale windows focus on broader temporal trends and context. Furthermore, we propose mutual attention pooling on

the stacked output $\hat{\mathbf{h}}_w \in \mathbb{R}^{c \cdot T \times d}$ of multiple scale window Transformers to integrate local and global insights:

$$W_{\text{window}_{i,j}} = \frac{\hat{\mathbf{h}}_i \hat{\mathbf{h}}_j}{\sum_{a=0}^{c \cdot T} \hat{\mathbf{h}}_i \hat{\mathbf{h}}_a}, \quad \hat{\mathbf{h}} = \text{WindowSum}(W_{\text{window}} \hat{\mathbf{h}}_w), \tag{5}$$

where $\text{WindowSum}(\cdot)$ is a summation function that applies on the window scale. The temporal outputs $\hat{\mathbf{h}} \in \mathbb{R}^{T \times d}$ are concatenated into the representation $\hat{\mathbf{H}}_{l-1}$ of all nodes (in Equation 2). This attention allows the model to adaptively prioritize relevant temporal scales based on inputs, ensuring flexibility in handling mixed irregular temporal patterns.

Learning Complex Traffic Behaviors with CHGAN

The frequent transitions of traffic flow along highway segments, local segments, and interweaving areas under overload scenarios result in complex traffic behaviors. We develop a coupled heterogeneous graph attention network (HTG) to model these interactions in a Transformer-based manner [Mao *et al.*, 2023]. CHGAN maps the MSWT's output $\hat{\mathbf{h}}_{l-1}^v \in \hat{\mathbf{H}}_{l-1}$ to the spatial output $\hat{\mathbf{h}}_l^* \in \mathbf{H}_l$ on $\mathcal{G}$, which is formulated as a message passing mechanism liking general GNNs, and the subscript $l-1$ is omitted:

$$\mathbf{h}_{\mathcal{N}(v)} = \text{Agg}_{u \in \mathcal{N}(v)}\big(\text{Attn}(u,v) \cdot \text{Msg}(u)\big), \quad \mathbf{h}^* = \text{Update}(\hat{\mathbf{h}}^v, \mathbf{h}_{\mathcal{N}(v)}), \tag{6}$$

where $\text{Agg}(\cdot)$, $\text{Attn}(\cdot)$, $\text{Msg}(\cdot)$, and $\text{Update}(\cdot)$ are the aggregation, attention, message, and update functions.

In CHGAN, the attention function $\text{Attn}(\cdot)$ generates the attention scores between two nodes by considering node types to learn important traffic behaviors across different segments. Specifically, for a target node $v \in \mathcal{V}$, we select its all k-hop neighbors set $\mathcal{N}(v)$. We then define a group-specific Query and Key projected function for each node type. The target node v and source node $u \in \mathcal{N}(v)$ are mapped into Query and Key vectors to calculate dot product as attention:

$$Q_v = \text{QLinear}_{\tau(v)}(\hat{\mathbf{h}}^v),\ K_u = \text{KLinear}_{\tau(u)}(\hat{\mathbf{h}}^u), \quad \alpha(\hat{\mathbf{h}}^u, \hat{\mathbf{h}}^v) = \frac{Q_v K_u^{\mathrm{T}}}{\sqrt{d}}. \tag{7}$$

Furthermore, we add the relative edge type bias to adjust the attention scores to enhance heterogeneous traffic interaction modeling. We assign a one-hot vector $h_{\phi(e_i)} \in \mathbb{R}^{|\mathcal{R}|}$ to each edge type, and the edge type feature is indexed by a learnable edge type matrix $H_{\mathcal{R}} \in \mathbb{R}^{|\mathcal{R}| \times d}$, i.e., $\hat{h}_{\phi(e_i)} = h_{\phi(e_i)} H_{\mathcal{R}}$. For node pair (u, v), we compute the average of edge type features along the path $(e_1, e_2, ..., e_k)$ from u to v, and a linear projection is employed to generate bias item:

$$\beta_{(u,v)} = \text{Linear}_{\text{bias}}\Big(\frac{1}{k}\sum_{i=1}^{k} \hat{h}_{\phi(e_i)}\Big), \tag{8}$$

the final attention score is calculated by:

$$\text{Attn}(u,v) = \text{Softmax}\big(\alpha(\hat{\mathbf{h}}^u, \hat{\mathbf{h}}^v) + \beta_{(u,v)}\big). \tag{9}$$

The message function $\text{Msg}(\cdot)$ extracts the source node's feature of target node v by a Value projection:

$$\text{Msg}(u) = \text{VLinear}_{\tau(u)}(\hat{\mathbf{h}}^u). \tag{10}$$

We eventually aggregate all source nodes' messages into the target node by $\text{Agg}(\cdot)$. Considering the neighbor proximity, traffic messages from source road segments at the high-hop pass into the target road segment will consume more than at the low-hop. Therefore, we design a simple decay function without trainable parameters in the aggregation stage:

$$\text{Agg}(\cdot) := \sum_{u \in \mathcal{N}(v)} \text{Attn}(u,v) \cdot \text{Msg}(u) \cdot e^{-\lambda(k_d - 1)}, \tag{11}$$

where λ is the decay factor, and k_d is the hop of a source node. Let the first output of $\text{Agg}(\cdot)$ be denoted by $\text{fwd}(h_{\mathcal{N}(v)})$ which incorporates the traffic flow features from all other nodes with different types and relations flowing to the target node. We flip the edge direction in HTG and put it into the above computation process to obtain the output $\text{bwd}(h_{\mathcal{N}(v)})$, which empowers the model with the ability to perceive the traffic flow features from all nodes reached by the target node's outgoing flow.

We use an activation linear projection combining the residual connection to update the target node's information from bidirectional sources:

$$\mathbf{h}^* = \text{Linear}\big(\sigma(\text{fwd}(\mathbf{h}_{\mathcal{N}(v)}) || \text{bwd}(\mathbf{h}_{\mathcal{N}(v)})\big) + \hat{\mathbf{h}}^v, \tag{12}$$

where $\mathbf{h}^* \in \mathbb{R}^{T \times d}$ is the output of $\text{Update}(\cdot)$ in target node v, and we denote the spatial representation of all nodes as $\mathbf{H}_l$ (in Equation 2). CHGAN improves the heterogeneous graph Transformer [Mao *et al.*, 2023] by introducing relative traffic behavior type learning with a hop decay function and bidirectional traffic flow-aware heterogeneous attention, effectively capturing complex traffic behaviors.

3.2 Traffic Data Augmentation with ATECL

The rarity of overload events leads to insufficient contextual data for training, requiring models to rely on normal traffic data. However, the temporal patterns and behaviors observed in normal traffic differ substantially from those during overload scenarios. Therefore, we devise an adaptive temporal enhancement contrastive learning (ATECL) strategy based on HSTM in this section, which enhances model generalization by bridging the gap between divergent temporal patterns and alleviates data sparsity by augmenting overload scenario data.

To enhance the generality of the model, we apply four augmentation operations $\mathcal{A} = \{Flip, Mask, Replace, Noise\}$ at the temporal level, where each operation serves to simulate different traffic conditions. *Flip:* we randomly flip $T \cdot \gamma$ positions by $max \pm x$ (using $+$ if $x \leq 0$, and $-$ otherwise) to simulate temporal shifts in traffic patterns; *Mask*: we randomly mask $T \cdot \gamma$ positions to 0 to force the model to learn temporal contextual dependencies; *Replace*: we randomly replace $T \cdot \gamma$ positions to noise to simulate fluctuating traffic conditions; *Noise*: we add the Gaussian noise to all positions to mimic unpredictable changes in traffic behavior. Here, γ represents the modification rate, x denotes a value at an arbitrary position within a temporal pattern, and max indicates the maximum flow.

Given a batch of N_b samples, we sample two augmentation operations $o_i, o_j \in \mathcal{A}$ for each temporal pattern to obtain $\mathbf{H}_0^{o_i}, \mathbf{H}_0^{o_j}$ and put them into the HSTM model to obtain the $2N_b$ outputs. Following the general contrastive learning task [Chen *et al.*, 2020], we enhance the similarity between positive pairs $(\mathbf{H}_i, \mathbf{H}_j)$ from the same temporal pattern and decrease the similarity between negative pairs $(\mathbf{H}_i, \mathbf{H}_k)$ from $2N_b - 2$ different temporal patterns:

$$\mathcal{L}_{CL} = -\log \frac{\exp\big(\mathrm{sim}(\mathbf{H}_i, \mathbf{H}_j)/\eta\big)}{\sum_{k=1}^{2N_b} \mathbb{I}_{k \neq i} \exp\big(\mathrm{sim}(\mathbf{H}_i, \mathbf{H}_k)/\eta\big)}, \quad (13)$$

where $\mathbb{I} = \{0, 1\}$ is a binary function outputting 1 if $k \neq i$ and η is a temperature parameter.

To further enhance the learning process and preserve common temporal knowledge across different augmentations, we introduce an adaptive temporal pattern $lv \in \mathbb{R}^{N \times T' \times d}$, which is concatenated with the augmented sequence:

$$\mathbf{H}_0^{o_i} = \mathbf{H}_0 || lv,\ lv \sim \mathcal{N}(0, I), \quad (14)$$

where the pattern is initialized to a normal distribution and updated during model training, acting as a memory to retain key temporal cues regardless of the specific augmentation applied. It helps the model avoid over-fitting to noise while ensuring generalization to real-world overload scenarios.

3.3 Loss Function

We employ a full connection layer to map the output $\mathbf{H}_l$ into a prediction traffic flow matrix $\hat{X}^{\mathcal{S}}_{t:t+H}$, and the ground truth matrix is $X^{\mathcal{S}}_{t:t+H}$. For the HIPO problem, we use the mean absolute error as the training objective loss defined by:

$$\mathcal{L}_{TR} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} |\hat{X}^{\mathcal{S}}_{t:t+H} - X^{\mathcal{S}}_{t:t+H}|_{ij}. \quad (15)$$

The total training loss is calculated by the weighted sum of regression loss and contrastive loss:

$$\mathcal{L} = \mathcal{L}_{TR} + \mu \mathcal{L}_{CL}, \quad (16)$$

where μ is a control factor.

4 Experiments

We conduct extensive experiments on two real-world datasets to evaluate our HST-WAVE's performance and answer the following research questions: **Q1)** How much improvement does the HST-WAVE achieve on the HIPO task compared to methods designed for normal traffic flow prediction? **Q2)** How much contribution does each component of the HST-WAVE make to the accuracy of addressing the HIPO task? **Q3)** How robust is the HST-WAVE on the segment with the highest traffic load compared to existing methods? **Q4)** How sensitive is the HST-WAVE to the different settings for hyper-parameters?

Data	Time range	#Nodes	#Normal/Overload
G56	06/10-10/03 2023	110	31,680/1,728
G60	06/01-10/03 2023	143	34,272/1,728

Table 1: Statistic details of G56 and G60. Their overload periods are from September 28 to October 3.

4.1 Experiment Settings

Datasets. We collect the traffic flow datasets on two highway networks during the normal and overload scenarios, including highways No. G56 and G60 are located in Hangzhou, China, and we consider the holiday periods to be overload scenarios. The traffic flow is obtained by recording the vehicle track points within the kilometer piles at a sampling frequency of 5 minutes, and the specific information is described in Table 1.

For both the G56 and G60 datasets, the time range for normal scenarios spans from June 10 (or June 1) to September 27, 2023, while the time range for overload scenarios is from September 28 to October 3, 2023 (the Mid-Autumn Festival and National Day in China). The highway network of G56 consists of 110 segment nodes, including 100 highway nodes (10 of which are interweaving area nodes) and 10 local segment nodes, while G60 comprises 143 segment nodes, including 130 highway nodes (13 of which are interweaving area nodes) and 13 local segment nodes. We select all data in the normal scenarios and two days of data in the overload scenarios as the training set, and the rest of the dataset (four days with 1152 time steps) in the overload scenarios as the test set. The historical 60-minute data is used to predict the traffic flow for the next 20, 40, and 60 minutes.

Metrics and Baselines. Following the previous works [Zhou *et al.*, 2024] in traffic flow prediction, Mean Absolute Errors (MAE), Root Mean Squared Errors (RMSE), and Mean Absolute Percentage Errors (MAPE) are adopted to measure the performance of methods. We select nine representative baselines for comparison with HST-WAVE: Historical Average (HA) [Jiang *et al.*, 2021], VAR [Lu *et al.*, 2016], DCRNN [Li *et al.*, 2018], STGCN [Yu *et al.*, 2018], Graph Wavenet (GWNet) [Wu *et al.*, 2019], Trafformer [Jin *et al.*, 2023a], PDFormer [Jiang *et al.*, 2023], STPGNN [Kong *et al.*, 2024], and DCST [Zhou *et al.*, 2024].

Model Settings. We implement our HST-WAVE by using PyTorch Lightning on the NVIDIA GeForce 3090 GPU with 24 GB memory. For the model's hyper-parameters, the hidden dimension d of HSTM is set to 64, the layers of MSWT and CHGAN are set to 2, the head of attention is set to 4, the hop decay factor λ is set to 0.5, the loss weight coefficient μ is set to 0.2, and the size T' of lv is set to 12. We use the Adam optimizer during the model training, the learning rate is set to 0.001, and the batch size is set to 64.

4.2 Comparison Results (RQ1)

Table 2 summarizes the experimental results of the HST-WAVE and baseline models for 20-, 40-, and 60-minute-ahead predictions on the G56 and G60 datasets. The results demonstrate that HST-WAVE consistently outperforms all baseline methods across three evaluation metrics in the HIPO task. Notably, at the 60-minute prediction horizon, HST-WAVE achieves substantial improvements over the best-performing existing methods, with MAE, RMSE, and MAPE improving by 11.1%, 11.0%, and 13.2% on G56, and by 17.7%, 11.5%, and 12.7% on G60, respectively.

HA and VAR perform the worst since they fail to handle complex spatial-temporal data. STPGNN and PDFormer achieve the second-best performance among all baseline

Data	Methods	20 min			40 min			60 min		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
G56	HA *(CIKM'21)*	33.280	44.295	0.183	34.119	45.061	0.194	35.794	46.381	0.213
	VAR *(TIIS'18)*	32.477	43.356	0.180	33.735	44.742	0.192	35.329	45.952	0.208
	DCRNN *(ICLR'18)*	30.681	40.209	0.178	32.324	42.742	0.186	34.357	45.299	0.196
	STGCN *(IJCAI'18)*	29.374	38.738	0.169	31.695	41.786	0.179	33.959	44.952	0.190
	GWNet *(IJCAI'19)*	31.011	40.862	0.179	32.153	42.537	0.185	34.393	45.446	0.196
	Trafformer *(AAAI'23)*	30.962	40.871	0.161	32.270	42.790	0.163	33.985	44.794	0.168
	PDFormer *(AAAI'23)*	30.209	40.257	<u>0.150</u>	31.765	42.261	<u>0.152</u>	33.109	43.958	<u>0.159</u>
	STPGNN *(AAAI'24)*	<u>28.272</u>	<u>37.175</u>	0.185	<u>30.142</u>	<u>39.576</u>	0.199	<u>32.040</u>	<u>41.994</u>	0.213
	DCST *(IJCAI'24)*	29.962	38.197	0.150	31.702	40.100	0.155	33.225	42.238	0.162
	HST-WAVE (ours)	**26.249**	**35.913**	**0.127**	**26.756**	**36.582**	**0.130**	**28.475**	**37.378**	**0.138**
G60	HA *(CIKM'21)*	25.053	35.283	0.201	26.398	36.426	0.218	28.644	38.296	0.228
	VAR *(TIIS'18)*	24.957	34.806	0.199	26.040	36.194	0.208	27.271	37.868	0.213
	DCRNN *(ICLR'18)*	21.144	28.391	0.188	22.069	30.726	0.206	24.153	30.795	0.208
	STGCN *(IJCAI'18)*	20.174	27.884	0.181	21.257	29.294	0.194	23.287	30.062	0.205
	GWNet *(IJCAI'19)*	20.021	27.956	0.183	21.616	29.147	0.194	23.872	30.785	0.206
	Trafformer *(AAAI'23)*	20.818	28.081	0.178	21.964	29.028	0.181	23.179	30.233	0.184
	PDFormer *(AAAI'23)*	20.094	27.414	0.166	21.287	28.504	0.167	22.689	29.283	0.179
	STPGNN *(AAAI'24)*	<u>18.842</u>	<u>24.206</u>	0.199	<u>20.403</u>	<u>27.388</u>	0.204	<u>21.764</u>	<u>28.449</u>	0.223
	DCST *(IJCAI'24)*	19.859	27.622	<u>0.159</u>	21.261	28.434	<u>0.164</u>	22.043	28.831	<u>0.173</u>
	HST-WAVE (ours)	**16.817**	**23.947**	**0.144**	**17.493**	**24.382**	**0.145**	**17.920**	**25.189**	**0.151**

Table 2: Performance comparison evaluated by MAE, MAPE, RMSE (lower is better) on G56 and G60 datasets. The best and second-best performances are highlighted in bold and underlined, respectively.

methods. STPGNN focuses on identifying key nodes in traffic graphs, which helps capture partial interaction information in interweaving areas. On the other hand, PDFormer leverages the relationships between similar nodes and temporal patterns, enabling it to adapt to varying traffic conditions.

HST-WAVE surpasses STPGNN and PDFormer due to its tailored design for the HIPO task. The MSWT module empowers it to address irregular temporal dependencies across diverse traffic conditions, while the CHGAN module equips the model to effectively capture complex transition relationships between road segments. Furthermore, the ATECL module significantly enhances the model's generalization ability in challenging overload scenarios.

4.3 Effectiveness of Component (RQ2)

We conduct ablation studies to evaluate the effectiveness of each component in HST-WAVE. The variants are as follows. 1) w/o HG: this variant replaces the CHGAN with the GCN. 2) w/o MS: this variant removes the multi-scale window convolutional operation. 3) w/o lv: this variant drops the adaptive temporal pattern. 4) w/o CL: this variant does not adopt the ATECL strategy.

Figure 3 presents the comparison of the above variants, which illustrate that: 1) w/o HG performs worse than HST-WAVE, which reflects the necessity of the CHGAN module in capturing the dynamic flow transitions among various type segments. 2) The performance of w/o MS also decreases, which confirms the effectiveness of the multi-scale window modeling mechanism in learning complex traffic patterns. 3) Removing the ATECL strategy (w/o CL) significantly degrades the model performance. This reason is that ATECL contributes to adapting to the irregular temporal patterns and data sparsity in overload scenarios. 4) w/o lv performs better than w/o CL in G56, but the opposite is valid on the G60. This suggests that lv in ATECL can mitigate the noise impact by preserving key temporal pattern knowledge.

4.4 Performance on Interweaving Areas (RQ3)

We verify the performance of the HST-WAVE on interweaving areas under overload scenarios, which are characterized by high traffic flow and frequent traffic behaviors. The results are shown in Figure 4, and we observe the following findings. HST-WAVE significantly outperforms PDFormer and STPGNN across all metrics on both datasets while exhibiting smaller error increments compared to the other methods. Although STPGNN achieves similar error increments to HST-WAVE on G60 (Figure 4 (d), (e), (f)), its overall performance remains substantially inferior to ours. Moreover, unlike STPGNN, our method does not rely on explicitly identifying key nodes to achieve superior performance in interweaving areas. This indicates that HST-WAVE attains high generality by effectively learning the complex and heterogeneous interactions present in interweaving areas.

4.5 Hyper-parameter Sensitivity (RQ4)

We select three important hyper-parameters of the HST-WAVE to analyze the effects on performances, including the weight coefficient μ of contrastive learning loss, the hidden dimension d of the model, and the length of the adaptive temporal pattern lv. Figure 5 shows the trend of MAE under different hyper-parameter settings on both datasets and default hyper-parameters in our experiments are marked by vertical

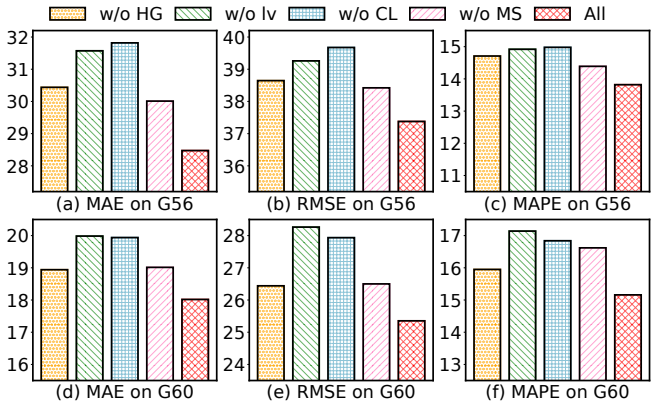

Figure 3: Effectiveness of each component on G56 and G60 for 60-minute-ahead prediction.

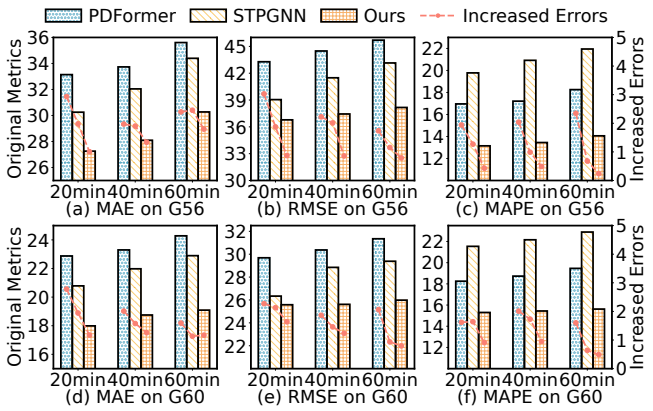

Figure 4: Performance on interweaving areas on the G56 and G60. The bar chart (left axis) shows performance in interweaving areas, and each point in the line chart (right axis) shows the error increment (metric deterioration) in interweaving areas versus all nodes.

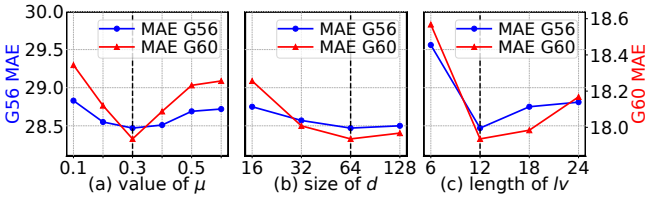

Figure 5: Hyper-parameter analysis on metric MAE on the G56 and G60 datasets for 60-minute-ahead prediction.

dashed lines. We have the following observations. 1) We can obtain the optimal results when setting $\mu = 0.3$, which strikes an effective balance between the traffic flow prediction loss and the enhancement contrastive loss. 2) The model's performance improves steadily as the d increases to 64, but larger dimensions offer no additional benefits and may increase the risk of over-fitting. 3) The optimal length for lv is set to 12, which is sufficient to capture key temporal cues and guide accurate predictions in overload scenarios.

5 Related Work

5.1 Traffic Prediction

Traffic prediction is a critical research area due to its significant role in traffic management and urban planning [Gomes *et al.*, 2023]. Early approaches relied on traditional statistical models like ARIMA [Min and Wynter, 2011] and Bayesian models [Wang *et al.*, 2014] but struggled with tackling complex spatial-temporal dependencies. Deep learning has exhibited superior potential in capturing these dependencies. For spatial modeling, Convolutional Neural Networks (CNNs) [Zhang *et al.*, 2020] and Graph Neural Networks (GNNs), including variations like Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) [Wu *et al.*, 2020; Ye *et al.*, 2021; Song *et al.*, 2022; Ji *et al.*, 2023; Kong *et al.*, 2024], are utilized to effectively represent road network structures and interactions. For temporal dependencies, CNNs and recurrent architectures like LSTM and GRU [Ma *et al.*, 2015] are used to extract sequential features.

Recent advances in traffic forecasting have been marked by the advent of Transformer-based approaches [Liu *et al.*, 2023; Jin *et al.*, 2023a; Jiang *et al.*, 2023; Zhou *et al.*, 2024], such as Trafformer [Jin *et al.*, 2023a], DCST [Zhou *et al.*, 2024], and PDFormer [Jiang *et al.*, 2023]. These approaches have notably enhanced prediction accuracy by simultaneously modeling spatial and temporal dependencies, setting a new benchmark in the field. Other studies [Lu *et al.*, 2018; Wang *et al.*, 2019b] focus on traffic prediction during holidays simply using parameter modeling of temporal patterns, which has difficulty handling complex spatial-temporal correlations, particularly traffic behaviors in interweaving areas.

5.2 Heterogeneous Graph and Time Series Augmentation

Heterogeneous graphs [Sun and Han, 2013] effectively model and integrate diverse types of entities and relationships in complex data, enabling more accurate representation learning. In recent years, attention mechanisms have been widely applied in the research of heterogeneous graph neural networks, such as HAN [Wang *et al.*, 2019a], HGT [Hu *et al.*, 2020], SlotGAT [Zhou *et al.*, 2023].

Time series augmentation techniques [Wen *et al.*, 2021; Zanella *et al.*, 2022; Sarkar *et al.*, 2020; Cheung and Yeung, 2021] aim to enhance model generalization on small or imbalanced datasets. They are typically applied to domain-specific temporal data or regular temporal patterns [Eldele *et al.*, 2021; Franceschi *et al.*, 2019], resulting in suboptimal performance when adapted to the HIPO problem.

6 Conclusion

In this paper, we investigated a novel problem: highway traffic flow prediction under overload scenarios, which is more challenging than ordinary traffic prediction tasks. We proposed a heterogeneous spatial-temporal graph network with adaptive contrastive learning, HST-WAVE, to address this issue. The framework incorporated MSWT and CHGAN modules within HSTM blocks to capture irregular temporal patterns and complex traffic behaviors, while the ATECL strategy improved generalization and mitigated data sparsity. Extensive experimental results demonstrated the effectiveness of our HST-WAVE on the HIPO problem.

Acknowledgments

This work is supported by the National Key R&D Program for the 14th-Five-Year Plan of China (2023YFC3804104 in 2023YFC3804100).

References

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1597–1607, 2020.

[Cheung and Yeung, 2021] Tsz-Him Cheung and Dit-Yan Yeung. MODALS: modality-agnostic automated data augmentation in the latent space. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[Cui *et al.*, 2020] Hua Cui, Gege Yuan, Ni Liu, Mingyuan Xu, and Huansheng Song. Convolutional neural network for recognizing highway traffic congestion. *Journal of Intelligent Transportation Systems*, 24(3):279–289, 2020.

[Eldele *et al.*, 2021] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 2352–2359, 2021.

[Franceschi *et al.*, 2019] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 4652–4663, 2019.

[Gomes *et al.*, 2023] Bernardo Gomes, José Coelho, and Helena Aidos. A survey on traffic flow prediction and classification. *Intelligent Systems with Applications*, 20:200268, 2023.

[Han *et al.*, 2024] Jindong Han, Weijia Zhang, Hao Liu, Tao Tao, Naiqiang Tan, and Hui Xiong. Bigst: Linear complexity spatio-temporal graph neural network for traffic forecasting on large-scale road networks. *Proceedings of the VLDB Endowment*, 17(5):1081–1090, 2024.

[Hu *et al.*, 2020] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the ACM Web Conference 2020*, pages 2704–2710, 2020.

[Ji *et al.*, 2023] Jiahao Ji, Jingyuan Wang, Chao Huang, Junjie Wu, Boren Xu, Zhenhe Wu, Junbo Zhang, and Yu Zheng. Spatio-temporal self-supervised learning for traffic flow prediction. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 4356–4364, 2023.

[Jiang *et al.*, 2021] Renhe Jiang, Du Yin, Zhaonan Wang, Yizhuo Wang, Jiewen Deng, Hangchen Liu, Zekun Cai, Jinliang Deng, Xuan Song, and Ryosuke Shibasaki. Dltraff: Survey and benchmark of deep learning models for urban traffic prediction. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pages 4515–4525, 2021.

[Jiang *et al.*, 2023] 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 37th AAAI Conference on Artificial Intelligence*, pages 4365–4373, 2023.

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

[Jin *et al.*, 2023b] Guangyin Jin, Lingbo Liu, Fuxian Li, and Jincai Huang. Spatio-temporal graph neural point process for traffic congestion event prediction. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 14268–14276, 2023.

[Kong *et al.*, 2024] Weiyang Kong, Ziyu Guo, and Yubao Liu. Spatio-temporal pivotal graph neural networks for traffic flow forecasting. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 8627–8635, 2024.

[Li *et al.*, 2018] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

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

[Lu *et al.*, 2016] Zheng Lu, Chen Zhou, Jing Wu, Hao Jiang, and Songyue Cui. Integrating granger causality and vector auto-regression for traffic prediction of large-scale wlans. *KSII Transactions on Internet and Information Systems*, 10(1):136–151, 2016.

[Lu *et al.*, 2018] Guoming Lu, Jiaxin Li, Jian Chen, Aiguo Chen, Jianbin Gu, and Ruiting Pang. A long-term highway traffic flow prediction method for holiday. In *Advanced Multimedia and Ubiquitous Engineering - MUE/FutureTech 2018*, volume 518, pages 153–159, 2018.

[Ma *et al.*, 2015] Xiaolei Ma, Zhimin Tao, Yinhai Wang, Haiyang Yu, and Yunpeng Wang. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, 54:187–197, 2015.

[Mao *et al.*, 2023] Qiheng Mao, Zemin Liu, Chenghao Liu, and Jianling Sun. Hinormer: Representation learning on heterogeneous information networks with graph transformer. In *Proceedings of the ACM Web Conference 2023*, pages 599–610, 2023.

[Martichoux, 2025] Alix Martichoux. Traffic has gotten way worse in these us cities, report finds. *The Hill*, 2025.

[Min and Wynter, 2011] Wanli Min and Laura Wynter. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies*, 19(4):606–616, 2011.

[Sarkar *et al.*, 2020] Anindya Sarkar, Anirudh Sunder Raj, and Raghu Sesha Iyengar. Neural data augmentation techniques for time series data and its benefits. In *Proceedings of the 19th IEEE International Conference on Machine Learning and Applications*, pages 107–114, 2020.

[Song *et al.*, 2022] Junho Song, Jiwon Son, Dong-Hyuk Seo, Kyungsik Han, Namhyuk Kim, and Sang-Wook Kim. ST-GAT: A spatio-temporal graph attention network for accurate traffic speed prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4500–4504, 2022.

[Sun and Han, 2013] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter*, 14(2):20–28, April 2013.

[Wang *et al.*, 2014] Jian Wang, Wei Deng, and Yuntao Guo. New bayesian combination method for short-term traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 43:79–94, 2014.

[Wang *et al.*, 2019a] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. Heterogeneous graph attention network. In *Proceedings of the World Wide Web Conference 2019*, pages 2022–2032, 2019.

[Wang *et al.*, 2019b] Zhenzhu Wang, Yishuai Chen, Jian Su, Yuchun Guo, Yongxiang Zhao, Weikang Tang, Chao Zeng, and Jingwei Chen. Measurement and prediction of regional traffic volume in holidays. In *2019 IEEE Intelligent Transportation Systems Conference*, pages 486–491, 2019.

[Wen *et al.*, 2021] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 4653–4660, 2021.

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

[Wu *et al.*, 2020] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 753–763, 2020.

[Ye *et al.*, 2021] Junchen Ye, Leilei Sun, Bowen Du, Yanjie Fu, and Hui Xiong. Coupled layer-wise graph convolution for transportation demand prediction. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 4617–4625, 2021.

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

[Zanella *et al.*, 2022] Rodrigo H. Zanella, Lucas A. de Castro Coelho, and Vinícius M. A. de Souza. TS-DENSE: time series data augmentation by subclass clustering. In *Proceedings of the 26th International Conference on Pattern Recognition*, pages 1800–1806, 2022.

[Zhang *et al.*, 2020] Junbo Zhang, Yu Zheng, Junkai Sun, and Dekang Qi. Flow prediction in spatio-temporal networks based on multitask deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):468–478, 2020.

[Zhou *et al.*, 2023] Ziang Zhou, Jieming Shi, Renchi Yang, Yuanhang Zou, and Qing Li. Slotgat: Slot-based message passing for heterogeneous graphs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 42644–42657, 2023.

[Zhou *et al.*, 2024] Yicheng Zhou, Pengfei Wang, Hao Dong, Denghui Zhang, Dingqi Yang, Yanjie Fu, and Pengyang Wang. Make graph neural networks great again: A generic integration paradigm of topology-free patterns for traffic speed prediction. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 2607–2615, 2024.