Spatio-Temporal Joint Graph Convolutional Networks for Traffic Forecasting

Chuanpan Zheng ©, Xiaoliang Fan ©, Senior Member, IEEE, Shirui Pan ©, Senior Member, IEEE, Haibing Jin ©, Zhaopeng Peng ©, Zonghan Wu ©, Cheng Wang ©, Senior Member, IEEE, and Philip S. Yu ©, Fellow, IEEE

Abstract—Recent studies have shifted their focus towards formulating traffic forecasting as a spatio-temporal graph modeling problem. Typically, they constructed a static spatial graph at each time step and then connected each node with itself between adjacent time steps to create a spatio-temporal graph. However, this approach failed to explicitly reflect the correlations between different nodes at different time steps, thus limiting the learning capability of graph neural networks. Additionally, those models overlooked the dynamic spatio-temporal correlations among nodes by using the same adjacency matrix across different time steps. To address these limitations, we propose a novel approach called Spatio-Temporal Joint Graph Convolutional Networks (STJGCN) for accurate traffic forecasting on road networks over multiple future time steps. Specifically, our method encompasses the construction of both pre-defined and adaptive spatio-temporal joint graphs (STJGs) between any two time steps, which represent comprehensible and dynamic spatio-temporal correlations. We further introduce dilated causal spatio-temporal joint graph convolution layers on the STJ to capture spatio-temporal dependencies from distinct perspectives with multiple ranges. To aggregate information from different ranges, we propose a multi-range attention mechanism. Finally, we evaluate our approach on five public traffic datasets and experimental results demonstrate that STJGCN is not only computationally efficient but also outperforms 11 state-of-the-art baseline methods.

Index Terms—Spatio-temporal, graph convolutional network, traffic forecasting.

I. INTRODUCTION

SPATIO-TEMPORAL data forecasting has received increasing attention from the deep learning community in recent years [1], [2], [3]. It plays a vital role in a wide range of applications, such as traffic speed prediction [4] and air quality inference [5]. In this paper, we study the problem of forecasting the future traffic conditions given historical observations on a road network.

Recent studies formulate traffic forecasting as a spatio-temporal graph modeling problem [4], [6], [7], [8], [9], [10], [11]. The basic assumption is that the state of each node is conditioned on its neighboring node information. Based on this, they construct a spatial graph with a pre-defined [4] or data-adaptive [7] adjacency matrix. In such a graph, each node corresponds to a location of interest (e.g., traffic sensor). The graph neural network [12] is applied on that graph to model the correlations among spatial neighboring nodes at each time step. To leverage the information from temporal neighboring nodes, they further connect each node with itself between adjacent time steps, which results in a spatio-temporal graph, as shown in Fig. 1(a). The 1D convolutional neural network [6] or recurrent neural network [4] is commonly used to model the correlations at each node between different time steps. By combining the spatial and temporal features, they are able to update the state of each node.

However, those spatio-temporal graphs do not explicitly reflect the correlations between different nodes at different time steps (e.g., the red dash lines in Fig. 1(b)). In such a graph, the information of spatial and temporal neighborhoods is captured through the spatial and temporal connections respectively, while the information of neighboring nodes across both spatial and temporal dimensions are not considered, which may restrict the learning ability of graph neural networks. For example, a traffic jam occurred at an intersection may affect not only current nearby roads (spatial neighborhoods) and its local future traffic condition (temporal neighborhoods), but also the downstream roads in next few hours (spatio-temporal neighborhoods). Thus, we argue that it is necessary to model the comprehensive correlations in the spatio-temporal data.

Another limitation of previous works is that they ignore the dynamic correlations among nodes at different time steps, as shown in Fig. 1(c). The road network distances among sensors (nodes) are commonly used to define the spatial graph [4], [6]. This pre-defined graph is usually static. Some researchers [7],

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
propose to learn a data-adaptive adjacency matrix, which is also unchanged over time steps. However, the traffic data exhibits strong dynamic correlations in the spatial and temporal dimensions, those static graphs are unable to reflect the dynamic characteristics of correlations among nodes. For example, the residence region is highly correlated to the office area during workday morning rush hours, while the correlation would be relatively weakened in the evening because some people might prefer to dining out before going home. Thus, it is crucial to model the dynamic spatio-temporal correlations for traffic forecasting.

This paper addresses these limitations from the following perspectives. First, besides the spatial and temporal connections, we further add the spatio-temporal connections between two time steps according to the spatio-temporal distances to define the spatio-temporal joint graph (STJG). In this way, the pre-defined STJG preserves comprehensive spatio-temporal correlations between any two time steps. Second, in order to adapt to the dynamic correlations among nodes, we suggest to explore an adaptive STJG, which is time-variant by encoding the time features. The adjacency matrix in this adaptive STJG is dynamic, changing over time steps. By constructing both the pre-defined and adaptive STJGs, we are able to preserve comprehensive and dynamic spatio-temporal correlations.

On these basis, we then develop the spatio-temporal joint graph convolution (STJGC) operations on both pre-defined and adaptive STJGs to simultaneously capture the spatio-temporal dependencies in a unified operation. We further design the dilated causal STJGC layers to extract multiple spatio-temporal ranges of information. Next, a multi-range attention mechanism is proposed to aggregate the information of different ranges. Finally, we apply independent fully-connected layers to produce the multi-step ahead prediction results. The whole framework is named as spatio-temporal joint graph convolutional networks (STJGCN), which can be learned end-to-end. To evaluate the efficiency and effectiveness of STJGCN, we conduct extensive experiments on five public traffic datasets. The experimental results demonstrate that our STJGCN is computationally efficient and achieves the best performance against 11 state-of-the-art baseline methods. Our main contributions are summarized as follows.

- We construct both pre-defined and adaptive spatio-temporal joint graphs (STJGs), which reflect comprehensive and dynamic spatio-temporal correlations.
- We design dilated causal spatio-temporal joint graph convolution layers on both types of STJG to model multiple ranges of spatio-temporal correlations.
- We propose a multi-range attention mechanism to aggregate the information of different ranges.
- We evaluate our model on five public traffic datasets, and experimental results demonstrate that STJGCN has high computation efficiency and outperforms 11 state-of-the-art baseline methods.

The rest of this paper is organized as follows. Section II reviews the related work. Section III presents the preliminary of this work. Section IV details the method of STJGCN. Section V compares STJGCN with state-of-the-art methods on five datasets. Finally, Section VI concludes this paper and draws future work.

II. RELATED WORK

A. Graph Convolutional Networks

Graph convolutional networks (GCNs) are successfully applied on various tasks (e.g., node classification [13], link prediction [14]) due to their superior abilities of handling graph-structured data [12]. There are mainly two types of GCN [15]: spatial GCN and spectral GCN. The spatial GCN performs convolution filters on neighborhoods of each node. Researchers in [16] propose a heuristic linear method for neighborhood selecting. GraphSAGE [17] samples a fixed number of neighbors for each node and aggregates their features. GAT [18] learns the weights among nodes via attention mechanisms. Researchers in [19] improve graph neural network architecture by exploiting correlation structure in the regression residuals. The spectral GCN defines the convolution in the spectral domain [20], which

![Fig. 1. The comprehensive and dynamic connections among nodes in graph-structured spatio-temporal data. There are three common scenarios: (a) Spatio-Temporal Graph: The node 2 at time step \( t \) can be influenced by nodes 1 and 3 at time step \( t-1 \) through the temporal connection. (b) Pre-defined Spatio-Temporal Joint Graph: The node 2 at time step \( t \) may also be affected by nodes 1 and 3 at time step \( t-1 \) through spatio-temporal connections. (c) Adaptive Spatio-Temporal Joint Graph: Compared with time step \( t-1 \), the connections among nodes 1, 2 and 3 exhibit strong dynamic characteristics at the time step \( t \). For instance, the connection between nodes 1 and 3 gets weakened, while the connection between nodes 2 and 3 becomes stronger. Both (b) and (c) scenarios have not been comprehensively explored in existing studies.](image-url)
is first introduced in [21]. ChebNet [22] reduces the computational complexity with fast localized convolution filters. In [13], researchers further simplify the ChebNet to a simpler form and achieve state-of-the-art performances on various tasks. Recently, a range of studies apply the GCN on time-series data and construct spatio-temporal graphs for traffic forecasting [4], [23], human action recognition [24], [25], etc.

B. Spatio-Temporal Forecasting

Spatio-temporal forecasting is an important research topic, which has been extensively studied for decades [26], [27], [28], [29], [30]. Recurrent neural networks (RNNs), especially the long short-term memory (LSTM) and gated recurrent unit (GRU) are successfully applied for modeling temporal correlations [31]. To capture the spatial dependencies, convolutional neural networks (CNNs) are introduced, which are restriction to process regular grid structures [32], [33], [34], [35], [36]. Recently, researchers apply graph neural networks to model the non-euclidean spatial correlations [37]. DCRNN [4] employs diffusion convolution to capture the spatial dependency and applies GRU to model the temporal dependency. STGCN [6] uses graph convolution and 1D convolution to model the spatial and temporal dependencies, respectively. Researchers in [38] study the effect of the order of spatial layers and temporal layers on STGCN model performance. Several works [8], [39], [40] introduce the attention mechanisms [41] into the spatio-temporal graph modeling to improve the prediction accuracy. AGSTN [42] proposes an attention adjustment mechanism to realize fluctuation modulation for learning time-evolving spatio-temporal correlation. Some studies consider more kinds of connections (e.g., semantic connection [43], edge interaction patterns [44]) to construct the spatial graph. The adjacency matrices in these models are usually pre-defined according to some prior knowledge (e.g., distances among nodes). Some researchers [7], [10] argue that the pre-defined adjacency matrix does not necessarily reflect the underlying dependencies among nodes, and propose to learn an adaptive adjacency matrix for graph modeling. However, both the pre-defined and adaptive adjacency matrices assume static correlations among nodes, which cannot adapt to the evolving systems (e.g., traffic network). Moreover, these graph-based methods do not explicitly model the correlations between different nodes at different time steps, which may restrict the learning ability of graph neural networks.

III. PRELIMINARY

Problem Definition. Suppose there are $N$ sensors (nodes) on a road network, and each sensor records $C$ traffic measurements (e.g., volume, speed) at each time step. Thus, the traffic conditions at time step $t$ can be represented as $X_t \in \mathbb{R}^{N \times C}$. The traffic forecasting problem aims to learn a function $f$ that maps the traffic conditions of historical $P$ time steps to next $Q$ time steps:

$$[X_{t-P+1}, X_{t-P+2}, \ldots, X_t] \xrightarrow{f} [X_{t+1}, X_{t+2}, \ldots, X_{t+Q}],$$

(1)

IV. METHODOLOGY

A. Framework Overview

Fig. 2 depicts the framework of our proposed Spatio-Temporal Joint Graph Convolutional Networks (STJGCN), which includes three modules. First, previous graph-based methods generally ignore the spatio-temporal connections and the dynamic correlations among nodes, so we thus propose the spatio-temporal joint graph (STJG) construction module to construct both pre-defined and adaptive STJGs, which preserve comprehensive and dynamic spatio-temporal correlations. Second, as the standard graph convolution operation models spatial correlations only, we propose the spatio-temporal joint graph convolution (STJGC) operation on both types of STJG to model the comprehensive and dynamic spatio-temporal correlations in a unified operation. Based on the STJGC, we further propose the dilated causal STJGC module to capture spatio-temporal dependencies within multiple neighborhood and time ranges. Finally, in the prediction module, we propose a multi-range attention mechanism to aggregate the information of different ranges, and apply fully-connected layers to produce the prediction results. We detail each module in the following subsections.

B. STJG Construction Module

In this module, we first pre-define the spatio-temporal joint graph (STJG) according to the spatio-temporal distances among nodes. While, the pre-defined graph may not reflect the underlying correlations among nodes [7], [10], we further propose to learn adaptive STJG. By constructing both types of STJG, we are able to represent comprehensive and dynamic spatio-temporal correlations among nodes.

1) Pre-Defined Spatio-Temporal Joint Graph: Previous studies [4], [6] for traffic forecasting on graphs usually define the spatial adjacency matrix based on pair-wise road network distances:

$$A_{i,j} = \exp\left(-\frac{\text{dist}(v_i, v_j)^2}{\sigma^2}\right),$$

(2)

where $\text{dist}(v_i, v_j)$ represents the road network distance from node $v_i$ to node $v_j$, $\sigma$ is the standard deviation of distances, and $A_{i,j}$ denotes the edge weight between node $v_i$ and node $v_j$. They construct the spatial graph at each time step, and then connect each node with itself between adjacent time steps to define the spatio-temporal graph. In such a graph, the connections between different nodes at different time steps are not incorporated, which may restrict its representation ability.

We propose to construct a spatio-temporal joint graph (STJG), which preserves comprehensive spatio-temporal correlations. The intuitive idea is to further connect different nodes between two time steps, as shown in Fig. 1(b). Thus, we modify (2) to be the STJG adjacency matrix, as:

$$A_{i,t-k,j,t} = \exp\left(-\frac{(k + 1) \cdot \text{dist}(v_i, v_j))^2}{\sigma^2}\right),$$

(3)

where $k$ is the time difference between two time steps. $A_{i,t-k,j,t}$ defines the edge weight between node $v_i$ at time step $t - k$ and
node $v_j$ at time step $t$, which decreases with the increase of spatio-temporal distance. When $k=0$, (3) degenerates to (2), which represents the spatial connections. If $i=j$, the STJG adjacency matrix defines the temporal connections at each node between two time steps. Otherwise, it represents the spatio-temporal connections between different nodes at different time steps. Thus, we are able to define a comprehensive spatio-temporal graph according to (3). Note that the STJG could be constructed between any two time steps, which makes it flexible to reveal multiple time-ranges of spatio-temporal correlations.

We filter the values smaller than a threshold $\delta_{adj}$ in the STJG adjacency matrix to eliminate weak connections and control the sparsity. As this adjacency matrix is conditioned on the time difference $k$, but irrelevant to a specific time step, we denote it as $A^{(k)} \in \mathbb{R}^{N \times N}$ in following discussions.

2) Adaptive Spatio-Temporal Joint Graph: Previous studies [7], [10] demonstrate that the pre-defined adjacency matrix may not reflect the underlying correlations among nodes, and propose adaptive ones. However, they only define the spatial graph, and it is unchanged over time steps. We propose to learn adaptive STJG adjacency matrices that could represent comprehensive and dynamic spatio-temporal correlations based on the latent space modeling algorithm [45].

a) Latent space modeling: Given a graph, we assume each node resides in a latent space with various attributes. The attributes of nodes and how these attributes interact with each other jointly determine the underlying relations among nodes. The nodes which are close to each other in the latent space are more likely to form a link. Mathematically, we aim to learn two matrices $U$ and $B$. Here, $U \in \mathbb{R}^{N \times d}$ denotes the $d$ latent attributes of the $N$ nodes, and $B \in \mathbb{R}^{d \times d}$ represents the attributes interaction patterns, which could be an asymmetric matrix for directed graph or symmetric matrix for undirected graph. The product of $U B U^\top$ could represent the connections among nodes.

b) Spatio-temporal embedding: We propose a spatio-temporal embedding to form the latent node attributes. We first randomly initialize a spatial embedding for each of the $N$ nodes, and then transform it to $d$ dimensions via fully-connected layers. To obtain time-varying node attributes and take periodic patterns in historical input data (i.e., morning rush hour) into account, we further encode the time information as the temporal embedding. At each time step, we consider two time features, i.e., time-of-day and day-of-week, which are encoded by one-hot coding and then be projected to $d$ dimensions using fully-connected layers. We then add the spatial and temporal embeddings together to generate the spatio-temporal embedding at each time step $t$, represented as $U_t \in \mathbb{R}^{N \times d}$, which can be updated during the training stage. The spatio-temporal embedding encodes both the node-specific and time-varying information, and it could mine periodic spatio-temporal patterns of historical data.

c) Adaptive STJG adjacency matrix: Based on the spatio-temporal embedding, we define the STJG adjacency matrix at time step $t$ according to the latent space modeling algorithm, as:

$$
\tilde{L}_t = \text{softmax}(\psi(U_t B U_t^\top)),$$

with

$$
\psi(x) = \begin{cases} 
    x, & \text{if } x \geq \delta_{adj} \\
    0, & \text{otherwise}
\end{cases},
$$

where $U_t \in \mathbb{R}^{N \times d}$ is the spatio-temporal embedding of $N$ nodes at time step $t$, $\psi(x)$ is used to eliminate the weights smaller than a threshold $\delta_{adj}$, and the softmax function is applied for normalization. $\tilde{L}_t \in \mathbb{R}^{N \times N}$ defines the spatial connections among $N$ nodes at time step $t$, which is dynamic, changing over time steps. In order to construct the connections between different time steps, we modify (4) as:

$$
\tilde{L}_{t-k:t} = \text{softmax}(\psi(U_{t-k} B U_{t-k}^\top)),$$

where $\tilde{L}_{t-k:t} \in \mathbb{R}^{N \times N}$ is the normalized STJG adjacency matrix between time steps $t-k$ and $t$. When $k=0$, (6) degenerates to (4), which describes the spatial graph at time step $t$. Thus, (6) is able to define the spatio-temporal joint graph between time steps $t-k$ and $t$ with comprehensive and dynamic spatio-temporal connections.
C. Dilated Causal STJGC Module

The standard graph convolution performs on spatial graphs to model spatial correlations only, thus propose the spatio-temporal joint graph convolution (STJGC) on both types of STJG to model spatio-temporal correlations in a unified operation. We further design dilated causal STJGC layers to capture multiple ranges of spatio-temporal dependencies, as shown in Fig. 2. In the following discussion, we first describe the STJGC operation in Section IV-C1, and then introduce the dilated causal STJGC layers in Section IV-C2.

1) Spatio-Temporal Joint Graph Convolution (STJGC):

Graph convolution is an effective operation for learning node information from spatial neighborhoods according to the graph structure, while the standard graph convolution performs on the spatial graph to model the spatial correlations only. In order to model the comprehensive and dynamic spatio-temporal correlations on the STJG, we propose the spatio-temporal joint graph convolution (STJGC) operations on both types of STJG.

a) Graph convolution: The graph convolution is defined as:

\[
Z = \phi(\hat{A}XW + b),
\]

(7)

Here, \(X \in \mathbb{R}^{N \times d_1}\) and \(Z \in \mathbb{R}^{N \times d_2}\) denote the input and output graph signals, \(W \in \mathbb{R}^{d_1 \times d_2}\) and \(b \in \mathbb{R}^{d_2}\) are learnable parameters, \(\phi(\cdot)\) is an activation function (e.g., ReLU [46]), \(\hat{A} = D^{-1/2}AD^{-1/2} \in \mathbb{R}^{N \times N}\) is the normalized adjacency matrix, where \(A\) is the adjacency matrix with self-loops, and \(D = \sum_j A_{i,j}\) is the degree matrix.

b) STJGC on pre-defined STJG: Consider the STJG between time steps \(t - k\) and \(t\), the information of each node at time step \(t\) comes from its spatial, temporal, and spatio-temporal neighborhoods:

\[
Z_{t}^{pd} = \phi(\hat{A}^{(k)}X_{t-k}W_{pd}^{t} + \hat{A}^{(0)}X_{t}W_{pd}^{0} + b_{pd}),
\]

(8)

where \(\hat{A}^{(k)}\) is the normalized pre-defined STJG adjacency matrix between time steps \(t - k\) and \(t\) (see (3)). In (8), \(\hat{A}^{(k)}X_{t-k}W_{pd}^{t}\) means we aggregate neighborhoods (both temporal and spatio-temporal) information from time step \(t - k\), and \(\hat{A}^{(0)}X_{t}W_{pd}^{0}\) means we aggregate the information from spatial neighborhoods at time step \(t\). Thus, by performing (8), we are able to model comprehensive spatio-temporal correlations between two time steps.

Furthermore, at time step \(t\), we propose to incorporate \(K\) (denoted as kernel size) time step information (e.g., \(t, t-1, \ldots, t-K+1\)) to update the node features. Specifically, we modify (8) as:

\[
Z_{t}^{pd} = \sum_{k=0}^{K-1} \phi(\hat{A}^{(k)}X_{t-k}W_{pd}^{t} + b_{pd}).
\]

(9)

In the case of a directed graph, we consider two directions of information propagation (i.e., forward and backward), corresponding to two normalized adjacency matrices:

\[
\hat{A}_{fw}^{(k)} = D_{fw}^{-1/2}A_{fw}^{(k)}D_{fw}^{-1/2} \quad \text{and} \quad \hat{A}_{bw}^{(k)} = D_{bw}^{-1/2}A_{bw}^{(k)}D_{bw}^{-1/2},
\]

where \(D_{fw} = \sum_j A_{fw,j}^{(k)}\) and \(D_{fw}^{-1}\) represent the out-degree and in-degree matrices, respectively. Thus, we transform (9) to:

\[
Z_{t}^{pd} = \sum_{k=0}^{K-1} \phi \left( \hat{A}_{fw}^{(k)}X_{t-k}W_{pd}^{t} + \hat{A}_{bw}^{(k)}X_{t-k}W_{pd}^{0} + b_{pd} \right),
\]

(10)

where \(X_{t-k} \in \mathbb{R}^{N \times d}\) and \(X_{t} \in \mathbb{R}^{N \times d}\) are the input graph signals at time steps \(t - k\) and \(t\), respectively, \(Z_{t}^{pd}\) denotes the updated feature at time step \(t\), \(W_{pd}^{t} \in \mathbb{R}^{d_{x} \times d}\), \(W_{pd}^{0} \in \mathbb{R}^{d_{x} \times d}\), and \(b_{pd} \in \mathbb{R}^{d}\) are learnable parameters.

By this design, our STJGC simultaneously models the information propagation from three kinds of connections (i.e., spatial, temporal, and spatio-temporal) in a unified operation.

c) STJGC on adaptive STJG: As the pre-defined STJG may not reflect the underlying correlations among nodes, we further propose STJGC on adaptive STJG. The computation is similar as that on pre-defined STJG:

\[
Z_{t}^{adt} = \sum_{k=0}^{K-1} \phi \left( \hat{L}_{t-k:t}X_{t-k}W_{adt}^{t} + b_{adt} \right),
\]

(11)

where \(\hat{L}_{t-k:t}\) is the normalized adaptive STJG adjacency matrix between time steps \(t - k\) and \(t\) (defined in (6)). Inspired by the bi-directional RNN [47], we consider both time directions of the information flow. Specifically, we compute two adaptive STJG adjacency matrices: \(\hat{L}_{t-k:t}\) and \(\hat{L}_{t:t-k}\), and modify (11) accordingly, as:

\[
Z_{t}^{adt} = \sum_{k=0}^{K-1} \phi \left( \hat{L}_{t-k:t}X_{t-k}W_{adt}^{t} + \hat{L}_{t:t-k}X_{t-k}W_{adt}^{t} + b_{adt} \right),
\]

(12)

where \(Z_{t}^{adt}\) is the updated feature at time step \(t\), which encodes the comprehensive and dynamic spatio-temporal correlations, \(W_{adt}^{t} \in \mathbb{R}^{d_{x} \times d}\), \(W_{adt}^{0} \in \mathbb{R}^{d_{x} \times d}\), and \(b_{adt} \in \mathbb{R}^{d}\) are learnable parameters.

d) Gating fusion: The pre-defined and adaptive STJGs represent the spatio-temporal correlations from distinct perspectives. To enhance the representation ability, we use a gating mechanism to fuse the features extracted on two types of STJG. Specifically, we define a gate to control the importance of two features as:

\[
G = \text{sigmoid}(W^{g}[Z_{t}^{pd}, Z_{t}^{adt}] + b^{g}),
\]

(13)

where \([\cdot, \cdot]\) denotes the concatenation operation, the sigmoid function is used to control the output lies in range \([0, 1]\), \(W^{g} \in \mathbb{R}^{2d_{x} \times d}\) and \(b^{g} \in \mathbb{R}^{d}\) are learnable parameters. The gate \(G \in \mathbb{R}^{N \times d}\) controls the information flow between pre-defined and adaptive STJGs in both node-wise and channel-wise. Based on the gate, we fuse two features as:

\[
Z_{t} = G \odot Z_{t}^{pd} + (1 - G) \odot Z_{t}^{adt},
\]

(14)

where \(\odot\) denotes the element-wise product. As a result, \(Z_{t} \in \mathbb{R}^{N \times d}\) represents the updated representation of \(N\) nodes at time step \(t\), which aggregates the information from their spatial, temporal, and spatio-temporal neighborhoods on both types of STJG.
2) Dilated Causal STJGC Layers: The STJGC operation is able to model the correlations in different time ranges by controlling the time difference $k$. In addition, different STJGC layers aggregate information within diverse neighborhood ranges. This makes it flexible to model the spatio-temporal correlations in multiple neighborhood and time ranges. The information in different ranges reveals distinct traffic properties. A small range uncovers the local dependency and a large range indicates the global dependency. Inspired by the dilated causal convolution [48], [49], which is able to capture diverse time-ranges of dependencies in different layers, we propose dilated causal STJGC layers to capture multiple ranges of spatio-temporal dependencies.

a) Dilated causal convolution: The dilated causal convolution operation slides over the input sequence by skipping elements with a certain time step (i.e., dilation factor $\gamma$), and it involves only historical information at each time step to satisfy the causal constraint. In this way, it models diverse time-ranges of dependencies in different layers.

b) Dilated causal STJGC: As illustrated in Fig. 3, we first transform the inputs into $d$ dimension space using fully-connected layers. Then we stack a couple of STJGC layers upon it in the dilated causal way. Different to the standard dilated causal convolution using 1D CNN, we use the STJGC in each layer to model the dynamic and comprehensive spatio-temporal correlations. Suppose the length of input graph signals is $P = 12$, we could stack four STJGC layers with kernel size $K = 2$ and dilation factor $\gamma = \{2, 4, 4, 4\}$ in each layer, respectively. The residual connections [50] are also applied in each STJGC layer at the corresponding output time steps. The number of STJGC layers, dilation factors and kernel size could be re-designed according to the length of input graph signals, in each layer to model the dynamic and comprehensive spatio-temporal correlations between distant nodes at distant time steps. Thus, it is essential to combine the multi-range information. In addition, the importance of different ranges could be diverse. We propose a multi-range attention mechanism to aggregate the information of different ranges extracted by the dilated causal STJGC layers, and then apply independent fully-connected layers to produce the multi-step ahead prediction results.

D. Prediction Module

In this module, we first propose a multi-range attention mechanism to aggregate the information of different ranges extracted by the dilated causal STJGC layers, and then apply independent fully-connected layers to produce the multi-step ahead prediction results.

1) Multi-Range Attention: As introduced in Section IV-C2, each STJGC layer captures different spatio-temporal ranges of dependencies. A small range uncovers the local dependency and a large range indicates the global dependency, e.g., the correlations between distant nodes at distant time steps. Thus, it is essential to combine the multi-range information. In addition, the importance of different ranges could be diverse. We propose a multi-range attention mechanism to aggregate the information of different ranges. Mathematically, we denote the hidden state of node $v_i$ at time step $t$ in $m$-th STJGC layer as $z_i^{(m)} \in \mathbb{R}^d$, the attention score is computed as:

$$s_i^m = v^\top \tanh(W^a z_i^{(m)} + b^a),$$
$$\alpha_i^m = \frac{\exp(s_i^m)}{\sum_{m=1}^M \exp(s_i^m)},$$

where $W^a \in \mathbb{R}^{d \times d}$, $b^a \in \mathbb{R}^d$, and $v \in \mathbb{R}^d$ are learnable parameters, $M$ is the number of STJGC layers, and $\alpha_i^m$ is the attention score, indicating the importance of $z_i^{(m)}$. Based on the attention scores, the output of the last STJGC layer covers the information from all input time steps.
scores, the multi-range information can be aggregated as:

\[ y_i = \sum_{m=1}^{M} \alpha_i^m z_i^{(m)} \]  \hspace{1cm} (17)

where \( y_i \) is the updated feature of node \( v_i \), which aggregates the information from multiple spatio-temporal ranges. The attention mechanism is conducted on all of the \( N \) nodes in parallel with shared learnable parameters, and produces an output as \( Y \in \mathbb{R}^{N \times d} \).

2) Independent Fully-Connected Layers: As the traffic of different time steps may exhibit different properties, it would be better to use different networks to generate the predictions at different forecasting horizons. We thus apply \( Q \) independent two-fully-connected layers upon \( Y \) to produce the \( Q \) time steps ahead prediction results:

\[ \hat{X}_{t+i} = \phi(YW_t^1 + b_1^1)W_2^1 + b_2^1, \]  \hspace{1cm} (18)

where \( \hat{X}_{t+i} \) denotes the prediction result of time step \( t+i \) (\( i = 1, 2, \ldots, Q \)), \( W_t^1 \in \mathbb{R}^{d \times d} \), \( b_1^1 \in \mathbb{R}^d \), \( W_2^1 \in \mathbb{R}^{d \times 1} \), and \( b_2^1 \in \mathbb{R} \) are the corresponding learnable parameters, \( \phi(\cdot) \) is an activation function.

3) Loss Function: The mean absolute error (MAE) loss is commonly used in the traffic forecasting problem [4], [7], [39]. In practice, the MAE loss optimizes all prediction values equally regardless of the size of the value, which leads to relatively non-ideal predictions for small values compared to the predictions of large values. The mean absolute percentage error (MAPE) loss is more relevant to the predictions of small values. Thus, we propose to combine the MAE loss and MAPE loss as our loss function:

\[ \mathcal{L}(\hat{X}_{t+i}; \Theta) = \frac{1}{Q} \left( \sum_{i=1}^{Q} \left( |\hat{X}_{t+i} - X_{t+i}| + \beta \cdot \frac{|\hat{X}_{t+i} - X_{t+i}|}{X_{t+i}} \cdot 100 \right) \right), \]  \hspace{1cm} (19)

where \( \beta \) is used to balance MAE loss and MAPE loss, \( \Theta \) denotes all learnable parameters in STJGCN.

E. Complexity Analysis

We further analyze the time complexity of the main components in each module in our STJGCN.

In the \textit{STJG construction module}, the computation mainly comes from the learning of adaptive STJG adjacency matrix (6). The time complexity is \( O(Nd^2 + N^2 d) \), where \( N \) denotes the number of nodes, \( d \) is the dimension of the spatio-temporal embedding. Regarding \( d \) as a constant, the time complexity turns to \( O(N^2) \), which is attributed to the pairwise computation of the \( N \) nodes’ embeddings. One concern is that the large-scaled node would result in a more expensive cost. To mitigate the scale problem, we suggest to only calculate the connected edges in adaptive STJG adjacency matrix according to a priori knowledge (i.e., pre-defined STJG).

In the \textit{dilated casual STJGC module}, the time complexity mainly depends on the computation of each STJGC operation ((10) and (12)), which incurs \( O(K(|E|d + Nd^2)) \) time complexity. Here, \( K \) is the kernel size, \( |E| \) denotes the number of edges in the graph, and \( d \) is the dimension of hidden states. The time complexity of STJGC mainly depends on \( |E| \), as each node aggregates information from its neighborhoods, whose number is equal to the edge number.

In the \textit{prediction module}, the time complexities of multi-range attention mechanism ((15), (16), and (17)) and independent fully-connected layers (18) are \( O(NMd + d^2) \) and \( O(QNd^2) \), respectively. Thus, the total time complexity of the prediction module is \( O(NMd + Qd^2) \), where \( M \) is the number of STJGC layers and \( Q \) is the number of time steps to be predicted. The time complexity is highly related to \( Q \), as we use \( Q \) independent fully-connected layers to produce the multi-step prediction results.

V. Experiments

A. Datasets

We evaluate our STJGCN on five highway traffic datasets: PeMSD3, PeMSD4, PeMSD7, PeMSD8 and Seattle-Loop. The previous four datasets are released in [8], [9]. These datasets are collected by the Caltrans Performance Measurement System (PeMS) from 4 districts in real time every 30 seconds. The raw traffic data is aggregated into 5-minute time interval. There are three kinds of traffic measurements in PeMSD4 and PeMSD8 datasets, including total flow, average speed, and average occupancy. In PeMSD3 and PeMSD7 datasets, only the traffic flow is recorded. Seattle-Loop is released in [51], [52], which is a highway speed dataset collected from 323 loop detectors in the Greater Seattle Area. The dataset contains 5-minute resolution traffic speed data for the entirety of 2015. Following previous studies [10], [11], [53], we predict the traffic flow in first four datasets, and traffic speed in last dataset. The summary statistics of five datasets are presented in Table I.

All datasets are normalized using the Z-Score method, and be split in chronological order with 60% for training, 20% for validation, and 20% for testing. The pair-wise road network distances are provided in the datasets, and we use them to construct the pre-defined STJG according to (3).

B. Experimental Setup

1) Evaluation Metrics: We adopt three widely used metrics for evaluation, i.e., mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE), which are defined as:

\[ MAE = \frac{1}{NQ} \sum_{i=1}^{N} \sum_{j=1}^{Q} |\hat{X}_{i,t+j} - X_{i,t+j}|, \]  \hspace{1cm} (20)

| Dataset | Time range               | Time interval | # Nodes |
|---------|--------------------------|---------------|---------|
| PeMSD3  | 1/Sept 2018 - 30/Nov 2018| 5-minute      | 358     |
| PeMSD4  | 1/Jan 2018 - 28/Feb 2018 | 5-minute      | 307     |
| PeMSD7  | 1/May 2017 - 31/Aug 2017 | 5-minute      | 883     |
| PeMSD8  | 1/Jul 2016 - 31/Aug 2016 | 5-minute      | 170     |
| Seattle-Loop | 1/Jan 2015 - 31/Dec 2015 | 5-minute      | 323     |
TABLE II
HYPERPARAMETER SETTINGS OF STJGCN ON FIVE DATASETS

| Dataset       | \(\Delta_{pdf} \) | \(\Delta_{adt} \) | \(d\) | \(K\) | \(\beta\) |
|---------------|------------------|------------------|------|------|--------|
| PeMSD3        | 0.5              | 0.5              | 64   | 2    | 0.1    |
| PeMSD4        | 0.5              | 0.5              | 64   | 3    | 1.0    |
| PeMSD7        | 0.9              | 0.7              | 64   | 2    | 0.5    |
| PeMSD8        | 0.3              | 0.3              | 64   | 2    | 1.5    |
| Seattle-Loop  | 0.5              | 0.3              | 64   | 2    | 0.1    |

\[
RMSE = \sqrt{\frac{1}{NQ} \sum_{i=1}^{N} \sum_{j=1}^{Q} (\hat{X}_{i,t+j} - X_{i,t+j})^2},
\]

\[
MAPE = \frac{1}{NQ} \sum_{i=1}^{N} \sum_{j=1}^{Q} \frac{|\hat{X}_{i,t+j} - X_{i,t+j}|}{X_{i,t+j}},
\]

where \(\hat{X}_{i,t+j}\) and \(X_{i,t+j}\) denote the prediction result and ground truth of node \(v_i\) at time step \(t + j\), respectively, \(N\) is the number of nodes, and \(Q\) is the number of time steps to be predicted.

2) Experimental Settings: The PeMSD3 and PeMSD7 datasets contain one traffic measurement (i.e., traffic flow). Thus, the dimensions of the input and output are \(C = 1\) and 1, respectively. The PeMSD4 and PeMSD8 datasets contain three traffic measurements (i.e., traffic flow, average speed, and average occupancy), and only the traffic flow is predicted in the experiments [10], [11]. Thus, the dimensions of the input and output are \(C = 3\) and 1, respectively. The Seattle-Loop dataset contains one traffic measurement (i.e., traffic speed). Thus, the dimensions of the input and output are \(C = 3\) and 1. Following previous studies [10], [11], [53], we use the traffic data of historical 12 time steps (\(P = 12\)) to forecast the next 12 time steps (\(Q = 12\)).

The core hyperparameters in STJGCN include the thresholds \(\Delta_{pdf}\) and \(\Delta_{adt}\) in pre-defined and adaptive STJG adjacency matrices respectively, the dimension \(d\) of hidden states, the kernel size \(K\) of each STJGC layer, and the threshold \(\beta\) in the loss function. We tune these hyperparameters on the validation set that achieve the best validation performance. We provide a parameter study in Section V-C3. The detailed hyperparameter settings of STJGCN on five datasets are presented in Table II.

The nonlinear activation function \(\phi(\cdot)\) in our STJGCN refers to the ReLU activation [46], and we also add a Batch Normalization [54] layer before each ReLU activation function.

We train our model using the Adam optimizer [55] with an initial learning rate 0.001 and batch size 64 on a NVIDIA Tesla V100 GPU card. We run the experiments for 200 epochs and save the best model that evaluated on the validation set. We run each experiment 5 times, and report the mean errors and standard deviations.

3) Baseline Methods: We compare STJGCN with 11 baseline methods, which could be divided into two categories. The first category is the time-series prediction models, including:
- VAR [56]: Vector Auto-Regressive is a traditional time-series model, which can capture pairwise relationships among all traffic series.
- FC-LSTM [57]: an encoder-decoder framework using long short-term memory (LSTM) with peephole for multi-step time-series prediction.
- SVR [58]: Support Vector Regression utilizes a linear support vector machine to perform regression.

The second category refers to the spatio-temporal graph neural networks, which are detailed as follows:
- DCRNN [4]: Diffusion Convolutional Recurrent Neural Network, which models the traffic as a diffusion process, and integrates diffusion convolution with recurrent neural network (RNN) into the encoder-decoder architecture.
- STGCN [6]: Spatio-Temporal Graph Convolutional Network, which employs graph convolutional network (GCN) to capture spatial dependencies and 1D convolutional neural network (CNN) for temporal correlations modeling.
- ASTGCN [8]: Attention based Spatio-Temporal Graph Convolutional Network that designs spatial and temporal attention mechanisms to capture spatial and temporal patterns, respectively.
- Graph WaveNet [7]: a graph neural network that performs diffusion convolution with both pre-defined and self-adaptive adjacency matrices to capture spatial dependencies, and applies 1D dilated causal convolution to capture temporal dependencies.
- STSFCN [9]: Spatio-Temporal Synchronous Graph Convolutional Network that designs spatio-temporal synchronous modeling mechanism to capture localized spatio-temporal correlations.
- AGCRN [10]: Adaptive Graph Convolutional Recurrent Network that learns data-adaptive adjacency matrix for graph convolution to model spatial correlations and uses gated recurrent unit (GRU) to model temporal correlations.
- GMAN [39]: Graph Multi-Attention Network is an encoder-decoder framework, which designs multiple spatial and temporal attention mechanisms in the encoder and decoder to model spatio-temporal correlations, and a transform attention mechanism to transform information from encoder to decoder.
- Z-GCNETs [11]: Time Zigzags at Graph Convolutional Networks that introduce the concept of zigzag persistence [59] into the graph convolutional networks for modeling the spatial correlations and use the GRU networks to capture the temporal dependencies.

C. Experimental Results

1) Overall Comparison: Table III presents the forecasting performance comparison of our STJGCN with 11 baseline methods. We observe that: 1) the time-series prediction models, including traditional approach (i.e., VAR), machine learning based method (i.e., SVR), and deep neural network (i.e., FC-LSTM) perform poorly as they only consider the temporal correlations. 2) Spatio-temporal graph neural networks generally achieve better performances as they further model the spatial correlations using graph neural networks. 3) Our STJGCN performs the best in terms of all metrics on all datasets (1.4%−7.7% improvement).
TABLE III
FORECASTING PERFORMANCE COMPARISON OF DIFFERENT MODELS ON FIVE DATASETS

| Dataset | Metrics | VAR | SVR | FC-LSTM | DCRNN | STJGCN | ASFJGCN | Graph WaveNet | STJGCN | AGCRN | GMAN | Z-GCNNet | STJGCN |
|---------|---------|-----|-----|---------|-------|--------|---------|---------------|--------|-------|------|--------|--------|
| PeMSD3  | MAE     | 19.72 | 19.77 | 19.95±0.39 | 17.74±0.33 | 19.78±0.67 | 18.87±0.42 | 13.67±0.08 | 15.74±0.09 | 18.10±0.06 | 19.52±0.09 | 19.60±0.07 | 14.92±0.10 | 15.76±0.07 |
| RMSE    | 32.38 | 32.78 | 33.83±0.46 | 29.64±0.47 | 33.87±1.14 | 30.71±1.02 | 24.62±0.14 | 24.39±0.24 | 28.55±0.28 | 24.33±0.19 | 27.90±0.86 | 23.79±0.41 |
| MAPE (%)| 20.50 | 23.04 | 19.95±0.51 | 16.83±0.93 | 17.33±5.94 | 18.95±1.08 | 15.72±0.23 | 15.46±0.72 | 15.02±0.28 | 15.19±0.25 | 15.51±1.17 | 14.81±1.16 |

TABLE IV
EFFECT OF SPATIO-TEMPORAL CONNECTIONS, DYNAMIC GRAPH MODELING, MULTIPLE-RANGE INFORMATION, AND INDEPENDENT FULLY-CONNECTED LAYERS

| Dataset | Metrics | STJGCN | w/o STC-pdf | w/o STC-adt | w/o STC | w/o dgm | w/o mr | w/o att | w/o idp |
|---------|---------|--------|-------------|-------------|---------|--------|------|-------|--------|
| PeMSD4  | MAE     | 18.81±0.06 | 18.99±0.14 | 19.07±0.10 | 19.36±0.09 | 19.70±0.06 | 19.03±0.09 | 18.97±0.09 | 18.69±0.89 |
| RMSE    | 30.35±0.09 | 30.63±0.23 | 30.71±0.13 | 30.80±0.10 | 31.47±0.05 | 30.79±0.08 | 30.56±0.12 | 30.46±0.10 |
| MAPE (%)| 11.92±0.04 | 12.02±0.07 | 12.07±0.06 | 12.27±0.08 | 12.39±0.07 | 13.10±0.03 | 11.96±0.02 | 11.95±0.02 |
| PeMSD8  | MAE     | 14.53±0.17 | 14.63±0.23 | 14.82±0.09 | 15.07±0.07 | 15.49±0.22 | 15.11±0.57 | 14.67±0.11 | 14.60±0.11 |
| RMSE    | 23.74±0.20 | 24.01±0.22 | 24.11±0.14 | 24.22±0.14 | 24.94±0.23 | 24.49±0.52 | 24.03±0.35 | 23.96±0.21 |
| MAPE (%)| 9.15±0.09 | 9.18±0.19 | 9.26±0.08 | 9.48±0.06 | 9.55±0.16 | 9.39±0.22 | 9.16±0.09 | 9.16±0.12 |

against the second best results). Compared with other graph-based methods, the advantages of our STJGCN are three-fold. First, STJGCN models comprehensive spatio-temporal correlations. Second, STJGCN is able to capture dynamic dependencies at different time steps. Third, STJGCN leverages the information of multiple spatio-temporal ranges.

2) Ablation Study: To better understand the effectiveness of different components in STJGCN, we conduct ablation studies on PeMSD4 and PeMSD8 datasets.

a) Effect of spatio-temporal connections: One difference between our STJGCN with normal spatio-temporal graph is that we explicitly add the spatio-temporal connections between different nodes at different time steps. To evaluate the effectiveness of this approach, we drop them separately/simultaneously from the pre-defined or/and adaptive STJG. These three variants of STJGCN are named as “w/o STC-pdf” (drop in pre-defined STJG), “w/o STC-adt” (drop in adaptive STJG), and “w/o STC” (drop in both types of STJG), respectively. The results in Table IV demonstrate that the introduction of spatio-temporal connections improves the performance as it helps the model to explicitly capture comprehensive spatio-temporal correlations.

b) Effect of dynamic graph modeling: To evaluate the effect of dynamic graph modeling, we conduct experiments of using shared fully-connected layers with Q units in the output layer to produce the Q time steps predictions. We name this variant of STJGCN as “w/o idp”, and present the experimental results in Table IV. We observe that STJGCN improves the performances by introducing independent learning parameters for multi-step prediction. A potential reason is that the traffic of different time steps may exhibit different properties, and using different networks to generate the predictions at different forecasting horizons could be beneficial.

c) Effect of multi-range information: To verify the effect of multi-range information, we design a variant of STJGCN, namely “w/o mr”, in which we do not combine multiple ranges of information but directly use the output of the last STJGCN layer to produce the predictions. The results in Table IV indicate the necessity of leveraging multi-range information. We further design a variant “w/o att” that directly adds the outputs of each STJGC layer together without using the multi-range attention mechanism, and it performs worse than STJGCN, showing that it is beneficial to distinguish the importance of different ranges of information.

d) Effect of independent fully-connected layers: In the prediction module, we use Q independent fully-connected layers to produce the multi-step predictions. To evaluate the effectiveness of this, we conduct experiments of using shared fully-connected layers with Q units in the output layer to produce the Q time steps predictions. We name this variant of STJGCN as “w/o idp”, and present the experimental results in Table IV. We observe that STJGCN improves the performances by introducing independent learning parameters for multi-step prediction. A potential reason is that the traffic of different time steps may exhibit different properties, and using different networks to generate the predictions at different forecasting horizons could be beneficial.

e) Effect of different STJGC adjacency matrix configurations: We further conduct experiments of using different STJGC adjacency matrix configurations to evaluate their effectiveness. As shown in Table V, the models with only pre-defined STJGC adjacency matrices (lines 3–4) achieve poor performances as they do not capture the underlying dependencies in the data. We observe that the models with only adaptive STJGC adjacency matrices (lines 5–6) could realize promising performances, which indicates that our model can also be used even if the graph structure is unavailable. By using both pre-defined and adaptive STJGC adjacency matrices (line 7), we could achieve better results. We further apply a gating fusion approach (Section IV-C1d) in STJGCN (line 8) and observe consistent improvement of

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE V  
EFFECT OF DIFFERENT STJG ADJACENCY MATRIX CONFIGURATIONS. THE TERM "GF" IN THE LAST LINE DENOTES THE GATING FUSION APPROACH  

| STJG adjacency matrix configuration | PeMSD4 | PeMSD8 |
|-----------------------------------|-------|-------|
| $A_{adj}^{(k)}$, $A_{base}^{(k)}$ | 24.64±0.05 38.21±0.02 15.70±0.08 | 18.52±0.10 29.24±0.18 11.35±0.08 |
| $[L_{t-k}, L_{t}]$ | 24.40±0.06 38.03±0.23 15.47±0.03 | 18.12±0.07 28.49±0.16 11.19±0.11 |
| $[L_{t-k}, L_{t-t}]$ | 19.39±0.12 31.60±0.23 12.38±0.08 | 15.93±0.15 25.87±0.23 9.98±0.07 |
| $[A_{adj}^{(k)}$, $A_{base}^{(k)}$, $L_{t-k}, L_{t-t}]$ | 19.35±0.13 31.47±0.16 12.34±0.14 | 15.42±0.15 24.80±0.32 8.95±0.14 |
| $[A_{adj}^{(k)}$, $A_{base}^{(k)}$, $L_{t-k}, L_{t-t}]$ + gf (ours) | 18.81±0.06 30.35±0.09 11.92±0.04 | 14.53±0.17 23.74±0.20 9.15±0.09 |

Fig. 4. Parameter study on the PeMSD4 dataset.

As shown in Figs. 4(c) and 5(c), increasing the number of hidden units could enhance the model’s expressive capacity. However, when it is larger than 64, the performance degrades significantly, as the model needs to learn more parameters and may suffer from the over-fitting problem.

Figs. 4(d) and 5(d) show that the model performs poorly when the kernel size equals to 1, as it captures only the spatial dependencies and does not consider the correlations in the temporal dimension. We can further observe that it is enough to aggregate the information from neighboring 2 or 3 time steps at each time step. When $K = 4$, the model’s performance degrades. It is possibly because that a node’s information at a time step may only correlated to the nodes at a limited number of neighboring time steps, and a large $K$ would introduce noises into the model. Thus, according to the validation loss, we set $K = 3$ and $K = 2$ on the PeMSD4 and PeMSD8 datasets, respectively.

In the parameter study of the threshold $\beta$ in the loss function, we report the validation MAE, RMSE, and MAPE instead of reporting the loss value, as the size of $\beta$ directly impacts the size of the loss value. As shown in Figs. 4(e), (g), 5(e), and (g), a larger $\beta$ means the model optimizes more on the MAE loss and less on the MAPE loss, and thus leads to smaller MAPE and larger MAE.

the predictive performances, as the gate is able to control the information flow between pre-defined and adaptive STJGs.

3) Parameter Study: We conduct a parameter study on five core hyperparameters in STJGCN on the PeMSD4 and PeMSD8 datasets, including the thresholds $\delta_{pdf}$ and $\delta_{adt}$ in the pre-defined and adaptive STJG adjacency matrices, respectively, the dimension $d$ of hidden states, the kernel size $K$ in the STJGC operation, and the threshold $\beta$ in the loss function. We change the parameter under investigation and fix other parameters in each experiment. Figs. 4 and 5 show the experimental results on the PeMSD4 and PeMSD8 datasets, respectively.

As shown in Figs. 4(a), (b), 5(a), and (b), the performance is not strongly sensitive to the sparsity of the STJG adjacency matrices, which we think is because the adaptive STJG adjacency matrix could adjust itself for aggregating the neighboring information during the training stage. While, in general, a more sparse adjacency matrix is beneficial to select the most related nodes for each node, and leads to better results. However, a too sparse graph may lose the connections between interrelated nodes, and thus degrades the performances. According to the validation loss, we set $\delta_{pdf} = \delta_{adt} = 0.5$ in the PeMSD4 dataset, and $\delta_{pdf} = 0.5$, $\delta_{adt} = 0.3$ in the PeMSD8 dataset.
Through a comprehensive consideration of the validation MAE, RMSE, MAPE and their standard deviations, we choose to use $\beta = 1.0$ and $\beta = 1.5$ in the PeMSD4 and PeMSD8 datasets, respectively.

4) Performance Comparison at Each Horizon: Figs. 6 and 7 present the forecasting performance comparison of our STJGCN with five representative baseline methods (i.e., Graph WaveNet, STSGCN, AGCRN, GMAN, and Z-GCNETs) at each prediction time step on the PeMSD4 and PeMSD8 datasets, respectively. We exclude other baseline methods due to their poorer performances, as shown in Table III. We can observe that Graph WaveNet performs well in the short-term (one or two time steps ahead) prediction. However, its performance degrades quickly with the increase of the forecasting horizon. The performance of GMAN degrades slowly when the predictions are made further into the future, and it performs well in the long-term (e.g., 12 time steps ahead) prediction, while still worse than STJGCN. In general, our model achieves the best performances at almost all horizons in terms of all three metrics on both datasets.

5) Model Size and Computation Time: We present the comparison of model size and computation time of our STJGCN with graph-based baseline methods in Table VI. The results in four PeMS datasets demonstrate the high computation efficiency of our model. In terms of the model size, STJGCN has fewer parameters than most of the baseline models. In the training phase, our model runs faster than other methods except for STGCN. In the inference stage, STGCN runs very slowly as it adopts an iterative way to generate multi-step predictions, while STJGCN and Graph WaveNet are the most efficient. By further considering the prediction accuracy (see Table III), our model shows superior ability in balancing predictive performances and time consumption as well as parameter settings.

The results in Seattle-Loop dataset show that our STJGCN compares favorably to baseline methods. In terms of the model...
size, STJGCN has not been affected by the larger amount of data, and still has fewer parameters than most of the baseline models. In the training phase, our STJGCN is faster than GMAN and Z-GCNETs. Other 6 baselines are more efficient than STJGCN but they show poor prediction performance (see Table III). In the inference stage, STJGCN is only slower than STSGCN and AGCRN, while both of which have worse prediction accuracy than our model (see Table III).

VI. CONCLUSION

We proposed STJGCN, which models comprehensive and dynamic spatio-temporal correlations and aggregates multiple ranges of information to forecast the traffic conditions over several time steps ahead on a road network. When evaluated on five public traffic datasets, STJGCN showed high computation efficiency and outperformed 11 state-of-the-art baseline methods. Our model could be potentially applied to other spatio-temporal data forecasting tasks, such as air quality inference and taxi demand prediction. We plan to investigate this as future works.

REFERENCES

[1] S. Wang, J. Cao, and P. S. Yu, “Deep learning for spatio-temporal data mining: A survey,” IEEE Trans. Knowl. Data Eng., vol. 34, no. 8, pp. 3681–3700, Aug. 2022.
[2] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. M. Choudhury, and K. Qin, “A survey on modern deep neural network for traffic prediction: Trends, methods and challenges,” IEEE Trans. Knowl. Data Eng., vol. 34, no. 4, pp. 1544–1561, Apr. 2022.
[3] C. Zheng, C. Wang, X. Fan, J. Qi, and X. Yan, “STPC-Net: Learn massive geo-sensory data as spatio-temporal point clouds,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 4, pp. 1544–1561, Apr. 2022.
[4] Y. Liu, R. Yu, C. Shahabi, and Y. Liu, “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting,” in Proc. Int. Conf. Learn. Representations, 2018, pp. 1–16.
[5] W. Cheng, Y. Shen, Y. Zhu, and H. Huang, “A neural attention model for urban air quality inference: Learning the weights of monitoring stations,” in Proc. AAAI Conf. Artif. Intell., 2018, pp. 2151–2158.
[6] B. Yu, H. Yin, and Z. Zha, “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting,” in Proc. Int. Joint Conf. Artif. Intell., 2018, pp. 3634–3640.
[7] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, “Graph wavenet for deep spatial-temporal graph modeling,” in Proc. Int. Joint Conf. Artif. Intell., 2019, pp. 1907–1913.
[8] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, “Attention based spatial-temporal graph convolutional networks for traffic flow forecasting,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 922–929.
[9] C. Song, Y. Lin, S. Guo, and H. Wan, “Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 914–921.
[10] W. Bai, L. Yao, C. Li, X. Wang, and C. Wang, “Adaptive graph convolutional recurrent network for traffic forecasting,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2020, Art. no. 1494.
[11] Y. Chen, I. Segovia-Dominguez, and Y. R. Gel, “Z-GCNETs: Time zigzags at graph convolutional networks for time series forecasting,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 1684–1694.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
23] S. Zhang, Y. Guo, P. Zhao, C. Zheng, and X. Chen, "A graph-based temporal attention framework for multi-sensor traffic flow forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 3104–3112, Jul. 2022.

24] S. Yuan, Y. Xiong, and D. Lin, "Spatial-temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3482–3489.

25] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12026–12035.

26] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "A comprehensive survey on traffic prediction," 2020, arXiv:2004.05555.

27] J. Sun, Z. Qian, L. Li, X. Yi, Y. Liang, and Y. Zheng, "Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 5, pp. 2348–2359, May 2022.

28] G. Gu et al., "Exploiting interpretable patterns for flow prediction in dockless bike sharing systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, pp. 640–652, Feb. 2022.

29] R. Jiang et al., "DeepCrowd: A deep model for large-scale citywide crowd density and flow prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 276–290, Jan. 2023.

30] S. Guo, Y. Lin, H. Wan, X. Li, and G. Cong, "Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 11, pp. 5415–5428, Nov. 2022.

31] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long-term short-memory neural network for traffic speed prediction using remote microwave sensor data," *Transp. Res. Part C: Emerg. Technol.*, vol. 54, pp. 187–197, 2015.

32] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1655–1661.

33] H. Yao et al., "Deep multi-view spatial-temporal network for taxi demand prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2588–2595.

34] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, "Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 33rd Innov. Appl. Conf. Intell. Conf. 9th AAAI Symp. Educ. Adv. Artif. Intell.*, 2019, Art. no. 695.

35] C. Zheng, X. Fan, C. Wen, L. Chen, C. Wang, and J. Li, "DeepSTD: Mining spatio-temporal disturbances of multiple context factors for citywide traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3744–3755, Sep. 2020.

36] J. Zhang, Y. Zheng, J. Sun, and D. Qi, "Flow prediction in spatio-temporal networks based on multitask deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 3, pp. 468–478, Mar. 2020.

37] J. Ye, J. Zhao, K. Ye, and C. Xu, "How to build a graph-based deep learning architecture in traffic domain: A survey," 2020, arXiv:2005.11691.

38] Y. H. Lau and R. C.-W. Wong, "Spatio-temporal graph convolutional networks for traffic forecasting: Spatial layers first or temporal layers first?", in *Proc. 29th Int. Conf. Adv. Geographic Inf. Syst.*, 2021, pp. 427–430.

39] C. Zheng, X. Fan, C. Wang, and J. Qi, "GMAN: A graph multi-attention network for traffic prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1234–1241.

40] X. Wang et al., "Traffic flow prediction via spatial temporal graph neural network," in *Proc. Web Conf.*, 2020, pp. 1082–1092.

41] A. Vassani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 9908–9909.

42] Y.-J. Lu and C.-T. Li, "AGSTN: Learning attention-adjusted spatio-temporal networks for short-term urban sensor value forecasting," in *Proc. IEEE Int. Conf. Data Mining*, 2020, pp. 1148–1153.

43] Y. Wang, H. Yin, H. Chen, T. Wo, J. Xu, and K. Zheng, "Origin-destination matrix prediction via graph convolution: A new perspective of passenger demand modeling," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 1227–1235.

44] W. Chen, L. Chen, Y. Xie, W. Cao, Y. Gao, and X. Feng, "Multi-range attentive bicomponent graph convolutional network for traffic forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 3529–3536.

45] D. Peng, C. Shahabi, U. Demiryurek, L. Zhu, R. Yu, and Y. Liu, "Latent space model for road networks to predict time-varying traffic," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1524–1534.

46] J. Ye, J. Zhao, K. Ye, and C. Xu, "How to build a graph-based deep learning architecture in traffic domain: A survey," 2020, arXiv:2005.11691.
Xiaoliang Fan (Senior Member, IEEE) received the PhD degree from the University Pierre and Marie Curie, France, in 2012. He is a senior research specialist with the Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, and Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China. His research interests include trustworthy AI and federated learning, spatio-temporal data mining, and services computing, etc. He has published more than 70 journals in the IEEE Transactions on Services Computing, IEEE Transactions on Mobile Computing, IEEE Transactions on Intelligent Transportation Systems, etc., and top conferences (AAAI, IJCAI, WWW, etc.) papers. His works are funded by NSFC and many industry collaborators. He is a senior member of the CCF.

Shirui Pan (Senior Member, IEEE) received the PhD degree in computer science from the University of Technology Sydney (UTS), Ultimo, NSW, Australia. He is currently a full professor with the School of Information and Communication Technology, Griffith University, Australia. His research interests include data mining and machine learning. To date, he has published more than 100 research papers in top-tier journals and conferences, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Neural Networks and Learning Systems, NeurIPS, ICML, and KDD. He is recognised as one of the AI 2000 AAAI/IJCAI Most Influential Scholars in Australia (2021).

Haibing Jin received the BSc degree in computer science and technology from Huazhong Agricultural University, Wuhan, China, in 2022. He is currently working toward the MAEng degree in computer science and technology with the Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, China. His research interests include spatio-temporal data representation learning and federated learning.

Zhao Peng Peng received the BSc degree in digital media technology from Shandong University, Weihai, China, in 2022. He is currently working toward the MAEng degree in computer science and technology with the Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, China. His research interests include spatio-temporal data representation learning and federated learning.

Zonghan Wu received the BS degree in systems science from the University of Shanghai for Science and Technology, and the MS degree in statistics from Linköping University. He is currently working toward the PhD degree in computer science with the University of Technology Sydney (UTS), Ultimo, NSW, Australia.

Cheng Wang (Senior Member, IEEE) received the PhD degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2002. He is currently a NanQiang professor with the School of Informatics, and director of Fujian Key Laboratory of Sensing and Computing for Smart Cities, both with Xiamen University, China. His research interests include remote sensing image processing, mobile LiDAR data analysis, and multi-sensor fusion. He has co-authored more than 150 papers in referred journals and top conferences including the IEEE Transactions on Geoscience and Remote Sensing, Pattern Recognition, IEEE Transactions on Intelligent Transportation Systems, AAAI, CVPR, IJCAI, and ISPRS-JPRS.

Philip S. Yu (Fellow, IEEE) received the PhD degree in electrical engineering from Stanford University, Stanford, CA, USA. He is currently a distinguished professor of computer science with the University of Illinois at Chicago, Chicago, IL, USA, where he is also the Wexler chair in information technology. He has published more than 830 articles in refereed journals and conferences. He holds or has applied for more than 300 U.S. patents. His research interests include Big Data, data mining, data streams, databases, and privacy. He is a fellow of the ACM. He received the ACM SIGKDD 2016 Innovation Award, the Research Contributions Award from the IEEE International Conference on Data Mining in 2003, and the Technical Achievement Award from the IEEE Computer Society in 2013.