ITC 3/55
Information Technology and Control
Vol. 55 / No. 3/ 2026
pp. 975-995
DOI 10.5755/j01.itc.55.3.44948

A Mode-Decomposition-Based Semantic Fusion Network with Graph-Guided Spatial Attention for Urban Load Demand Forecasting

Received 2026/05/05 | Accepted after revision 2026/07/20

HOW TO CITE: Huang, X., Hu, K., Hu, Z., Yu, C. (2026). A Mode-Decomposition-Based Semantic Fusion Network with Graph-Guided Spatial Attention for Urban Load Demand Forecasting. *Information Technology and Control*, 55(3), 975-995. https://doi.org/10.5755/j01.itc.55.3.44948

A Mode-Decomposition-Based Semantic Fusion Network with Graph-Guided Spatial Attention for Urban Load Demand Forecasting

Xiaoyu Huang
School of Electrical and Information Engineering, Tianjin University, Tianjin, 300072, China;
e-mail: xiaoyv_h19@tju.edu.cn

Kunshu Hu
School of Electrical and Information Engineering, Tianjin University, Tianjin, 300072, China;
e-mail: kunshu_hu2019@tju.edu.cn

Zexi Hu
Foshan Nanhai Power Supply Bureau of Guangdong Power Grid Co., Ltd., Foshan 528000, China;
e-mail: huzexi1127@163.com

Chenyue Yu*
National University of Singapore, Singapore; e-mail: e1143627@u.nus.edu

Corresponding author: e1143627@u.nus.edu

Accurate urban load demand forecasting is critical for the economic dispatch and reliable operation of power systems. However, urban load forecasting remains challenging because forecasting models must extract reliable representations from entangled multi-scale dynamics, temporal semantic dependencies, and complex spatiotemporal correlations. To address these issues, this paper proposes a mode-decomposition-based semantic fusion network with graph-guided spatial attention (MDSFGA) for urban load demand forecasting. Specifically, a multi-scale spatiotemporal encoder is developed to decompose historical observations into globally frequency-consistent mode components across spatial nodes and to construct corresponding spatiotemporal representations through branch-wise feature extraction and an adaptive mode feature fusion

module. Within this encoder, a semantic-aware temporal extractor incorporates temporal semantic context into recurrent dependency modeling. At the same time, a graph-guided spatial attention module combines mode-specific graph priors with spatial self-attention to jointly model local structural constraints and global spatial dependencies. Experiments on four public load datasets demonstrate that MDSFGA outperforms representative baselines. Notably, MDSFGA achieves a MAPE of 15.65% on the CELP-VA dataset, with a 7.83% improvement compared with the second-best baseline.

KEYWORDS: Urban load demand forecasting; Spatiotemporal representation learning; Multivariate variational mode decomposition; Graph-guided spatial attention

Introduction

With the advancement of smart grids, load demand forecasting has garnered significant attention in recent years [45]. Load demand forecasting reflects variations in power supply and demand and facilitates the formulation of rational generation and transmission plans, enabling more efficient real-time power dispatch [43]. Accurate load demand forecasting helps ensure the stability and reliability of power systems while reducing generation costs.

Early studies mainly relied on traditional statistical and machine learning approaches [7, 14, 18]. However, these methods are typically based on linear assumptions or require handcrafted features, which may constrain their forecasting accuracy. In contrast, deep learning models can automatically extract salient features from raw data, avoiding potential biases introduced by manual feature engineering [1, 33]. In practice, electricity consumption nodes within the same community often exhibit similar load patterns due to shared weather conditions, geographical proximity, holiday schedules, and electricity pricing policies [45]. Motivated by this observation, some studies have combined convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to capture spatiotemporal dependency patterns in the data [15, 21, 36]. However, electricity consumption nodes in urban systems are typically distributed across irregular spatial layouts rather than regular grids, making CNN-based approaches less effective for modeling such spatial relationships. Consequently, subsequent studies have integrated temporal modeling methods with graph neural networks (GNNs), leading to the emergence of spatiotemporal graph neural networks (STGNNs) [2, 20, 39]. These models capture spatiotemporal dynamics through message-passing mechanisms. In addition, some studies have employed attention mechanisms to model spatial relationships [22, 23]. Collectively, these approaches can better exploit spatiotemporal dependencies, enabling more precise forecasting. Despite the notable progress of recent methods, the following three challenges remain:

First, urban load series usually exhibit long-term trends, periodic variations, and high-frequency fluctuations simultaneously. Since these patterns are entangled across different temporal scales, direct encoding of raw sequences often makes it difficult for deep models to identify the dominant dynamics [47]. Existing studies have attempted to alleviate this issue by constructing multi-scale representations in the time domain [25, 27, 35, 40] or by introducing time-frequency analysis techniques [11, 38, 46]. However, these methods are primarily designed for individual spatial nodes and are less suitable for urban systems characterized by coupled multi-node dynamics.

Second, urban load demand is not determined solely by historical observations. Temporal semantic factors, such as time of day and day of the week, play an important role in reflecting load dynamics. Relying only on historical observations makes it difficult to adequately characterize regular patterns such as intraday peaks and weekday-weekend differences [31]. Although many temporal modeling methods learn sequential dependencies effectively, the explicit integration of temporal semantic context into dependency modeling remains insufficiently explored.

Third, spatial correlations among nodes are highly complex [32]. For instance, the temporary closure of a commercial complex may lead to increased electricity demand in proximal and functionally related stores. Recent studies have adopted dynamic graph-based GNN approaches [17, 19] or spatial attention mechanisms [22, 23]. However, GNN-based meth-

ods tend to rely heavily on local graph structures and may overlook global correlations, whereas attention-based methods may aggregate information from irrelevant nodes. Balancing local structural constraints with global dependency aggregation remains a challenging problem.

To address the above issues, this paper proposes a mode-decomposition-based semantic fusion network with graph-guided spatial attention (MDSFGA) for urban load demand forecasting. The proposed model is built upon an encoder–decoder architecture and is designed to learn robust spatiotemporal representations for direct multi-step forecasting. Specifically, MDSFGA incorporates a multi-scale spatiotemporal encoder (MSTE), which employs multivariate variational mode decomposition (MVMD) to decouple historical observations into multi-scale spatiotemporal mode components with globally consistent frequencies across spatial nodes. By combining a multi-branch network with an adaptive mode feature fusion module, MSTE mitigates representation ambiguity caused by entangled multi-scale patterns and provides multi-scale features for prediction. Within this encoder, each branch consists of a cascaded semantic-aware temporal extractor (SATE) and a graph-guided spatial attention (GSA) module. SATE incorporates temporal semantic information into temporal dependency modeling through a fusion-gated recurrent unit, enhancing the representation of intraday and intraweek load regularities. GSA combines correlation graph priors derived from mode components with spatial self-attention to jointly model local structural constraints and global dependency aggregation. The main contributions are summarized as follows:

- We address the limitations of node-wise multi-scale modeling by developing a globally consistent spatiotemporal encoding scheme that decouples entangled dynamics across spatial nodes.
- We design a semantic-aware temporal extractor with fusion-gated recurrent units that explicitly incorporates time-of-day and day-of-week cues into temporal dependency learning.
- We propose a graph-guided spatial attention module that integrates mode-specific correlation priors with spatial self-attention to jointly model local structural constraints and global spatial dependencies.

2. Related Work

Load demand forecasting has been a subject of extensive research over the past decades. Early studies typically relied on traditional statistical or machine learning methods, such as the Auto-Regressive Integrated Moving Average (ARIMA) model [18], Vector Auto-Regression (VAR) [14], or Support Vector Regression (SVR) [7]. These approaches generally depend on linear assumptions or handcrafted feature engineering, which limit their ability to capture the intrinsic nonlinear characteristics of the data [28]. In contrast to conventional machine learning methods, deep learning models can automatically extract complex features from large-scale datasets, demonstrating superior representation and modeling capabilities [1]. For temporal dependency modeling, deep learning has primarily evolved along three representative technical paradigms. The first is based on Temporal Convolutional Networks (TCNs) [3], which utilize dilated causal convolutions to expand the receptive field and efficiently capture local temporal patterns. The second paradigm consists of Recurrent Neural Networks (RNNs) [21], which model sequential dependencies through recurrent mechanisms. The third paradigm is built upon self-attention architectures, typified by the Transformer [33], which explores global temporal dependencies and demonstrates pronounced advantages in modeling long-range interactions.

As urban load infrastructures represent complex spatiotemporal systems composed of multiple functional regions with strong inter-regional interactions, conventional models often struggle to characterize their underlying patterns. Consequently, recent research on urban load forecasting has gradually shifted from purely temporal modeling toward the joint exploration of multi-scale temporal dynamics and spatiotemporal dependencies.

2.1. Temporal Multi-scale Modeling

The evolution patterns of spatiotemporal data are inherently multi-scale. While fine-grained scales capture transient fluctuations within the system, coarse-grained scales are more effective in uncovering macroscopic periodicities or long-term trends. Consequently, an in-depth exploration of these multi-scale dynamic characteristics is widely regarded as a critical avenue for enhancing forecast-

ing performance [12]. Current methods typically process raw sequences directly in the time domain to construct multi-scale representations. Specifically, MICN [35] employs a multi-scale average pooling module to separate trend and seasonal components. Building upon similar decomposition principles, LSTNN [34] introduces down-sampling and patched sampling strategies to fit these components separately, enhancing accuracy by modeling them independently before aggregation in the output layer. To further leverage multi-resolution information, TimeMixer [37] utilizes bottom-up and top-down mixing strategies on subseries. Other approaches, such as PatchTST [25], segment time series into local patches and leverage temporal attention to explore inter-patch correlations. Furthermore, ST-Norm [8] facilitates feature extraction by integrating temporal and spatial normalization into deep residual architectures, explicitly isolating high-frequency signals and local spatial variations.

Beyond time-domain processing, other approaches project time series into the frequency domain to reveal their inherent multi-scale structures. For instance, TimesNet [38] reconfigures one-dimensional sequences into two-dimensional tensors via the Discrete Fourier Transform (DFT), capturing variations both within and across periods. StemGNN [4] combines the Graph Fourier Transform with the DFT to capture spatial correlations and periodic features simultaneously. To better handle non-stationary signals, STWave [11] utilizes the multi-level Discrete Wavelet Transform (DWT) to decouple series into trend and event components, while WPMixer [24] assigns these wavelet components to resolution-specific branches. Moving beyond fixed-basis transforms, Variational Mode Decomposition (VMD) [10] offers an adaptive mechanism to decompose series into multiple mode components across different frequency bands. Recent works like TF-VMD [6] and VMD-ConvTSMixer [46] have successfully combined VMD with self-attention and multi-scale convolutions to achieve fine-grained spatiotemporal feature extraction.

Despite these advancements, two limitations persist in the context of urban load forecasting. First, most decomposition-based methods decouple each node independently, which may lead to inconsistent frequency representations across the urban system, weakening system-level spatiotemporal modeling. Second, existing methods primarily focus on pattern decomposition or frequency transformation, while the explicit integration of temporal semantic context into dependency modeling remains relatively underexplored. Given that urban load dynamics are heavily influenced by semantic cues, there is a clear need to jointly consider globally consistent multi-scale extraction and semantic-aware temporal modeling.

2.2. Spatiotemporal Dependency Modeling

Spatial dependency refers to the intricate interdependencies among disparate spatial nodes. Traditional studies characterize these relationships using predefined graph structures, typically constructed based on geographical distance [42] or Dynamic Time Warping (DTW) similarity [13] to capture static spatial correlations. To move beyond fixed topologies, data-driven approaches have been developed to infer latent spatial dependencies during training. For instance, AGCRN [2] generates node-adaptive parameters to construct customized graphs, while GWNet [39] integrates a learnable adaptive adjacency matrix with bidirectional diffusion graph convolutions and stacked dilated causal convolutions to jointly model spatial and temporal dynamics.

However, urban systems are inherently dynamic, where transient events can sharply escalate local loads and alter the underlying patterns of the electricity supply network. Static graphs often lack the flexibility to capture such evolving spatiotemporal relationships. To circumvent the rigidity of static representations, recent research has transitioned toward dynamic graph modeling. DGCRN [19] incorporates a hypernetwork into a Gated Recurrent Unit (GRU) to generate time-varying adjacency matrices, which are fused with static graphs for fine-grained modeling. Similarly, Auto-DSTSGN [17] utilizes an automated search mechanism to identify optimal combinations of spatial graphs, while GDGCN [41] employs a tensor-based constructor to learn interval-specific dependencies. Furthermore, ASTGCN [16] leverages a dot-product attention mechanism to compute spatial correlation matrices from input features, dynamically adjusting the adjacency matrix within the graph convolution framework.

Despite these advancements, significant challenges persist. Message-passing-based GNNs typically rely

on local graph structures and often require multiple stacked layers to expand the receptive field, which increases computational overhead and may lead to over-smoothing issues [9]. In contrast, attention mechanisms provide a natural global receptive field, enabling the effective capture of long-range dependencies. Consequently, these mechanisms have been widely adopted to learn complex spatial correlations, as seen in architectures like STAEformer [22] and iTransformer [23], which apply Transformer layers across the spatial dimension. However, purely attention-based methods may inadvertently aggregate information from irrelevant nodes. This underscores the need for a mechanism capable of jointly exploiting local constraints and global dependency aggregation for accurate urban load forecasting.

3. Problem Definition

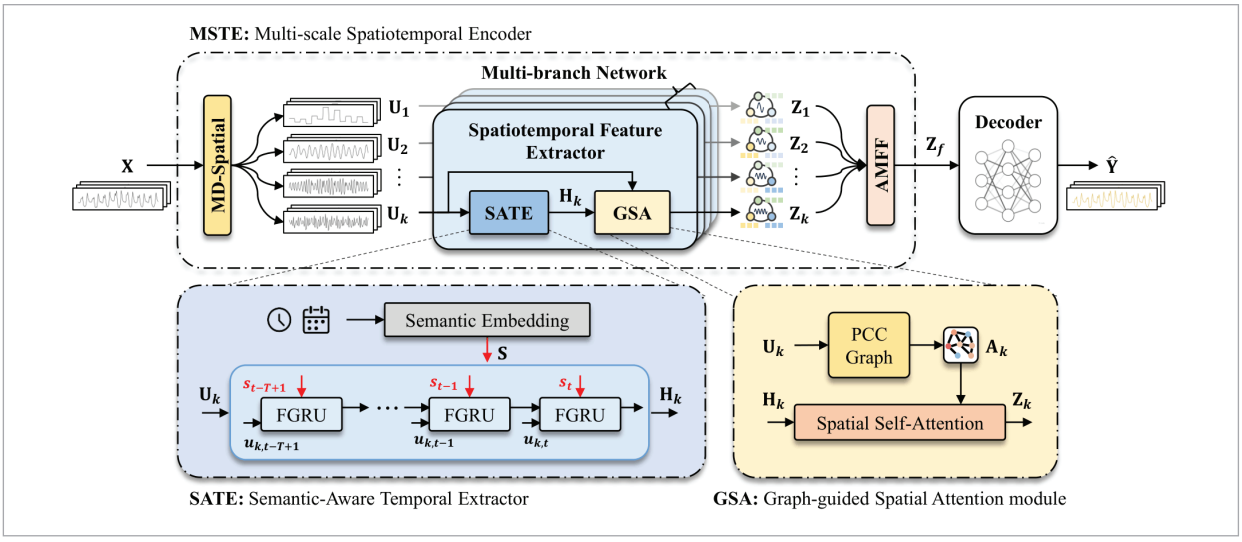
An urban load system comprising multiple energy-consuming nodes can be characterized as a spatiotemporal system. Given a historical input window, we denote the multi-node load observation sequence as $X \in R^{T_h \times N}$, where T_h is the look-back horizon and N represents the number of spatial nodes. In addition to the historical load values, we introduce the corresponding temporal semantic information $S \in R^{T_h \times N \times d_s}$ to describe the semantic context associated with the input window. In this work, S is derived from timestamps which include hour-of-day and day-of-week information to represent intraday and intraweek regularities.

The urban load demand forecasting task is formulated as learning a mapping function $f_\theta : (X, S) \rightarrow Y$, where $Y \in R^{T_f \times N}$ denotes the predicted future load demand over the forecasting horizon T_f.

4. Methodology

As illustrated in Figure 1, MDSFGA employs an encoder-decoder architecture comprising a multi-scale spatiotemporal encoder (MSTE) and a lightweight MLP-based decoder that maps integrated representations to multi-step outputs. The MSTE serves as the core representation learning module and is structured into three sequential procedures. First, it applies mode decomposition with globally consistent frequencies across spatial nodes to decouple the historical observations X into multi-scale spatiotemporal mode components $\{U\}_{k=1}^{K}$, mitigating the representation ambiguity arising from entangled multi-scale dynamics. Second, these mode components are processed by a multi-branch spatiotemporal feature extraction architecture. Within each branch, a cascaded semantic-aware temporal extractor (SATE) and a graph-guided spatial attention (GSA) module function as the core feature

Figure 1
The overall framework of the proposed MDSFGA.



extractors, facilitating the integration of temporal semantic context into dependency modeling while jointly capturing local structural constraints and global spatial dependencies. Third, the branch-wise features $\{Z\}_{k=1}^{K}$ are adaptively aggregated by the adaptive mode feature fusion (AMFF) module, which prevents the dominance of any single mode and generates a robust spatiotemporal representation. Ultimately, the spatiotemporal features Z_f are decoded into the final prediction results through a lightweight MLP decoder.

4.1. Multi-scale Spatiotemporal Encoder

Urban load dynamics are shaped by regular patterns across multiple temporal scales. The effective separation of these dominant patterns is essential for accurate forecasting [12]. Unlike existing multi-scale methods that typically decompose or process each node sequence independently [11, 24, 37], the proposed multi-scale spatiotemporal encoder (MSTE) emphasizes globally consistent decomposition across spatial nodes. This approach enables the construction of multi-scale representations at the system level rather than merely at the node level. To this end, MSTE is structured into three sequential stages: (1) globally consistent mode decomposition, (2) multi-branch spatiotemporal feature extraction, and (3) adaptive mode feature fusion.

Firstly, if mode decomposition is performed independently for each spatial node, the same mode may correspond to disparate central frequencies across nodes due to local distribution variations. Such inconsistency weakens the alignment of multi-scale representations at the system level and constrains the efficacy of spatiotemporal modeling. To address this issue, MSTE incorporates multivariate variational mode decomposition (MVMD) [29] to enforce frequency alignment across spatial nodes during the decomposition process.

Specifically, after obtaining the node-wise center frequencies and amplitude spectra at the (m+1)-th ADMM iteration, MVMD performs a cross-node frequency alignment step:

$$\omega_{k,m+1} = \frac{\sum_{i=1}^{N} \omega_{k,m+1}^{(i)} \left|\tilde{u}_{k,m+1}^{(i)}\right|^2}{\sum_{i=1}^{N} \left|\tilde{u}_{k,m+1}^{(i)}\right|^2}, \quad (1)$$

where $\left|\tilde{u}_{k,m+1}^{(j)}\right|^2$ denotes the amplitude spectrum of the k-th mode at node i. This operation aligns the center frequency of corresponding modes across different spatial nodes based on their spectral energy distribution, enhancing the global consistency of the decomposed mode components.

The globally consistent decomposition process can be concisely expressed as:

$$(U_1, U_2, \cdots, U_K) = MVMD_{spa}(X, K), \quad (2)$$

where K denotes the predefined number of decomposition modes. Each mode component $U_k \in R^{T_h \times N}$ characterizes system dynamics at a specific frequency ω_k.

In implementation, MVMD is applied at the input-window level rather than being precomputed over the whole time series. Each forecasting sample with the historical window $W_{in} = [t-T_h+1,\ldots,t-1,t]$ is independently decomposed into mode components. This decomposition process only uses the observations within the look-back horizon T_h. The same window-level decomposition protocol is consistently adopted during training, validation, and testing, ensuring that the mode components are generated without using future information.

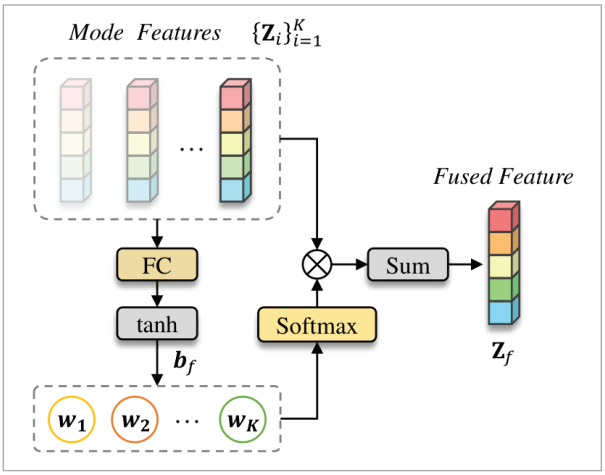
Secondly, MSTE adopts a multi-branch spatiotemporal feature extraction strategy. This architecture allows the model to encode heterogeneous dynamics at different temporal scales in separate representation subspaces, precluding premature feature mixing. The process is formulated as:

$$Z_k = FE_k(U_k; \theta_k), \quad (3)$$

where $FE_k(\cdot)$ denotes the spatiotemporal feature extractor of the k-th branch, comprising a cascaded SATE and a GSA module. θ_k are parameters in this extractor. $Z_k \in R^{N \times d}$ is the latent representation.

Finally, given that mode features contribute unequally to the forecasting task, MSTE employs an adaptive mode feature fusion (AMFF) module to consolidate branch-wise representations. A simplistic averaging strategy implicitly assumes equal importance for all mode components, which may attenuate discriminative information from more informative frequency branches. Conversely, direct concatenation increases the computational load on the decoder and may intro-

Figure 2
Illustration of the Adaptive Mode Feature Fusion (AMFF) module.



duce redundant mode information. To mitigate these issues, AMFF determines a feature-dependent importance score for each mode-specific branch representation via a learnable nonlinear mapping function, as demonstrated in Figure 2.

Specifically, for each node n, the fusion weight of the i-th mode is adaptively generated from the current branch feature. Let $z_i^{(n)}$ be the feature vector, the fusion process is defined as:

$$w_i^{(n)} = b_f^{\top} \tanh\left(FC\left(z_i^{(n)}\right)\right) \tag{4}$$

$$\gamma_i^{(n)} = \frac{\exp\left(w_i^{(n)}\right)}{\sum_{j=1}^{K} \exp\left(w_j^{(n)}\right)} \tag{5}$$

$$z_f^{(n)} = \sum_{i=1}^{K} \gamma_i^{(n)} z_i^{(n)}, \tag{6}$$

where $b_f \in R^d$ is a learnable parameter vector. The fused representation $Z_f \in R^{N\times d}$ is then fed into the MLP decoder for direct multi-step forecasting.

4.2. Semantic-aware Temporal Extractor

Urban load dynamics are shaped not only by historical observations but also by temporal semantic context, such as intraday and intraweek regularities. Consequently, effective temporal dependency modeling should account for semantic cues, rather than relying exclusively on numerical historical values.

Specifically, the semantic input used in SATE is derived from the timestamp of each historical time step. For a given time step τ, we first extract its hour-of-day index and day-of-week index. SATE introduces learnable hourly embedding vectors $T^H \in R^{24\times d_s}$ and daily embedding vectors $T^D \in R^{7\times d_s}$. The semantic embedding s_τ is constructed by concatenating the hourly embedding τ^H, the daily embedding τ^D, and a learnable node identifier E:

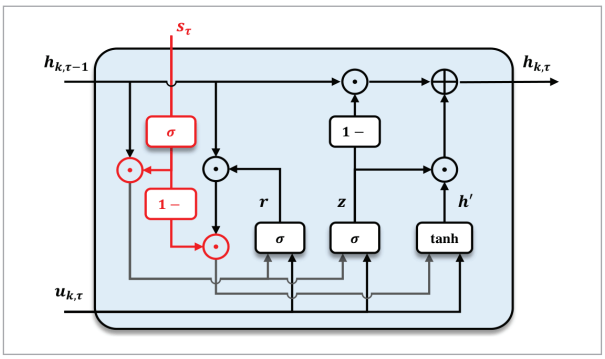
$$s_\tau = Concat\left(E, \tau^H, \tau^D\right). \tag{7}$$

These embeddings are jointly optimized during training to capture latent intraday and intraweek regularities. Therefore, the semantic input construction only involves timestamp parsing and embedding lookup. The resulting semantic representation encodes intraday and intraweek temporal regularities while retaining node-level contextual heterogeneity.

Conventional GRU-based sequence modeling propagates temporal dependencies through hidden states. However, GRU lacks the mechanism for explicitly differentiating how temporal semantic contexts influence the current update. To address this limitation, SATE introduces a fusion-gated recurrent unit (FGRU), as illustrated in Figure 3, where the semantic embedding is transformed into a gating signal and subsequently used to modulate historical memory before recurrent update decisions are finalized.

Specifically, given the semantic embedding, a gate signal is first derived as:

Figure 3
Structure of the Fusion-Gated Recurrent Unit (FGRU).



$$s = \sigma(W_s s_\tau + b_s) \tag{8}$$

$$G\left(h_{k,\tau-1}; s\right) = h_{k,\tau-1} \odot s \tag{9}$$

where ⊙denotes the Hadamard product, $\sigma(\cdot)$ represents the sigmoid activation function. The semantic context is utilized to selectively emphasize the historical information most relevant to the current temporal condition. The semantically filtered historical state is then employed to compute the reset and update gates:

$$r = \sigma\left(\left[u_{k,\tau} || G\left(h_{k,\tau-1}; s\right)\right] W_r + b_r\right) \tag{10}$$

$$z = \sigma\left(\left[u_{k,\tau} || G\left(h_{k,\tau-1}; s\right)\right] W_z + b_z\right), \tag{11}$$

where || denotes concatenation, $u_{k,\tau}$ is the current input of the k-th mode. Consequently, the gate computation is no longer contingent solely on the raw historical state, but rather on semantically filtered memory that is cognizant of the current temporal context. Moreover, the FGRU defines a reset-modulated historical state as:

$$H\left(h_{k,\tau-1}; s, r\right) = \left[(1 - s) \odot h_{k,\tau-1}\right] \odot r, \tag{12}$$

where $(1–s)\odot h_{k,\tau-1}$ represents the information not directly emphasized by the semantic gate but remains potentially relevant for capturing transient temporal variations. Based on this state, the candidate hidden state and the updated hidden state are computed as:

$$h' = \tanh\left(\left[u_{k,\tau} || H\left(h_{k,\tau-1}; s, r\right)\right] W + b\right) \tag{13}$$

$$h_{k,\tau} = (1 - z) \odot h_{k,\tau-1} + z \odot h'. \tag{14}$$

As a result, temporal dependencies are encoded in a context-aware manner rather than through conventional hidden-state recursion alone. The recursive temporal encoding process of SATE can be summarized as:

$$H_{k,\tau} = \mathrm{FGRU}\left(U_{k,\tau}, H_{k,\tau-1}, S_\tau; \theta_{kf}\right), \tag{15}$$

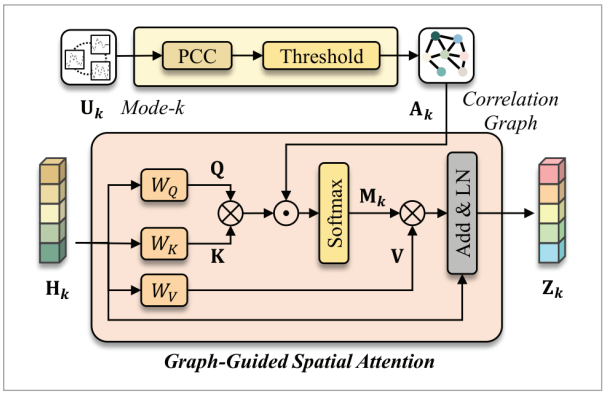
where $H_{k,\tau}$ is the updated hidden state. The initial state $H_{k,\tau}$ is defined as a zero tensor. To enhance generalization and computational efficiency, SATE shares the parameters θ_{kf} across all spatial nodes.

4.3. Graph-guided Spatial Attention Module

While spatial self-attention inherently provides a global receptive field for modeling long-range interactions, it may inadvertently aggregate spurious information from irrelevant nodes in the absence of structural constraints. To mitigate this limitation, we propose a graph-guided spatial attention (GSA) module, as illustrated in Figure 4. The central premise is to construct a dynamic mode-specific correlation graph before each mode component and subsequently utilize this to guide the spatial self-attention mechanism. This design facilitates the joint modeling of local structural constraints and global dependency aggregation. The GSA module involves two synergistic stages: correlation graph construction and graph-guided attention aggregation.

To characterize local structural correlations among spatial nodes, we employ the Pearson correlation coefficient (PCC) [30] to quantify pairwise similarity within each mode. Notably, the graph prior is constructed not from the raw mixed sequences but from the mode component U_k of the current historical input window W_{in}. Since the mode component represents a band-limited signal centered around a specific frequency, computing the PCC on this component reflects mode-specific synchronous fluctuations while mitigating the sensitivity of correlation estimation to transient outliers. As such, for each forecasting sample and each mode, GSA generates a corresponding dynamic mode-specific graph prior A_k. The PCC graph is defined as:

Figure 4
Illustration of the Graph-guided Spatial Attention (GSA) module.



$$r_{ij}^{k}=\frac{\sum_{t\in W_{in}}\left(u_{k,t}^{(i)}-\mu_{k}^{(i)}\right)\left(u_{k,t}^{(j)}-\mu_{k}^{(j)}\right)}{\sqrt{\sum_{t\in W_{in}}\left(u_{k,t}^{(i)}-\mu_{k}^{(i)}\right)^{2}\sum_{t\in W_{in}}\left(u_{k,t}^{(j)}-\mu_{k}^{(j)}\right)^{2}}}, \quad (16)$$

where $\mu_k^{(i)}$ and $\mu_k^{(j)}$ denote the mean values of the mode component. In urban load forecasting, positive correlations usually indicate synchronous fluctuations, shared demand patterns, or similar temporal consumption regularities among nodes. In contrast, negative correlations may reflect opposite variation trends; directly using them as attention priors may guide the model to aggregate inconsistent information. Therefore, we retain positive and sufficiently strong correlations to provide reliable structural constraints for spatial information aggregation. A threshold $\in_c$ is applied to sparsify the graph topology:

$$a_{ij}^{k}=\begin{cases} r_{ij}^{k}, & if\ r_{ij}^{k}>\epsilon_{c} \\ 0, & otherwise \end{cases} \quad (17)$$

$$A_{k}=\begin{pmatrix} 0 & a_{12}^{k} & \cdots & a_{1N}^{k} \\ a_{21}^{k} & 0 & & a_{2N}^{k} \\ \vdots & & \ddots & \vdots \\ a_{N1}^{k} & a_{N2}^{k} & \cdots & 0 \end{pmatrix} \quad (18)$$

where $A_k \in R^{N\times N}$ serves as the mode-specific graph prior. The graph is recomputed for each input window and each mode during both training and inference. Subsequently, GSA combines the correlation graph with spatial self-attention.

Formally, node features are linearly projected to obtain the query (Q), key (K), and value (V) representations:

$$Q=H_kW_Q,\quad K=H_kW_K,\quad V=H_kW_V \quad (19)$$

$$M_k=Softmax\left(\frac{QK^{\top}}{\sqrt{d}}\odot A_k\right) \quad (20)$$

$$Z_k=LN(M_kV+H_k), \quad (21)$$

where $M_k \in R^{N\times N}$ denotes the graph-guided attention matrix, Layer normalization (LN) and residual connections are employed to stabilize training. The self-attention mechanism captures deep semantic dependencies among node features. By modulating spatial self-attention with A_k, the model preserves the global receptive field of self-attention while introducing local structural constraints derived from mode-specific synchronous load variations.

4.4. Prediction Decoder

Rather than utilizing recursive decoding, the proposed framework adopts a direct multi-step forecasting strategy to mitigate error accumulation as the forecasting horizon extends. To derive future load demand predictions from the fused spatiotemporal representation Z_l, we employ a lightweight MLP-based decoder as the prediction head. The decoding process is formulated as:

$$Z_l=GELU\left(FC(Z_{l-1})\right)+Z_{l-1} \quad (22)$$

$$\hat{Y}=Linear(Z_L), \quad (23)$$

where L denotes the number of decoder layers and $GELU(\cdot)$ represents the Gaussian Error Linear Unit activation function.

5. Experiments and Analysis

5.1. Datasets and Metrics

We evaluate the proposed model on four public Commercial Electric Load Profiles (CELP) datasets [26], specifically covering Virginia (VA), Vermont (VT), Washington (WA), and West Virginia (WV). The key characteristics are summarized in Table 1.

Table 1
Summary statistics of the datasets.

Datasets	VA	VT	WA	WV
Nodes	182	52	377	143
Time Range	2012/1/1 - 2012/12/31			
Sampling Rate	1 hour			
Time Steps	8760			
Unit	kWh			

According to the metadata provided with the CELP datasets, each node corresponds to a load profile associated with a specific electricity-consuming entity. Therefore, the spatial correlations considered in this work refer to interdependencies among load nodes with different consumption patterns, building functions, human activity rhythms, and regional usage characteristics.

Forecasting performance is assessed using three widely adopted metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). Lower values signify superior forecasting performance. They are defined as follows:

$$MAE = \frac{1}{M}\sum_{m=1}^{M}\left|\hat{Y}_m - Y_m\right| \tag{24}$$

$$RMSE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{Y}_m - Y_m\right)^2} \tag{25}$$

$$MAPE = \frac{1}{M}\sum_{m=1}^{M}\left|\frac{\hat{Y}_m - Y_m}{Y_m}\right| \times 100\%, \tag{26}$$

where M denotes the total number of evaluation samples, while $\hat{Y}_m$ and Y_m represent the model prediction and ground-truth, respectively.

Each dataset is partitioned into training, validation, and test sets at a ratio of 7:1:2. All input sequences are standardized via Z-score normalization. To avoid information leakage, the mean and standard deviation are calculated only from the training set and then applied to the validation and test sets. After forecasting, the predicted values are inverse-transformed to the original scale before computing MAE, RMSE, and MAPE.

5.2. Baselines

To comprehensively evaluate the effectiveness of MDSFGA, we compare it against 16 representative forecasting methods across five categories: **Traditional methods** (SARIMA and LSTM), **GNN-based methods** (StemGNN[4], STGCN[42], AGCRN[2] and GWNet[39]), **Transformer-based methods** (PatchTST[25], iTransformer[23] and STAEformer[22]), **MLP-based methods** (STID[31] and BLSTF[5]), **Decomposition-based methods** (DLinear[44], STNorm[8], TimeMixer[37], STWave[11] and WPMixer[24]). All baseline models are implemented using the hyperparameter settings recommended in their original papers to ensure a fair comparison.

5.3. Implementation Details

The primary hyperparameters for the proposed MDSFGA are summarized in Table 2. MDSFGA is implemented in PyTorch 2.0.0. The L1 loss is employed as the training objective. All computations are conducted on an NVIDIA GeForce RTX 3080 GPU with CUDA 12.2.

Table 2
Main hyperparameter settings of MDSFGA.

Hyperparameter	Value
Number of decomposition modes K	4
Encoder hidden dimension d	96
Semantic embedding dimension d_s	4
Threshold of the PCC graph $\grave{O}_c$	0.6
Decoder MLP layers L	5
Dropout rate	0.1
Learning rate	0.001
Batch size	64
Epoch number	100
Optimizer	Adam

All preprocessing and auxiliary computations are performed consistently across the training, validation, and test stages. For each forecasting sample, MVMD is applied only to the current historical input window to generate the mode components, without accessing the future forecasting horizon. Based on these window-level mode components, the PCC graph in GSA is recomputed for each sample and each mode.

5.4. Performance Comparison Analysis

Multi-step load forecasting was first conducted with the horizon T_f set to 24 for all compared methods. To account for stochasticity, each experiment was repeated five times, and the average values of each evaluation metric are reported in Table 3.

Table 3
Performance comparison of various models on the CELP datasets, where the best and second-best results are highlighted in **bold** and underlined, respectively.

Dataset	CELP-VA			CELP-VT			CELP-WA			CELP-WV		
Metrics	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
SARIMA	108.54	301.61	30.13	149.68	380.08	38.65	94.51	319.64	31.05	124.35	343.68	32.48
LSTM	92.31	239.68	26.99	121.64	330.53	31.55	82.96	253.94	26.72	102.69	295.13	28.46
DLinear	93.47	243.85	23.66	120.86	318.73	26.03	80.99	230.54	19.35	99.67	287.66	23.95
StemGNN	91.56	258.19	29.41	122.09	356.16	35.94	81.81	265.48	27.83	106.36	300.17	32.70
STGCN	83.82	242.65	24.62	106.78	305.56	28.91	72.89	237.47	23.50	96.68	281.66	25.75
AGCRN	83.58	260.98	24.44	110.88	330.92	29.52	63.27	231.99	19.85	92.59	290.39	21.41
PatchTST	79.48	212.50	20.74	103.86	275.35	23.89	67.45	200.03	17.46	92.53	247.03	21.24
WPMixer	78.87	210.41	20.37	101.83	275.51	23.16	69.21	202.16	17.51	98.25	256.89	22.51
TimeMixer	76.41	209.73	19.35	98.01	274.81	22.50	61.61	197.26	14.94	90.14	249.09	19.77
iTransformer	73.93	212.81	19.23	93.75	265.24	21.32	65.34	232.17	15.56	94.98	258.97	21.54
STNorm	75.47	224.41	21.42	98.83	294.41	25.38	63.25	224.72	19.40	91.53	282.05	23.78
GWNet	69.49	216.16	20.36	91.93	291.89	24.36	61.28	228.56	18.48	85.69	272.78	22.65
STID	68.94	188.70	18.28	85.17	227.91	19.36	57.95	181.94	14.58	79.82	222.50	18.69
STWave	70.75	186.14	19.18	97.88	244.31	23.06	55.63	170.28	16.29	79.49	218.83	19.27
STAEformer	65.95	190.85	17.03	80.39	224.93	17.90	47.57	162.37	12.90	73.96	224.46	16.97
BLSTF	64.47	184.76	16.98	79.49	218.76	17.55	47.02	161.77	12.59	73.42	220.31	16.54
MDSFGA(ours)	**59.80**	**173.94**	**15.65**	**74.48**	**206.92**	**16.72**	**46.52**	**157.48**	**12.28**	**71.97**	**212.58**	**16.36**

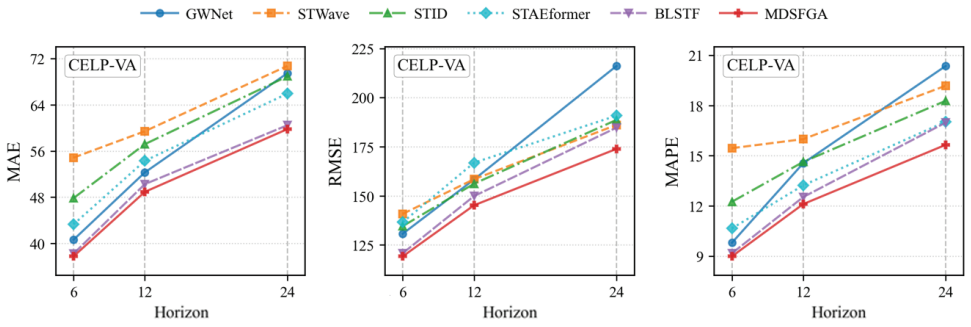
The experimental results show that MDSFGA achieves the best overall performance across the evaluated CELP datasets. Specifically, compared with the second-best results on each dataset, MDSFGA reduces MAPE by 7.83%, 4.73%, 2.46%, and 1.09% on CELP-VA, CELP-VT, CELP-WA, and CELP-WV, respectively. These results indicate that MDSFGA can more effectively characterize complex spatiotemporal evolution patterns involved in urban load demand forecasting under the 24-step forecasting horizon setting. A closer examination of the comparative results yields the following observations:

1 *Comparison with decomposition-based methods:* MDSFGA achieves better performance than representative decomposition-based forecasting methods, such as STWave and TimeMixer. This suggests that, compared with conventional node-wise decomposition strategies, the proposed globally consistent multi-scale representation mechanism helps construct more coherent and discriminative spatiotemporal features.

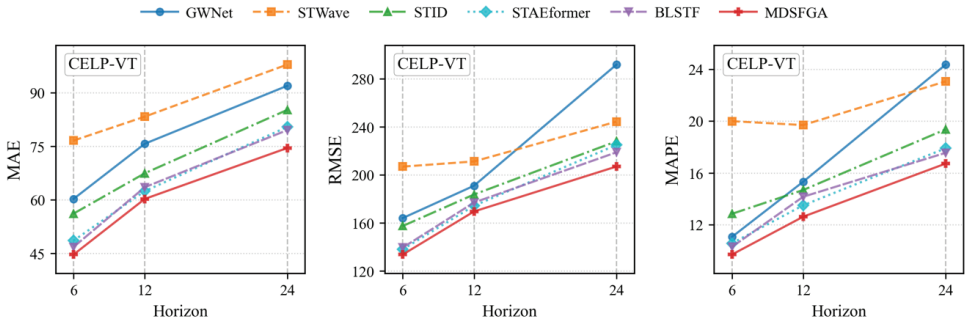
2 *Comparison with GNNs and Transformer-based methods:* MDSFGA achieves better forecasting performance than GNN-based methods (e.g., STGCN and GWNet) as well as Transformer-based architectures (e.g., iTransformer and STAEformer). These results suggest that relying solely on local graph propagation or global attention may be insufficient for modeling complex urban spatial dependencies. In contrast, combining local structural constraints with global dependency aggregation appears to be more effective for spatial modeling in this forecasting setting.

Figure 5

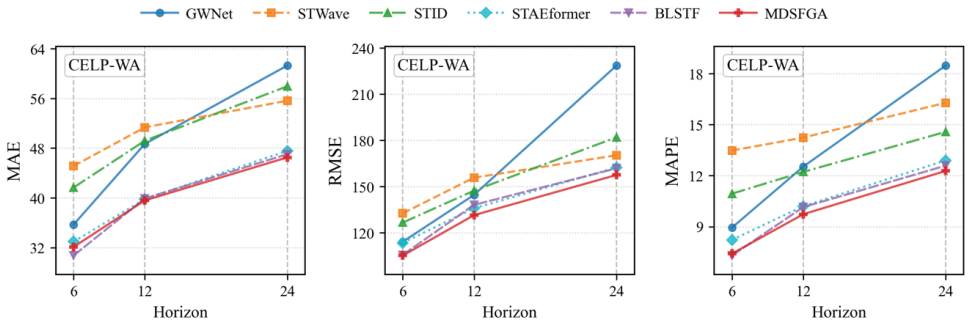
Performance comparison under different prediction horizons. Subplots (a)-(d) illustrate the performance of the models across four datasets.



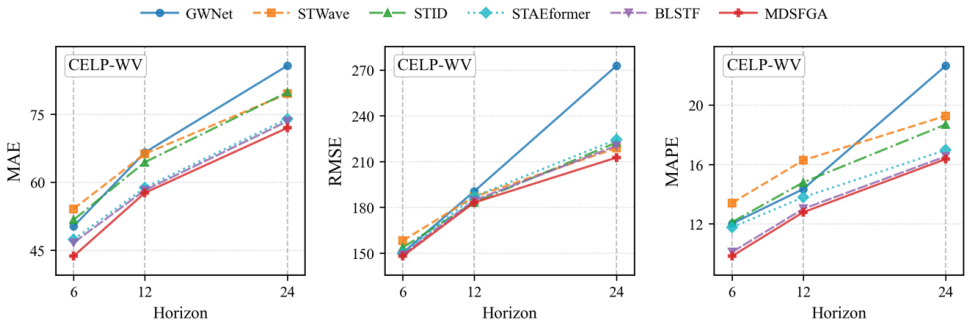
(a) Model performance comparison under different prediction horizons at CELP-VA dataset



(b) Model performance comparison under different prediction horizons at CELP-VT dataset



(c) Model performance comparison under different prediction horizons at CELP-WA dataset



(d) Model performance comparison under different prediction horizons at CELP-WV dataset

3 *Comparison with MLP-based methods:* This performance gain may be attributed to the decoupling of multi-frequency dynamic patterns before prediction. The learned spatiotemporal representations provide a more informative feature basis for the lightweight MLP decoder, supporting more accurate forecasting outputs.

To further evaluate the performance of the proposed framework across different forecasting horizons, we conduct comparative experiments with lengths of 6, 12, and 24. The performance comparisons against representative baseline models are reported in Figure 5(a)-(d).

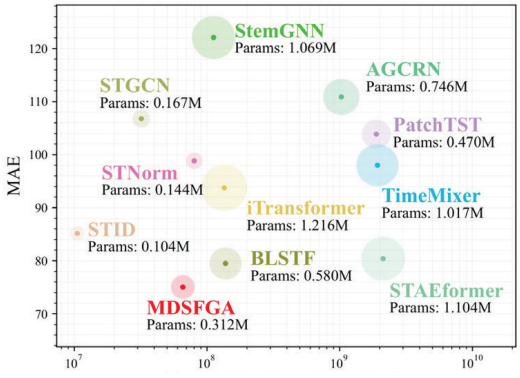
The results across varying forecasting horizons suggest that MDSFGA maintains relatively stable performance as the forecasting horizon extends. While forecasting complexity increases for all compared methods as the horizon expands from 6 to 24, the proposed framework exhibits a comparatively stable error growth trajectory across different datasets. This indicates that the learned spatiotemporal representations provide a robust foundation for multi-step forecasting.

Specifically, under the 6-step forecasting horizon setting, MDSFGA achieves the lowest MAE on the CELP-VA, CELP-VT, and CELP-WV datasets. On CELP-WA, it attains the lowest RMSE and remains highly competitive in terms of MAE and MAPE. As the horizon extends to 12 and 24, MDSFGA outperforms representative baselines in most scenarios. For instance, on the CELP-VT dataset at a 24-step forecasting horizon setting, the MAE of MDSFGA is lower than that of GWNet, STWave, and BLSTF. These results suggest that the proposed method can maintain relatively stable spatiotemporal representations across extended forecasting spans.

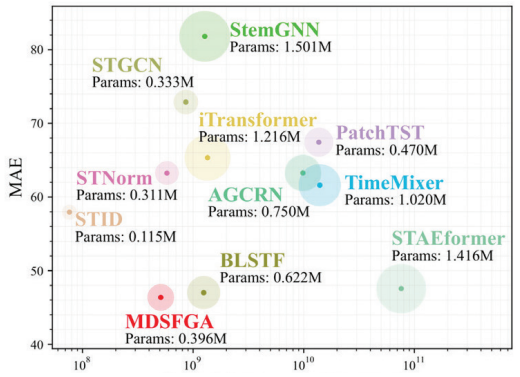
To further evaluate the computational efficiency of MDSFGA, we select CELP-VT and CELP-WA for comparison, as they feature the smallest (52) and largest (377) node counts among the four datasets, respectively. For several representative baseline models, we report both the floating-point operations (FLOPs) and parameter counts. As illustrated in Figure 6(a)-(b), the vertical axis represents the MAE under the 24-step forecasting horizon setting, the horizontal axis shows FLOPs on a logarithmic scale, and the bubble size represents the parameter count.

In terms of neural network inference cost, the FLOPs of MDSFGA are 6.6×10^7 on CELP-VT, which is lower than those of computationally intensive baselines such as PatchTST and STAEformer. When evaluated on CELP-WA, the FLOPs increase to 5.1×10^8 while remaining within a moderate range. Compared with STID, which exhibits low model complexity but considerably weaker forecasting accuracy, MDSFGA provides stronger predictive performance. Furthermore, in contrast to STAEformer, MDSFGA maintains competitive accuracy with a lower computational burden. Compared with the strongest competing baseline, BLSTF, MDSFGA achieves a lower MAE on CELP-VT while maintaining lower FLOPs, indicating a favorable efficiency-performance balance. Overall, these results suggest that MDSFGA delivers improved forecasting accuracy without requiring a large number of parameters or high computational overhead.

Figure 6
Comparison of model efficiency on (a) CELP-VT and (b) CELP-WA. The bubble size qualitatively indicates the parameter scale.



(a) Model FLOPs at CELP-VT dataset

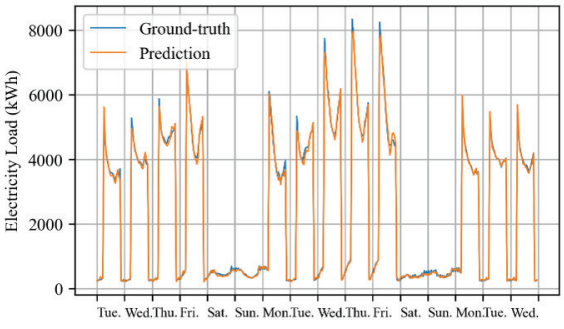


(b) Model FLOPs at CELP-WA dataset

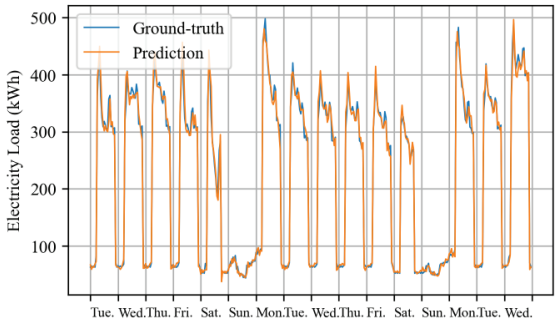
5.5. Visualization Analysis

Figure 7

Visualization of prediction results and ground truth from CELP-VT. Subplot (a) and (b) illustrate model prediction on two random select nodes, respectively.



(a) Visualization of prediction results and ground truth at Node 21.



(b) Visualization of prediction results and ground truth at Node 43.

To provide a qualitative assessment of how the proposed framework captures temporal semantic context, we select two nodes (21 and 43) from the CELP-VT dataset. The displayed time steps are selected from a prediction segment in the test set, and the predicted trajectories are compared with the corresponding ground truth, as shown in Figure 7(a)-(b). The predicted curves generally follow the daily load fluctuations of both nodes, capturing typical peak and trough patterns. Together with the weekday and weekend annotations in the figure, the visualization suggests that MDSFGA captures some differences in temporal consumption patterns across weekdays and weekends, in addition to short-term numerical variations. This provides qualitative support for the role of timestamp-derived semantic cues in helping SATE capture intraday and intraweek regularities.

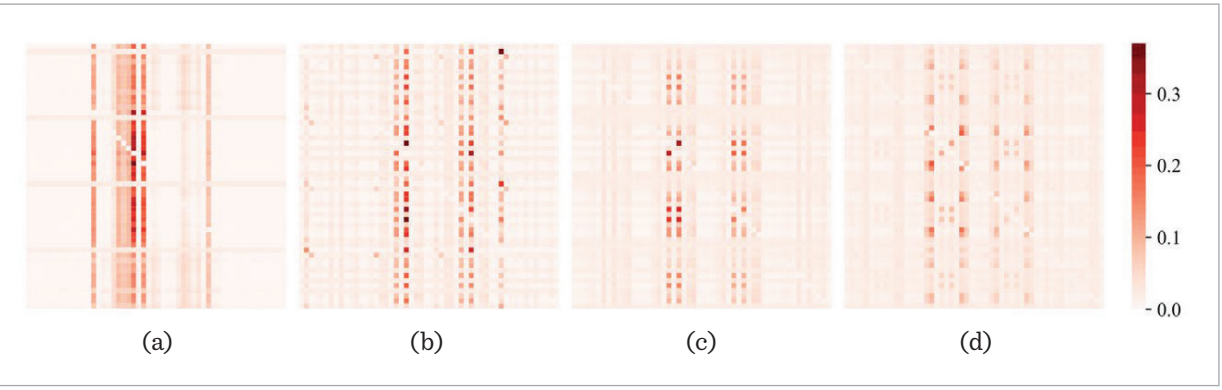
To further examine the internal spatial modeling dynamics of MDSFGA, we visualize the attention maps generated by the GSA module at a randomly selected time step on the CELP-VT dataset, as shown in Figure 8(a)-(d).

The four subplots are arranged in ascending order of frequency. In each attention map, both the horizontal and vertical axes represent spatial nodes, and the node order follows the metadata-based ordering used in the dataset. Each row indicates the attention distribution from one source node to other target nodes, and darker colors correspond to larger attention weights.

At the low-frequency scale, the attention pattern is relatively homogeneous, suggesting that spatial nodes

Figure 8

Visualization of the attention maps M_k generated by the GSA modules on CELP-VT with K=4. Subplots (a) to (d) are arranged in ascending order of frequency.



may share broader long-term variations. At medium-frequency scales, more concentrated patterns can be observed, indicating stronger dependencies among certain groups of nodes. At the high-frequency scale, the attention map becomes more localized, which may correspond to short-term fluctuations affecting a smaller subset of nodes. Overall, this visualization provides qualitative support for the ability of the proposed multi-scale decomposition and graph-guided spatial attention mechanism to characterize spatial dependency patterns with different granularities across frequency branches.

5.6. Ablation Study

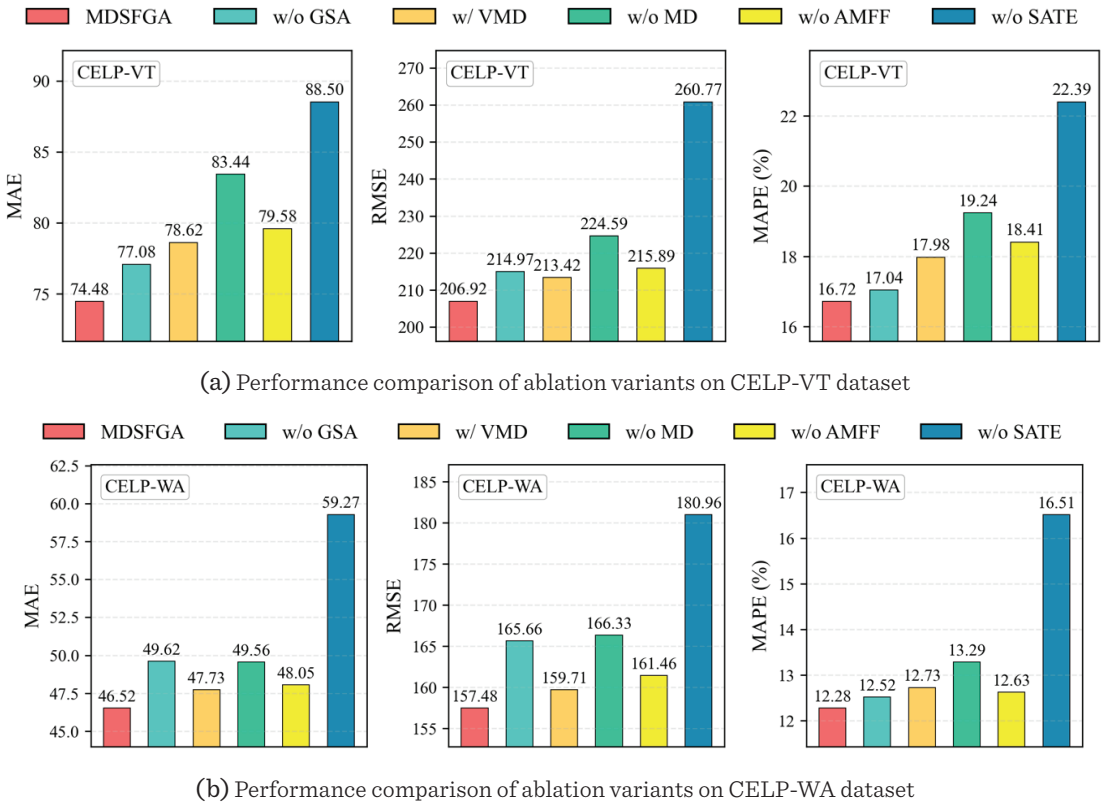
To quantitatively evaluate the contribution of each core component, we design several ablation variants and evaluate them on the CELP-VT and CELP-WA datasets. The ablation variants are defined as follows:

1. **w/ VMD:** Replaces the globally consistent MVMD with independent node-wise VMD.
2. **w/o MD:** Removes the mode decomposition strategy and extracts features directly from the raw sequences.
3. **w/o AMFF:** Replaces the adaptive mode feature fusion module with simple averaging across mode-specific branch features.
4. **w/o SATE:** Removes external temporal semantic information, degenerating the temporal extractor into a standard GRU.
5. **w/o GSA:** Replaces the graph-guided spatial attention module with standard spatial self-attention.

Figure 9(a)-(b) shows that removing multi-scale decomposition (w/o MD) or replacing MVMD with independent node-wise decomposition (w/ VMD) leads to a decline in forecasting accuracy. As sug-

Figure 9

Performance comparison of ablation variants. Subplot (a) and (b) illustrate ablation results on CELP-VT and CELP-WA, respectively.



(a) Performance comparison of ablation variants on CELP-VT dataset

(b) Performance comparison of ablation variants on CELP-WA dataset

gested by the attention-map visualization in Figure 8(a)-(d), different frequency bands exhibit distinct spatial patterns. Without global consistency, the frequency-specific components may not be effectively aligned across spatial nodes, which can weaken system-level spatiotemporal representation learning. These results support the effectiveness of globally consistent multi-scale decomposition in MDSFGA.

The comparison with the w/o AMFF variant further shows the contribution of adaptive mode feature fusion. When AMFF is replaced with simple averaging, the model treats all mode-specific branch features as equally important. This fixed averaging strategy cannot adjust the contribution of different frequency branches according to node-level and input-dependent feature variations. In contrast, AMFF assigns feature-dependent fusion weights to different modes, enabling the model to emphasize more discriminative mode representations for each input. The performance gap between MDSFGA and w/o AMFF variant suggests that adaptive mode fusion contributes to more effective multi-scale representation aggregation.

Moreover, when the model lacks temporal semantic context (w/o SATE), the forecasting performance deteriorates. Without semantic-aware temporal encoding, the model is less capable of distinguishing intraday dynamics and weekday-weekend variations than the complete MDSFGA. These results support the effectiveness of explicitly incorporating timestamp-derived semantic context into temporal dependency modeling. Finally, when the graph prior constraint is removed (w/o GSA), the forecasting performance also declines. This suggests that, without the PCC-based dynamic correlation graph derived from mode components, pure spatial self-attention may be more prone to aggregating redundant or irrelevant node information, weakening the representation of intrinsic spatiotemporal dependencies.

To further examine the influence of different similarity measures on graph-guided spatial attention, we conduct an additional comparison on CELP-VT and CELP-WA. The model variants are as follows:

1 The self-attention variant is identical to the w/o GSA setting in the ablation study.

2 The soft-DTW-based variant transforms the pairwise soft-DTW distance into a similarity score within [0,1] using an RBF kernel, and the same thresholding method is applied to sparsify the graph prior.

3 The MI-based variant uses mutual information scores to generate graph-prior weights, and only scores larger than the threshold are retained.

4 The PCC-based variant corresponds to the complete MDSFGA, where positive Pearson correlations larger than the threshold are used to construct the dynamic mode-specific graph prior.

The results are summarized in Table 4. Compared with the w/o GSA variant, all similarity-guided variants achieve lower forecasting errors on both datasets, suggesting that introducing a graph prior can provide useful structural constraints for spatial attention. Among the alternative similarity measures, the soft-DTW-based variant performs competitively, especially on CELP-VT, where it obtains a slightly lower MAPE than MDSFGA. This suggests that shape-based similarity can capture the node-pair relationships. However, soft-DTW requires a relatively large amount of computation. The MI-based variant also improves over the w/o GSA variant, suggesting that nonlinear dependency information can be useful, although its estimation may be less stable under limited input-window samples.

Table 4
Impact of different similarity measurement methods on forecasting performance.

Dataset	CELP-VT		
Metrics	**MAE**	**RMSE**	**MAPE**
Self-attention (w/o GSA)	78.62	213.42	17.98
soft-DTW + Attention	74.79	206.93	**16.52**
MI + Attention	75.49	208.21	17.36
PCC + Attention (MDSFGA)	**74.48**	**206.92**	16.72
Dataset	CELP-WA		
Self-attention (w/o GSA)	47.73	159.71	12.73
soft-DTW + Attention	46.64	158.01	12.35
MI + Attention	47.01	158.29	12.43
PCC + Attention (MDSFGA)	**46.52**	**157.48**	**12.28**

Overall, MDSFGA achieves the lowest MAE and RMSE on CELP-VT and the best results across all three metrics on CELP-WA. These results suggest that PCC is a simple and effective choice for measur-

ing synchronous fluctuations within the same modal frequency band. By combining PCC-based graph priors with spatial self-attention, MDSFGA can help preserve local structural constraints while allowing global dependency aggregation, achieving a favorable balance between graph-guided regularization and flexible spatial modeling.

5.7. Hyperparameter Sensitivity

To assess the sensitivity of MDSFGA to key hyperparameters, we investigate the influence of the number of decomposition modes and the threshold for PCC-based graph construction on forecasting performance.

As illustrated in Figure 10, the forecasting performance follows a trajectory of rapid gains followed by gradual saturation as K increases. Specifically, as it increases from 1 to 4, all three metrics decline substantially and consistently. This suggests that multi-scale mode decomposition can help separate mode patterns distributed across different frequency bands, reducing feature extraction complexity and improving spatiotemporal representation quality. When K greater than 4, the performance gains become marginal, whereas the parameter count scales linearly with the number of mode branches.

The threshold $\grave{O}_c$ governs the sparsity of the PCC-based correlation graph in the GSA module, determining which node pairs are retained as structural priors based on their fluctuation similarities. As shown in Figure 11, the model attains peak performance on both datasets when c lies within the range of 0.6 to 0.7. If $\in_c$ is set too low, the constructed graph may introduce excessive redundant connections among less relevant nodes, potentially incorporating spurious noise during feature aggregation. Conversely, if $\in_c$ is set too high, the PCC graph becomes overly sparse, which may lead to the loss of essential spatial dependencies.

5.8. Generalization Validation

To further evaluate the general spatiotemporal forecasting ability of MDSFGA beyond urban load

Figure 10

Impact of the number of decomposition modes on forecasting performance.

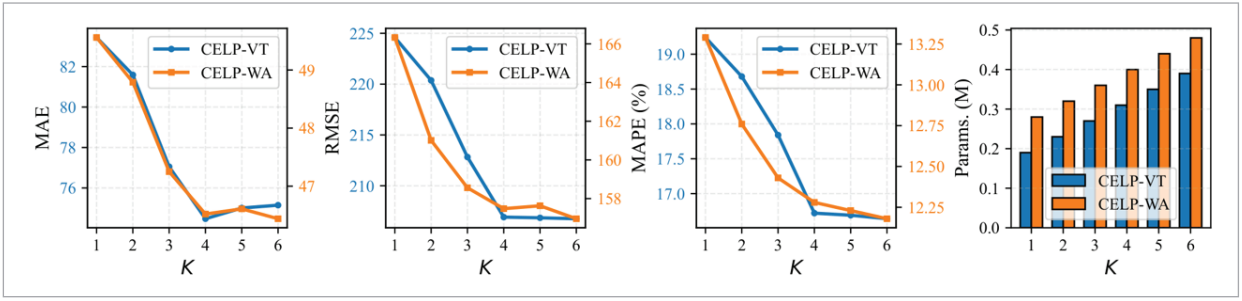


Figure 11

Impact of the threshold for PCC-based graph construction on forecasting performance.

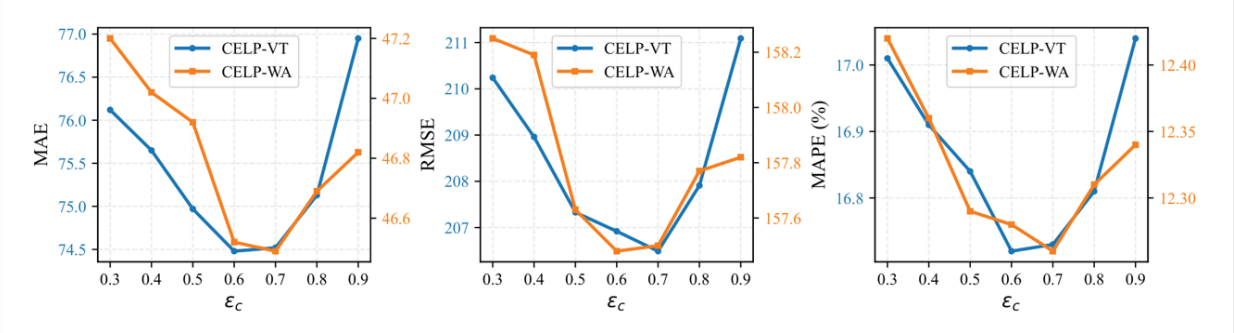


Table 5
Statistics of the generalization validation dataset and 12-step forecasting performance.

Dataset	PEMS-BAY		
Time Range	2017.01 - 2017.05		
Sampling Rate	5 min		
Nodes	325		
Unit	km/h		
Metrics	**MAE**	**RMSE**	**MAPE**
StemGNN	2.57	5.88	6.55
STGCN	2.01	4.56	4.74
AGCRN	1.99	4.64	4.72
PatchTST	2.01	4.65	4.81
STNorm	1.96	4.70	4.70
GWNet	2.02	4.70	4.85
STWave	1.91	4.36	4.48
STID	1.89	4.40	4.47
STAEformer	1.90	4.31	4.44
TimeMixer	1.91	4.36	4.46
iTransformer	1.90	4.33	4.43
WPMixer	1.91	4.35	4.48
BLSTF	1.88	4.30	4.42
MDSFGA(ours)	**1.83**	**4.25**	**4.32**

datasets, we conduct an additional cross-domain validation on PEMS-BAY, which is used here to examine whether the proposed spatiotemporal modeling mechanism remains effective. The statistics of PEMS-BAY and the one-hour forecasting results are reported in Table 5. MDSFGA achieves the lowest MAE, RMSE, and MAPE among the compared methods. Compared with the strongest competing baseline, BLSTF, MDSFGA reduces MAE from 1.88 to 1.83, RMSE from 4.30 to 4.25, and MAPE from 4.42 to 4.32. It also outperforms other recent spatiotemporal forecasting models such as STAEformer, STID, STWave, iTransformer, and WPMixer. These results suggest that the combination of mode decomposition, semantic temporal modeling, and graph-guided spatial attention may also provide useful spatiotemporal representations in a cross-domain multi-node forecasting task.

6. Conclusion

We propose MDSFGA, an encoder-decoder network for urban load demand forecasting. The proposed multi-scale spatiotemporal encoder (MSTE) decouples historical observations into globally frequency-consistent mode components across spatial nodes, facilitating the construction of spatiotemporal representations through branch-wise feature extraction and adaptive fusion. Within this framework, the semantic-aware temporal encoder (SATE) incorporates temporal semantic context into recurrent dependency modeling to strengthen the representation of intraday and intraweek load regularities. Simultaneously, the graph-guided spatial attention (GSA) module combines mode-specific graph priors with spatial self-attention to jointly model local structural constraints and global spatial dependencies. Experiments suggest the effectiveness and efficiency of MDSFGA.

Despite these encouraging results, several limitations and practical applicability conditions should be noted. The current MDSFGA mainly relies on multi-node historical load observations and timestamp-derived semantic information, such as hour-of-day and day-of-week cues. It does not explicitly incorporate richer exogenous variables, such as meteorological variables, electricity pricing, social events, or holiday indicators. Therefore, the proposed model is most suitable for scenarios where multi-node historical load data are sufficiently available, timestamps are reliable, and the forecasting task mainly depends on historical consumption regularities and spatiotemporal correlations among nodes. In more complex urban energy systems where abrupt external events or weather-sensitive demand changes play a major role, incorporating additional exogenous variables may further improve forecasting robustness.

Several avenues remain for future research. First, as the performance of the multi-scale encoder depends partly on the predefined number of decomposition modes, developing more adaptive or automated decomposition strategies could further enhance the flexibility across different scenarios. Second, the

design of semantic embeddings can be extended beyond timestamp cues by integrating meteorological conditions, electricity pricing, social events, and other context-aware variables. Such extensions may further improve the practical applicability of MDSFGA in complex power-system forecasting environments.

Data Sharing Agreement

The CELP datasets supporting this study are publicly available at: https://data.openei.org/submissions/153 or https://drive.google.com/file/d/1OiO-Eibj3gjXkfNQjJSWQKkxaUI-axx.

References

1. Aslam, S., Herodotou, H., Mohsin, S.M., Javaid, N., Ashraf, N., Aslam, S. A Survey on Deep Learning Methods for Power Load and Renewable Energy Forecasting in Smart Microgrids. Renewable and Sustainable Energy Reviews 2021, 144, 110992. https://doi.org/10.1016/j.rser.2021.110992
2. Bai, L., Yao, L., Li, C., Wang, X., Wang, C. Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting. Proceedings of the Advances in Neural Information Processing Systems, 2020, 17804-17815. https://dl.acm.org/doi/10.5555/3495724.3497218
3. Bai, S., Kolter, J.Z., Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. ArXiv:1803.01271, 2018. https://arxiv.org/abs/1803.01271
4. Cao, D., Wang, Y., Duan, J., Zhang, C., Zhu, X., Huang, C., Tong, Y., Xu, B., Bai, J., Tong, J., Zhang, Q. Spectral Temporal Graph Neural Network for Multivariate Time-Series Forecasting. Proceedings of the Advances in Neural Information Processing Systems, 2020, https://dl.acm.org/doi/10.5555/3495724.3497215
5. Chen, J., Pan, S., Peng, W., Xu, W. Bilinear Spatiotemporal Fusion Network: An Efficient Approach for Traffic Flow Prediction. Neural Networks 2025, 187. https://doi.org/10.1016/j.neunet.2025.107382
6. Chen, Y., Huang, J. Traffic Flow Prediction Model Based on Variational Mode Decomposition and Transformer. Proceedings of SPIE - The International Society for Optical Engineering, 2024, 13018. https://doi.org/10.1117/12.3023983
7. Chen, Y., Xu, P., Chu, Y., Li, W., Wu, Y., Ni, L., Bao, Y., Wang, K. Short-Term Electrical Load Forecasting Using the Support Vector Regression Model to Calculate the Demand Response Baseline for Office Buildings. Applied Energy 2017, 195, 659-670. https://doi.org/10.1016/j.apenergy.2017.03.034
8. Deng, J., Chen, X., Jiang, R., Song, X., Tsang, I.W. ST-Norm: Spatial and Temporal Normalization for Multi-variate Time Series Forecasting. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2021, 269-278. https://doi.org/10.1145/3447548.3467330
9. Di Giovanni, F., Giusti, L., Barbero, F., Luise, G., Liò, P., Bronstein, M. On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology. Proceedings of the International Conference on Machine Learning, 2023, 202, 7865-7885. https://dl.acm.org/doi/10.5555/3618408.3618720
10. Dragomiretskiy, K., Zosso, D. Variational Mode Decomposition. IEEE Transactions on Signal Processing 2014, 62, 531-544. https://doi.org/10.1109/TSP.2013.2288675
11. Fang, Y., Qin, Y., Luo, H., Zhao, F., Xu, B., Zeng, L., Wang, C. When Spatio-Temporal Meet Wavelets: Disentangled Traffic Forecasting via Efficient Spectral Graph Attention Networks. Proceedings of the 2023 IEEE 39th International Conference on Data Engineering, 2023, 517-529. https://doi.org/10.1109/ICDE55515.2023.00046
12. Fang, Y., Qin, Y., Luo, H., Zhao, F., Zheng, K. STWave+: A Multi-Scale Efficient Spectral Graph Attention Network with Long-Term Trends for Disentangled Traffic Flow Forecasting. IEEE Transactions on Knowledge and Data Engineering 2024, 36, 2671-2685. https://doi.org/10.1109/TKDE.2023.3324501
13. Fang, Z., Long, Q., Song, G., Xie, K. Spatial-Temporal Graph ODE Networks for Traffic Flow Forecasting. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2021, 364-373. https://doi.org/10.1145/3447548.3467430
14. García-Ascanio, C., Maté, C. Electric power demand forecasting using interval time series: A comparison between VAR and iMLP. Energy Policy, 2010, 38, 715-725. https://doi.org/10.1016/j.enpol.2009.10.007
15. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T. Recent Advances in Convolutional Neural Networks. Pattern Recognition 2018, 77, 354-377. https://doi.org/10.1016/j.patcog.2017.10.013

16. Guo, S., Lin, Y., Feng, N., Song, C., Wan, H. Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting. Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019, 922-929. https://doi.org/10.1609/aaai.v33i01.3301922

17. Jin, G., Li, F., Zhang, J., Wang, M., Huang, J. Automated Dilated Spatio-Temporal Synchronous Graph Modeling for Traffic Prediction. IEEE Transactions on Intelligent Transportation Systems 2023, 24, 8820-8830. https://doi.org/10.1109/TITS.2022.3195232

18. Lee, C. M., Ko, C. N. Short-term load forecasting using lifting scheme and ARIMA models. Expert Systems with Applications, 2011, 38, 5902-5911. https://doi.org/10.1016/j.eswa.2010.11.033

19. Li, F., Feng, J., Yan, H., Jin, G., Yang, F., Sun, F., Jin, D., Li, Y. Dynamic Graph Convolutional Recurrent Network for Traffic Prediction: Benchmark and Solution. ACM Transactions on Knowledge Discovery from Data 2023, 17. https://doi.org/10.1145/3532611

20. Li, M., Zhu, Z. Spatial-Temporal Fusion Graph Neural Networks for Traffic Flow Forecasting. Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021, 4189-4196. https://doi.org/10.1609/aaai.v35i5.16542

21. Lipton, Z. C., Berkowitz, J., Elkan, C. A Critical Review of Recurrent Neural Networks for Sequence Learning. ArXiv:1506.00019, 2015. https://arxiv.org/abs/1506.00019

22. Liu, H., Dong, Z., Jiang, R., Deng, J., Deng, J., Chen, Q., Song, X. Spatio-Temporal Adaptive Embedding Makes Vanilla Transformer SOTA for Traffic Forecasting. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, 4125-4129. https://doi.org/10.1145/3583780.3615160

23. Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. Proceedings of the 12th International Conference on Learning Representations, ICLR, 2024. https://openreview.net/forum?id=JePfAI8fah

24. Murad, M.M.N., Aktukmak, M., Yilmaz, Y. WPMixer: Efficient Multi-Resolution Mixing for Long-Term Time Series Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence 2025, 39, 19581-19588. https://doi.org/10.1609/aaai.v39i18.34156

25. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J. A Time Series is Worth 64 Words: Long-Term Forecasting with Transformers. Proceedings of the 11th International Conference on Learning Representations, ICLR, 2023. https://openreview.net/forum?id=Jbdc0vTOcol

26. Ong, S., Clark, N. Commercial and Residential Hourly Load Profiles for All TMY3 Locations in the United States. Open Energy Data Initiative, National Renewable Energy Laboratory 2014. https://doi.org/10.25984/1788456

27. Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y. N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting. Proceedings of the 8th International Conference on Learning Representations, ICLR, 2020. https://openreview.net/forum?id=r1ecqn4YwB

28. Rai, A., Shrivastava, A., Jana, K.C. Differential Attention Net: Multi-Directed Differential Attention Based Hybrid Deep Learning Model for Solar Power Forecasting. Energy 2023, 263, 125746. https://doi.org/10.1016/j.energy.2022.125746

29. Rehman, N.U., Aftab, H. Multivariate Variational Mode Decomposition. IEEE Transactions on Signal Processing 2019, 67, 6039-6052. https://doi.org/10.1109/TSP.2019.2951223

30. Schober, P., Schwarte, L.A. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia and Analgesia 2018, 126, 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

31. Shao, Z., Zhang, Z., Wang, F., Wei, W., Xu, Y. Spatial-Temporal Identity: A Simple yet Effective Baseline for Multivariate Time Series Forecasting. Proceedings of the 31st ACM International Conference on Information and Knowledge Management, 2022, 4454-4458. https://doi.org/10.1145/3511808.3557702

32. Shao, Z., Zhang, Z., Wei, W., Wang, F., Xu, Y., Cao, X., Jensen, C.S. Decoupled Dynamic Spatial-Temporal Graph Neural Network for Traffic Forecasting. Proceedings of the VLDB Endowment 2022, 15, 2733-2746. https://doi.org/10.14778/3551793.3551827

33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. Attention Is All You Need. Proceedings of the Advances in Neural Information Processing Systems, 2017, 5999-6009. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

34. Wang, D., Guo, G., Ouyang, T., Yu, D., Zhang, H., Li, B., Jiang, R., Xu, G., Deng, S. A Lightweight Spatio-Temporal Neural Network with Sampling-based Time Series Decomposition for Traffic Forecasting. IEEE Transactions on Intelligent Transportation Systems 2025, 26, 8682-8693. https://doi.org/10.1109/TITS.2025.3552010

35. Wang, H., Peng, J., Huang, F., Wang, J., Chen, J., Xiao, Y. MICN: Multi-Scale Local and Global Context Modeling for Long-Term Series Forecasting. Proceedings of the 11th International Conference on Learning Representations, ICLR, 2023. https://openreview.net/forum?id=zt53IDUR1U

36. Wang, S., Cao, J., Yu, P.S. Deep Learning for Spatio-Temporal Data Mining: A Survey. IEEE Transactions on Knowledge and Data Engineering 2022, 34, 3681-3700. https://doi.org/10.1109/TKDE.2020.3025580

37. Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J.Y., Zhou, J. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. Proceedings of the 12th International Conference on Learning Representations, ICLR, 2024. https://openreview.net/forum?id=7oLshfEIC2

38. Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. Proceedings of the 11th International Conference on Learning Representations, ICLR, 2023. https://doi.org/10.48550/arXiv.2210.02186

39. Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C. Graph WaveNet for Deep Spatial-Temporal Graph Modeling. Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019, 1907-1913. https://doi.org/10.24963/ijcai.2019/264

40. Wu, Z., Wang, Y., Zhang, L. MSSTN: Multi-Scale Spatial Temporal Network for Air Pollution Prediction. Proceedings of the IEEE International Conference on Big Data, 2019, 1547-1556. https://doi.org/10.1109/BigData47090.2019.9005574

41. Xu, Y., Han, L., Zhu, T., Sun, L., Du, B., Lv, W. Generic Dynamic Graph Convolutional Network for Traffic Flow Forecasting. Information Fusion 2023, 100. https://doi.org/10.1016/j.inffus.2023.101946

42. Yu, B., Yin, H., Zhu, Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018, 3634-3640. https://doi.org/10.24963/ijcai.2018/505

43. Yu, S., He, B., Fang, L. Multi-step short-term forecasting of photovoltaic power utilizing TimesNet with enhanced feature extraction and a novel loss function. Applied Energy 2025, 388. https://doi.org/10.1016/j.apenergy.2025.125645

44. Zeng, A., Chen, M., Zhang, L., Xu, Q. Are Transformers Effective for Time Series Forecasting? Proceedings of the AAAI Conference on Artificial Intelligence 2023, 37, 11121-11128. https://doi.org/10.1609/aaai.v37i9.26317

45. Zhao, H., Wu, Y., Ma, L., Pan, S. Spatial and Temporal Attention-Enabled Transformer Network for Multivariate Short-Term Residential Load Forecasting. IEEE Transactions on Instrumentation and Measurement 2023, 72, 1-11. https://doi.org/10.1109/TIM.2023.3305655

46. Zhang, Y., Zhong, K., Xie, X., Huang, Y., Han, S., Liu, G., Chen, Z. VMD-ConvTSMixer: Spatiotemporal Channel Mixing Model for Non-Stationary Time Series Forecasting. Expert Systems with Applications 2025, 271, 126535. https://doi.org/10.1016/j.eswa.2025.126535

47. Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., Jin, R. FEDformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting. Proceedings of the 39th International Conference on Machine Learning, 2022, 27268-27286. https://proceedings.mlr.press/v162/zhou22g/zhou22g.pdf

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).