# Hybrid Spatial-Temporal Graph Neural Networks with Adaptive Graph Framelets for Traffic Forecasting

Jun Xie[1], Chenghao Qiu[1], Ming Li[2,3], Xiaosheng Zhuang[1]
[1]*Department of Mathematics*, *City University of Hong Kong*, Hong Kong, China
[2]*Zhejiang Key Laboratory of Intelligent Education Technology and Application*,
*Zhejiang Normal University*, Jinhua, China
[3] *Zhejiang Institute of Optoelectronics*, Jinhua, China
junxie9@cityu.edu.hk, chenghqiu2-c@my.cityu.edu.hk, mingli@zjnu.edu.cn, xzhuang7@cityu.edu.hk

***Abstract*—Accurate traffic forecasting is crucial for intelligent transportation systems (ITS). Given the complex spatio-temporal dynamics of urban traffic, graph neural networks (GNNs) have been widely adopted to capture spatio-temporal dependencies. However, existing spatio-temporal GNNs (STGNNs) struggle to model heterogeneous temporal patterns—such as long-term trends and short-term fluctuations—and fail to capture multiscale spatial correlations under graph heterophily. To bridge these gaps, we propose the HSTGNN-Framelet, a hybrid STGNN with adaptive graph framelets. Temporally, a discrete wavelet transform (DWT) layer decouples traffic signals into low-frequency trends and high-frequency events, which are subsequently processed by a dual-channel learner employing temporal attention for trends and dilated convolution for events. Spatially, the embeddings of the nodes construct an adaptive adjacency matrix that drives the hierarchical clustering of Paris to generate permutation-equivariant Haar graph framelets. These framelets enable multiscale spectral decomposition, effectively capturing critical high-frequency oscillations typically lost in traditional low-pass graph convolutional networks (GCNs). Experiments on three real-world datasets demonstrate that HSTGNN-Framelet achieves state-of-the-art performance, delivering the best or near-best results across all forecasting horizons and reducing long-term prediction error accumulation. Our results highlight the importance of multiscale feature extraction for characterizing spatio-temporal traffic flow.**

***Index Terms*—Traffic forecasting, graph neural networks, graph framelets, wavelet transform, spatio-temporal modeling, permutation equivariance**

## I. INTRODUCTION

Urban traffic systems generate a huge volume of time-series data from road sensors, providing critical information for traffic management, route planning, and congestion mitigation. Accurate traffic forecasting enables proactive decision-making in intelligent transportation systems (ITS) [1], [2]. Since traffic forecasting is fundamentally a spatio-temporal modeling problem on graphs, graph neural networks (GNNs) have recently achieved remarkable progress in this domain.

Despite significant advances in spatio-temporal graph neural networks (STGNNs) [3]–[5], traffic data exhibit complex patterns in both spatial and temporal domains, which pose severe challenges to GNNs. Temporally, traffic flow exhibits heterogeneous characteristics, including long-term trends (e.g., morning/evening rush hours, weekly cycles) and short-term fluctuations (e.g., random events, accidents). Spatially, the road network forms a non-Euclidean graph structure where traffic states propagate along roads with time-varying connection strengths [6]–[8]. Two fundamental challenges remain inadequately addressed:

**Heterogeneous Temporal Pattern Modeling.** Traffic time series comprise multiple components with distinct frequencies and temporal dynamics. Long-term trends evolve gradually and maintain correlations across distant time steps, while short-term events are volatile and exhibit strong correlations only within consecutive time steps. Existing methods typically process all components through a unified architecture, failing to distinguish their unique contributions, which leads to sub-optimal long-term forecasting performance [9].

**Multiscale Spatial Correlation Capture under Heterophily.** The strength of spatial relationships between road segments changes over time. Moreover, real-world traffic networks exhibit heterophily — neighboring nodes may have significantly different traffic patterns (e.g., a major arterial road adjacent to a side street). Traditional graph convolutional networks act as low-pass filters, smoothing signals across edges and losing critical high-frequency information. While adaptive graph-based methods learn static node embeddings, they cannot capture multiscale spatial dependencies where both low-frequency (smooth) and high-frequency (oscillatory) components coexist.

This paper addresses these challenges by integrating two recent advances: (i) the multi-resolution analysis via graph framelets proposed by Li et al. [10], which provide permutation-equivariant sparse representations of graph signals with provable tight frame properties; and (ii) the hybrid

X. Zhuang was supported in part by the Research Grants Council of Hong Kong (Project no. CityU 11302023, CityU 11301224, and CityU 11300825) and the National Natural Science Foundation of China (NSFC 12471400). M. Li acknowledged the support from the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C03262), and the Jinhua Science and Technology Plan (No. 2023-3-003a).

979-8-3195-1509-4/26/$31.00 ©2026 IEEE
DOI 10.1109/ICKG70389.2026.00033

spatio-temporal architecture from Wang et al. [9], which decouples temporal patterns and utilizes adaptive/dynamic graphs. Our key insight is that graph framelets, originally designed for heterophilous node classification, are naturally suited for traffic forecasting because they can extract multi-scale spatial features while preserving permutation equivariance.

We propose the **Hybrid Spatio-Temporal Graph Neural Network with Adaptive Graph Framelets (HSTGNN-Framelet)**, which includes the following contributions:

1) **Data-Driven Adaptive Graph Framelet Generation:** Learnable node embeddings are leveraged to construct an adaptive adjacency matrix $A_{\text{adp}}$ that captures latent spatial semantics. This matrix serves as input to the Paris hierarchical clustering, generating data-driven multiscale partitions that yield permutation-equivariant Haar graph framelets. To the best of our knowledge, this is the first work to generate graph framelets from learned adaptive graphs for traffic forecasting.
2) **Multi-Resolution Temporal-Spatial Decomposition with Wavelet Transform:** Discrete wavelet transforms (DWT) are applied to decouple both temporal and spatial traffic signals into multi-frequency components. Temporally, the DWT separates trends (low-frequency) from events (high-frequency). Spatially, framelet-based DWT decomposes graph signals across scales for multi-resolution graph learning. The **Bures-Wasserstein discrepancy** is added as a regularization term in the loss function, promoting consistent frequency representations across scales.
3) **Multi-Channel Framelet Feature Fusion:** A comprehensive multi-channel architecture integrates three complementary feature types: (i) original features, (ii) multi-hop spatial aggregations, and (iii) multiscale framelet decompositions of aggregated features. A learnable attention mechanism is then introduced to adaptively fuse these channels.

Extensive experiments on three real-world traffic datasets demonstrate that HSTGNN-Framelet achieves state-of-the-art performance on both short- and long-term forecasting tasks.

## II. RELATED WORK

### A. Traffic Forecasting Methods

Early statistical approaches, including Historical Average (HA), ARIMA [11], and Kalman filtering, assumed linear relationships and stationary distributions, limiting their effectiveness for non-linear, non-stationary traffic data. Machine learning methods, such as Support Vector Regression (SVR) and k-Nearest Neighbors (k-NN), improved forecasting performance but still heavily relied on manual feature engineering.

The advent of deep learning brought significant advances to this domain. Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks [12], and Gated Recurrent Units (GRUs) [13] were introduced to capture temporal dependencies effectively. However, these methods largely ignored the spatial dependencies inherent in road networks.

### B. Graph Neural Networks for Traffic Forecasting

Graph Convolutional Networks (GCNs) [14] extended convolutional operations to graph-structured data. In traffic forecasting, a road network is naturally represented as a graph, with nodes representing sensors and edges representing road connections.

Spatio-Temporal GNNs (STGNNs) combine graph convolutions with temporal modules to capture joint dependencies. For instance, DCRNN [6] models traffic diffusion as a bidirectional random walk, while STGCN [7] integrates graph convolutions with gated temporal convolutions. GWNET [8] introduces adaptive graph learning coupled with dilated causal convolutions, and AGCRN [4] adaptively learns node embeddings to capture node-specific patterns. Furthermore, ASTGCN [3] incorporates spatial and temporal attention mechanisms, PGCN [15] progressively updates the adjacency matrices, and STPGNN [16] identifies pivotal nodes to enhance spatial dependency modeling.

Recognizing that traffic correlations evolve dynamically, several recent works have proposed dynamic graph generation mechanisms. For example, Han et al. [17] learn time-specific adjacency matrices, and Lan et al. [18] compute spatio-temporal-aware distances. However, these methods either fail to capture multiscale spatial dependencies or lack permutation equivariance.

### C. Graph Wavelets and Framelets

Wavelet theory provides powerful multi-resolution analysis tools for decomposing signals at different scales. Classical wavelet transforms have been successfully extended to graphs through both spectral approaches [19] and spatial (Haar-type) approaches [20]–[22]. Li et al. [10] recently proposed binary Haar graph framelets with provable mathematical properties, *permutation equivariance*, *sparsity*, and *efficiency*. These properties are highly desirable for modeling complex traffic data.

While Li et al. [10] focused primarily on heterophilous node classification, we demonstrate that graph framelet transforms are equally valuable for spatio-temporal traffic forecasting. The multiscale spatial decomposition complements traditional multi-hop aggregations, providing a robust alternative to high powers of adjacency matrices, which often lead to dense matrices and oversmoothing. Moreover, we propose generating hierarchical partitions from *learned adaptive graphs*, making the framelet transform data-driven rather than purely topology-based.

## III. PRELIMINARIES

### A. Problem Formulation

A traffic network is represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$, where $\mathcal{V}$ is the vertex set of $N = |\mathcal{V}|$ nodes (sensors), $\mathcal{E}$ is the edge set, and $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix. At each time step $t$, the traffic state is observed as a graph signal $X_t \in \mathbb{R}^{N \times F}$, where $F$ denotes the dimension of node feature.

Given $T$ historical observations $\mathcal{X} = [X_{t-T+1}, \ldots, X_t] \in \mathbb{R}^{T \times N \times F}$, the objective of the traffic forecasting task is to

learn a mapping function $\mathcal{F}$ that predicts the traffic states for the next $S$ future time steps:

$$\tilde{Y} = \mathcal{F}(\mathcal{X}; A, \Theta) \in \mathbb{R}^{S \times N \times F}, \tag{1}$$

where $\Theta$ represents the learnable parameters, and $\tilde{Y} = [\tilde{X}_{t+1}, \ldots, \tilde{X}_{t+S}]$ denotes the predicted multistep traffic.

### B. Binary Haar Graph Framelets

We briefly review the essential theoretical foundations of binary Haar graph framelets introduced by Li et al. [10].

**Hierarchical Clustering:** A $K$-hierarchical clustering $\mathcal{P}_K = \{\mathcal{V}_1, \ldots, \mathcal{V}_K\}$ is a sequence of partitions where each $\mathcal{V}_j := \{s_{\mathbf{\Lambda}} \subset \mathcal{V} : \dim(\mathbf{\Lambda}) = j\}$ partitions $\mathcal{V}$ into distinct clusters, and $\mathcal{V}_{j+1}$ sequentially refines $\mathcal{V}_j$.

**Tight Frame Construction:** Following the parent-children relationship in [10], for each cluster $s_{\mathbf{\Lambda}}$ with $L_{\mathbf{\Lambda}}$ children that encode position information, the scaling vector is defined as $p_{\mathbf{\Lambda}} = \frac{1}{\sqrt{L_{\mathbf{\Lambda}}}}\mathbf{1}$, which implies equal weights, and the framelet matrix $B_{\mathbf{\Lambda}}$ is constructed with rows $w_{\ell_1,\ell_2}$, where $(w_{\ell_1,\ell_2})_t = \frac{1}{\sqrt{L_{\mathbf{\Lambda}}}}$ for $t = \ell_1$, $-\frac{1}{\sqrt{L_{\mathbf{\Lambda}}}}$ for $t = \ell_2$, and 0 otherwise. Consequently, the collection $\mathcal{F}_{j_0}(\mathcal{P}_K)$ from Corollary 1 of [10] forms a tight frame of $L_2(\mathcal{G})$.

### C. Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) decomposes a time series into low-frequency signal approximation and high-frequency detail coefficients [9]. For a level-2 DWT decomposition, the process operates as follows:

- Level 1: The input signal is decomposed into $\{A_1, D_1\}$, where $A_1$ represents the approximation coefficients and $D_1$ denotes the detail coefficients.
- Level 2: The approximation component $A_1$ is further decomposed into $\{A_2, D_2\}$.

## IV. METHODOLOGY

### A. Overall Architecture

Figure 1 illustrates the overall architecture of HSTGNN-Framelet, which consists of four main components:

1) **Temporal Decoupling Layer (TDL):** Decomposes input traffic signals into low-frequency (trend) and high-frequency (event) components using the level-2 discrete wavelet transform (DWT).
2) **Dual-Channel Learner (DCL):** Processes the trend and event components through separate channels: temporal attention is utilized for long-range trends, and dilated temporal convolution is applied for short-range events, which are subsequently integrated via a gated fusion unit.
3) **Hybrid Graph Learning with Framelets (HGL-Framelet):** Extracts multiscale spatial features with dynamic graph framelet transforms (capturing time-varying multiscale structures).
4) **Output Layer:** Integrates multi-layer representations via residual connections to predict the next $S$ future time steps.

### B. Temporal Decoupling Layer

The TDL applies a level-2 Haar DWT along the temporal dimension to decompose the input traffic signals into low-frequency trends and high-frequency events [9]. For each time series, the wavelet decomposition yields approximation and detail coefficients. By zeroing out the high-frequency detail coefficients and applying the inverse transform, we reconstruct the stable trend component $\mathcal{X}_L \in \mathbb{R}^{T \times N \times D}$ for node feature hidden dimension $D$. Conversely, setting the low-frequency approximation coefficients to zero yields the volatile event component $\mathcal{X}_H \in \mathbb{R}^{T \times N \times D}$. These two decoupled components are then projected into a high-dimensional feature space via learnable linear layers.

### C. Dual-Channel Learner

The DCL processes the trend and event streams through specialized dual-channel [9]. The trend channel employs an attention mechanism to capture long-range dependencies effectively:

$$E = \text{Softmax}\left(V \cdot \sigma\big(\text{Conv}_1(X_l)^\top \cdot W \cdot \text{Conv}_2(X_l) + b\big)\right), \tag{2}$$

where $X_l \in \mathbb{R}^{T \times N \times D}$ is the trend representation, $\text{Conv}_1$ and $\text{Conv}_2$ are $1 \times 1$ convolutions that compress the feature and node dimensions respectively, $W \in \mathbb{R}^{N \times D}$, $V, b \in \mathbb{R}^{T \times T}$ are learnable parameters. The Softmax function is applied along the time dimension, yielding temporal attention scores $E \in \mathbb{R}^{T \times T}$.

For the trend representation $\mathcal{X}_L$, we get the output by temporal attention as

$$X_{\text{att}} = \mathcal{X}_L \times_{\text{t}} E, \tag{3}$$

where $\times_{\text{t}}$ denotes tensor–matrix multiplication along the time dimension.

Concurrently, the event channel utilizes dilated causal convolutions to model short-range, volatile temporal patterns:

$$X_{\text{conv}}[t] = \sum_{k=0}^{K_t - 1} f[k] \cdot \mathcal{X}_H[t-k], \tag{4}$$

where $f \in \mathbb{R}^{K_t}$ is the kernel. We stack $L = 4$ layers with $K_t = 4$, yielding a receptive field of $L(K_t - 1) + 1 = 13$ time steps. Finally, a gated fusion mechanism adaptively combines the outputs of both channels:

$$g = \sigma(X_{\text{conv}} W_{tc} + X_{\text{att}} W_{ta} + b_g), \tag{5}$$

$$X_T = g \odot X_{\text{conv}} + (1 - g) \odot X_{\text{att}}, \tag{6}$$

where $W_{tc} \in \mathbb{R}^{D \times D}, W_{ta} \in \mathbb{R}^{D \times D}$ are learnable parameters, $b_g \in \mathbb{R}^D$ is the bias vector, and $\odot$ denotes the Hadamard product.

### D. Data-Driven Graph Framelet Generation

To capture latent spatial semantics beyond the physical topology, we learn node embeddings to construct an adaptive adjacency matrix:

$$A_{\text{adp}} = \text{Softmax}(\text{ReLU}(E_1 E_2^T)), \tag{7}$$

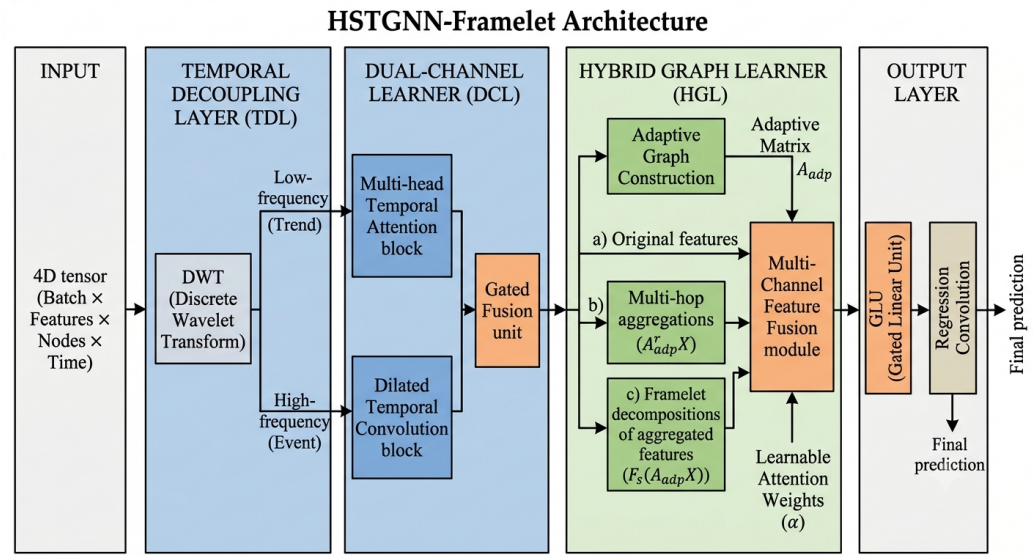


Fig. 1. Overall architecture of HSTGNN-Framelet. TDL decouples trends/events via DWT; DCL processes each frequency component; HGL fuses adaptive graph features (Eq. (7)) and framelet transforms via learnable weights $\alpha$. Training is end-to-end: the residual link (Eq. (9)) and loss (Eq. (11)) backpropagate gradients to jointly update the adaptive graph, attention weights, and all parameters.

where $E_1, E_2 \in \mathbb{R}^{N\times\beta}$ denote learnable node embedding matrices with node hidden dimension $\beta$. Taking $A_{\text{adp}}$ as the input, Paris hierarchical clustering is performed to generate data-driven multiscale partitions with $h := \max_{\dim(\mathbf{\Lambda})>0} L_{\mathbf{\Lambda}}$. Given the binary Haar graph framelet system $\mathcal{F}_{j_0}(\mathcal{P}_K)$, we use $\Phi_j = (\Phi_{\mathbf{\Lambda}})_{\dim(\mathbf{\Lambda})=j} \in \mathbb{R}^{N_j\times N}$ and $\Psi_j = (\Psi_{(\mathbf{\Lambda},m)})_{\dim(\mathbf{\Lambda})=j,m\in[M_{\mathbf{\Lambda}}]} \in \mathbb{R}^{M_j\times N}$ to be the matrix representations of the scaling vectors and framelet vectors at scale $j$, respectively, where $[M_{\mathbf{\Lambda}}] := \{1, 2, \ldots, M_{\mathbf{\Lambda}}\}$. We denote $F_0(M) := \Phi_1^\top \Phi_1 M$ and $F_j(M) := \Psi_j^\top \Psi_j M$, $1 \le j \le K-1$.

### E. Multi-Channel Feature Fusion

To fully exploit the multi-resolution spatial characteristics, we construct three types of complementary feature channels: (i) original features $X_T$; (ii) multi-hop aggregations; (iii) multiscale framelet decompositions of aggregated features $F_j(A_{\text{adp}}X_T)$. These diverse channels are dynamically integrated via a learnable attention-based fusion mechanism:

$$X_{\text{fuse}} = \sum_{i=1}^{n_C} \alpha_i \cdot \mathfrak{N}(X_i W_i), \tag{8}$$

where $\alpha_i$ are trainable attention weights satisfying $\alpha_i \in (0,1)$ and $\sum_i \alpha_i = 1$, $\mathfrak{N}(\cdot)$ is the row normalization operation, and $W_i \in \mathbb{R}^{d_i\times D}$. $X_i \in \mathbb{R}^{T\times N\times d_i}$ is the corresponding channel for each $1 \le i \le n_C$. The total number of nultichannel is defined as $n_C = 1 + r + K$, $\{X_T, A_{\text{adp}}X_T, A^2_{\text{adp}}X_T, ..., A^r_{\text{adp}}X_T, F_0(A_{\text{adp}}X_T), F_1(A_{\text{adp}}X_T), ..., F_{K-1}(A_{\text{adp}}X_T)\}$.

### F. Output layer and loss function

The following equation defines the calculation for the output of the spatio-temporal layers:

$$X_{out} = X_T + X_{\text{fuse}}. \tag{9}$$

Here, $X_{out}$ constitutes the output derived from the hybrid graph layer. Next, multi-step forecasting is realized through the output layer as:

$$\tilde{Y} = F_{\text{out}} X_{\text{out}}, \tag{10}$$

interpreting $F_{\text{out}}$ as a fully connected network.

For optimizing the model, the loss function is formulated by augmenting the mean absolute error $L_{\text{MAE}}$ with the Bures-Wasserstein discrepancy $L_{\text{Dist}}$ [23] for regularization purposes:

$$L = \gamma L_{\text{Dist}} + (1-\gamma) L_{\text{MAE}}, \tag{11}$$

where the weighting factor $\gamma = 0.1$.

## V. EXPERIMENTS

### A. Datasets

We evaluate the proposed HSTGNN-Framelet model on three real-world public traffic datasets: PEMS04, PEMS08, and Urban-core, with their statistics summarized in Table I.

TABLE I
DATASET STATISTICS

| Dataset | Nodes | Interval | Samples | Features |
|---|---|---|---|---|
| Urban-core | 304 | 5 min | 8640 | Speed |
| PEMS04 | 307 | 5 min | 16992 | Speed, Flow, Occupancy |
| PEMS08 | 170 | 5 min | 17856 | Speed, Flow, Occupancy |

All datasets are split into training, validation, and test sets with ratios of 70%, 10%, and 20%, respectively.

TABLE II
REGULAR FORECASTING RESULTS OF DIFFERENT MODELS ON PEMS04, URBAN-CORE, AND PEMS08.

| Data | Models | 15 min | | | 30 min | | | 60 min | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MAE | RMSE | MAPE(%) | MAE | RMSE | MAPE(%) | MAE | RMSE | MAPE(%) | Rank(↑) |
| PEMS04 | STGCN | 18.78 | 29.93 | 13.33 | 19.70 | 31.38 | 13.94 | 21.55 | 33.88 | 13.92 | |
| | DCRNN | 18.79 | 29.65 | 13.16 | 20.05 | 31.51 | 14.08 | 20.19 | **31.64** | 14.24 | |
| | GWNET | 18.17 | 28.98 | 13.14 | 19.18 | 30.54 | 13.72 | 21.04 | 33.14 | 15.12 | |
| | ASTGCN | 19.17 | 30.61 | 12.92 | 20.32 | 32.39 | 13.90 | 23.31 | 36.73 | 16.06 | |
| | AGCRN | 19.11 | 31.40 | 12.65 | 19.84 | 32.71 | 13.10 | 21.46 | 35.38 | 14.08 | |
| | Ada-STNet | 17.81 | 28.68 | **12.38** | 18.66 | 30.02 | 13.09 | 20.06 | 32.01 | 14.29 | |
| | PGCN | 18.02 | 28.85 | 12.96 | 19.02 | 30.35 | 13.68 | 20.76 | 32.70 | 14.96 | |
| | STPGNN | 17.70 | 28.56 | **12.38** | 18.50 | 29.88 | 12.83 | 20.12 | 32.12 | 14.14 | |
| | HSTGNN | 17.67 | **28.55** | 12.77 | **18.43** | **29.73** | 13.51 | **19.90** | 31.75 | 14.38 | |
| | **HSTGNN-Framlet** | **17.66** | 28.64 | 12.45 | 18.50 | 30.05 | **12.82** | 19.92 | 32.14 | **13.74** | |
| Urban-core | STGCN | 2.44 | 3.74 | 9.54 | 2.66 | 3.99 | 10.63 | 2.88 | 4.26 | 11.75 | |
| | DCRNN | 2.43 | 3.73 | 9.53 | 2.65 | 3.99 | 10.68 | 2.84 | 4.23 | 11.62 | |
| | GWNET | 2.40 | 3.70 | 9.31 | 2.60 | 3.93 | 10.38 | 2.75 | 4.12 | 11.17 | |
| | ASTGCN | 2.53 | 3.93 | 9.91 | 2.72 | 4.19 | 10.85 | 2.90 | 4.37 | 11.93 | |
| | AGCRN | 2.48 | 3.81 | 9.76 | 2.64 | 4.00 | 10.57 | 2.84 | 4.25 | 11.52 | |
| | Ada-STNet | 2.40 | 3.70 | 9.36 | 2.61 | 3.94 | 10.43 | 2.78 | 4.15 | 11.28 | |
| | PGCN | 2.41 | 3.72 | 9.37 | 2.63 | 3.98 | 10.46 | 2.83 | 4.21 | 11.44 | |
| | STPGNN | 2.40 | 3.71 | **9.27** | 2.60 | 3.94 | **10.30** | 2.76 | 4.11 | 11.13 | |
| | HSTGNN | **2.39** | **3.69** | 9.29 | **2.58** | **3.92** | 10.32 | 2.73 | 4.10 | **11.06** | |
| | **HSTGNN-Framlet** | 2.42 | 3.70 | 9.50 | 2.60 | **3.92** | 10.40 | **2.72** | **4.07** | 11.09 | |
| PEMS08 | STGCN | 14.55 | 22.72 | 9.38 | 15.62 | 24.55 | 9.84 | 17.86 | 27.56 | 11.14 | |
| | DCRNN | 14.21 | 22.19 | 9.19 | 15.18 | 24.01 | 9.79 | 16.97 | 26.90 | 10.94 | |
| | GWNET | 13.68 | 21.61 | 8.80 | 14.65 | 23.43 | 9.39 | 16.42 | 26.25 | 10.75 | |
| | ASTGCN | 16.37 | 25.31 | 9.98 | 18.33 | 28.28 | 10.97 | 22.44 | 34.11 | 13.24 | |
| | AGCRN | 14.78 | 23.27 | 9.49 | 15.77 | 25.01 | 10.17 | 17.52 | 27.69 | 11.25 | |
| | Ada-STNet | 13.48 | 21.51 | 8.78 | 14.33 | 23.24 | 9.36 | 15.69 | 25.51 | **10.25** | |
| | PGCN | 13.44 | 21.53 | 8.84 | 14.25 | 23.12 | 9.42 | 15.54 | 25.31 | 10.28 | |
| | STPGNN | 13.26 | 21.47 | 8.77 | 14.14 | 23.39 | 9.16 | 15.70 | 25.82 | 10.27 | **3rd** |
| | HSTGNN | 13.11 | **21.36** | **8.51** | 13.94 | 23.09 | 9.16 | 15.47 | 25.50 | 10.40 | **2nd** |
| | **HSTGNN-Framlet** | **12.95** | 21.39 | 8.67 | **13.70** | **22.96** | **9.14** | **15.08** | **25.11** | 10.29 | **1st** |

### *B. Baseline Methods*

To verify the effectiveness of HSTGNN-Framelet, we compare it against a wide range of baselines, including: Historical Average (HA), ARIMA [11], STGCN [7], DCRNN [6], GWNET [8], ASTGCN [3], AGCRN [4], Ada-STNet [5], PGCN [15], and STPGNN [16].

### *C. Implementation Details*

The model is implemented with a batch size of 64 and trained with the Adam optimizer at a learning rate of 0.001. $r = 5$ is the number of multi-hop aggregation steps, and $h = 8$. A level-2 decomposition is applied in the TDL. Three widely adopted metrics are utilized for evaluation: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

### *D. Main Results and Discussion*

Table II presents the evaluation results for short-term predictions (15, 30, and 60 minutes, corresponding to Horizons 3, 6, and 12). The key findings are summarized as follows:

- **HSTGNN-Framelet achieves state-of-the-art performance** on PEMS08 and Urban-core across all horizons, while maintaining highly competitive results on PEMS04. For instance, on PEMS08 (60 min), our model achieves an MAE of 15.08, outperforming HSTGNN (15.47), PGCN (15.54), and Ada-STNet (15.69) by 2.5%, 3.0%, and 3.9%, respectively. On Urban-core (60 min), our model yields consistent improvements, reducing the MAE to 2.72 compared to HSTGNN (2.73) and GWNET (2.75).
- **Adaptive graph approaches consistently outperform static methods.** Models with adaptive or dynamic graph generation (e.g., Ada-STNet, PGCN, STPGNN, and HSTGNN-Framelet) demonstrate a clear advantage over static structural approaches (e.g., STGCN, DCRNN, and ASTGCN). Specifically, on PEMS08 (60 min), HSTGNN-Framelet improves upon STGCN (17.86) by 15.6% and GWNET (16.42) by 8.2% in terms of MAE.
- **Framelet transforms yield vital complementary information.** Compared to its vanilla backbone HSTGNN (which lacks framelet layers), HSTGNN-Framelet secures steady improvements on PEMS08 (15.08 vs. 15.47) and Urban-core (2.72 vs. 2.73). On PEMS04, although the MAE is comparable (19.92 vs. 19.90), our model substantially improves prediction accuracy, with MAPE (13.74% vs. 14.38%).
- **The dual-channel temporal decoupling layer is highly effective.** On PEMS04 at the 60-minute horizon, HSTGNN-Framelet achieves the lowest MAPE (13.74%) compared to HSTGNN (14.38%) and STPGNN (14.14%), demonstrating the distinct advantages of separating low-frequency global trends from high-frequency localized events.

Table III evaluates the long-term forecasting capabilities

TABLE III
LONG-TERM FORECASTING RESULTS OF DIFFERENT MODELS ON PEMS08 AND URBAN-CORE.

| Data | Models | 3 hours | | | 4 hours | | | 5 hours | | | Rank(↑) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MAE | RMSE | MAPE(%) | MAE | RMSE | MAPE(%) | MAE | RMSE | MAPE(%) | |
| Urban-core | STGCN | 2.92 | 4.31 | 12.06 | 2.94 | 4.34 | 12.18 | 3.04 | 4.47 | 12.71 | |
| | DCRNN | 2.88 | 4.30 | 11.80 | 2.90 | 4.31 | 11.90 | 2.95 | 4.37 | 12.16 | |
| | GWNET | 2.90 | 4.33 | 11.79 | 2.91 | 4.34 | 11.84 | 2.97 | 4.41 | 12.19 | |
| | ASTGCN | 3.27 | 4.72 | 13.65 | 3.36 | 4.83 | 14.12 | 3.58 | 5.06 | 15.34 | |
| | AGCRN | 3.02 | 4.44 | 12.39 | 3.05 | 4.47 | 12.47 | 3.10 | 4.51 | 12.73 | |
| | Ada-STNet | 2.97 | 4.37 | 12.28 | 2.99 | 4.39 | 12.37 | 3.06 | 4.47 | 12.68 | |
| | PGCN | 2.85 | 4.23 | 11.79 | 2.87 | 4.25 | 11.80 | 2.97 | 4.37 | 12.21 | |
| | STPGNN | 2.83 | 4.19 | 11.77 | 2.86 | 4.20 | 11.81 | 2.98 | 4.38 | 12.24 | |
| | HSTGNN | **2.79** | **4.13** | 11.63 | **2.83** | **4.15** | 11.65 | 2.95 | 4.33 | 12.07 | |
| | **HSTGNN-Framlet** | 2.82 | 4.21 | **11.49** | **2.83** | 4.22 | **11.60** | **2.85** | **4.23** | **11.60** | |
| PEMS08 | STGCN | 21.32 | 33.03 | 14.70 | 23.05 | 35.35 | 15.69 | 24.88 | 37.88 | 17.19 | |
| | DCRNN | 20.48 | 31.73 | 13.73 | 21.38 | 33.03 | 14.56 | 23.12 | 35.22 | 16.35 | |
| | GWNET | 21.27 | 35.06 | 13.75 | 21.54 | 35.76 | 13.96 | 22.01 | 36.38 | 14.45 | |
| | ASTGCN | 29.32 | 42.91 | 19.76 | 31.63 | 46.27 | 21.19 | 36.10 | 52.47 | 25.35 | |
| | AGCRN | 21.29 | 33.84 | 14.50 | 22.53 | 35.78 | 15.52 | 24.73 | 38.79 | 17.15 | |
| | Ada-STNet | 19.52 | 31.30 | 12.77 | 20.81 | 32.92 | 13.89 | 22.30 | 34.71 | 14.97 | |
| | PGCN | 19.82 | 31.99 | 13.42 | 20.86 | 33.60 | 14.17 | 22.30 | 35.59 | 15.66 | |
| | STPGNN | 19.56 | 31.79 | 13.03 | 20.46 | 33.29 | 13.93 | 21.92 | 35.21 | 15.17 | **3rd** |
| | HSTGNN | 18.09 | 29.47 | 12.81 | 18.53 | 30.33 | 13.34 | 19.37 | 31.29 | 14.55 | **2nd** |
| | **HSTGNN-Framlet** | **17.26** | **28.75** | **12.01** | **17.67** | **29.41** | **12.44** | **18.84** | **30.81** | **13.69** | **1st** |

for extended horizons up to 5 hours (Horizons 36, 48, and 60). HSTGNN-Framelet demonstrates remarkable robustness against error accumulation:

- On PEMS08 (5-hour horizon), HSTGNN-Framelet achieves an MAE of 18.84, delivering performance gains of 2.7%, 14.1%, and 15.5% over HSTGNN (19.37), STPGNN (21.92), and PGCN/Ada-STNet (22.30), respectively. This underscores the power of multiscale framelet spectral decomposition in modeling long-range dependencies.
- **Long-term error accumulation is significantly mitigated.** Evaluating the error growth from 3 hours to 5 hours on PEMS08, HSTGNN-Framelet exhibits the slowest degradation (+9.2%: 17.26→18.84) compared to PGCN (+12.5%: 19.82→22.30) and Ada-STNet (+14.2%: 19.52→22.30), showing a strong capability to restrict multi-step error propagation.
- On Urban-core (5-hour horizon), our model achieves an MAE of 2.85, outperforming HSTGNN/DCRNN (2.95) and PGCN/GWNET (2.97). The error growth from 3 hours to 5 hours is exceptionally low (+1.1%: 2.82→2.85), compared to HSTGNN (+5.7%: 2.79→2.95), validating the stability provided by the trend-event temporal decoupling scheme.

### *E. Ablation Study*

To quantitatively evaluate the contribution of each core component, we conduct ablation studies on PEMS08 and Urban-core at the 60-minute forecasting horizon. Three ablated variants are evaluated: (i) *w/o Adaptive Graph* (using fixed topology for framelets), (ii) *w/o Multi-Channel Fusion* (using single input features), and (iii) *w/o DistDF Loss* (removing the structural regularization loss). Table IV reveals the following insights:

TABLE IV
ABLATION STUDY RESULTS ON PEMS08 AND URBAN-CORE (60 MIN)

| Data | Models | 60 min | | |
|---|---|---|---|---|
| | | MAE | RMSE | MAPE(%) |
| PEMS08 | HSTGNN-Framelet | 15.08 | 25.11 | 10.29 |
| | w/o Adaptive Graph | 15.58 | 25.43 | 10.59 |
| | w/o Multi-Channel | 15.72 | 25.59 | 10.74 |
| | w/o $L_{\text{Dist}}$ | 15.69 | 25.53 | 10.66 |
| Urban-core | HSTGNN-Framelet | 2.72 | 4.07 | 11.09 |
| | w/o Adaptive Graph | 2.79 | 4.15 | 11.45 |
| | w/o Multi-Channel | 2.81 | 4.16 | 11.41 |
| | w/o $L_{\text{Dist}}$ | 2.79 | 4.19 | 11.38 |

- **Removing the multi-channel fusion mechanism causes the most severe performance degradation**, leading to an average MAE increase of 3.8% across datasets (4.2% on PEMS08 and 3.3% on Urban-core). This confirms that integrating original features, multi-hop structural aggregations, and multiscale framelet reconstructions is critical for robust representation.
- **The adaptive graph framelet is essential for capturing spatial heterophily.** Replacing the adaptive data-driven graph with a fixed structure degrades performance by 3.3% on PEMS08 and 2.5% on Urban-core to MAE, validating that learned hierarchical partitions fit dataset-specific profiles better than predefined topologies.
- **The Bures-Wasserstein discrepancy ensures stable spectral decomposition.** Eliminating the Bures-Wasserstein discrepancy regularization results in an average performance drop of 3.3% (4.0% on PEMS08 and 2.6% on Urban-core) to MAE, demonstrating its critical role in regularizing multiscale framelet representations.

## VI. CONCLUSION

This paper introduced HSTGNN-Framelet, a novel hybrid spatio-temporal graph neural network architecture that

embeds permutation-equivariant Haar graph framelets into a temporally decoupled dual-channel framework. Extensive evaluations on three real-world datasets demonstrate that our model establishes new state-of-the-art benchmarks for both short- and long-term traffic forecasting. On PEMS08 (60 min), HSTGNN-Framelet achieves an MAE of 15.08, surpassing the vanilla HSTGNN by 2.5% and static methods (STGCN) by 15.6%. For long-term 5-hour predictions, its advantage further expands, achieving an MAE of 18.84 and demonstrating the slowest error propagation rate among all compared models. Quantitative ablation experiments further validate the collective and individual contributions of temporal decoupling, adaptive framelet generation, and multi-channel feature fusion.

**Limitations and Future Work:** Despite its superior accuracy, HSTGNN-Framelet exhibits certain limitations that open up directions for future research: (i) The integration of DWT and multiscale framelet decompositions introduces higher computational costs during training compared to lightweight models like STGCN. We plan to explore lightweight coefficient pruning and sparse matrix operations; (ii) the Paris clustering algorithm requires empirical tuning of hierarchical partition depths. Future work will investigate adaptive wavelet selection via learnable filter banks and self-configuring graph rewiring methods. However, our evaluation focuses on moderate-scale graphs. The generalizability to larger graphs (e.g., thousands of nodes) or those with varying sparsity remains unexplored. Future work will validate on diverse benchmarks (e.g., METR-LA, NYC Taxi) to assess robustness across varied topologies.

## REFERENCES


[1] R.-G. Cirstea, T. Kieu, C. Guo, B. Yang, and S. J. Pan, "Enhancenet: Plugin neural networks for enhancing correlated time series forecasting," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 1739–1750.

[2] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.

[3] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 922–929.

[4] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," *Advances in neural information processing systems*, vol. 33, pp. 17 804–17 815, 2020.

[5] X. Ta, Z. Liu, X. Hu, L. Yu, L. Sun, and B. Du, "Adaptive spatio-temporal graph neural network for traffic forecasting," *Knowledge-based systems*, vol. 242, p. 108199, 2022.

[6] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *International Conference on Learning Representations*, 2018.

[7] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 3634–3640.

[8] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling," *arXiv preprint arXiv:1906.00121*, 2019.

[9] P. Wang, L. Feng, Y. Zhu, and H. Wu, "Hybrid spatial–temporal graph neural network for traffic forecasting," *Information Fusion*, vol. 118, p. 102978, 2025.

[10] J. Li, R. Zheng, H. Feng, M. Li, and X. Zhuang, "Permutation equivariant graph framelets for heterophilous graph learning," *IEEE Transactions on neural networks and learning systems*, vol. 35, no. 9, pp. 11 634–11 648, 2024.

[11] H. Yang, X. Li, W. Qiang, Y. Zhao, W. Zhang, and C. Tang, "A network traffic forecasting method based on sa optimized arima–bp neural network," *Computer Networks*, vol. 193, p. 108102, 2021.

[12] Y. Liu, H. Zheng, X. Feng, and Z. Chen, "Short-term traffic flow prediction with conv-lstm," in *2017 9th international conference on wireless communications and signal processing (WCSP)*. IEEE, 2017, pp. 1–6.

[13] W. Shu, K. Cai, and N. N. Xiong, "A short-term traffic flow prediction model based on an improved gate recurrent unit neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 16 654–16 665, 2021.

[14] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[15] Y. Shin and Y. Yoon, "Pgcn: Progressive graph convolutional networks for spatial–temporal traffic forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 7633–7644, 2024.

[16] W. Kong, Z. Guo, and Y. Liu, "Spatio-temporal pivotal graph neural networks for traffic flow forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 8, 2024, pp. 8627–8635.

[17] L. Han, B. Du, L. Sun, Y. Fu, Y. Lv, and H. Xiong, "Dynamic and multi-faceted spatio-temporal deep learning for traffic speed forecasting," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 547–555.

[18] S. Lan, Y. Ma, W. Huang, W. Wang, H. Yang, and P. Li, "Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting," in *International conference on machine learning*. PMLR, 2022, pp. 11 906–11 917.

[19] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and computational harmonic analysis*, vol. 30, no. 2, pp. 129–150, 2011.

[20] M. Gavish, B. Nadler, and R. R. Coifman, "Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning," in *27th International Conference on Machine Learning (ICML)*, 2010, pp. 367–374.

[21] M. Li, Z. Ma, Y. G. Wang, and X. Zhuang, "Fast haar transforms for graph neural networks," *Neural Networks*, vol. 128, pp. 188–198, 2020.

[22] X. Zheng, B. Zhou, Y. G. Wang, and X. Zhuang, "Decimated framelet system on graphs and fast g-framelet transforms," *Journal of Machine Learning Research*, vol. 23, no. 18, pp. 1–68, 2022.

[23] H. Wang, L. Pan, Y. Lu, Z. Chu, X. Li, S. He, Z. Chen, H. Li, Q. Wen, and Z. Lin, "Distdf: Time-series forecasting needs joint-distribution wasserstein alignment," in *International Conference on Learning Representations*, 2026.