Electrical Power and Energy Systems 160 (2024) 110074

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

Enhancing multivariate, multi-step residential load forecasting with spatiotemporal graph attention-enabled transformer

Pengfei Zhao [a], Weihao Hu [a], Di Cao [a,b,*], Zhenyuan Zhang [a], Wenlong Liao [c], Zhe Chen [d], Qi Huang [a,e]

[a] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
[b] Institute of Electronic and Information Engineering of UESTC in Guangdong, Dongguan, China
[c] Wind Engineering and Renewable Energy Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland
[d] Department of Energy Technology, Aalborg University, Aalborg, Denmark
[e] School of Information Engineering, Southwest University of Science and Technology, Mianyang, China

ARTICLE INFO

Keywords:
Residential load forecasting
Spatiotemporal modeling
Deep neural network

ABSTRACT

Short-term residential load forecasting (STRLF) holds great significance for the stable and economic operation of distributed power systems. Different households in the same region may exhibit similar consumption patterns owing to the analogous environmental parameters. Incorporating the spatiotemporal correlations can enhance the load forecasting performance of individual households. To this end, a spatiotemporal graph attention (STGA)-enabled Transformer is proposed for multivariate, multi-step residential load forecasting in this paper. Specifically, the multiple residential loads are cast to a graph and a Transformer with a graph sequence-to-sequence (Seq2Seq) structure is employed to model the multi-step load forecasting problem. Gated fusion-based STGA blocks are embedded in the encoder and decoder of the Transformer to extract dynamic spatial correlations and non-linear temporal patterns among multiple residential loads. A transform attention block is further designed to transfer historical graph observations into future graph predictions and alleviate the error propagation between the encoder and decoder. The embedding of multiple attention modules in the Seq2Seq framework allows us to capture the spatiotemporal correlations between residents and achieve confident inference of load values several steps ahead. Numerical simulations on residential data from three different regions demonstrate that the developed Transformer method improves multi-step load forecasting by 14.7% at least, compared to the state-of-the-art benchmarks.

1. Introduction

As our societies increasingly rely on diverse and intermittent sources of energy, optimizing the balance between electric supply and demand has gained paramount importance [1]. Short-term residential load forecasting (STRLF) emerges as a critical facet of this optimization process, enabling utilities and energy stakeholders to anticipate electricity consumption patterns at a granular level [2]. The ability to forecast electric short-term residential loads not only enhances the efficiency of energy distribution but also facilitates the integration of renewable energy sources and the implementation of demand-side management strategies [3,4]. However, this task is facing increasing challenges. The inherent complexity of human behavior, the variability of energy demand patterns, and the impact of external factors such as weather conditions pose significant obstacles to achieving accurate short-term residential load forecasts [5].

Traditional electric load forecasting methods, mainly designed for system-level applications, can be broadly categorized into two groups: statistical methods and machine learning methods. Statistical techniques, exemplified by exponential smoothing [6] and autoregressive integrated moving averaging (ARIMA) [7], concentrate on discerning the periodicity and trends within load time series [8]. These methods have found extensive application in load forecasting. On the other hand, machine learning methods provide increased flexibility for capturing non-linear relationships [9,10]. Recent years have witnessed the ascendancy of deep neural networks [11,12], including recurrent neural networks (RNNs) [4,13–16], convolutional neural networks (CNNs) [2,17], deep residual neural networks [18–20], and Transformer

* Corresponding author.
E-mail address: caodi@uestc.edu.cn (D. Cao).

https://doi.org/10.1016/j.ijepes.2024.110074
Received 4 January 2024; Received in revised form 3 April 2024; Accepted 3 June 2024
Available online 13 June 2024

frameworks based on attention mechanisms [21,22]. These deep-learning methodologies are now preferred choices for short-term load forecasting.

Contrary to system-level load demand forecasting, predicting the load demand of individual households is more challenging due to the high volatility nature of residential loads. Ref. [4] addresses the STRLF problem using long-short-term memory neural networks (LSTM). Another proposed approach in [23] involves an RNN method based on a pooling mechanism for forecasting the load of individual households. In [24], researchers propose to use Markov-chain mixture distribution model for very-short term residential load forecasting. Ref. [2] utilizes a CNN with a squeeze-and-excitation block to capture micrometeorological data, while Ref. [25] introduces a CNN and gated recurrent units (GRU) framework that seamlessly integrates into a mixture density network, enabling the direct prediction of probability density functions. A Transformer model with auto-correlation (Autoformer) [26] is employed to capture the temporal dependency of the residential load series. To address the challenge of limited training samples, transfer learning-based STRLF methods [20] are also proposed.

In practice, residents in the same region may exhibit similar consumption patterns due to the analogous environmental parameters. However, the aforementioned studies ignore the spatial correlations between different households, which, if incorporated, could improve the forecasting performance of individual households [27]. To this end, various spatiotemporal (ST) forecasting methods have been proposed in the literature. Historically, methods for ST forecasting have primarily relied on statistical and machine learning techniques. Ye et al. [28] utilize kernel density estimation and adaptive k-means methodologies to aggregate spatial data. Gilanifar et al. [29] propose a Bayesian spatiotemporal Gaussian process model to capture the relevance between different residential units. Zhao et al. [30] leverage spatial correlation information from multiple wind farms to construct a correlation-constrained sparsity-controlled vector autoregressive model, further solved by a constrained mixed integer nonlinear programming method. However, the random and intricate consumption behaviors of households render ST dependencies in residential load data challenging to address. As a result, traditional statistical and machine learning methods, constrained by their learning capabilities, struggle to grasp the complex mapping relationships inherent in STRLF tasks, leading to suboptimal accuracy in forecasting results [31].

In recent years, deep learning methods have gained significant attention for their ability to automatically extract feature representations from raw ST data. Numerous architectures combining convolutional layers with RNNs have been introduced for ST prediction. Jalalifar et al. [32] utilize a convolutional long short-term memory (CovLSTM) neural network for wind power forecasting, integrating the spatial information capturing ability of CNNs with the temporal dependency modeling capabilities of LSTM. Similarly, a CovLSTM-based encoder-decoder structure is proposed for photovoltaic generation forecasting [33]. Khodayar et al. [34] present a graph-driven LSTM combined with graph dictionary learning for behind-the-meter load and PV power forecasting. A convolutional gated recurrent unit (CovGRU) incorporating variational Bayesian inference is introduced for probabilistic wind speed forecasting [35]. However, CNN-based models are limited to considering absolute spatial relationships within a two-dimensional Euclidean space and require grid-like data such as images, which may not align well with the irregularities often found in typical residential points on a map.

In contrast, graph neural networks (GNNs) facilitate forecasting by iteratively updating node representations based on information from neighboring nodes, making them more suitable for graph-like data and offering an alternative approach for handling ST data. Huang et al. [36] utilize a graph convolutional network (GCN) for wide-area multiple bus load forecasting, where the adjacency matrix is computed based on geography and grid topology using maximal information coefficient values. Refs. [37,38] employ a self-adaptive adjacency matrix-based GCN to capture spatiotemporal relationships among multiple residential loads. Refs. [39,40] utilize multi-graph-based GCN to capture spatial correlations, while incorporating temporal convolutional networks (TCN) [39] and LSTM networks [40] to capture temporal patterns for electric vehicle charging station demand forecasting. Chen et al. [41] introduce a dynamic adjacency matrix to acquire time-varying spatial weight allocations for wind speed nodes, followed by employing GCNs and GRUs to capture spatial and temporal patterns, respectively. However, load correlations among residential units are highly intricate and subject to dynamic changes influenced by many factors such as electricity price, weather conditions, and geographical locations. GNNs typically operate on a static graph structure, which may not be well-suited for scenarios involving spatial load nodes. Although dynamic adjacency matrices for GNNs have been proposed to address the issue of fixed graph structures, they may still fail to capture the dynamic relevancies found in real-world scenarios. Moreover, the introduction of dynamic adjacency matrices presents new challenges related to computational complexity and the potential for overfitting [42].

Attention mechanisms allow models to dynamically focus on different parts of the input data, capturing complex dependencies more effectively [43]. This flexibility is advantageous for ST load forecasting tasks where the relationships between residential units are not easily captured by fixed graph structures. Many attention mechanism-enabled transformer structures have been proposed for time series forecasting. For instance, models like Informer [44], Autoformer [45], Pyraformer [46], and FEDformer [47] have exhibited success in capturing long-term dependencies compared to GNN-based methods [35]. However, these transformer models just focus on temporal dependencies in time series and ignore the spatial correlations among multiple variables. To incorporate spatial information into the transformer, several variations with spatiotemporal graph attention have been introduced. For instance, Yu et al. [48] introduce a spatiotemporal graph transformer network for trajectory prediction. This model comprises separate spatial transformers and temporal transformers designed to capture complex ST interactions. Wang et al. [49] propose a synchronous spatiotemporal graph transformer that integrates the benefits of attention mechanisms and graphs for traffic data prediction. However, these transformers with ST graph attention are not yet sufficiently mature for ST residential load forecasting. Three unsolved challenges arise when applying transformer models to ST residential load forecasting:

(1) Attention Modules: Designing attention modules capable of effectively capturing dynamic spatial correlations and nonlinear temporal dependencies in residential load series, characterized by high volatility and stochasticity, is a significant challenge.
(2) Integration of Temporal and Spatial Information: Current transformer models with ST attention typically handle spatial and temporal information separately, and then merely concatenate them in the output layer. However, this approach may not seamlessly integrate spatial and temporal factors, potentially resulting in inaccurate and unreliable predictions.
(3) Error Propagation: Errors in the encoder can propagate to the decoder through attention mechanisms and subsequent decoder layers [50]. Therefore, it is necessary to mitigate error propagation between the encoder and decoder.

To this end, a spatiotemporal graph attention mechanism (STGA)-enabled transformer network for short-term multi-resident, multi-step load forecasting is proposed in this paper. Unlike traditional GCN or CNN-based spatiotemporal load forecasting methods, the proposed model fully leverages the advantages of the attention mechanism to capture spatiotemporal factors. This allows the proposed method to bridge the gap between modeling long-range dependencies in load series and capturing dynamic spatial correlations among multiple households. The main contributions of this paper are as follows:

(1) A novel *spatiotemporal graph attention-enabled Transformer framework* is proposed for multivariate residential load forecasting. The historical data of multiple residential loads are cast to graph signals and a spatiotemporal graph attention model is employed to capture the dynamic spatial and temporal correlations among resident units. The multiple pieces of spatiotemporal information are further seamlessly integrated by gated fusion units. The graph attention mechanism allows the proposed method to automatically focus on the information most beneficial for forecasting each individual residential load, thus achieving robustness against unrelated spatial information. This differentiates it from GCN-based spatiotemporal models, which require predefining the graph structure that may not match the dynamic graph structures in real residential load scenarios. The proposed method also differs from traditional transformer-based time series forecasting models, which disregard spatial correlations among multiple households.

(2) A *Transformer with graph Seq2Seq structure* is designed for multi-step residential load forecasting. The graph Seq2Seq model enables the proposed method to efficiently capture the complex temporal correlations in multi-step forecasting tasks. The designed transform attention module further allows the transfer of historical graph observations to future predictions and the alleviation of error propagation. This differentiates from typical multi-step forecasting methods that may suffer from obvious performance degradation with the increasing number of forecasting steps.

The remainder of this paper is organized as follows: A detailed description of the proposed method is shown in Section II. The case study is presented in Section III and Section IV concludes this paper.

2. Methodology

2.1. Problem statement

Let $\mathscr{G} = (\mathscr{V}, \mathscr{E}, \mathscr{A})$ denote a weighted graph representing households with spatial and temporal relevance, where $\mathscr{V}$ is a set of $N = |\mathscr{V}|$ nodes of distributed power systems; $\mathscr{E}$ is a set of edges expressing connectivity between nodes; $\mathscr{A}_{v_i,v_j} \in \mathscr{A} \in \mathbb{R}^{N\times N}$ represents weighted adjacency matrix between nodes v_i and v_j. In the STRLF task, $\mathscr{A}$ can be represented as correlations between different residential units (different from traditional spatiotemporal methods that predetermine $\mathscr{A}$, this paper employ spatial attention mechanism dynamically learn $\mathscr{A}$). The load demands of multiple houses at time t are denoted as a graph signal $X_t \in \mathbb{R}^{N\times d}$, where d denotes the feature dimension of each node. Given historical load observations of W time steps at time t, denoted as $\mathscr{X} = (\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}, \ldots, \boldsymbol{X}_{t_W}) \in \mathbb{R}^{W\times N\times d}$, our goal is to predict multiple residential loads of next H time steps, denoted as $\widehat{\boldsymbol{Y}} = (\widehat{\boldsymbol{X}}_{t_{W+1}}, \widehat{\boldsymbol{X}}_{t_{W+2}}, \ldots, \widehat{\boldsymbol{X}}_{t_{W+H}}) \in \mathbb{R}^{H\times N}$.

There are three challenges in the multivariate and multi-step residential load forecasting problem: (1) the strong stochasticity make it difficult to achieve accurate forecasting of residential load; (2) the spatiotemporal correlations between different residents are difficult to capture owing to its dynamic characteristics and the complex relationship between residents; (3) the complex temporal correlations in multi-step forecasting renders the problem challenging to solve. To this end, this paper proposes an STGA mechanism-enabled Transformer for the multivariate and multi-step residential load forecasting problem.

2.2. The proposed residential load forecasting method

The following part will provide a detailed description of the proposed model, a graphic illustration of which is shown in Fig. 1. It can be observed that our proposed method is an encoder-decoder architecture, which is extremely advantageous for multi-step load forecasting. Both the encoder and decoder consist of several STGA blocks with skip connections, and each STGA block contains spatial attention and temporal attention modules, as well as gated fusion units. There is a transform attention block between the encoder and decoder. In addition, a spatiotemporal (ST) embedding block is also added to the STGA block and transform attention block.

(1) Spatiotemporal Embedding

The spatial information of households plays a crucial role in the prediction model. In this paper, we address the challenge of incorporating spatial information by embedding the spatial characteristics of each household into a vector, thus preserving the graph structure of future loads. Since the location information of households is typically unavailable in practice, one-hot coding method is adopted to encode the spatial information. Specifically, the embedded spatial information for each household at a certain step can be represented as a vector $e^S \in \mathbb{R}^{\psi}$. Therefore, the spatial embedding of all N households for $W + H$ time steps is $\boldsymbol{E}^S \in \mathbb{R}^{(W+H)\times N\times\psi}$. Then, $\boldsymbol{E}^S$ is fed into two fully connected (FC) layers to co-train the pre-learned spatial features with the whole model.

The spatial embedding, although effective in expressing static location information for each household, is limited in capturing dynamic relationships between multiple households as time changes. To address this limitation, temporal embedding is utilized, encoding time information from each time step into a vector. Incorporating temporal

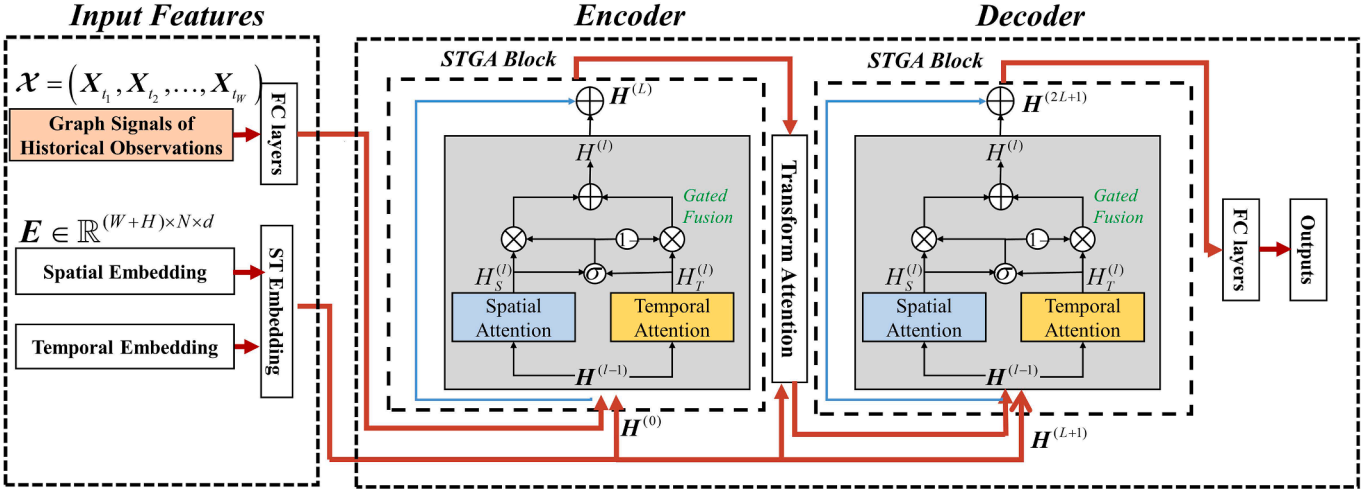

Fig. 1. The overall framework of the proposed method.

embedding enables us to capture evolving relationships between households over time, enhancing the model's ability to capture temporal dynamics in the data. One-hot coding method is employed to encode the hour-of-day and day-of-week indexes as temporal embedding. Consequently, the embedded temporal information for each household at a certain step can be represented as $e^T \in \mathbb{R}^{\tau}$. The temporal embedding of all N households for $W+H$ time steps is denoted as $\boldsymbol{E}^T \in \mathbb{R}^{(W+H)\times N\times \tau}$. Then, a two-layer FC is employed to co-train temporal embedding.

To represent time-variant multiple residential loads, the spatial and temporal embedding is concatenated, denoted as spatiotemporal (ST) embedding $\boldsymbol{E}^{ST} = [\boldsymbol{E}^S, \boldsymbol{E}^T] \in \mathbb{R}^{(W+H)\times N\times(\psi+\tau)}$. $\boldsymbol{E}^{ST}$ is then passed through a FC layer, becomes $\boldsymbol{E} = \mathbb{R}^{(W+H)\times N\times d}$. The ST embedding block contains location and temporal information for each residential load at each time step, which will be used for the STGA and transform attention blocks.

(2) Spatiotemporal Graph Attention Block

It can be seen from Fig. 1 that a STGA block includes spatial attention, temporal attention and a gated fusion unit. The input of l-th STGA block is denoted as $\boldsymbol{H}^{(l-1)}$, where the hidden state of node v_i at time step t_j is $h^{l-1}_{v_i,t_j}$. $\boldsymbol{H}^{(l-1)}$ passes through spatial attention module and temporal attention module, producing $\boldsymbol{H}_S^{(l)}$ and $\boldsymbol{H}_T^{(l)}$, where the hidden state of node v_i at time step t_j is represented as $hs^l_{v_i,t_j}$ and $ht^l_{v_i,t_j}$, respectively. $\boldsymbol{H}_S^{(l)}$ and $\boldsymbol{H}_T^{(l)}$ are adaptively fused using gated fusion unit, obtaining final output of STGA block, denoted as $\boldsymbol{H}^{(l)}$.

i) Spatial Attention: Due to shared conditions such as temperature and holidays, multiple households in the same region may show similar consumption patterns. Unlike traditional CNN or GCN-based spatiotemporal modeling methods, this paper develops a spatial attention mechanism that adaptively extracts correlations between multiple households. The core behind this mechanism is to dynamically assign weights to different nodes (households) at different time steps, allowing the model to focus on the most relevant and influential nodes for accurate forecasting. Fig. 2 depicts the spatial attention module. For node v_i at time step t_j, the output of spatial attention is calculated through the weighted sum of all nodes:

$$hs^{(l)}_{v_i,t_j} = \sum_{v\in\mathscr{V}} \alpha_{v_i,v} \cdot h^{(l-1)}_{v,t_j} \tag{1}$$

where $\mathscr{V}$ denotes the set of all nodes; $\alpha_{v_i,v}$ denotes the attention score which indicates the importance of node v to v_i. Note that the summation of $\sum_{v\in\mathscr{V}}\alpha_{v_i,v}$ equals to 1: $\sum_{v\in\mathscr{V}}\alpha_{v_i,v} = 1$. The attention score is learned using both consumption patterns at a certain time step and ST embedding. The dot-product attention is utilize in this work [43], and $\alpha_{v_i,v}$ can be calculated as:

$$s_{v_i,v} = \frac{\left\langle h^{(l-1)}_{v_i,t_j} \middle\| e_{v_i,t_j}, h^{(l-1)}_{v,t_j} \middle\| e_{v,t_j} \right\rangle}{\sqrt{2d}} \tag{2}$$

$$\alpha_{v_i,v} = \frac{\exp\left(s_{v_i,v}\right)}{\sum_{v_r\in\mathscr{V}}\exp\left(s_{v_i,v_r}\right)} \tag{3}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product; e_{v_i,t_j} denotes ST embedding of node v_i at time t_j; $\|$ denotes the concatenate operator; $2d$ denotes the dimension of $h^{(l-1)}_{v_i,t_j}\|e_{v_i,t_j}$. Multi-head attention mechanism is utilized in this paper [43], which is achieved by concatenating K parallel attention mechanism. Therefore, Eqs. (1)–(3) become:

$$s^{(k)}_{v_i,v} = \frac{\left\langle f^{(k)}_{s,1}\left(h^{(l-1)}_{v_i,t_j}\|e_{v_i,t_j}\right), f^{(k)}_{s,2}\left(h^{(l-1)}_{v,t_j}\|e_{v,t_j}\right)\right\rangle}{\sqrt{d_1}} \tag{4}$$

$$\alpha^{(k)}_{v_i,v} = \frac{\exp\left(s^{(k)}_{v_i,v}\right)}{\sum_{v_r\in\mathscr{V}}\exp\left(s^{(k)}_{v_i,v_r}\right)} \tag{5}$$

$$hs^{(l)}_{v_i,t_j} = \|_{k=1}^{K}\left\{\sum_{v\in\mathscr{V}}\alpha^{(k)}_{v_i,v}\cdot f^{(k)}_{s,3}\left(h^{(l-1)}_{v,t_j}\right)\right\} \tag{6}$$

where d_1 equals to d/K, $f^{(k)}_{s,1}(\cdot)$, $f^{(k)}_{s,2}(\cdot)$ and $f^{(k)}_{s,3}(\cdot)$ represent different learnable projects, denoted as:

$$f(x) = \mathrm{ReLU}(x\mathbf{W}+\mathbf{b}) \tag{7}$$

where $\mathbf{W}$ and $\mathbf{b}$ denote the learnable parameters; ReLU is activation function.

ii) Temporal Attention: The load patterns of residential units at specific time steps exhibit a strong dependency on their past observations, and these relationships evolve nonlinearly over time. Traditional spatiotemporal methods often rely on RNN and TCN to model temporal dependencies and fall short in capturing long-term temporal dependencies in load series. In contrast, this paper introduces a temporal attention mechanism to effectively capture these nonlinear temporal patterns. The essence of this mechanism lies in dynamically assigning varying weights to different previous observations at different time steps. Fig. 3 shows a graphical representation of the temporal attention module. For node v_i at time t_j, the output of temporal attention is the weighted sum of a specific number of recent load observations:

$$ht^{(l)}_{v_i,t_j} = \sum_{t\in N_{t_j}}\beta_{t_j,t}\cdot h^{(l-1)}_{v_i,t} \tag{8}$$

where $\mathscr{N}_{t_j}$ denotes the set of time steps prior to time t_j; $\beta_{t_j,t}$ denotes the attention score which indicates the importance of time t to t_j. The summation of $\sum_{t\in N_{t_j}}\beta_{t_j,t}$ is also equals to 1: $\sum_{t\in N_{t_j}}\beta_{t_j,t} = 1$. Similar to spatial attention, the temporal attention score $\beta_{t_j,t}$ can be calculated as:

$$u_{t_j,t} = \frac{\left\langle h^{(l-1)}_{v_i,t_j}\|e_{v_i,t_j}, h^{(l-1)}_{v_i,t}\|e_{v_i,t}\right\rangle}{\sqrt{2d}} \tag{9}$$

$$\beta_{t_j,t} = \frac{\exp\left(u_{t_j,t}\right)}{\sum_{t_r\in\mathscr{N}_{t_j}}\exp\left(u_{t_j,t_r}\right)} \tag{10}$$

This procedure enables the model to learn the relevance between different time steps by considering both historical load patterns and ST

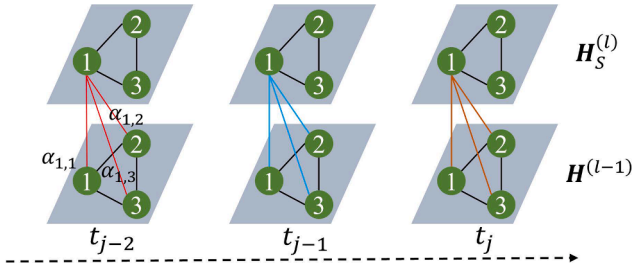

Fig. 2. A graphical illustration of the spatial attention module.

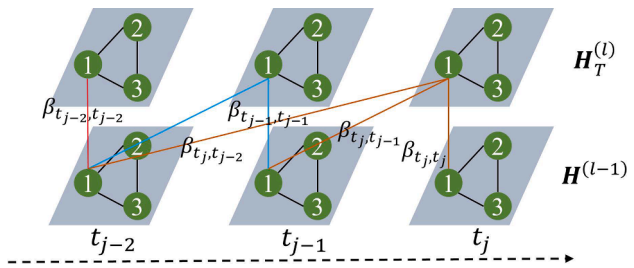

Fig. 3. A graphical illustration of the temporal attention module.

embedding. Similar to spatial attention, temporal attention employs multi-head attention by concatenating parallel temporal attention mechanisms [43], denoted as:

$$u_{t_j,t}^{(k)} = \frac{\left\langle f_{t,1}^{(k)}\left(h_{v_i,t_j}^{(l-1)} \| e_{v_i,t_j}\right), f_{t,2}^{(k)}\left(h_{v_i,t}^{(l-1)} \| e_{v_i,t}\right)\right\rangle}{\sqrt{d_1}} \tag{11}$$

$$\beta_{t_j,t}^{(k)} = \frac{\exp\left(u_{t_j,t}^{(k)}\right)}{\sum_{t_r \in \mathcal{N}_{t_j}} \exp\left(u_{t_j,t_r}^{(k)}\right)} \tag{12}$$

$$ht_{v_i,t_j}^{(l)} = \|_{k=1}^{K}\left\{\sum_{t \in \mathcal{N}_{t_j}} \beta_{t_j,t}^{(k)} \cdot f_{t,3}^{(k)}\left(h_{v_i,t}^{(l-1)}\right)\right\} \tag{13}$$

where $u_{t_j,t}^{(k)}$ represents the relevance between time t and t_j; $\beta_{t_j,t}^{(k)}$ denotes attention score of k-th head.

(iii) Gated Fusion: As mentioned earlier, both temporal and spatial information play significant roles in determining future load demands. To effectively integrate these two sources of information, a gated fusion unit is introduced. This unit is specifically designed to adaptively combine and fuse the learned temporal and spatial information. The output of spatial and temporal attention mechanism are $\boldsymbol{H}_S^{(l)}$ and $\boldsymbol{H}_T^{(l)}$, fused by:

$$\boldsymbol{H}^{(l)} = \xi \odot \boldsymbol{H}_S^{(l)} + (1-\xi) \odot \boldsymbol{H}_T^{(l)} \tag{14}$$

where:

$$\xi = \sigma\left(H_S^{(l)} \boldsymbol{W}_{z,1} + H_T^{(l)} \boldsymbol{W}_{z,2} + \boldsymbol{b}_z\right) \tag{15}$$

where $\boldsymbol{W}_{z,1} \in \mathbb{R}^{D\times D}$, $\boldsymbol{W}_{z,2} \in \mathbb{R}^{D\times D}$ and $\boldsymbol{b}_z \in \mathbb{R}^D$ are learnable weights and bias; $\sigma(\cdot)$ denotes sigmoid activation function; ξ is the gated unit; $\odot$ denotes element-wise product. It can be seen that gated fusion unit can achieve an adaptive combination of learned spatial and temporal information.

(3) Transform Attention Block

The constrained learning capacity of traditional spatiotemporal approach poses a significant challenge in making multi-step load forecasts for residential loads. In response, this paper introduces a transform attention block positioned between the encoder and decoder. This innovative block is designed to adaptively select crucial features from historical observations [51] and mitigate error propagation between the encoder and decoder. Fig. 4 gives a graphical depiction of the transform attention module. Similar to Eqs. (4)–(6), the relevance between historical steps $t(t_1, \ldots t_W)$ and forecast steps $t_j(t_{W+1}, \ldots, t_{W+H})$ can be calculated as:

$$\lambda_{t_j,t}^{(k)} = \frac{f_{tr,1}^{(k)}\left(e_{v_i,t_j}\right), f_{tr,2}^{(k)}\left(e_{v_i,t}\right)}{\sqrt{d}} \tag{16}$$

Then it is normalized by:

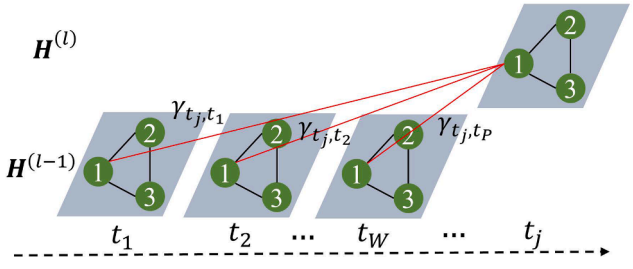

Fig. 4. A graphical illustration of the transform attention module.

$$\gamma_{t_j,t}^{(k)} = \frac{\exp\left(\lambda_{t_j,t}^{(k)}\right)}{\sum_{t_W}^{t_r=t_1} \exp\left(\lambda_{t_j,t_r}^{(k)}\right)} \tag{17}$$

By the operation of the transform attention block, the encoded historical load patterns is transformed to decoded forecasting loads by adaptively selecting importance features from historical observations using $\gamma_{t_j,t}^{(k)}$:

$$h_{v_i,t_j}^{(l)} = \|_{k=1}^{K}\left\{\sum_{t_W}^{t=t_1} \gamma_{t_j,t}^{(k)} \cdot f_{tr,3}^{(k)}\left(h_{v_i,t}^{(l-1)}\right)\right\} \tag{18}$$

The learnable parameters in Eqs. (13)–(15) are shared by all the households and all the time steps.

(4) Information Flow of the Proposed Model

As shown in Fig. 1, our proposed model is an encoder-decoder framework. To fully utilize prior knowledge and preserve the graph structure, ST embedding is fed forward through three modules: the encoder block, the transform attention block, and the decoder block for joint training. Input features $\mathcal{X} \in \mathbb{R}^{W\times N\times d}$ are fed forward through L STGA blocks to learn the dynamic spatial and temporal factors between multiple residential units, and produces $\boldsymbol{H}^{(L)} \in \mathbb{R}^{W\times N\times d}$. Following the encoder, a transform attention block is used to transfer $\boldsymbol{H}^{(L)}$ to prediction loads, producing $\boldsymbol{H}^{(L+1)} \in \mathbb{R}^{H\times N\times d}$. The integration of transform attention block allows the proposed model to transfer historical observations into future predictions and alleviate the problem of error propagation. Next, $\boldsymbol{H}^{(L+1)}$ passes through L decoder STGA blocks, and becomes $\boldsymbol{H}^{(2L+1)} \in \mathbb{R}^{H\times N\times d}$. Finally, $\boldsymbol{H}^{(2L+1)}$ is squeezed into a 2D matrix and fed into a FC layer, obtaining the final predictions $\widehat{\boldsymbol{Y}} \in \mathbb{R}^{H\times N}$, where H eques to the forecasting steps and N eques to the total number of households. Our proposed method allows for end-to-end training via gradient descent by minimizing the regression loss function.

2.3. Implement of the proposed method

The flowchart of the proposed method is depicted in Fig. 5 and comprises the following detailed implementation steps:

(1) Data Collection: Historical observations for various households and time variables, such as time-of-day and day-of-week indices, are collected.
(2) Handling Outliers and Missing Values: Raw data may contain some outliers and missing values due to the manipulation of meter readings or packet loss, which have negative impact on the training of the forecasting model. Therefore, the specific thresholds are set for each area load to identify outliers. The outliers are, then, converted to Nan values that are replaced by the values calculated by linear interpolation method.
(3) Feature Normalization: All datasets undergo normalization using the mean and standard deviation of the training set, calculated as follows:

$$\widehat{\mathscr{X}} = \frac{\mathscr{X} - mean(\mathscr{X}_{train})}{std(\mathscr{X}_{train})} \tag{19}$$

where $\widehat{\mathscr{X}}$ and $\mathscr{X}$ denote normalized and raw data, respectively; $mean(\mathscr{X}_{train})$ and $std(\mathscr{X}_{train})$ represent the mean and standard deviation of training set, respectively.
(4) Data Division: The selected datasets are divided into training, validation, and test sets in an 8:1:1 ratio.
(5) Model Training:
- Hyperparameters Tuning: The training set is used to train the model parameters, while the validation set is utilized to fine-tune hyperparameters using a grid search method until the desired accuracy is achieved.

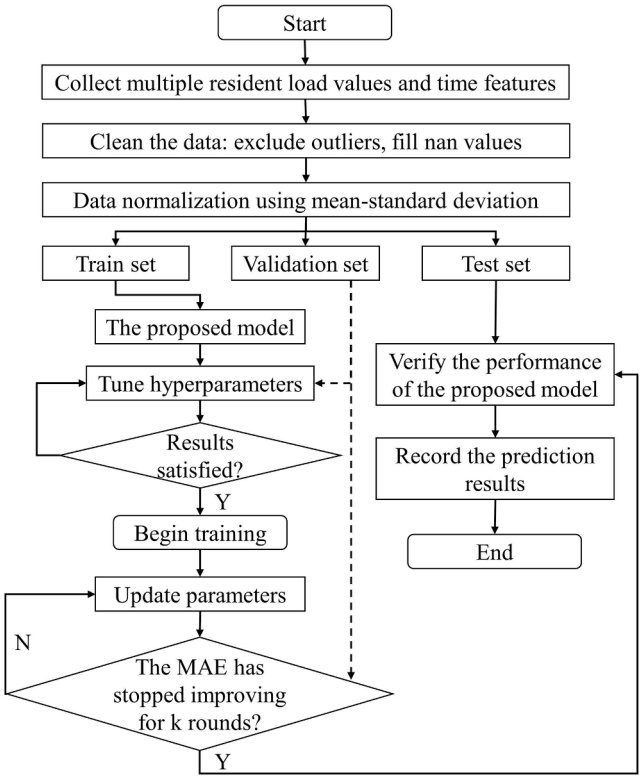

Fig. 5. Flowchart of the proposed method.

- Formal Training: An early stopping strategy is implemented to improve training efficiency. Specifically, if the evaluation metric on the validation set does not improve after k rounds, the training process is terminated, and the model parameters are saved.

(6) Result recording: Trained model parameters are loaded, evaluation indicators are calculated, and forecasting values are saved for the test set.

3. Case study

3.1. Dataset description

Publicly available datasets from three regions are used to validate the effectiveness of the proposed method. Three datasets are collected from two open websites.

(1) ***Open EI Residential Load Data*** [52]: This dataset contains hourly residential load demands from a variety of cities and states across the United States. For this study, one-year load data from New York (NY) and Texas (TX) in 2012 are selected. The New York dataset comprises 24 individual households, while the Texas dataset comprises 60 individual households.

(2) ***SGSC Project Dataset*** [53]: The dataset contains electricity consumption data from over 10,000 individual households in New South Wales (NSW), Australia, spanning the years 2010 to 2014 with a 30-minute time resolution. For this study, 64 household load datasets covering the period from January 1, 2013, to December 31, 2013, are selected.

The statistical characteristics of three datasets are shown in Table 1, including house number, minimum, maximum and standard deviation of target datasets. Fig. 6 depicts the load profiles of eight randomly selected houses in three areas. It can be seen that load profiles in the same city are obviously very similar, and correlations between multiple houses change dynamically over time. Note that that each area contains anomalous households, such as House 7 in TX and House 1 in NSW, and that these anomalous households may have a negative impact on the spatiotemporal model. The proposed Transformer model with attention mechanism is specifically designed to capture these dynamic spatial correlations and non-linear temporal patterns. In addition, it can be seen the load profile of the NSW dataset is more volatile than that of the NY and TX datasets.

Table 1
Statistical characteristics of three datasets.

Area	House Number	Min (kWh)	Max (kWh)	Mean (kWh)	Std. (kWh)
NY	24	1.196	71.863	15.166	11.720
TX	60	1.118	46.499	8.749	6.069
NSW	64	0.001	6.478	0.468	0.571

3.2. Benchmarks & hyperparameters

Various models served as benchmarks in the case study, encompassing the following algorithms: fully connected neural network (FCNN), LSTM [4], CNN-GRU [25], original Transformer, Autoformer [26], and self-adaptive Graph WaveNet (Ada-GWN) [37]. Specifically, FCNN, LSTM, and CNN-GRU represent the three most popular deep frameworks for short-term residential load forecasting. Transformer and Autoformer stand out as two state-of-the-art Seq2Seq frameworks designed to capture long-term temporal dependencies. Ada-GWN represents the state-of-the-art spatiotemporal modeling approach, employing GCN to capture spatial correlations.

For hyperparameters, FCNN consists of three hidden layers with 32, 64 and 32 hidden units. The settings for LSTM, CNN-GRU, and Ada-GWN are consistent with references [4,25,37], respectively. The hyperparameter configurations for the original Transformer and Autoformer methods align with those specified in ref. [26]. The candidate hyperparameters of the proposed method are outlined in Table 2. The validation set is utilized to fine-tune the hyperparameters of all models, and grid-search is used to determine the optimal values. The Adam optimizer is used to train the proposed model for 100 epochs, with an initial learning rate set to 0.001. To be fair, an early stopping strategy is employed to enhance the training efficiency of all methods. To evaluate the performance of the various methods, two widely accepted indicators, namely the mean absolute percentage error (MAPE) and mean absolute error (MAE), are utilized in the simulation. The average MAPE and MAE values are calculated by averaging the individual MAPE and MAE values for all households in a specific area. All simulations are conducted on a desktop computer equipped with an Intel Core i7-9750H CPU @ 2.60 GHz and 16 GB RAM.

3.3. Comparison with state-of-the-art

(1) Single-Step Forecasting: The results achieved by different methods on NSW dataset are shown in Table 3, which are averaged value of five repeated experiments. In the table, the term "single-resident forecasting" indicates training a distinct model for each resident. On the other hand, "multi-resident forecasting" involves training a single global model that considers ST data from all residents collectively. Owing to the strong volatility of single residential load, the FCNN method fails to capture the load consumption pattern of each resident. Its MAPE on single resident forecasting task is 49.04 %, which is very large. When the recurrent neural network is employed to capture the temporal correlations, the LSTM and CNN-GRU methods achieve better performance than the FCNN method. But their forecasting errors are still relatively high. By contrast, the attention mechanism-informed Seq2Seq structure allows the capture of the long-term temporal dependencies. This enables the Transformer, Autoformer, and the proposed method to outperform other benchmarking methods. When the historical data of

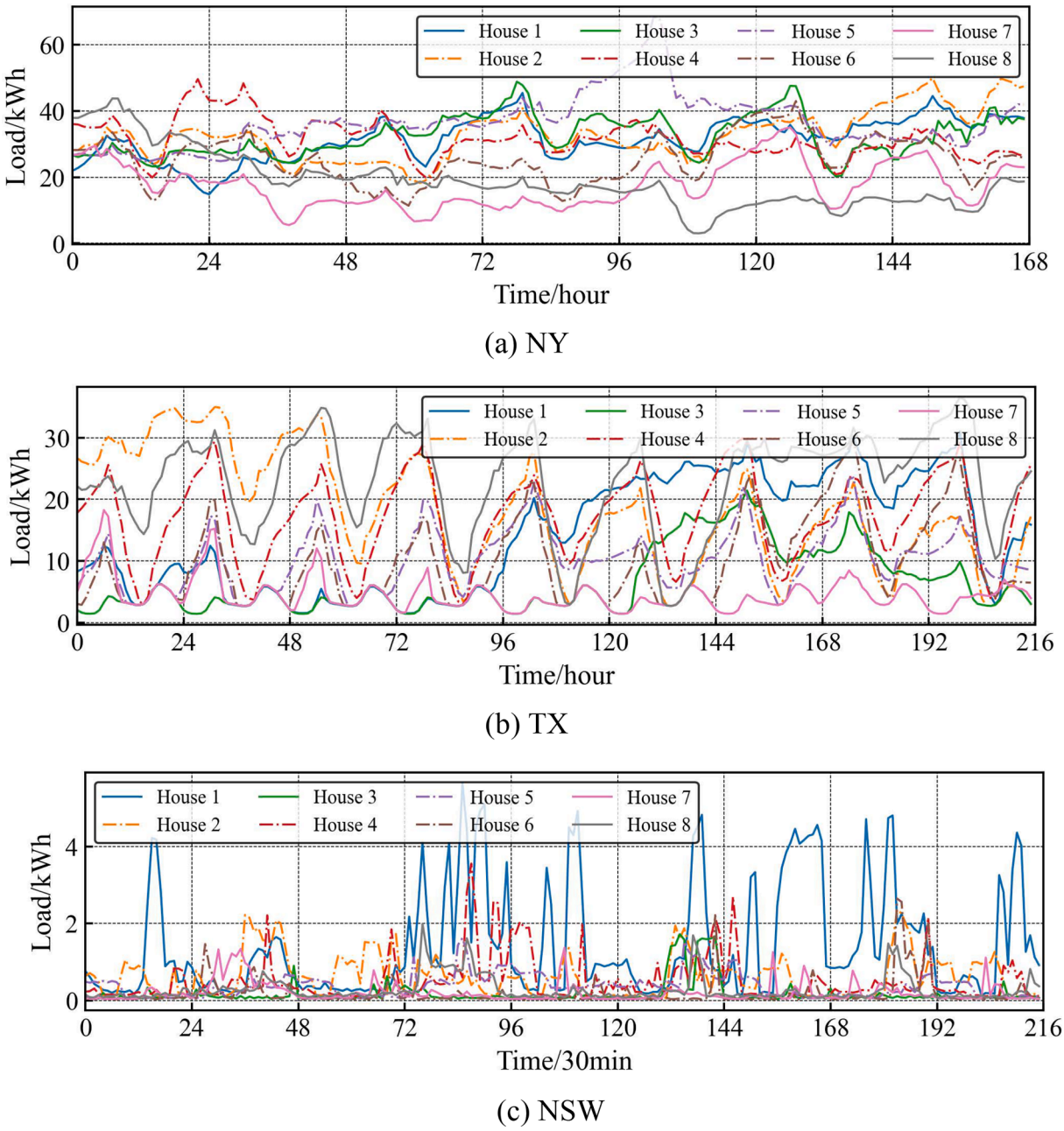

Fig. 6. Load profiles of eight randomly selected houses in three areas.

Table 2
Hyperparameters.

Hyper-parameters	Candidates
Hidden units of FC layers	[50,100,150]
Attention heads	8
Encoder layer	[1,2,4]
Decoder layer	[1,2,4]
Representation dimension (d)	[32,64,128]
History length	[6,12,24,72,168]

Table 3
Single-step forecasting results of various methods on NSW dataset.

Methods	*Single-resident forecasting*		*Multi-resident forecasting*	
	MAPE	*MAE*	*MAPE*	*MAE*
FCNN	49.04 %	0.171	58.01 %	0.228
LSTM [4]	44.31 %	0.148	51.41 %	0.174
CNN-GRU [25]	42.30 %	0.144	43.30 %	0.151
Transformer	41.39 %	0.147	45.56 %	0.154
Autoformer [26]	**37.56 %**	**0.135**	39.52 %	0.143
Ada-GWN [37]	45.21 %	0.156	35.15 %	0.137
Proposed	38.21 %	0.141	**31.45 %**	**0.124**

multiple residents are utilized by the forecasting model, the performance for FCNN method become even worse. The LSTM, Transformer, and Autoformer methods also suffer from obvious performance degradation owing to the lack of spatial feature extraction capability. By contrast, the Ada-GWN method employs a self-adaptive adjacency matrix-based graph neural network to explore the hidden spatial dependencies between residents. This allows it to achieve obvious performance enhancement when the data of multiple residents are used. The benefit of considering spatial correlations is observed here. Although the CNN-GRU method also applies convolution network to

extract the spatial correlations, it suffers from slight performance degradation owing to the negative impact of spatial uncorrelated information. Different from the convolution operation used by Ada-GWN and CNN-GRU methods that are sensitive to the spatial information, the proposed method employs STGA model to capture both the spatial and temporal correlations between residents and attentive to the most relevant features. This allows the proposed method to better exploit the spatial correlations to enhance its performance on each individual resident. As a result, it obtains performance enhancement when multiple residential data are utilized. The results demonstrate the superiority of the proposed method over other learning-based methods in exploiting the spatial correlations.

Further tests are carried out utilizing the NY and TX dataset to further evaluate the performance of the proposed method. The performances for various methods on the two datasets are listed in Tables 4 and 5, respectively. That the proposed method outperforms other forecasting methods on single residential load forecasting task. The STGA module further allows it to extend it lead when the data of multiple residents are used. The results are consistent with those observed in Table 3 and demonstrate the superiority of the proposed method.

The computational time for various methods are presented in Table 6. It can be seen from the table that the calculation efficiency of FCN, LSTM, and CNN-GRU methods are relatively high. Owing to the complex network structure, the calculation time of Transformer, Autoformer, Ada-GWN, and the proposed methods are higher than that of the FCN, LSTM, and CNN-GRU methods. Compared with the Transformer, Autoformer, and Ada-GWN methods, the proposed method can achieve better performance with less training time. Its calculation time is reduced by 32.12 %, 34.44 %, and 17.59 % when compared with that of the Transformer, Autoformer, and Ada-GWN methods, respectively. Note that the calculation time refers to the offline training time for the learning-based methods. When the training procedure is completed, they can provide the forecasting values in few milliseconds. This can meet the practical requirement of system operator and is one of the main advantage of the learning-based method.

(2) Multi-Step Forecasting: Tests are carried out to evaluate the multi-step forecasting performance of the proposed method. The multi-step forecasting performance for different methods are listed in Table 7. It can be observed from the table that MAPE and MAE values obtained by the FCNN, LSTM and CNN-GRU methods on NY dataset when the forecasting step is set to 3 are much larger than that achieved on single-step forecasting tasks. More obvious performance degradation is observed for these methods when increasing the forecasting step to 6 and 24. This demonstrates the difficulty in capturing the inherent temporal correlations on multi-step forecasting tasks. In contrast, owing to the strong time feature series modeling capability of Seq2Seq model, the Transformer and Autoformer methods achieve better performance than the FCNN, LSTM, and CNN-GRU methods. The Ada-GWN method also performs well in multi-step load forecasting. However, it suffers from obvious performance degradation when increasing the number of forecasting step. The disadvantage of convolution operations on modeling long-range dependencies is observed here. Different from the aforementioned methods, the proposed method employs a transformer network with graph Seq2Seq structure to learn the complex temporal correlations in multi-step forecasting task. This allows the proposed method to outperform other benchmarking methods by a large margin when the forecasting step is set to 3 and further extends its lead when increasing the forecasting step to 6 and 24. Similar phenomenon can be observed in the forecasting task of TX and NSW dataset, where the MAPEs for the proposed method outperforms other method by 15.80 %, 29.74 % and 16.39 % at least when the forecasting step is set to 3, 6 and 24, respectively. The results illustrate the advantage of the proposed Seq2Seq-based method on the multi-step forecasting task.

To further evaluate the multi-step forecasting performance for different methods in detail, the MAPE values obtained by various methods on each moment when the forecasting step is set to 6 are shown in Fig. 7. It can be seen from the figure that the MAPE values obtained by the baseline models increase rapidly as the number of steps increases. The MAPE values achieved by the Ada-GWN method increase at a faster rate than that of the Autoformer method. This observation further illustrates the disadvantage of graph neural networks in capturing long-term temporal dependencies. By contrast, the increase of MAPE for the proposed method is much gentle than other methods. It achieves relatively stable performance and outperforms other methods by a large margin on the last two moments. The advantage of the proposed method can also be observed on TX dataset. The results further demonstrate the superiority of the proposed method over other methods on multi-step forecasting tasks.

The actual and predicted values of various methods on two randomly selected households from the TX dataset, with the number of forecasting steps set to 24, are depicted in Fig. 8 to assess their performances comprehensively. The forecasting errors for various benchmarking methods are relatively large, highlighting the challenge of residential load forecasting. In contrast, the proposed method exhibits superior performance compared to other methods, providing accurate forecasting values for different residents. These results further underscore the effectiveness of the proposed approach.

Table 4
Single-step forecasting results of various methods on NY dataset.

Methods	*Single-resident forecasting*		*Multi-resident forecasting*	
	MAPE	*MAE*	*MAPE*	*MAE*
FCNN	5.74 %	1.28	11.90 %	2.71
LSTM [4]	5.72 %	1.24	6.35 %	1.44
CNN-GRU [25]	5.67 %	1.29	6.74 %	1.50
Transformer	5.28 %	1.17	6.05 %	1.33
Autoformer [26]	**4.63 %**	1.07	5.89 %	1.26
Ada-GWN [37]	5.01 %	1.14	4.53 %	0.97
Proposed	4.64 %	**1.05**	**4.38 %**	**0.89**

Table 5
Single-step forecasting results of various methods on TX dataset.

Methods	*Single-resident forecasting*		*Multi-resident forecasting*	
	MAPE	*MAE*	*MAPE*	*MAE*
FCNN	9.21 %	0.83	34.56 %	3.48
LSTM [4]	8.85 %	0.80	28.85 %	2.88
CNN-GRU [25]	8.94 %	0.87	15.67 %	1.20
Transformer	7.85 %	0.78	13.98 %	1.14
Autoformer [26]	7.09 %	0.70	11.25 %	1.04
Ada-GWN [37]	6.95 %	0.75	6.50 %	0.58
Proposed	**6.87 %**	**0.67**	**6.12 %**	**0.54**

Table 6
Time usage of different methods.

Methods	Training Time (s)		Test time (s)
	NY/TX	NSW	
FCNN	29.73	26.17	0.02
LSTM	232.16	168.12	0.11
CNN-GRU	296.48	259.23	0.17
Transformer	1407.60	957.90	0.11
Autoformer	1464.87	1066.26	0.13
Ada-GWN	1159.58	856.44	0.10
Proposed	955.50	698.44	0.11

3.4. Sensitivity and impact analysis of function modules

Ablation tests are performed to evaluate the impact of specific modules on the performance of the proposed method. In this test, the comparative methods include model B-E, which are formulated by removing the temporal attention module, the spatial attention module,

Table 7
Multi-step forecasting results of various methods.

Area	Methods	Forecasting steps = 3		Forecasting steps = 6		Forecasting steps = 24	
		MAPE	MAE	MAPE	MAE	MAPE	MAE
NY	FCNN	19.86 %	4.05	22.39 %	4.85	39.41 %	8.27
	LSTM [4]	9.88 %	2.08	16.95 %	3.50	28.65 %	6.29
	CNN-GRU [25]	10.51 %	2.18	15.25 %	3.19	25.75 %	6.05
	Transformer	9.11 %	1.88	13.01 %	2.74	22.58 %	4.03
	Autoformer [26]	8.75 %	1.75	11.26 %	2.46	18.46 %	3.57
	Ada-GWN [37]	8.04 %	1.58	12.44 %	2.57	20.34 %	3.48
	Proposed	6.77 %	1.35	8.74 %	1.86	13.56 %	2.64
TX	FCNN	52.72 %	4.31	58.24 %	4.79	67.53 %	6.31
	LSTM [4]	46.05 %	3.87	55.02 %	4.57	59.55 %	5.05
	CNN-GRU [25]	24.26 %	2.14	34.35 %	3.44	48.56 %	4.24
	Transformer	15.87 %	1.29	28.54 %	2.58	41.62 %	3.64
	Autoformer [26]	14.95 %	1.21	21.45 %	1.77	39.83 %	3.58
	Ada-GWN [37]	13.88 %	1.13	22.69 %	1.92	33.38 %	3.41
	Proposed	8.95 %	0.75	14.99 %	1.17	23.52 %	1.96
NSW	FCNN	79.47 %	0.26	82.21 %	0.27	109.3 %	0.41
	LSTM [4]	68.88 %	0.24	75.41 %	0.28	86.97 %	0.34
	CNN-GRU [25]	55.24 %	0.22	61.80 %	0.23	72.55 %	0.28
	Transformer	58.40 %	0.17	60.25 %	0.24	68.33 %	0.27
	Autoformer [26]	52.38 %	0.18	55.98 %	0.19	67.48 %	0.25
	Ada-GWN [37]	47.43 %	0.17	52.24 %	0.18	64.32 %	0.24
	Proposed	40.44 %	0.14	44.22 %	0.15	53.78 %	0.18

the encoder-decoder module, and the ST embedding block, respectively. The MAE of single-step and six-step load forecasting on the TX and NY datasets are shown in Fig. 9. It can be observed from the figure that obvious performance degradation can be observed for the proposed method when removing the spatial or temporal attention module. This illustrates the huge impact of both modules on the performance of the proposed method. When the encoder-decoder is replaced by six STGA modules, model D achieves the worst performance on the multi-step forecasting tasks. The superiority of Seq2Seq model in multi-step forecasting task is observed here. The performance obtained by model E is close to that of the proposed method, which demonstrates that the ST embedding block can further enhance the performance of the proposed spatial–temporal-attention-enabled transformer network. The results demonstrate the effectiveness of different modules of the proposed method.

3.5. Robustness to spatially uncorrelated and load anomalous households

Tests are carried out to evaluate the spatial correlation capture ability of the proposed method. Five cases are considered in this test: (1) Case 1, where the dataset only contains 10 houses in NY; (2) Case 2, where the dataset contains 10 houses in NY and 3 houses in TX; (3) Case 3, where the dataset contains 10 houses in NY and 6 houses in TX; (4) Case 4, where the dataset contains 10 houses in NY and House 1 in NSW; (5) Case 5, where the dataset contains 10 houses in NY and House 2 in NSW. The Pearson correlation coefficient (PCC) heat map of different houses is shown in Fig. 10. It can be seen that PCCs between the residents in NY dataset are greater than 0.7, indicating a strong correlation. By contrast, the PCCs between the residents in NY, TX, and NSW datasets are relatively small, which demonstrates the weak correlations. Therefore, Case 2–5 simulate the conditions when the training set contains some uncorrelated residents.

The average MAPE/MAE values of NY dataset achieved by different methods under three cases are listed in Table 8. It is clear that comparison models suffer from obvious performance degradation when uncorrelated households are contained in dataset. This reveals that sensitivity of typically method to the spatial information. Instead of capturing the spatial correlations using convolution network, the proposed method employs spatial attention module to adaptively extract spatial knowledge. This allows the proposed method to be attentive to

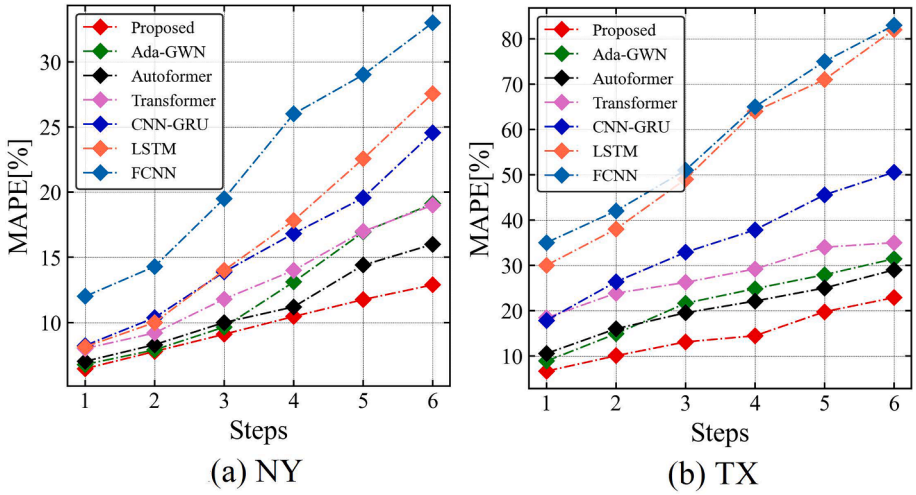

Fig. 7. MAPEs at each step.

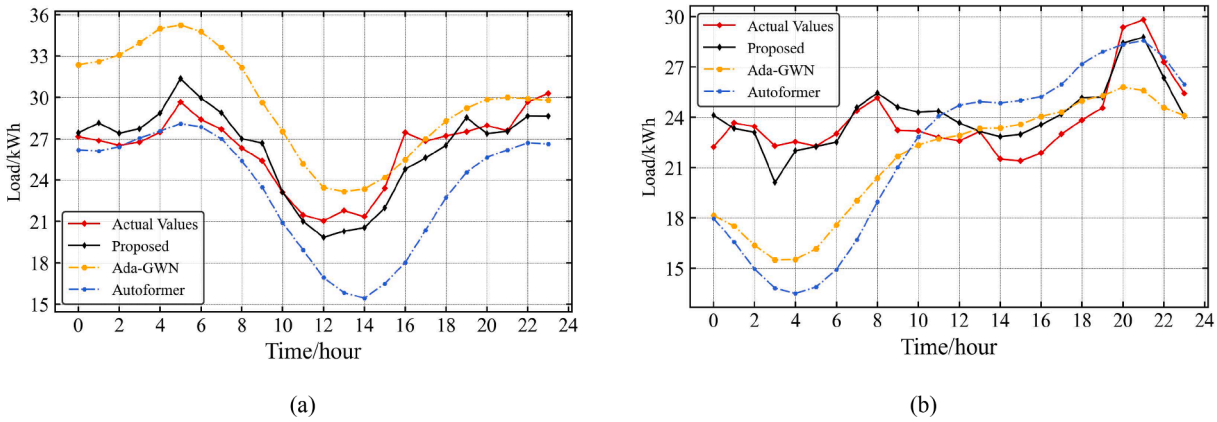

Fig. 8. Performance of 24-step load forecasting of two randomly selected houses in TX dataset.

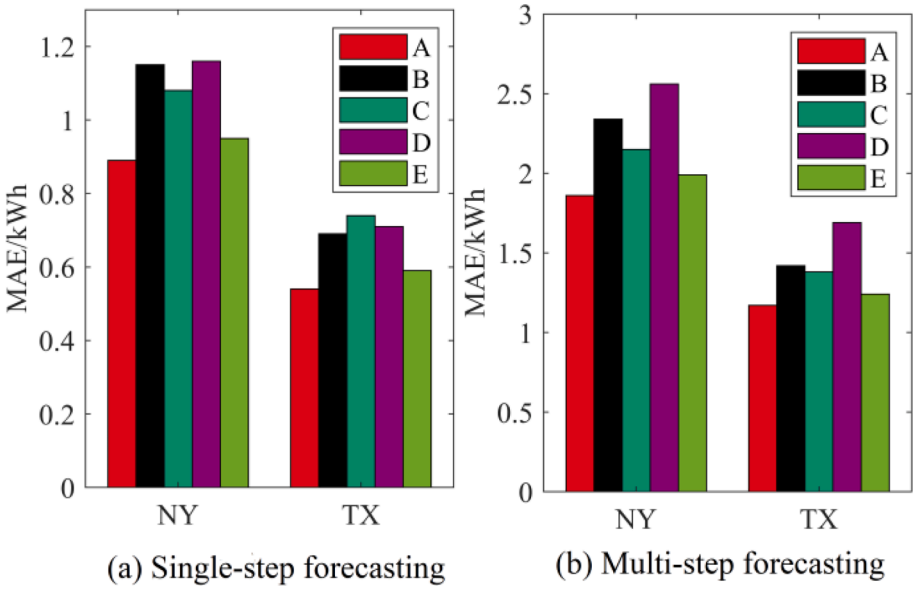

Fig. 9. Average MAE of various models for single-step and six-step forecasting on NY and TX datasets. Models: (A) whole model, (B) model without temporal attention mechanism, (C) model without spatial attention mechanism, (D) model without encoder-decoder structure, (E) model without ST embedding block.

the information most related to the forecasting task and mitigate the negative impact of uncorrelated spatial information. As a result, only slight performance degradation is observed for the proposed method under Cases 2–5. The results illustrate the robustness of the proposed method against uncorrelated spatial information.

3.6. Trade-off between forecasting accuracy and computational demand

Given input history length W, the number of households N, representational dimension d. The computational complexity of the proposed method can be denoted as $\mathcal{O}\left((W^2+N^2)d+(W+N)d^2\right)$. In this context, the main factors influencing the model complexity are the history length, the number of households, and the representational dimension. For a specific spatiotemporal load prediction task, the number of households is fixed. Therefore, the history length (H) and the representational dimension (R) are considered. Besides, the number of encoder and decoder layers (L) has a dominant influence on the volume of model parameters. Hence, this hyperparameter is also taken into account. Load data from TX are utilized to conduct one-hour-ahead forecasts for analyzing the trade-off between forecasting accuracy and computational demand. The training and testing time, MAPEs and MAEs with different hyperparameters are listed in Table 9. It can be observed from the table that: (1) As the length of historical observations increases, both the training and testing times of the model progressively rise. The model attained its highest prediction accuracy when the length of lagged observations inputted to the model is 6. Beyond this point, the model's accuracy gradually declines with further increases in the history length. Therefore, it is crucial to select the appropriate historical value length to strike a balance between forecasting accuracy and computational demand for the specific task at hand. (2) As the number of encoder and decoder layers, as well as the representational dimension, increases, the computational time of the model also increases. The highest accuracy of the model is achieved when the number of layers is set to 2, and the representational dimension is set to 64. Hence, it is also crucial to select the suitable encoder and decoder layers, as well as representational dimensions, to achieve a trade-off between forecasting accuracy and computational demand for the specific task at hand.

In summary, this experiment provides a trade-off between forecasting accuracy and computational demand. If the model seeks to strike a balance between computational efficiency and prediction accuracy, it could be beneficial to adjust the history length, the number of transformer layers, and the representational dimension accordingly.

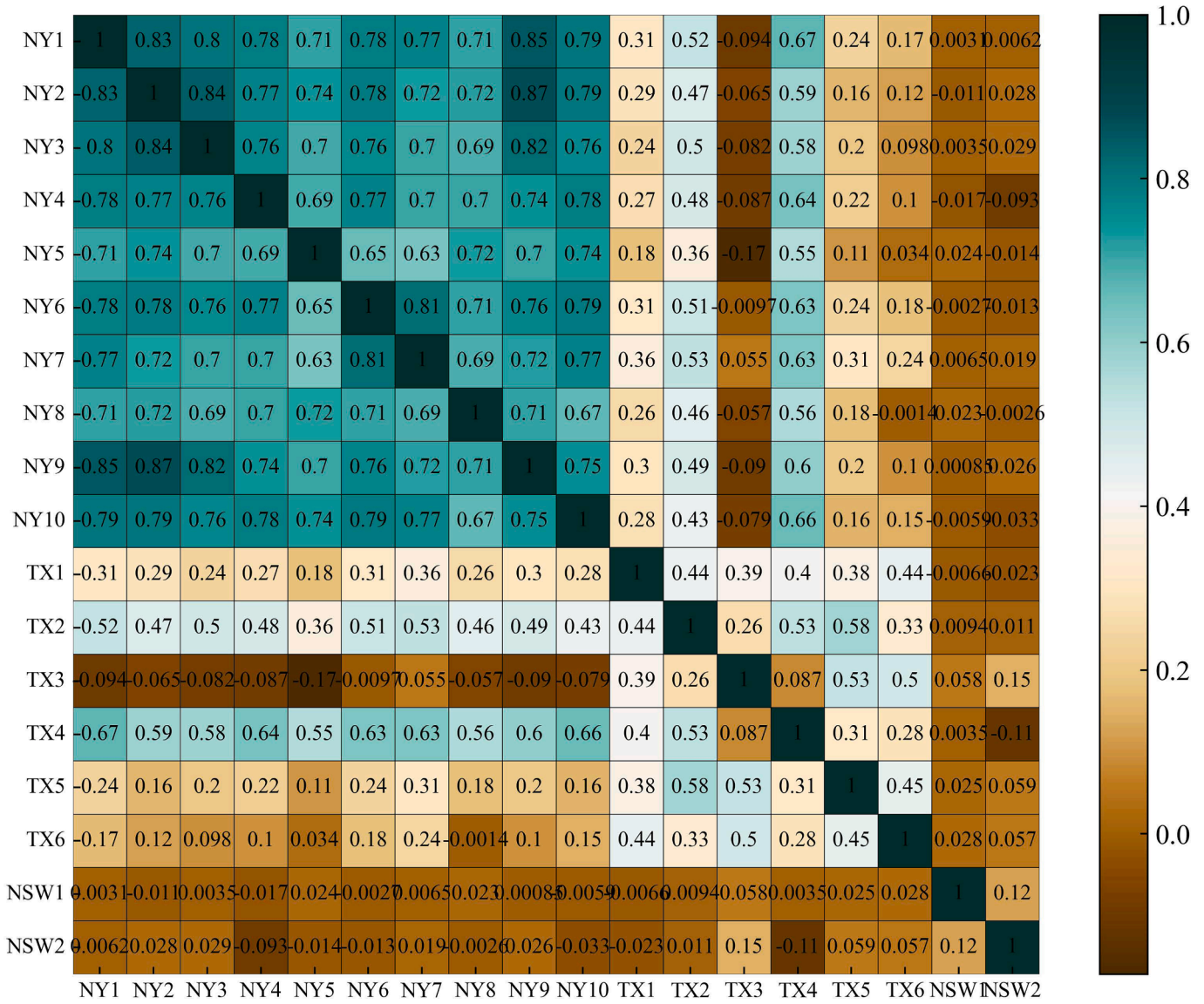

Fig. 10. PCC heat map of different houses.

Table 8
MAPEs/MAEs achieved by different methods under three cases.

Methods	Case 1	Case 2	Case 3	Case 4	Case5
FCNN	8.13	11.73	13.77	8.97	8.74
	%/2.11	%/3.08	%/3.65	%/2.35	%/2.28
LSTM [4]	6.85	8.91	9.98	7.76	7.31
	%/1.75	%/2.33	%/2.51	%/2.15	%/1.94
CNN-GRU	5.21	6.51	8.98	5.98	5.45
[25]	%/1.37	%/1.70	%/2.38	%/1.56	%/1.46
Transformer	5.02	7.03	9.74	5.96	5.67
	%/1.33	%/1.84	%/2.58	%/1.50	%/1.48
Autoformer	4.92	6.44	9.16	5.53	5.06
[26]	%/1.31	%/1.68	%/2.35	%/1.43	%/1.32
Ada-GWN	4.89	5.97	7.63	5.01	5.08
[37]	%/1.27	%/1.51	%/2.01	%/1.30	%/1.35
Proposed	4.55	4.90	5.15	4.61	4.59
	%/1.19	%/1.28	%/1.35	%/1.20	%/1.19

Table 9
Trade-off between forecasting accuracy and computational demand.

H	L	R	Offline training	Online testing	MAPE	MAE
3	2	64	10.79 min	0.069 s	6.33 %	0.63
6	2	64	15.93 min	0.113 s	6.12 %	0.54
12	2	64	30.59 min	0.248 s	8.46 %	0.70
24	2	64	106.90 min	1.045 s	9.41 %	0.82
72	2	64	177.18 min	1.183 s	11.43 %	0.97
168	2	64	279.01 min	2.105 s	12.87 %	1.05
6	1	64	8.69 min	0.055 s	8.40 %	0.84
6	2	64	15.93 min	0.113 s	6.12 %	0.54
6	4	64	24.15 min	0.151 s	6.68 %	0.61
6	2	32	11.75 min	0.083 s	6.34 %	0.59
6	2	64	15.93 min	0.113 s	6.12 %	0.54
6	2	128	25.92 min	0.193 s	7.25 %	0.61

3.7. Performance under unusual load patterns

Further tests are conducted to assess the adaptability and reliability of the proposed method under extreme weather conditions. Load data during flood periods are collected from distributed substations in the Brisbane, Queensland, Australia. It is reported that the 2022 eastern Australia floods were one of the nation's most severe recorded flood disasters. The floods occurred from late February to early May in South East Queensland, the Wide Bay-Burnett, and parts of coastal New South Wales. In South-East Queensland alone, more than 20,000 homes were

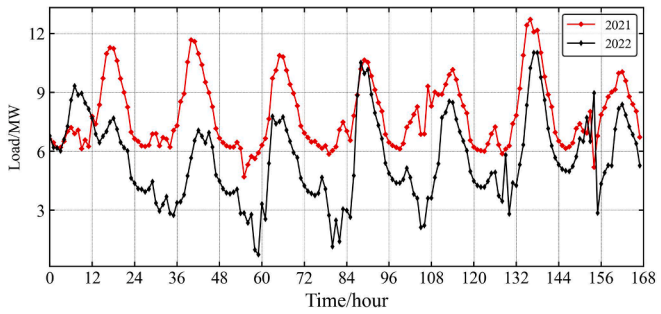

Fig. 11. Load curve comparison between the flood period and the normal period.

inundated, and power outages affected over 51,000 properties [54,55]. All public transport services were shut down for several days. These significant public events caused marked changes in consumption behavior, leading to noticeable alterations in electric load patterns. Fig. 11 illustrates the load curves from February 25, 2022, to March 2, 2022 (flood period), and February 25, 2021, to March 2, 2021 (normal period). It is evident that the load profiles experience significant drops during floods. Such dramatic drops may adversely affect the accuracy of the load forecasting model. Therefore, simulations on this dataset can check the adaptability and reliability of the proposed method.

Load data from three low-voltage substations in the Brisbane flood region are collected for simulation. This dataset covers the period from January 1, 2020, to May 30, 2022, with a one-hour resolution. The MAPEs for six-step forecasting of various methods during both the normal period and flood period are listed in Table 10. It is apparent from the table that the prediction accuracy of all methods declines during the flood period. This suggests that floods modify the load profiles, thus adversely affecting the load forecasting model. However, by employing temporal attention to model nonlinear temporal dependencies and spatial attention to adaptively leverage load information within the same region to improve forecasting performance, the proposed method outperforms other benchmarks in five out of six cases. These results reveal the adaptability and reliability of the proposed method under unusual load patterns.

4. Discussions

Main achievements: This article proposed a transformer with graph spatiotemporal attention for STRLF. By employing temporal and spatial attention mechanisms to capture nonlinear temporal patterns and dynamic spatial correlations among multiple households, respectively, the proposed method achieved superior performance in terms of accuracy and robustness compared to several state-of-the-art load forecasting models. Numerical experiments on three datasets revealed the following: (1) The proposed method is capable of extracting complex spatial and temporal information from multiple households and is attentive to the information that is most beneficial to the forecasting of each individual resident. The obtained results show more than 3 % improvement in single-step forecasting and 15 % improvement in multi-step forecasting compared to FCNN, LSTM, CNN-GRU, Transformer, Autoformer, and Ada-GWN models. (2) The proposed method exhibits robustness to spatially uncorrelated households and anomalous load patterns. Even when anomalous houses are included in the dataset, the performance degradation of the proposed method is more slight compared to other benchmarks.

Applications: Our model can facilitate effective demand-side management strategies, enabling utilities to incentivize consumers to shift electricity consumption to off-peak hours or participate in demand response programs, which can alleviate peak demand pressures and reduce overall electricity costs. Besides, the predicted loads provided by our model can support the integration of renewable energy sources by providing insights into demand fluctuations, allowing for better coordination of renewable generation with consumer demand. Moreover, our model can aid in tariff design and revenue forecasting, enabling utilities to develop pricing structures that reflect actual usage patterns and ensure financial sustainability.

Limitations and future works: Although the proposed method showcases effectiveness and robustness across multiple dimensions, two limitations have been identified: (1) According to the analysis presented in Section 3.6 and the results listed in Table 9, the computational complexity of the proposed method is $\mathscr{O}\left((W^2+N^2)d+(W+N)d^2\right)$, which is relatively high. It is noted that as the length of the historical observations increases, both offline training and online testing times significantly lengthen. In the future, we intend to address this limitation by modifying the attention module to reduce computational complexity. One potential approach involves designing a hierarchical attention mechanism to transform the computational complexity into a linear one. (2) As per the analysis presented in Section 3.7 and the findings listed in Table 10, the proposed method exhibits performance degradation under extreme weather conditions. Therefore, there is a necessity to develop more efficient algorithms to further enhance forecasting performance under unusual conditions. In our future work, we intend to design adaptive learning technologies, such as generalized additive models and Kalman filters, to swiftly adapt to new electricity consumption patterns.

5. Conclusions

This paper proposed a novel spatiotemporal graph attention mechanism-based encoder-decoder framework for short-term multivariate residential load forecasting. The proposed framework is capable of capturing dynamic spatial correlations as well as nonlinear temporal patterns among multiple households. Extensive comparative experiments on real residential load datasets have led to the following conclusions: (1) the proposed method is capable of extracting complex spatial and temporal information from multiple households. In comparison to other state-of-the-art benchmark models, the proposed method produced more accurate and robust forecasting results. (2) The proposed graph Seq2Seq model is capable of providing more accurate and reliable multi-step residential load forecasting than conventional forecasting models. (3) The proposed method is robust to some spatial uncorrelated and load anomalous households.

Table 10
MAPEs Obtained by Various Methods for Six-Step Forecasting.

Methods	Substation A		Substation B		Substation C	
	Normal period	Extreme weather	Normal period	Extreme weather	Normal period	Extreme weather
FCNN	9.174 %	13.583 %	6.940 %	8.925 %	5.143 %	5.479 %
LSTM [4]	8.394 %	12.029 %	6.587 %	8.915 %	5.193 %	5.348 %
CNN-GRU [25]	7.963 %	11.584 %	5.875 %	9.254 %	4.876 %	5.541 %
Transformer	8.557 %	11.386 %	6.035 %	8.433 %	4.512 %	6.736 %
Autoformer [26]	7.430 %	10.977 %	5.561 %	7.457 %	3.987 %	5.535 %
Ada-GWN [37]	7.158 %	9.973 %	5.877 %	7.925 %	4.057 %	4.541 %
Proposed	6.331 %	7.431 %	5.054 %	6.658 %	3.795 %	4.676 %

CRediT authorship contribution statement

Pengfei Zhao: Writing – original draft, Visualization, Software, Resources, Methodology, Investigation. **Weihao Hu:** Writing – review & editing, Supervision, Resources. **Di Cao:** Writing – review & editing, Visualization, Supervision, Software, Resources. **Zhenyuan Zhang:** Writing – review & editing, Investigation. **Wenlong Liao:** Writing – review & editing, Investigation. **Zhe Chen:** Supervision, Project administration. **Qi Huang:** Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022B1515250001, in part by the National Natural Science Foundation of China under Grant 52277083, in part by China Postdoctoral Science Foundation under Grant 2023M730495, in part by Sichuan Province Innovative Talent Funding Project for Postdoctoral Fellows under Grant BX202210, and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515110939.

References

[1] Yang Z, Zhu R, Liao W. Minkowski distance based pilot protection for tie lines between offshore wind farms and MMC. IEEE Transactions on Industrial Informatics 2024;20(6):8441–52.
[2] Cheng L, Zang H, Xu Y, Wei Z, Sun G. Probabilistic residential load forecasting based on micrometeorological data and customer consumption pattern. IEEE Trans Power Syst 2021;36(4):3762–75.
[3] Ji Y, Buechler E, Rajagopal R. Data-driven load modeling and forecasting of residential appliances. IEEE Trans Smart Grid 2020;11(3):2652–61.
[4] Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y. Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Trans Smart Grid 2019;10(1):841–51.
[5] Panda SK, Ray P, Salkuti SR. A review on short-term load forecasting using different techniques. Lect Notes Electr Eng 2022:433–54.
[6] Sadaei HJ, Enayatifar R, Abdullah AH, Gani A. Short-term load forecasting using a hybrid model with a refined exponentially weighted fuzzy time series and an improved harmony search. Int J Electr Power Energy Syst 2014;62:118–29.
[7] Luzia Ruan, Rubio L, Velasquez CE. Sensitivity analysis for forecasting Brazilian electricity demand using artificial neural networks and hybrid models based on Autoregressive Integrated Moving Average. Energy 2023;274:127365.
[8] Hong T, Wang P, Willis H. A naïve multiple linear regression benchmark for short term load forecasting. In: IEEE Power & Energy Society General Meeting, Detroit, Michigan, USA; 2011.
[9] Venkataramana Veeramsetty D, Chandra R, Salkuti SR. Short term active power load forecasting using machine learning with feature selection. Lect Notes Electr Eng 2022:103–24.
[10] Salkuti SR. Short-term electrical load forecasting using hybrid ANN–DE and wavelet transforms approach. Electr Eng 2018;100(4):2755–63.
[11] Reddy SS, Momoh JA. Short term electrical load forecasting using back propagation neural networks. In: 2014 North American power symposium (NAPS). IEEE; 2014. p. 1–6.
[12] Reddy SS, Jung C-M. Short-term load forecasting using artificial neural networks and wavelet transform. Int J Appl Eng Res 2016;11(19):9831–6.
[13] Lin J, Ma J, Zhu J, Chen Y. Short-term load forecasting based on LSTM networks considering attention mechanism. Int J Electr Power Energy Syst 2022;137:107818.
[14] Wang Y, Gan D, Sun M, Zhang N, Lu Z, Chen K. Probabilistic individual load forecasting using pinball loss guided LSTM. Appl Energy 2019;235:10–20.
[15] Wei J, Wu X, Yang T, Jiao R. Ultra-short-term forecasting of wind power based on multi-task learning and LSTM. Int J Electr Power Energy Syst 2023;149:109073.
[16] Venkataramana Veeramsetty D, Rakesh Chandra D, Salkuti SR. Short-term electric power load forecasting using factor analysis and long short-term memory for smart cities. Int J Circ Theory Appl 2020;49(6):1678–703.
[17] Khan N, Haq IU, Khan SU, Rho S, Lee MY, Baik SW. DB-Net: A novel dilated CNN based multi-step forecasting model for power consumption in integrated local energy systems. Int J Electr Power Energy Syst 2021;133:107023.
[18] Zhao P, et al. Geometric loss-enabled complex neural network for multi-energy load forecasting in integrated energy systems. IEEE Trans Power Syst 2023. https://doi.org/10.1109/tpwrs.2023.3345328.
[19] Zhao P, Cao D, Wang Y, Chen Z, Hu W. Gaussian process-aided transfer learning for probabilistic load forecasting against anomalous events. IEEE Trans Power Syst 2023;38(3):2962–5.
[20] Zhang Z, Zhao P, Wang P, Lee W-J. Transfer learning featured short-term combining forecasting model for residential loads with small sample sets. IEEE Trans Ind Appl 2022;58(4):4279–88.
[21] Zhao P, Hu W, Cao D, Zhang Z, Huang Y, Dai L, et al. Probabilistic multienergy load forecasting based on hybrid attention-enabled transformer network and Gaussian process-aided residual learning. IEEE Transactions on Industrial Informatics 2024;20(6):8379–93.
[22] Faustine A, Pereira L. FPSeq2Q: fully parameterized sequence to quantile regression for net-load forecasting with uncertainty estimates. IEEE Trans Smart Grid 2022;13(3):2440–51.
[23] Shi H, Xu M, Li R. Deep learning for household load forecasting—a novel pooling deep RNN. IEEE Trans Smart Grid 2018;9(5):5271–80.
[24] Munkhammar J, Widén J. Very short term load forecasting of residential electricity consumption using the Markov-chain mixture distribution (MCM) model. Appl Energy 2021;282:116180.
[25] Afrasiabi M, Mohammadi M, et al. Deep-based conditional probability density function forecasting of residential loads. IEEE Trans Smart Grid 2020;11(4):3646–57.
[26] Jiang Y, et al. Very short-term residential load forecasting based on deep-autoformer. Appl Energy 2022;328:120120.
[27] Melo JD, Carreno EM, Padilha-Feltrin A. Multi-agent simulation of urban social dynamics for spatial load forecasting. IEEE Trans Power Syst 2012;27(4):1870–8.
[28] Ye C, Ding Y, Wang P, Lin Z. A data-driven bottom-up approach for spatial and temporal electric load forecasting. IEEE Trans Power Syst 2019;34(3):1966–79.
[29] Mostafa Gilanifar H, Wang KS, Ozguven EE, Arghandeh R. Multitask Bayesian spatiotemporal Gaussian processes for short-term load forecasting. IEEE Trans Ind Electron 2020;67(6):5132–43.
[30] Zhao Y, Ye L, Pinson P, Tang Y, Lu P. Correlation-constrained and sparsity-controlled vector autoregressive model for spatio-temporal wind power forecasting. IEEE Trans Power Syst 2018;33(5):5029–40.
[31] Wang S, Cao J, Yu PS. Deep learning for spatio-temporal data mining: a survey. IEEE Trans Knowl Data Eng 2022;34(8):3681–700.
[32] Jalalifar R, Delavar MR, Ghaderi SF. SAC-ConvLSTM: a novel spatio-temporal deep learning-based approach for a short term power load forecasting. Exp Syst Appl 2024;237:121487.
[33] Chai S, Xu Z, Jia Y, Wong WK. A robust spatiotemporal forecasting framework for photovoltaic generation. IEEE Trans Smart Grid 2020;11(6):5370–82.
[34] Mahdi Khodayar G, Liu JW, Kaynak O, Khodayar ME. Spatiotemporal behind-the-meter load and PV power forecasting via deep graph dictionary learning. IEEE Trans Neural Networks Learn Syst 2021;32(10):4713–27.
[35] Liu Y, et al. Probabilistic spatiotemporal wind speed forecasting based on a variational Bayesian deep learning model. Appl Energy 2020;260:114259.
[36] Huang N, Wang S, Wang R, Cai G, Liu Y, Dai Q. Gated spatial-temporal graph neural network based short-term load forecasting for wide-area multiple buses. Int J Electr Power Energy Syst 2023;145:108651.
[37] Lin W, Wu D, Boulet B. Spatial-temporal residential short-term load forecasting via graph neural networks. IEEE Trans Smart Grid 2021;12(6):5373–84.
[38] Wu D, Lin W. Efficient residential electric load forecasting via transfer learning and graph neural networks. IEEE Trans Smart Grid 2023;14(3):2423–31.
[39] Shi J, Zhang W, Bao Y, Gao DW, Wang Z. Load forecasting of electric vehicle charging stations: attention based spatiotemporal multi-graph convolutional networks. IEEE Trans Smart Grid 2023. Early access.
[40] Kim HJ, Kim MK. Spatial-temporal graph convolutional-based recurrent network for electric vehicle charging stations demand forecasting in energy market. IEEE Trans Smart Grid 2024. Early access.
[41] Chen Z, Zhang B, Du C, Meng W, Meng A. A novel dynamic spatio-temporal graph convolutional network for wind speed interval prediction. Energy 2024:130930.
[42] Zhu J, Yan Y, Zhao L, Heimann M, Akoglu L, Koutra D. Beyond homophily in graph neural networks: current limitations and effective designs. Adv Neural Inf Proces Syst 2020;33:7793–804.
[43] Vaswani A, et al. Attention is all you need. Adv Neural Inf Proces Syst 2017;30.
[44] Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, Zhang W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 35, No. 12; 2021. p. 11106–15.
[45] Wu H, Xu J, Wang J, Long M. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. Adv Neural Inf Proces Syst 2021;34:22419–30.
[46] Liu S, et al. Pyraformer: low-complexity pyramidal attention for long-range time series modeling and forecasting. In: Proc ICLR; 2022.
[47] Zhou T, Ma Z, Wen Q, Wang X, Sun L, Jin R. FEDformer: frequency enhanced decomposed transformer for long-term series forecasting. PMLR 2022:27268–86.
[48] Yu C, Ma X, Ren J, Zhao H, Yi S. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In: Proc Eur Conf Comput Vis. Springer; 2020. p. 507–23.
[49] Wang T, et al. Synchronous spatiotemporal graph transformer: a new framework for traffic data prediction. IEEE Trans Neural Networks Learn Syst 2023;34(12):10589–99.

[50] Li Z, et al. Text compression-aided transformer encoding. IEEE Trans Pattern Anal Mach Intell 2022;44(7):3840–57.
[51] Zheng C, Fan X, Wang C, Qi J. GMAN: a graph multi-attention network for traffic prediction. Proc AAAI 2020;34(01):1234–41.
[52] OPENEI, Mar. 2023. [Online]. https://openei.org/datasets/files/961/pub/RESIDENTIAL_LOAD_DATA_E_PLUS_OUTPUT/HIGH/.
[53] Smart Grid, Smart City, Australian Govern, Australia, Canberra, ACT, Australia, Mar. 2023. [Online]. https://data.gov.au/data/dataset/smart-grid-smart-city-customer-trial-data.
[54] 2022 eastern Australia floods. https://en.wikipedia.org/wiki/2022_eastern_Australia_floods.
[55] Major flood disaster in Brisbane, 20,000 houses inundated. ABC news. 1 March 2022. [Online]. https://www.abc.net.au/news/2022-03-01/major-flood-disaster-in-brisbane,-20,000-houses-flooded/13775742.