Information Fusion 105 (2024) 102229

Contents lists available at ScienceDirect

Information Fusion

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

# TC-GCN: Triple cross-attention and graph convolutional network for traffic forecasting

Lei Wang [a], Deke Guo [b], Huaming Wu [c,*], Keqiu Li [a], Wei Yu [d,*]

[a] College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[b] Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China
[c] Center for Applied Mathematics, Tianjin University, Tianjin 300072, China
[d] School of International Business, Zhejiang Yuexiu University of Foreign Languages, Shaoxing 312069, China

ARTICLE INFO

*Keywords:*
Cross-attention
Graph convolutional network
Dynamic graph modeling
Traffic forecasting

ABSTRACT

With the rapid development of urbanization, increasingly more data are being acquired by intelligent transportation systems (ITSs), which is of great significance for traffic flow forecasting. Efficient intelligent traffic management systems (TMSs) depend on accurately forecasting traffic flows as well as reasonably researching and judging traffic states. However, the current prediction models used in transportation forecasting tasks are traditional temporal- and spatial-dimensional prediction approaches. Especially when faced with road channelization, the numbers of lanes and lane types increase significantly, causing significant increases in the dimensionality and complexity of traffic data, and research models utilizing the cross-interaction relationships among multiple dimensions have not been considered. This has led to unsatisfactory prediction results in complex traffic environments. Therefore, it is very important to explore the exchange–correlation of complex dimensional data and the mining of hidden attributes. This paper presents a method for performing multidimensional cross-attention and spatiotemporal graph convolution. This method fully considers cross-information and constructs an attention cross-view between every pair of dimensions among the channel, time and space domains to model the cross-dimensional dependencies of traffic data. We innovatively propose a triple cross-attention and graph convolutional network (TC-GCN), which can achieve further improved traffic forecasting performance. The TC-GCN is verified on two real-world traffic datasets, namely, METR-LA and PEMS-BAY, and the experimental results are compared with those of multiple advanced baselines, showing that the proposed approach outperforms most of these baselines, which proves the effectiveness of the method proposed in this paper.

## 1. Introduction

With the continuous advancement of social change and development, the construction of urban transportation infrastructure, comprising expressways, transportation hubs and high-speed railways, has inaugurated an era characterized by accelerated advancement. To alleviate the traffic congestion problem that follows, intelligent transportation systems (ITSs) have become an indispensable comprehensive technology for the development of urban transportation [1]. ITS is an important development direction in the field of transportation, as the problems of traffic congestion and frequent accidents become increasingly prominent with the growth in the number of vehicles and the population. Therefore, traffic flow condition prediction, road network optimization, and traffic management have become extremely important. An accurate traffic forecasting model helps with route planning and scheduling when traveling in daily life and assists traffic managers in adjusting their traffic strategies, which can reduce the burden of traffic congestion [2]. In response to the traffic pressure imposed on modern transportation, different lanes are usually designated in an actual road network.

Furthermore, at critical intersections and pivotal road junctures, it becomes imperative to incorporate intersection-channel interaction data. This inclusion introduces additional dimensions and intricacies to the already complex traffic data landscape. As a result, prediction models must exhibit heightened accuracy, possess the capacity to discern insights from multiple dimensions, and demonstrate adeptness in managing substantial volumes of sparsely available data. For the purpose of enhancing the overall traffic efficiency of urban trunk roads and optimizing their road capacity, the implementation of synchronized

* Corresponding authors.
*E-mail addresses:* wanglei2019@tju.edu.cn (L. Wang), guodeke@gmail.com (D. Guo), whming@tju.edu.cn (H. Wu), keqiu@tju.edu.cn (K. Li), weiyu@zyufl.edu.cn (W. Yu).

https://doi.org/10.1016/j.inffus.2024.102229
Received 22 November 2023; Received in revised form 30 December 2023; Accepted 2 January 2024
Available online 5 January 2024

traffic signal patterns, referred to as "green waves", emerges as a potent solution. During periods of low road utilization, the traffic signal lights situated at adjacent intersections along the urban trunk road are interlinked, facilitating unified dispatching and cohesive linkage control. As traffic flows converge upon each intersection, the duration spent waiting at red signals diminishes, and green lights expediently illuminate. This achievement is made possible by meticulously configuring the time intervals between the activation and deactivation states of the signal lights. However, achieving seamless signal light control necessitates proficient traffic flow forecasting and comprehensive state analysis. These prerequisites underscore the crucial role of effective traffic forecasting and state assessment in enabling the harmonized regulation of signal lights.

The prevailing approaches for traffic prediction can be categorized into two distinct groups based on the underlying technologies employed, namely, non-learning methods and learning methods. Moreover, within the domain of learning methods, a further bifurcation can be made, distinguishing between traditional learning approaches and deep learning techniques [1]. Non-learning methods are mainly based on statistics. Their core idea is usually to assume that the future forecasting data have the same distribution as the past data, and they use mathematical statistics techniques and multiple linear regression models to study multiple random variables. The correlations between the variables are established, as is a functional relationship model between the variables [3]. For example, in the historical average (HA) method, support vector regression (SVR) [4] and other methods, the model parameters mainly depend on the settings provided by experts in related fields and are not suitable for dealing with complex dynamic time series data. Therefore, these methods often fail to achieve satisfactory results in the face of nonlinear data such as traffic flow data. Due to the popularity of machine learning, traditional machine learning methods are also widely used in traffic flow forecasting. The feature representations between the data obtained in this manner exhibit a better nonlinear fitting effect. However, it is difficult for traditional machine learning methods to establish the spatiotemporal dependencies in traffic, and their ability to mine complex spatiotemporal patterns is still limited [5]; thus, their prediction results still have much room for improvement.

In recent years, deep learning-based methods have emerged as mainstream approaches for traffic flow prediction [6]. Their core idea typically involves the utilization of various deep learning techniques to learn the temporal and spatial characteristics inherent in traffic forecasting. For instance, graph convolutional networks (GCNs) [7] excel at capturing the spatial topologies within traffic networks, while recurrent neural networks (RNNs) [8] are better at capturing features in the temporal dimension. Studies have shown that hybrid models can compensate for the shortcomings of individual models, enabling them to better handle high-dimensional data, capture complex nonlinear relationships, and achieve improved prediction accuracy. These aspects make them more suitable for short-term traffic flow prediction at this stage [9]. For instance, Li et al. [10] proposed a diffusion convolutional RNN (DCRNN) model, wherein an RNN is employed to encapsulate the temporal features of traffic data, while diffuse convolution is utilized to capture the spatial features. The spatiotemporal GCN (ST-GCN) model [11] preserves the topology of traffic by means of spatiotemporal graph convolution. STTN [12] exploits both dynamic-directed spatial correlations and long-term temporal correlations, resulting in enhanced accuracy for long-term traffic prediction. This is accomplished through the integration of a self-attention mechanism, which effectively captures bidirectional temporal interdependencies across multiple time intervals.

Unfortunately, within the context of traffic data processing, post-channeling lane data tends to possess diminished granularity and more discrete attributes. This characteristic results in discontinuous spatiotemporal samples within lane data, presenting a challenge in establishing precise spatiotemporal relationships within models. In addition, channeled lane data also have complex hidden attributes spanning multiple dimensions, all intricately interconnected. Therefore, corresponding measures need to be taken during the modeling process to overcome these obstacles. As illustrated in Fig. 1, the congestion state caused by traffic accidents at detection point No. 1 during time $t_1$ can wield influence over No. 3 at time $t_2$ in both spatial and temporal dimensions. This necessitates an information exchange not only between No. 1 and No. 3 simultaneously but also across differing time points. In the realm of traffic prediction, it becomes imperative to model cross-information dynamics encompassing temporal and feature dimensions, temporal and spatial dimensions, and spatial and feature dimensions. To sum up, acknowledging and effectively incorporating the cross-influences among the diverse dimensions of traffic data is of paramount significance during the modeling of spatiotemporal and feature dimensions for accurate traffic forecasting. Moreover, the cross-influence between the various dimensions of the traffic data encountered when modeling the spatiotemporal and feature dimensions is extremely important for traffic forecasting.

To solve the aforementioned challenges, this paper introduces a comprehensive end-to-end framework termed the Triple Cross-Attention and GCN (TC-GCN). This framework operates seamlessly across the spatiotemporal and feature dimensions, addressing both the spatiotemporal dependency intricacies inherent in traffic prediction and the cross-dependence predicament across the three dimensions. By establishing a cross-information perspective for the spatiotemporal and feature dimensions, TC-GCN not only captures the intricate spatiotemporal dependencies in traffic prediction but also effectively addresses the cross-dependence intricacies present among these three dimensions. Remarkably, the TC-GCN framework represents the pioneering effort in modeling the intricate cross-dependencies within traffic data in the realm of traffic forecasting. The combination of Cross Attention and GCN aims to extract both cross and spatial information behind traffic spatiotemporal data simultaneously.

The main contributions of this paper are four-fold:

- TC-GCN: A framework capturing time/space/feature dimension interrelations;
- Addresses spatiotemporal information misalignment in traffic analysis;
- Features tailored readout for spatiotemporal and time/feature perspectives;
- Introduces time block mechanism for dynamic temporal data preservation.

The remainder of this paper is structured as follows: In Section 2, we concentrate on the existing research landscape pertinent to our study. Section 3 outlines the TC-GCN model devised for traffic forecasting. The experimental outcomes and their analysis are presented in Section 4. Lastly, Section 5 offers concluding remarks.

## 2. Related work

The predicament of traffic prediction holds a significant stature within the domains of spatiotemporal data mining and ITSs. substantial advancements have been realized in handling profoundly nonlinear data, particularly in the field of deep learning. The traffic data considered in this study is structured in the form of 3D tensors, encompassing temporal, spatial, and feature dimensions. The principal advancements aligned with this paper are detailed as follows.

Traffic prediction is a kind of spatiotemporal prediction task [7]. Within the domain of traffic research, traffic data are typically regarded as multivariate time series, encompassing diverse metrics like traffic speed, flow, and capacity across various monitoring points within a road network. The traffic flow prediction can be divided into traditional methods and deep learning-based methods into large categories, among which deep learning-based methods can be further refined into time-dependent models and spatial-dependent models, as shown in Table 1.

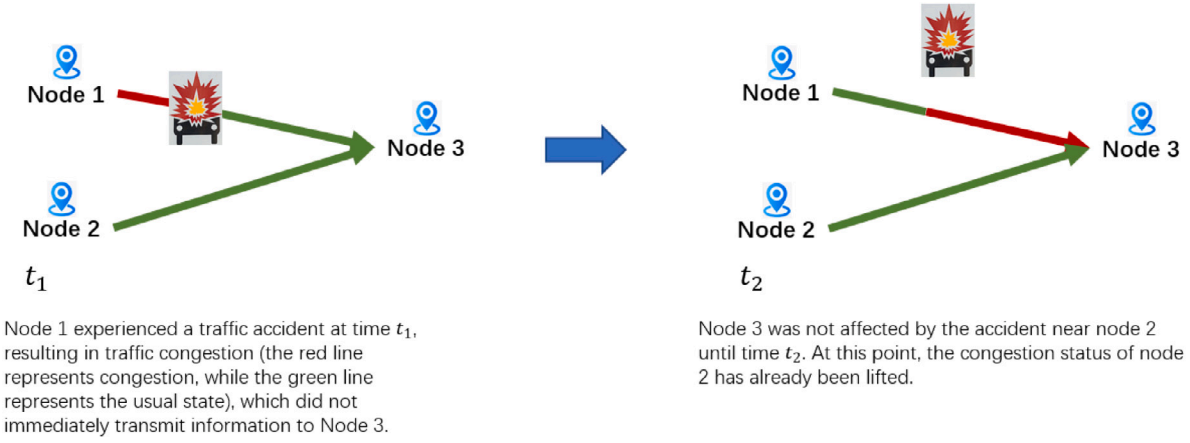



**Fig. 1.** The cross-information dependencies in the temporal and spatial dimensions.

**Table 1**
The summary of the traffic prediction methods.

| Category | | Approaches |
|---|---|---|
| Traditional methods | | [13], [14], [15], [16], [17], [18], [19] |
| Deep learning-based methods | Temporal dimension models | [8], [11], [12], [20], [21], [22], [23], [24], [25] |
| | Spatial dependence models | [3], [26], [27], [28], [29], [30], [31], [32], [33] |

Concurrently, spatial relationships exist among road segments, facilitating the representation of inter-road spatial associations as graphs, where individual road segments assume the role of nodes, and the edges embody their spatial correlations. This intricate interplay necessitates the effective capture of both temporal and spatial insights. The specific overview is as follows.

Firstly, the early research work on traffic flow prediction is traditional methods, which are mainly based on regression models, such as the ARIMA model [13] and the non-parametric regression model. These studies take into account the temporal correlation of future traffic flows in regions with historical data, but lack consideration of spatial correlation between regions. In addition, some scholars used machine learning algorithms, such as support vector machines (SVM) [14], gradient lifting regression trees [15], linear regression models [16], etc., to make predictions based on features extracted from multi-source data (such as POI data [17], weather data, etc.). Although these methods lack the deep consideration of the relationship between time and space and have a large room for improvement, they still have a wide range of applications in practice.

Secondly, the temporal dimension, invariably addressed through recurrent neural networks (RNNs), holds a pivotal role in influencing traffic forecasting outcomes [8,11,12,20,21]. Leveraging an extension of the fully connected LSTM (FC-LSTM) network, the convLSTM architecture [12] was based on LSTM and convolutional operations, thus enhancing its efficacy in extracting features from spatiotemporal data. Time dimension processing essentially involves modeling the trend and periodicity of spatiotemporal traffic data. In recent years, the rapid advancements in transformers [22] have led to the emergence of numerous long-time series prediction techniques, such as those detailed in [23–25]. However, it is imperative to note that these methodologies solely address temporal dependencies.

Thirdly, the utilization of graph neural networks is prevalent in processing spatial information. For instance, Wu et al. [32] proposed a novel graph neural network architecture, known as Graph-WaveNet, for spatiotemporal graph modeling. Their approach incorporates adaptive dependency matrices learned through node embeddings, facilitating the model's capacity to capture concealed spatial dependencies inherent in the input data. DCRNN [33] modeled the traffic flow as a diffusion process on a directed graph and introduced a diffusion-based convolutional recursive neural network, This framework, entrenched within deep learning, serves as an effective paradigm for traffic prediction, seamlessly merging spatial and temporal dependencies inherent in traffic flows. Additionally, GMAN [30] utilized a multigraph attention-based depth network that performs attention operations across both spatial and temporal dimensions.‘

However, the aforementioned spatiotemporal-dependent models tend to independently model temporal and spatial aspects, often disregarding the crucial cross-information interactions between these dimensions.

## 3. Methods

The fundamental objective of TC-GCN is to enhance the accuracy of traffic prediction by fostering a cross-dimensional approach that effectively models the information embedded within both the spatiotemporal and feature dimensions of the data.

### 3.1. Problem definition

**Definition 1** (*Road Network* $G$)**.** A traffic network is defined as a weighted undirected graph denoted by $G = (V, E, A)$, serving to depict the inherent topological configuration of its road infrastructure. Here, $V = \{v_0, \ldots, v_S\}$ represents a set of $S$ monitoring nodes, while $E$ corresponds to the ensemble of edges signifying interrelationships. The adjacency matrix $A \in \mathbb{R}^{S\times S}$ is employed to quantitatively express the degree of connection strength across these nodes.

**Definition 2** (*Feature Matrix* $X$)**.** The traffic data inherent in the road network $G$ is considered as the attribute features of the nodes within $V$. This is succinctly represented by the matrix $X \in \mathbb{R}^{C\times P\times S}$, where $C$ stands for the quantity of node attribute features, encompassing parameters like traffic speed, traffic flow, and traffic density. Furthermore, $P$ signifies the duration of the historical time series, while $S = |V|$ denotes the count of sensor nodes present in the network.

**Definition 3** (*Cross-Attention View*)**.** Let a three-dimensional structured tensor be denoted by $X \in \mathbb{R}^{d_1\times d_2\times d_3}$, with $d_1$, $d_2$, and $d_3$ representing the three dimensions. In this context, the cross-attention view linking dimensions $d_1$ and $d_2$ is precisely defined as:

$$V_{d_1 \rightleftarrows d_2} = r(X) \in \mathbb{R}^{d_1\times d_2}, \tag{1}$$

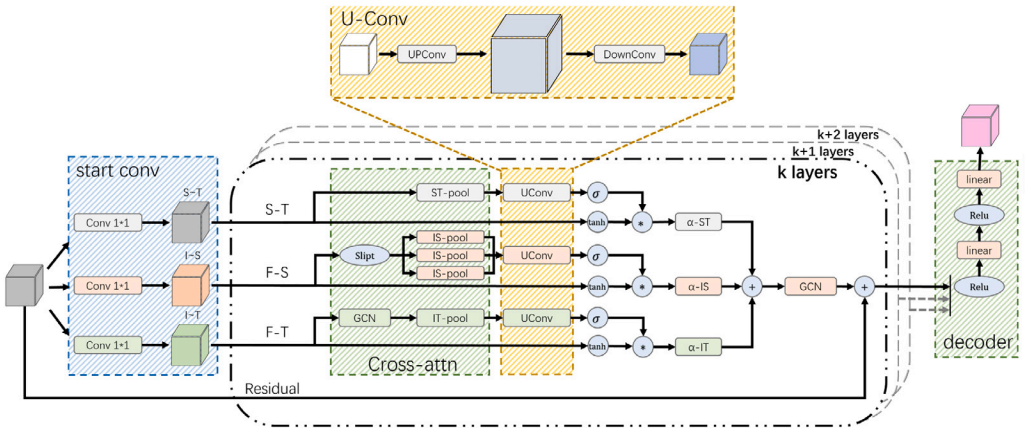



**Fig. 2.** The proposed TC-GCN framework. The core composition of the TC-GCN model involves an $L$-layer cross-information module comprised of ST blocks. A light blue background represents the starting convolution, an orange background represents u-convolution, and a green background represents the decoder.

where $r(X)$ stands for the primary content mapping function. It is important to note that identical functions can be established in the definitions of $V_{d_1 \leftrightarrows d_3}$ and $V_{d_2 \leftrightarrows d_3}$.

$V_{d_1 \rightleftarrows d_2}$ characterizes the mutual dependence existing between the dimensions $d_1$ and $d_2$ within the context of the three-dimensional tensor $X$. Within the scope of traffic data $X \in \mathbb{R}^{C \times P \times S}$, this concept possesses evident physical significance. As shown in Fig. 1, $V_{S \leftrightarrows P}$ represents spatiotemporal cross-information, and then $V_{S \leftrightarrows P}[n_3, t_1] = 0$ and $V_{S \leftrightarrows P}[n_3, t_2] \neq 0$. The primary aim of the TC-GCN model is to effectively capture the cross-information interactions within diverse dimensions of traffic data through the prism of a cross-attention view, thereby augmenting the accuracy of spatiotemporal traffic flow prediction. Evident from the earlier definition, the pivotal facet of the cross-attention view resides in the strategic design of $r(X)$.

Given historical traffic data denoted as $X = \{x_0, \ldots, x_P\} \in \mathbb{R}^{C \times S \times P}$ for each road intersection, the future traffic data can be predicted as follows:

$$\hat{Y} = f_\theta(X), \ X \in \mathbb{R}^{C \times P \times S}, \tag{2}$$

where $P$ signifies the length of the historical time window for traffic flow, $S$ stands for the number of traffic detection nodes, and $C$ represents the traffic attributes of each node, encompassing parameters such as speed and flow. $\hat{Y} = \{x_{P+1}, \cdots, x_{P+Q}\}$ represents the predicted output generated by the model $f_\theta$, where $\theta$ is the model parameter.

During the training process, the ground-truth value $Y$ corresponding to $\hat{Y}$ is provided. The training objective is to progressively minimize the disparity between $\hat{Y}$ and $Y$.

### 3.2. The TC-GCN framework

As illustrated in Fig. 2, the core composition of the TC-GCN model involves an $L$-layer cross-information module comprised of ST blocks. This module is designed to intricately amalgamate insights originating from the intricate intersection of the three dimensions intrinsic to traffic spatiotemporal data. These dimensions include spatiotemporal intersection, feature space intersection, and feature time intersection information. The process of intersecting each pair of these dimensions is orchestrated through the utilization of the cross-attention view elucidated in Definition 3. To effectively harness the spatial information ingrained within the data, TC-GCN systematically addresses the spatial dependency challenge prevalent within traffic data. This is executed through the integration of the diffusion convolution methodology [10] within the framework of ST blocks.

### 3.3. The cross-information contained in the ST blocks

Similar to the existing deep learning-based traffic prediction models [5,34], the TC-GCN initiates its process by standardizing the input data $X$. The computation formula for this standardization is expressed as follows:

$$X_{norm} = \frac{X - \mu}{\sigma}, \tag{3}$$

where $\mu$ and $\sigma$ denote the mean and variance of $X$, respectively. The incorporation of these standardization operations renders the TC-GCN deep learning model more amenable to effective training.

Simultaneously, we execute a two-dimensional convolution involving the temporal and spatial dimensions of the normalized $X_{norm}$, employing a kernel size of $k = 1 \times 1$. The original feature dimensionality $C$ is retained as the channel of the convolution, which is standardized to $F$. The computational process is illustrated by:

$$X_s = conv2D_s\left(X_{norm}\right) \in \mathbb{R}^{F \times P \times S}, \tag{4}$$

where $conv2D_s$ is a two-dimensional convolution and $F$ is the output channel.

In fact, following the operation elucidated by Eq. (4), we transform the original attribute dimension $C$ into the standardized feature dimension, thereby generating $X_s$. Subsequently, this processed data $X_s$ serves as input to the ST blocks and diffusion convolution modules within every layer of the TC-GCN. For a given layer $l \in 1, \ldots, L-1$, the input is denoted as $X^{(l)}$, while the output stemming from the application of the ST blocks and diffusion convolution is termed $X^{(l+1)}$. It is crucial to note that the initialization point is $X^{(0)} = X_s$. With this groundwork established, let us delve into a more comprehensive elaboration of the ST blocks.

The traffic data exhibits three interdependent dimensions characterized by intricate relationships. For instance, congestion at a specific detection node can progressively propagate and influence neighboring nodes over time. This phenomenon is not merely confined to local effects but can exert influence on the broader traffic state, owing to the intricate interplay of traffic space relationships. As a result, achieving an effective alignment and interplay of the time, space, and feature dimensions emerges as a critical imperative in the domain of traffic forecasting.

The conventional fusion of the three underlying dimensions in spatiotemporal traffic prediction overlooks the inherent structural insights ingrained within the data, as highlighted in works such as [14,35]. Typically, this approach flattens the traffic data to facilitate attention operations, thereby disregarding the intricate cross-intersection that exists within each physical dimension of the spatiotemporal traffic data. This technique, essentially confined within a single dimension, does

not fully capture the interplay between dimensions. To address this limitation, we design a structured data manipulation approach within the ST blocks. The cross-information extraction module is primarily constructed upon three efficient cross-attention views. This section proceeds to elaborate extensively on these three modules facilitating information intersection.

#### 3.3.1. The cross-attention view $V_{S\leftrightarrows P}$

The responsiveness of road traffic to accidents is notably influenced by spatiotemporal information. The urgency of addressing incidents varies significantly between remote road sections during nighttime and central road sections during morning hours, leading to diverse emergency treatment pressures. Consequently, urban traffic management places a particular emphasis on intersections and peak hours. Given the strong interdependence within spatiotemporal information, we undertake the task of capturing the cross-information between the spatial dimension $S$ and temporal dimension $T$ through the lens of spatiotemporal modeling. This approach seeks to compensate for the information loss incurred due to the irregular sampling of data obtained via current technological means.

The derivation of the cross-attention view $V_{S\leftrightarrows T}$ involves a three-step process:

- **Feature Dimension Readout** ($R_F$): This step entails the extraction of pivotal information from the input data $X^{(l)}$ along the feature dimension.
- **Mapping Function** ($M$): A mapping function, denoted as $M$, employs convolutional operations to produce the cross-attention view $V_{S\leftrightarrows T}$.
- **Gating Mechanism** ($G$): An additional gating mechanism, referred to as $G$, is introduced to filter the information and generate the final output.

These three successive steps collaborate to generate the desired cross-attention view $V_{S\leftrightarrows T}$, effectively capturing the interdependence between the spatial and temporal dimensions in the data.

Initially, to extract cross-information from the traffic data, it is imperative to tap into the insights embedded within the statistical feature dimension. To achieve this, we employ a readout operation that sifts through the feature dimension, capturing its crucial attributes. Subsequently, this data is mapped onto the spatiotemporal cross-view. The mathematical representation of this process is as follows:

$$\begin{aligned} X_{rf}^{(l)} &= R_F(X^{(l-1)}) \\ &= concate(R_{fmax}(X^{(l-1)}), R_{fmix}(X^{(l-1)}), R_{fmean}(X^{(l-1)})), \end{aligned} \tag{5}$$

where the operation $concat(\cdot)$ signifies the concatenation of tensors across the channels of $X^{(l-1)} \in \mathbb{R}^{F\times P\times S}$. The symbols $R_{fmax}(X^{(l-1)})$, $R_{fmix}(X^{(l-1)})$ and $R_{fmean}(X^{(l-1)})$ correspond to readout operations employed to compute the maximum, minimum, and mean values along the feature dimension $F$, respectively. Eq. (5) is essentially an extension of the pooling operation. Importantly, this formulation retains the spatiotemporal dimensions while effectively extracting critical insights from the feature dimension. This strategic approach enhances the efficiency of the computation involved in calculating the cross-attention view.

Next, the TC-GCN undertakes a depth transformation on the readout feature $X_{rf}^{(l)} \in \mathbb{R}^{3\times Q\times S}$ to make it adaptively structured. To achieve this, a two-dimensional convolution is employed, characterized by a progressive increase and subsequent decrease in values. Within the TC-GCN framework, this operation is referred to as the U-D Conv operation. The mathematical expression capturing this process is as follows:

$$V_{S\leftrightarrows P}^{(l)} = \sigma\left(conv2D_{down}(\sigma(conv2D_{up}(X_{rf}^{(l)})))\right), \tag{6}$$

where $conv2D_{up}$ denotes a two-dimensional convolution operation with a kernel size of $1\times 1$, utilizing input and output channels both of size 3, and incorporating a latent hyperparameter such that $latent > 3$. Meanwhile, $conv2D_{down}$ represents a two-dimensional convolution operation with kernels of size $1\times 1$ and latent input and output channels both of size 1. The symbol $\sigma$ signifies the activation function. The U-D Conv structure closely resembles the U-Net architecture commonly employed in image segmentation, and it is adept at generating the cross-attention view using a limited number of parameters. Subsequent experiments validate its superior performance in comparison to ordinary mapping structures.

Lastly, Eq. (6) illustrates that $V_{S\leftrightarrows T}^{(l)}$ embodies the cross-information between time and space. This cross-information is subsequently employed to filter the output information in subsequent computations. The dynamics of this information filtering procedure are captured by:

$$X_{fp}^{(l)} = conv2D\left(X^{(l-1)}\right) \in \mathbb{R}^{F\times S\times P}, \tag{7}$$

$$X_{ST}^{(l)} = \tanh\left(X_{fp}^{(l)}\right) \odot \mathrm{softmax}\left(V_{S\leftrightarrows T}^{(l)}\right), \tag{8}$$

where the operation of $\mathrm{softmax}$ standardization is applied along the last dimension of $V_{S\leftrightarrows T}^{(l)}$, and the symbol $\odot$ signifies the Hadamard tensor convolution operation [36]. It can be seen from Eq. (8) that $softmax\left(V_{S\leftrightarrows T}^{(l)}\right)$ exhibits similarities to the attention coefficient, but different from the original attention coefficient. The conventional attention coefficient matrix relinquishes the structural information inherent in the original data, transitioning instead into a cross-information attention coefficient that captures interactive attention dynamics.

#### 3.3.2. The cross-attention view $V_{F\leftrightarrows S}$

Similar to $V_{S\leftrightarrows T}$, we can obtain the cross-information between the space and feature dimensions, which represents the relationship strength between the feature and space dimensions in traffic data.

This section provides a detailed description of the cross-information between the spatial and feature dimensions. Initially, we reposition the time axis of $X^{(l)}$ to the channel position, as follows:

$$X_{ot}^{(l)} = Rot_t\left(X^{(l-1)}\right) \in \mathbb{R}^{Q\times F\times S}. \tag{9}$$

Following the rotation operation $Rott$, the procedure for acquiring the cross-information between the feature and space dimensions is fundamentally analogous to that of $VS\leftrightarrows T$. This process similarly entails three steps: $R_P$, $m$, and feature filtering.

Nevertheless, the temporal readout function differs from the feature readout function due to the ordered nature of the feature dimension. If the temporal readout function is similar to Eq. (5), it would lead to the loss of temporal order information. To retain the temporal order information, we devise the temporal readout function $R_P$, wherein the temporal dimension is partitioned into $p$ patches, as follows:

$$X_1^{(l)}, \ldots, X_p^{(l-1)} = split(X_{ot}^{(l)}). \tag{10}$$

Subsequently, the maximum, minimum, and average information is extracted from the $p$ time patches using the following formula:

$$\begin{aligned} X_{rt}^{(l)} &= R_P\left(X_1^{(l)}, \ldots, X_p^{(l-1)}\right) \\ &= concat(P_{tmax}(X_i^{(l-1)}),\ P_{tmix}(X_i^{(l-1)}), P_{tmean}(X_i^{(l-1)}))|_{i=1}^{p} \in \mathbb{R}^{3p\times F\times S}. \end{aligned} \tag{11}$$

Similar to Eq. (6), the operator $m$ ultimately derives a cross-attention view encompassing spatial and feature information, as expressed by:

$$\begin{aligned} V_{F\leftrightarrows S}^{(l)} &= m(X_{rt}^{(l)}) \\ &= \sigma(conv2D_{down}(\sigma(conv2D_{up}(X_{rt}^{(l)})))), \end{aligned} \tag{12}$$

where $conv2D_{up}$ and $conv2D_{down}$ possess the same definitions as in Eq. (6), but they apply to the input and output channels, respectively.

The output of the information filtering process and Eq. (8) for the spatial and feature information are expressed as follows:

$$X_{tp}^{(l)} = conv2D(X^{(l-1)}), \tag{13}$$

$$X_{SF}^{(l)} = \tanh(X_{tp}^{(l)}) \odot sigmoid(V_{S\leftrightarrows F}^{(l)}). \tag{14}$$

#### 3.3.3. The cross-attention view $V_{F \leftrightarrows P}$

As known to all, time and features constitute two pivotal dimensions within traffic data. Consequently, the TC-GCN model must also acquire the cross-information between these dimensions.

The objective of this section is to derive the interaction between the feature and temporal dimensions in spatiotemporal traffic data. Obtaining the feature-temporal interaction relies on the spatial information readout $R_S$. Similar to Eq. (9), we begin by rotating the output $X^{(l)}$ of the ST blocks to reposition the dimension to be read (spatial dimension) as the channel dimension for further processing. In mathematical terms:

$$X_{os}^{(l)} = Rot_t(X^{(l)}) \in \mathbb{R}^{S \times F \times P}. \tag{15}$$

The cross-attention view $V_{F \leftrightarrows P}$ follows the same principle as described in Eqs. (5) and (6), and it is formulated as follows:

$$X_{rs}^{(l)} = R_S(X^{(l-1)}) = concate(R_{smax}(X^{(l-1)}), R_{smix}(X^{(l-1)}), R_{smean}(X^{(l-1)})), \tag{16}$$

$$V_{F \leftrightarrows P}^{(l)} = \sigma(conv2D_{down}(\sigma(conv2D_{up}(X_{rs}^{(l)})))), \tag{17}$$

where the symbol representations are the same as those in Eq. (6).

The input of the final obtained feature-time interlayer is as follows:

$$X_{sp}^{(l)} = conv2D(X^{(l-1)}), \tag{18}$$

$$X_{FQ}^{(l)} = \tanh(X_{sp}^{(l)}) \odot sigmoid(V_{F \leftrightarrows P}^{(l)}). \tag{19}$$

### 3.4. Cross-information fusion

After obtaining the various sets of cross-information $X_{ST}^{(l)}, X_{SF}^{(l)}$ and $X_{FQ}^{(l)}$, we need to fuse them, and to make the fusion process adjustable, the TC-GCN adopts a trainable weighted method for information fusion, and its formula is expressed as follows:

$$X_{cross}^{(l)} = \alpha_{ST}^{(l)} X_{ST}^{(l)} + \alpha_{FS}^{(l)} X_{FS}^{(l)} + \alpha_{FT}^{(l)} X_{FT}^{(l)}, \tag{20}$$

where $\alpha_{ST}^{(l)}, \alpha_{FS}^{(l)}$ and $\alpha_{FT}^{(l)}$ are the trainable parameters of the TC-GCN, which represent the adaptive sum of pairwise interaction results for three dimensions of traffic spatiotemporal data.

### 3.5. Spatial diffusion-based GCNs

Spatial information embedded within traffic data, specifically the interconnections among traffic detection points denoted as $V$, carries significant relevance for spatiotemporal prediction [1]. In this context, this component amalgamates spatial information by leveraging the cross-information $X_{cross}^{(l)}$. To capture spatial information within the TC-GCN framework, the WaveNet's diffusion convolution technique [32] is adopted. This entails the application of the subsequent graph convolution formulation:

$$\begin{aligned} X^{(l)} &= GCN(X_{cross}^{(l)}) \\ &= \sum_{k=0}^{K} A^k X_{cross}^{(l)} W_{k_1} + (A^T)^k X_{cross}^{(l)} W_{k_2} + A_{apt}^k X_{cross}^{(l)}, \end{aligned} \tag{21}$$

where $A$ signifies the distance-weighted adjacency matrix of the traffic network, serving as a pre-established spatial relationship. Its structure adheres to the representation outlined in the literature [2]. Additionally, $A^T$ stands for the transpose of this matrix, and $K$ denotes the order of the GCN. The expression $A_{adp}$ corresponds to the ensuing adaptive spatial relationship:

$$A_{adp} = softmax(ReLU(E_1 E_2)), E_1, E_2^T \in \mathbb{R}^{S \times d}, \tag{22}$$

where $E_1$ and $E_2$ symbolize the trainable parameters, while $d$ represents a hyperparameter. Eq. (22) essentially adaptively obtains the spatial relationships in traffic data through representation learning.

Following several iterations of cross-information and spatial information extraction, the output $X^{(l)}, l \in \{0, \ldots, L-1\}$ from each layer is harmonized. This culminates in the derivation of the final output of the TC-GCN, which can be expressed as:

$$\hat{Y} = f(X^{(0)}, \ldots, X^{(L-1)}) = \sum_{l=0}^{L-1} X^{(l)}. \tag{23}$$

### 3.6. The loss function of the TC-GCN

Since the Huber loss is less sensitive to outliers when contrasted with the squared error loss, the TC-GCN model deliberately adopts the former as its designated loss function. The expressions for Huber's loss functions are outlined as follows:

$$L_{Huber}(\hat{Y}, Y) = \begin{cases} \frac{1}{2}(Y - \hat{Y})^2, & \|(Y - \hat{Y})\| \le \delta, \\ \delta|Y - \hat{Y}| - \frac{1}{2}\delta^2, & \text{otherwise.} \end{cases} \tag{24}$$

## 4. Experiments and results

In this section, we embark on the validation process to ascertain the effectiveness of the proposed TC-GCN model. This validation is carried out through a combination of comparative baseline experiments and detailed component analyses, encompassing diverse methods. The assessment is conducted across two distinct real datasets, thus providing a robust evaluation of the model's performance and capabilities.

In particular, we address the following research questions:

- **RQ1:** How does our proposed TC-GCN method compare in performance to various state-of-the-art methods?
- **RQ2:** How do the fundamental components of the TC-GCN contribute to the network's resilience?
- **RQ3:** We employ the U-D Conv operation as presented in Eqs. (6), (12), and (17). To what extent does its incorporation contribute to effectiveness and performance enhancement?
- **RQ4:** In Eqs. (5), (11) and (16), we employ max, min, and *mean* functions to extract statistical information. To what degree does the utilization of these strategies contribute to the overall effectiveness of the approach?
- **RQ5:** In Eq. (10), what rationale underlies the division of the temporal dimension into $p$ patches?

### 4.1. Datasets

This paper undertakes the validation of the proposed TC-GCN framework using two publicly available transportation network datasets, namely METR-LA and PEMS-BAY, obtained from an open-source code provided in a prior publication [32]. METR-LA comprises traffic speed statistics acquired from 207 sensors located on Los Angeles County highways, covering a span of four months (from Mar. 1 to Mar. 7, 2012). The traffic speed was aggregated every 5 min and the adjacency matrix was calculated by the distance between sensors in the traffic networks. On the other hand, PEMS-BAY encompasses traffic speed data collected over a six-month duration (from January 1, 2017 to May 31, 2017) from 325 sensors across the Bay Area with a sampling interval was 5 min. The 325 × 325 adjacency matrix was built according to the spatial relationship between roads. The dataset is partitioned chronologically, allocating 70% of the samples for training, 10% for validation, and the remaining 20% for testing purposes.

### 4.2. Metrics

In order to conduct an equitable comparison of the aforementioned baselines with the model introduced in this paper, a set of three widely employed traffic forecasting error metrics is employed to assess their performance. These metrics encompass the mean squared error (MSE), mean absolute error (MAE), and root-mean-square error (RMSE). Their respective formulations are defined as follows, as delineated in prior research [31]:

- Mean squared error

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left\|Y_t - \hat{Y}_t\right\|^2, \tag{25}$$

- Mean absolute error

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left\|Y_i - \hat{Y}_i\right\|, \tag{26}$$

- Root-mean-square error

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)}. \tag{27}$$

### 4.3. Experimental setup

The experimental procedures detailed in this paper are executed on a Linux server equipped with a GeForce RTX 2080Ti GPU boasting 12 GB of memory. The dataset is consistently partitioned into training, validation, and test sets with a proportion of 7:3:1. Following 50 training epochs, the optimal model established on the validation set is applied for evaluation on the test set. The adaptive moment estimation (Adam) optimizer is employed, utilizing a learning rate of 0.001. Considering the balance between accuracy and speed, in the initial convolution specified by Eq. (4), the count of output channels is set to 32 ($F = 32$). The parameter *latent* featured in Eqs. (6), (12), and (17) is determined as 32. For Eq. (10), a patch size of 3 is chosen and the other patch size accuracies can be found in Table 4. Lastly, the latent dimensionality $d$ within Eq. (22) is established at 10 because of the simplicity of operations.

### 4.4. Baselines

A comprehensive validation process is conducted on the TC-GCN model, encompassing an array of rigorous assessments. The verification baselines encompass both statistical methodologies and data-driven approaches.

- HA [37]: It refers to the Historical Average model;
- SVR [14]: It denotes the Support Vector Regression model;
- FNN: It is a network that employs matrix decomposition principles for supervised learning. It acquires the embedding layer to capture continuous dense features through this approach;
- GCN [37]: It stands for Graph Convolutional Network, a model adept at leveraging the information intrinsic to the data as well as the interrelations existing among the data points;
- GRU [38]: It corresponds to a Gated Recurrent Unit, constituting a form of recurrent neural network;
- DCRNN [33]: It represents Diffusion Convolutional Recurrent Neural Network, which harnesses a bidirectional random walk mechanism to encapsulate spatial dependencies within a decoder-structured network;
- AGCRN [39]: It denotes Adaptive Graph Convolutional Recurrent Network, characterized by two modules and a recurrent network. This configuration facilitates the automatic capture of spatiotemporal correlations inherent in traffic flow sequences;
- Graph-WaveNet [32]: It encompasses a learning paradigm grounded in adaptive dependency matrices and node embeddings. The model adeptly captures latent spatial dependencies intrinsic to the provided data;
- GMAN [30]: It refers to Graph Multi-Attention Network, a model featuring spatial, temporal, and translation attention layers. This architecture forecasts traffic conditions at diverse time intervals and various locations within the framework of a road network graph.

### 4.5. Experimental results

#### 4.5.1. Forecasting performance comparison: RQ1

Tables 2 and 3 present a comprehensive comparative analysis of various methods across different prediction horizons: 15-minute (horizon 3), 30-minute (horizon 6), and 1-hour (horizon 12) ahead predictions. In these tables, the superior performance is indicated by underlined numbers, while suboptimal performance is marked with numbers accompanied by asterisks.

According to Table 2 and Table 3, the following conclusions can be drawn.

- The TC-GCN consistently achieves the highest accuracy across both datasets in the majority of instances. Notably, it demonstrates a substantial performance superiority over temporal models like DCRNN, GMAN, and AGCRN, which are prominent spatiotemporal deep learning models.
- Upon contrasting the statistical approaches (HA and SVR) with the data-driven methodologies, a discernible trend emerges: data-driven methods exhibit superior performance compared to their statistical counterparts. This observation holds true except for the GCN model, which falls under the data-driven umbrella and primarily caters to spatial dimensions. This trend substantiates the fundamental temporal nature of traffic data, underscoring the challenge of fulfilling the inherent a priori assumptions associated with statistical methods.
- The TC-GCN method put forth in this study closely approximates and consistently outperforms the WaveNet and GMAN methods. This performance advantage is attributed to the TC-GCN's capacity to capture cross-information, thereby enhancing its predictive capabilities.

#### 4.5.2. Effect of each component: RQ2

It can be seen from Fig. 2 that the three cross-attention-based branches (S-T, S-F, and F-T) form the central elements of the TC-GCN architecture. To verify their effectiveness, comprehensive experiments were executed on the METR-LA and PEMS-BAY datasets. The assessment encompasses input and prediction intervals ranging from 15 min (horizon 7) to 1 h (horizon 12), with the metric of interest being the Mean Absolute Error ($MAE$). The ensuing list enumerates the comparative methods for reference.

- w/o ST: In this configuration, the S-T branch depicted in Fig. 2 is excluded from the TC-GCN architecture. As a result, Eq. (20) is adjusted to take the form:

$$X^{(l)}_{cross} = \alpha^{(l)}_{FS} X^{(l)}_{FS} + \alpha^{(l)}_{FT} X^{(l)}_{FT}. \tag{28}$$

- w/o FS: In this configuration, the F-S branch is removed from the TC-GCN architecture. As a result, Eq. (20) is adjusted to take the form:

$$X^{(l)}_{cross} = \alpha^{(l)}_{ST} X^{(l)}_{ST} + \alpha^{(l)}_{FT} X^{(l)}_{FT}. \tag{29}$$

- w/o FT: In this configuration, the F-T branch is omitted from consideration. Consequently, Eq. (20) is adjusted to take the form:

$$X^{(l)}_{cross} = \alpha^{(l)}_{ST} X^{(l)}_{ST} + \alpha^{(l)}_{FS} X^{(l)}_{FS}. \tag{30}$$

The experimental outcomes are visually presented in Fig. 3. The findings conclusively indicate that, in the majority of instances, the accuracy exhibited by the TC-GCN surpasses that of the comparative methods. This substantiates the discernible positive impact of each subbranch in enhancing the model's accuracy.

#### 4.5.3. The effect of the U-D Conv operation: RQ3

The U-D convolution serves as a pivotal process in the cross-attention generation in Eqs. (6), (12) and (17). In order to establish the efficacy of the U-D Conv operation and its consequential impact

**Table 2**
Comparison among the experimental results obtained by the baselines on the METR-LA dataset.

| | 15 min | | | 30 min | | | 1 h | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | MSE | MAE | RMSE | MSE | MAE | RMSE | MSE | MAE | RMSE |
| HA | 75.0750 | 3.5416 | 8.6646 | 118.0912 | 45 365 | 10.8670 | 189.2791 | 6.2397 | 13.7579 |
| SVR | 69.9264 | 3.2971 | 8.3622 | 103.4865 | 4.0887 | 10.1728 | 154.7231 | 5.3561 | 12.4388 |
| FNN | 30.2496 | 3.1718 | 5.4928 | 37.9989 | 3.4167 | 6.1505 | 48.5412 | 3.7574 | 6.9500 |
| GCN | 75.6262 | 5.9177 | 8.6913 | 84.1558 | 6.1783 | 9.1580 | 99.7477 | 6.6400 | 9.9458 |
| GRN | 32.2965 | 2.9487 | 5.6340 | 43.5997 | 3.2976 | 6.5040 | 59.5445 | 3.8622 | 7.5601 |
| DCRNN | 27.1907 | 2.7468 | 5.1444 | 34.5166 | 3.0530 | 5.7685 | 53.3642 | 3.8106 | 7.1273 |
| AGCRN | 22.8343 | 2.5461 | 4.7443 | 30.7892 | 2.8235 | 5.4909 | 40.6558 | 3.1556 | 6.3031 |
| WaveNet | 20.7290* | 2.4506* | 4.5158* | 27.6090* | 2.7127* | 5.1911* | 37.9661* | 3.0443 | 6.0721 |
| GMAN | 21.3438 | 2.4709 | 4.5924 | 29.6966 | 2.7689 | 5.4081 | 42.0717 | 3.1755 | 6.4405 |
| TC-GCN | 20.4548 | 2.4457 | 4.4852 | 27.2302 | 2.6816 | 5.1566 | 37.1123 | 3.0495* | 6.0867* |

**Table 3**
Comparison among the experimental results obtained by the baselines on the PEMS-BAY dataset.

| | 15 min | | | 30 min | | | 1 h | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | MSE | MAE | RMSE | MSE | MAE | RMSE | MSE | MAE | RMSE |
| HA | 9.5010 | 1.3975 | 3.0824 | 18.7981 | 1.8557 | 4.3357 | 36.4378 | 2.5788 | 6.0364 |
| SVR | 7.007 | 1.225 | 2.6471 | 12.9692 | 1.5577 | 4.7611 | 22.6681 | 2.0309 | 4.7611 |
| FNN | 12.5717 | 1.9202 | 3.5405 | 16.3261 | 2.0714 | 4.0326 | 20.5379 | 2.2472 | 4.5272 |
| GCN | 49.4765 | 3.5329 | 7.0337 | 50.9295 | 3.6209 | 7.1356 | 53.8557 | 3.7717 | 7.3351 |
| GRN | 6.6762 | 1.2248 | 2.5052 | 11.9706 | 1.5151 | 3.3147 | 18.3235 | 1.8596 | 4.1208 |
| DCRNN | 5.6616 | 1.1554 | 2.3093 | 10.8403 | 1 5103 | 3.1453 | 22.9179 | 2.2366 | 4.5324 |
| AGCRN | 5.5342 | 1.1453 | 2.2895 | 8.7044 | 1.3459 | 2.8546 | 13.1818 | 1 5827 | 3.5308 |
| WaveNet | 5.0028* | 1.0973* | 2.1781* | 8.5086* | 1.3273* | 2.8162* | 13.3827* | 1 5973* | 3.5446* |
| GMAN | 5.1039 | 1.1140 | 2.2036 | 8.4563 | 1.3228 | 2.8239 | 13.1639 | 1.6048 | 3.5521 |
| TC-GCN | 4.9199 | 1.0916 | 2.159871 | 8.1990 | 1.3025 | 2.7664 | 13.0277 | 1.5688 | 3.4982 |

on the enhanced accuracy of the model, an investigative approach is undertaken. This approach involves the replacement of the U-shaped convolution module with a conventional convolution module, aptly named 'wo uconv'. Subsequently, the $MSE$ results are systematically assessed for horizons ranging from 3 (15 min) to 12 (60 min), leveraging data derived from two distinct datasets. The resulting outcomes are graphically portrayed in Fig. 4.

It can be seen from Fig. 4 that after replacing the U-D Conv module with a common convolution module, the overall accuracy of the TC-GCN on the METR-LA dataset decreases, especially for long sequence prediction horizons (10 to 12). This means that the U-D Conv module can effectively model the hidden features contained in long sequences. On both datasets, the TC-GCN outperforms the variants that remove this module in most cases, which shows that U-D Conv has a certain representation learning ability and can encode data features.

#### 4.5.4. The effect of the readout function: RQ4

To undertake a deeper analysis regarding the impact of the readout function on cross-attention generation, this study delves into the implementation of distinct methodologies, namely, minimum pooling, average pooling, and maximum pooling, on the observed information. The deliberate selection of diverse pooling techniques, when employed in tandem, facilitates the encapsulation of distinctive features across various strata and levels of granularity. The outcomes of these experimental endeavors are visually conveyed in Fig. 5, where it becomes evident that, in the majority of instances, a combination of maximum, minimum, and average pooling yields optimal results. Consequently, the TC-GCN adopts the amalgamated approach of maximum, minimum, and average pooling as its designated readout function, as corroborated by the evidence stemming from the experimental results.

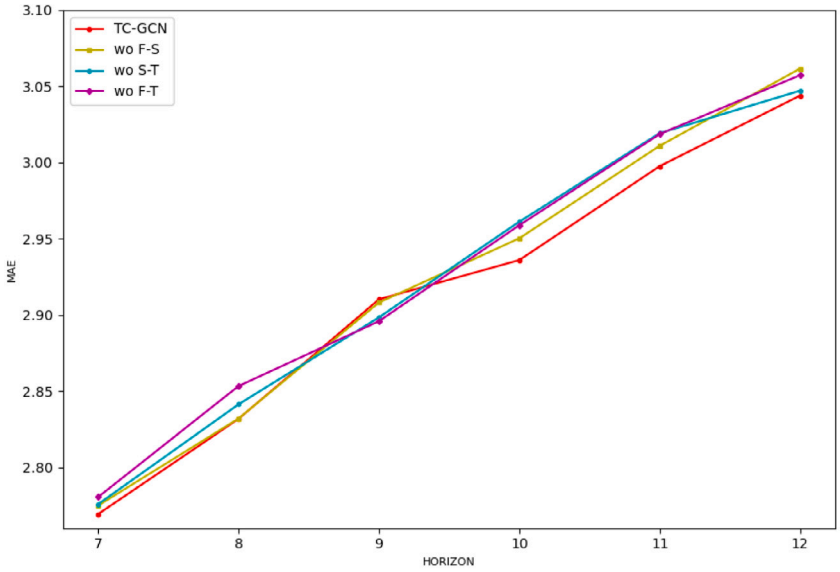
**Table 4**
Effects of different splits on the TC-GCN performance (MSE) in terms of temporal dimension pooling.

| $k$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| META-LA | 37.7868 | 38.0069 | **37.6408** | 37.6410 | 38.0069 |
| PEMS-BAY | 13.6040 | 13.4574 | **13.0277** | 14.1727 | 14.0857 |

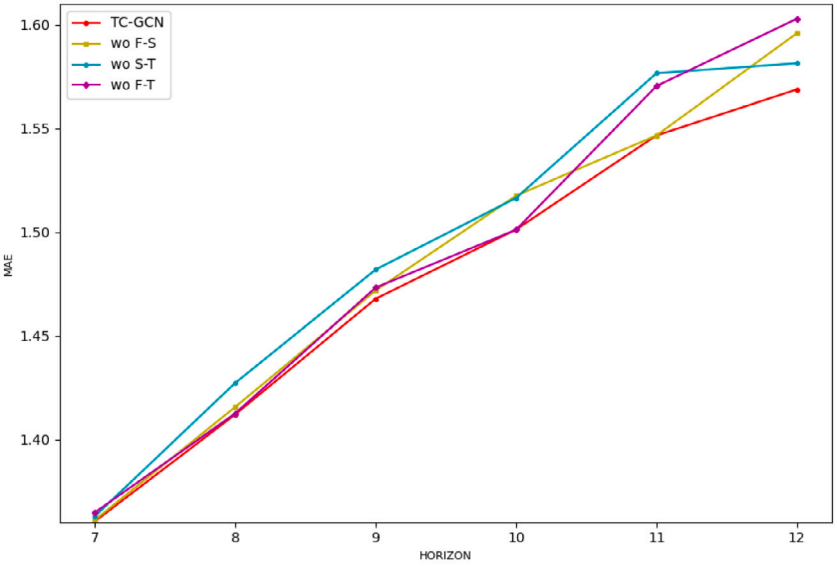
#### 4.5.5. The effect of the $R_P$ function: RQ5

Drawing from the insights garnered in the preceding experiments, it becomes increasingly clear that the operational scope of TC-GCN within a 12-step temporal horizon engenders a substantial amplification in the overall model accuracy. Given this observation, our endeavor aims to provide robust validation for the effectiveness of the $R_P$ module and to gauge the impact of partition count on model accuracy. Thus, we deliberate on the configuration of the partition parameter, adopting values of $k = 1, 2, \ldots, 5$, all the while focusing on the context of the aforementioned 12-step horizon. The outcomes stemming from this endeavor are comprehensively outlined in Table 4.

The experimental results reveal that the model attains its maximum accuracy when $k = 3$. Given the inherent chronological order of time, segmenting the input data facilitates the model's acquisition of long-term temporal dependencies, including periodic variations. In this way, the model achieves higher accuracy. When $k > 3$, the accuracy of the model experiences a diminishing trend. This decline can be attributed to the potential overutilization of data segmentation, leading to a dilution of temporal contextual dependencies. The time-splitting module allows the model to extract data features of varying granularities and hierarchies through the partitioning of time intervals with differing lengths.




(a) METR-LA dataset

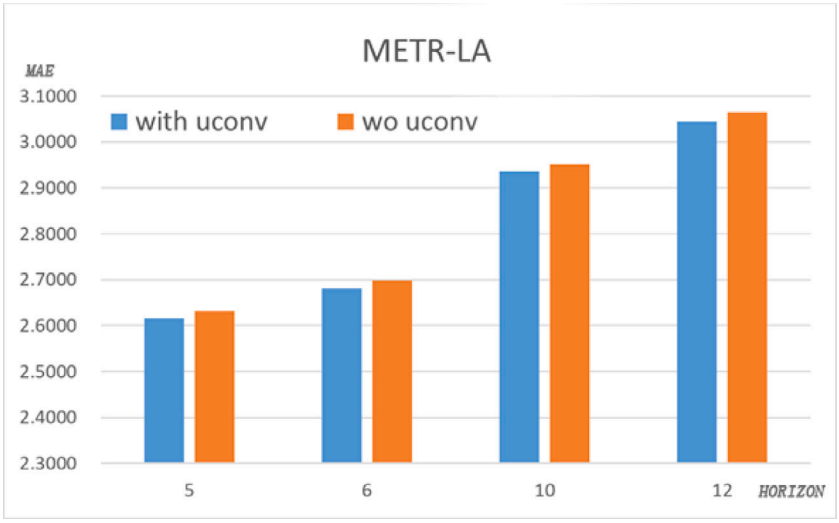



(b) PEMS-BAY dataset

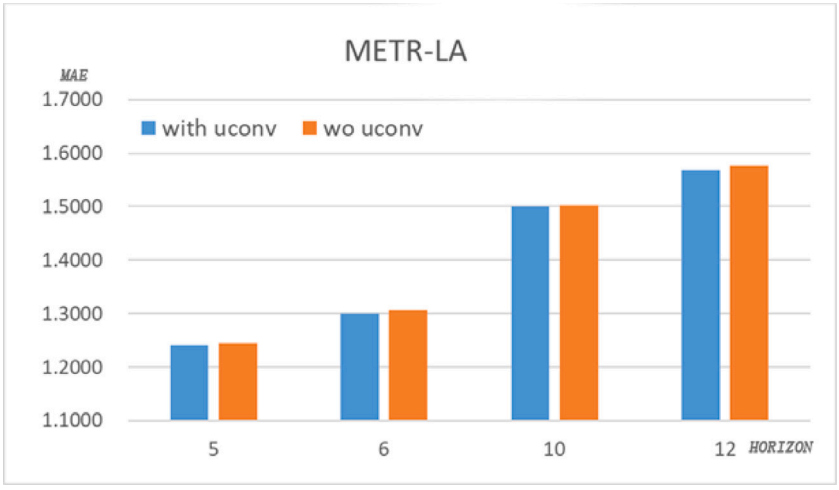
**Fig. 3.** Effect of each component in METR-LA and PEMS-BAY datasets.

## 5. Conclusion

This paper introduces a novel TC-GCN model that employs a three-dimensional modeling framework encompassing time, space, and feature map depth. The model adeptly captures the cross-latitude features and dependencies between these three dimensions, enabling proficient predictions for both short-term and long-term traffic data. A notable facet of this approach is the novel utilization of pooling combinations, effectively extracting information distributed across diverse granularities within the dataset. A pivotal advancement lies in the incorporation of a graph convolution module. This module seamlessly amalgamates temporal hierarchical convolution with spatial self-attention mechanisms. Impressively, this integration preserves spatial structure information while dynamically emulating intricate nonlinear spatial connections among nodes. The experimental results obtained on two real-world datasets show that the performance of the proposed TC-GCN model is superior to that of the mainstream models over both long and short durations. Our future endeavors




(a) METR-LA dataset




(b) PEMS-BAY dataset

**Fig. 4.** The effect of the presence or absence of U-D Conv on the predictive performance achieved by the TC-GCN model on the METR-LA and PEMS-BAY datasets.

will revolve around the application of the model to real-world urban traffic scenarios prevalent in contemporary times. This approach will involve a deliberate incorporation of the distinct characteristics exhibited by modern urban traffic, with the overarching goal of fulfilling the complex requirements associated with accurate traffic flow prediction.

## CRediT authorship contribution statement

**Lei Wang:** Writing – original draft, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition. **Deke Guo:** Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Huaming Wu:** Writing – review & editing, Project administration, Investigation, Funding acquisition, Conceptualization. **Keqiu Li:** Writing – review & editing, Supervision, Conceptualization. **Wei Yu:** Writing – review & editing, Investigation.

## 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

Data will be made available on request.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62071327), Tianjin Intelligent Manufacturing Special Fund Project, China (No. 20211093) and Tianjin Science and Technology Planning Project, China (No. 22ZYQYSN00110 and 22ZYYYJC00020).

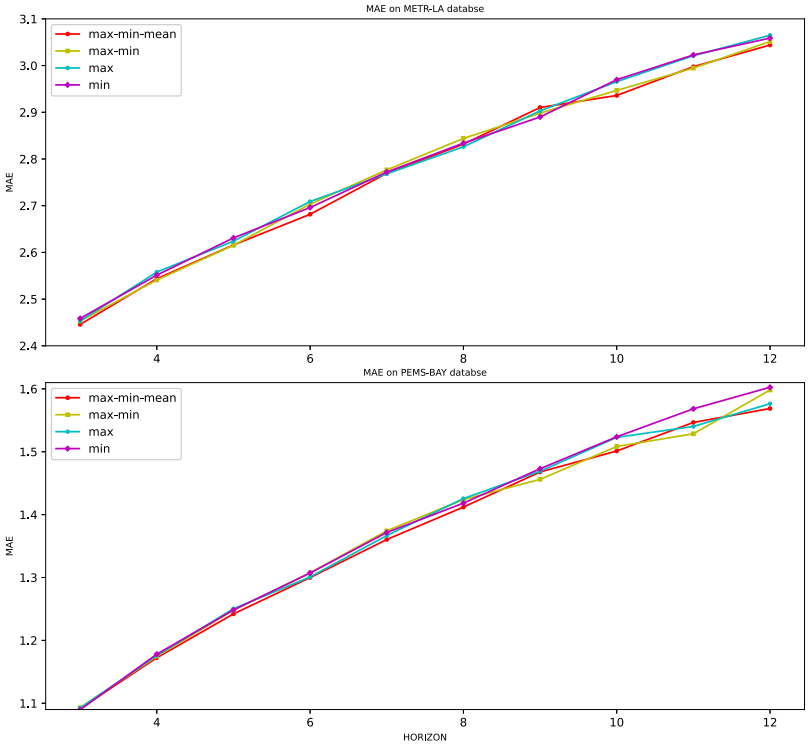



**Fig. 5.** Influences of different combination readout strategies, where "min–max-mean" denotes that the minimum, maximum and mean pooling functions are used; "max–min" signifies that the maximum and minimum functions are used; "max" and "min" represent the maximum and minimum functions are used, respectively.

## References


[1] H. Yuan, G. Li, A survey of traffic prediction: From spatio-temporal data to intelligent transportation, Data Sci. Eng. 6 (1) (2021) 63–85.

[2] L. Bai, L. Yao, C. Li, X. Wang, C. Wang, Adaptive graph convolutional recurrent network for traffic forecasting, in: Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 17804–17815.

[3] H. Huang, Dynamic modeling of urban transportation networks and analysis of its travel behaviors, Chin. J. Manag. 2 (1) (2005) 18–22.

[4] E. Zivot, J. Wang, Vector autoregressive models for multivariate time series, in: Modeling financial time series with S-PLUS®, Springer, 2006, pp. 385–429.

[5] Y. Wang, J. Zhang, H. Zhu, M. Long, J. Wang, P.S. Yu, Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9154–9162.

[6] A.A. Kashyap, S. Raviraj, A. Devarakonda, S.R. Nayak K, S. KV, S.J. Bhat, Traffic flow prediction models–A review of deep learning techniques, Cogent Eng. 9 (1) (2022) 2010510.

[7] J. Wang, W. Wang, W. Yu, X. Liu, K. Jia, X. Li, M. Zhong, Y. Sun, Y. Xu, STHGCN: A spatiotemporal prediction framework based on higher-order graph convolution networks, Knowl.-Based Syst. 258 (2022) 109985.

[8] A. Graves, Long short-term memory, in: Supervised Sequence Labelling with Recurrent Neural Networks, Springer, 2012, pp. 37–45.

[9] G. Atluri, A. Karpatne, V. Kumar, Spatio-temporal data mining: A survey of problems and methods, ACM Comput. Surv. 51 (4) (2018) 1–41.

[10] Y. Li, R. Yu, C. Shahabi, Y. Liu, Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, 2017, arXiv preprint arXiv:1707.01926.

[11] B. Yu, H. Yin, Z. Zhu, Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting, 2017, arXiv preprint arXiv:1709.04875.

[12] M. Xu, W. Dai, C. Liu, X. Gao, W. Lin, G.-J. Qi, H. Xiong, Spatial-temporal transformer networks for traffic flow forecasting, 2020, arXiv preprint arXiv:2001.02908.

[13] D. Piccolo, A distance measure for classifying ARIMA models, J. Time Ser. Anal. 11 (2) (1990) 153–164.

[14] W. Zhang, C. Zhang, F. Tsung, Transformer based spatial-temporal fusion network for metro passenger flow forecasting, in: 2021 IEEE 17th International Conference on Automation Science and Engineering, CASE, IEEE, 2021, pp. 1515–1520.

[15] X. Huang, L. Li, Location method of garden air pollution source based on gradient lifting regression tree algorithm, Int. J. Environ. Technol. Manag. 26 (6) (2023) 445–456.

[16] D. Li, Predicting short-term traffic flow in urban based on multivariate linear regression model, J. Intell. Fuzzy Systems 39 (2) (2020) 1417–1427.

[17] T. Zhang, L. Sun, L. Yao, J. Rong, et al., Impact analysis of land use on traffic congestion using real-time traffic and POI, J. Adv. Transp. 2017 (2017).

[18] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang, Z. Wang, Prediction of urban human mobility using large-scale taxi traces and its applications, Front. Comput. Sci. 6 (2012) 111–121.

[19] J. Zhang, Y. Zheng, D. Qi, Deep spatio-temporal residual networks for citywide crowd flows prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, no. 1, 2017.

[20] V. Shepelev, I. Slobodin, Z. Almetova, D. Nevolin, A. Shvecov, A hybrid traffic forecasting model for urban environments based on convolutional and recurrent neural networks, Transp. Res. Procedia 68 (2023) 441–446.

[21] S. Leelavathy, M. Nithya, R. Dhaya, S. Muthuselvan, R. Kanthavel, K. Rajakumari, Effective traffic model for intelligent traffic monitoring enabled deep RNN algorithm for autonomous vehicles surveillance systems, in: 2023 4th International Conference on Smart Electronics and Communication, ICOSEC, IEEE, 2023, pp. 1191–1200.

[22] N. Wu, B. Green, X. Ben, S. O'Banion, Deep transformer models for time series forecasting: The influenza prevalence case, 2020, arXiv preprint arXiv:2001.08317.

[23] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, Y. Wang, Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction, Sensors 17 (4) (2017) 818.

[24] B. Lim, S.Ö. Arık, N. Loeff, T. Pfister, Temporal fusion transformers for interpretable multi-horizon time series forecasting, Int. J. Forecast. 37 (4) (2021) 1748–1764.

[25] Y. Seo, M. Defferrard, P. Vandergheynst, X. Bresson, Structured sequence modeling with graph convolutional recurrent networks, in: International Conference on Neural Information Processing, Springer, 2018, pp. 362–373.

[26] F. Li, J. Feng, H. Yan, G. Jin, F. Yang, F. Sun, D. Jin, Y. Li, Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution, ACM Trans. Knowl. Discov. Data 17 (1) (2023) 1–21.

[27] K. Lee, W. Rhee, DDP-GCN: Multi-graph convolutional network for spatiotemporal traffic forecasting, Transp. Res. C 134 (2022) 103466.

[28] R. Fu, Z. Zhang, L. Li, Using LSTM and GRU neural network methods for traffic flow prediction, in: 2016 31st Youth Academic Annual Conference of Chinese Association of Automation, YAC, IEEE, 2016, pp. 324–328.

[29] H. Wang, R. Zhang, X. Cheng, L. Yang, Hierarchical traffic flow prediction based on spatial-temporal graph convolutional network, IEEE Trans. Intell. Transp. Syst. 23 (9) (2022) 16137–16147.

[30] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, vol. 29, 2016.

[31] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, H. Li, T-GCN: A temporal graph convolutional network for traffic prediction, IEEE Trans. Intell. Transp. Syst. 21 (9) (2019) 3848–3858.
[32] Z. Wu, S. Pan, G. Long, J. Jiang, C. Zhang, Graph wavenet for deep spatial-temporal graph modeling, 2019, arXiv preprint arXiv:1906.00121.
[33] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.
[34] J. Smith, K.F. Wallis, A simple explanation of the forecast combination puzzle, Oxf. Bull. Econ. Stat. 71 (3) (2009) 331–355.
[35] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, W. Zhang, Informer: Beyond efficient transformer for long sequence time-series forecasting, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, no. 12, 2021, pp. 11106–11115.
[36] J. Wang, W. Wang, X. Liu, W. Yu, X. Li, P. Sun, Traffic prediction based on auto spatiotemporal multi-graph adversarial neural network, Physica A 590 (2022) 126736.
[37] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.
[38] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, 2014, arXiv preprint arXiv:1409.1259.
[39] W. Zhang, T. Du, J. Wang, Deep learning over multi-field categorical data, in: European Conference on Information Retrieval, Springer, 2016, pp. 45–57.