# ISTNet: Inception Spatial Temporal Transformer for Traffic Prediction

Chu Wang, Jia Hu, Ran Tian(✉), Xin Gao, and Zhongyu Ma

College of Computer Science and Engineering, Northwest Normal University, Lanzhou, China
{tianran,2019221875,mazybg}@nwnu.edu.cn

**Abstract.** As a typical problem in spatial-temporal data learning, traffic prediction is one of the most important application fields of machine learning. The task is challenging due to (1) Difficulty in synchronizing modeling long-short term temporal dependence in heterogeneous time series. (2) Only spatial connections are considered and a mass of semantic connections are ignored. (3) Using independent components to capture local and global relationships in temporal and spatial dimensions, resulting in information redundancy. To this end, we propose Inception Spatial Temporal Transformer (ISTNet). First, we design an Inception Temporal Module (ITM) to explicitly graft the advantages of convolution and max-pooling for capturing the local information and attention for capturing global information to Transformer. Second, we consider both spatially local and global semantic information through the Inception Spatial Module (ISM), and handling spatial dependence at different granular levels. Finally, the ITM and ISM brings greater efficiency through a channel splitting mechanism to separate the different components as the local or a global mixer. We evaluate ISTNet on multiple real-world traffic datasets and observe that our proposed method significantly outperforms the state-of-the-art method.

**Keywords:** Traffic prediction · Spatial-Temporal data prediction · attention mechanism

## 1 Introduction

Many countries have recently boosted the construction of smart cities to help realize the intelligent management of cities. Among them, traffic prediction is the most important part of smart city construction, and accurate traffic prediction can help relevant departments to guide vehicles reasonably [4], thereby avoiding traffic congestion and improving highway operation efficiency.

Traffic prediction is the classical spatial-temporal prediction problem, it is challenging due to the complex intra-dependencies (i.e., temporal correlations within one traffic series) and inter-dependencies (i.e., spatial correlations among

C. Wang and J. Hu—Equal contribution.


X. Wang et al. (Eds.): DASFAA 2023, LNCS 13943, pp. 414–430, 2023.
https://doi.org/10.1007/978-3-031-30637-2_27

multitudinous correlated traffic series) [1]. Traditional methods Vector Auto-Regressions (VARs) [12], Auto-Regressive Integrated Moving Average (ARIMA) [7] rely on the assumption of smoothness and fails to capture the complex spatial-temporal patterns in large-scale traffic data. With the rise of deep learning, researchers utilize convolutional neural networks (CNN) [24] and graph convolutional networks (GCN) [6] to capture spatial correlations based on grid structure and non-euclidean structure, respectively. For temporal correlations researchers model the local and global correlations of time series using CNN and recurrent neural networks (RNN) [3], respectively.

To joint spatial-temporal relationship, recent studies formulate the traffic prediction as a spatial temporal graph modeling problem, STGNN [20] and DGCRN [9] integrate GCN into RNN, STSGCN [18] constructs local spatial-temporal graphs by adjacent time steps, and STFGNN [10] constructs adaptive spatial-temporal fusion graphs using dynamic time regularization (DTW). While having shown the effectiveness of introducing the graph structure of data into a model, but there is still a lack of satisfactory progress in long-term traffic prediction, mainly due to the following two challenges. First, the current study is weak in modeling temporal dependence. Researchers typically capture local and global temporal dependence by using CNNs alone or the attention mechanism, with the former requiring stacking multiple layers to capture long sequences, and the latter's preference for global information jeopardizing its ability to capture local correlation. Some studies complement CNNs with attention by serial or parallel approaches, which can lead to information loss and information redundancy. It is a challenging problem to model local and global information synchronously and maintain high computational efficiency.

Second, there is a powerful correlation between traffic conditions on adjacent roads in a traffic network, and we can calculate the distance weights between nodes by a threshold Gaussian kernel function. However, traffic networks are spatially heterogeneous, for example, traffic conditions near schools in different areas are similar, but their distance weights may be zero, and this potential global spatial correlation is important. Some studies construct dynamic graph by feature matrix, but this relationship is not robust because the feature matrix changes with time. Simultaneous modeling of local and global spatial correlations is difficult with guaranteed computational efficiency.

To address the aforementioned challenges, we propose a Inception Spatial Temporal Transformer (ISTNet) to perform traffic prediction task. The core of ISTNet is the ST-Block, which contains Inception Temporal Module and Inception Spatial Module, Inception Temporal Module aims to augment the perception capability of ISTNet in the temporal dimension by capturing both global and local information in the data. To this end, the Inception Temporal Module first splits the input feature along the channel dimension, and then feeds the split components into local mixer and global mixer respectively. Here the local mixer consists of a max-pooling operation and a convolution operation, while the global mixer is implemented by pyramidal attention. Inception Spatial Module and Inception Temporal Module make the same segmentation in

the channel dimension and feed into local mixer (local GCN) and global mixer (global GCN), respectively. In this way, ISTNet can effectively capture local and global information on the corresponding channel, thereby learning more comprehensive characterization. Further, ISTNet stacks a plurality of ST-Blocks with residual connections and generates multi-step prediction results at one time via an attention mechanism. We conducted extensive experiments on four real-world datasets and the experiments demonstrated that ISTNet achieves state-of-the-art performance. In summary, we summarize the contributions of this work as follows:

- We propose the Inception Temporal Module to model local and global temporal correlations, which grafts the merit of CNNs for capturing local information and attention for capturing global information to Transformer.
- We propose the Inception Spatial Module to model local and global spatial correlations, and construct a self-adaptive adjacency matrix that preserves hidden spatial dependencies. Inception Spatial Module can tackle spatial dependencies of nodes' information extracted by Inception Temporal Module at different granular levels.
- We propose the Inception Spatial Temporal Transformer (ISTNet) to perform the traffic prediction task, which stacks multiple ST-Blocks with residual connections and controls the ratio of local and global information by the frequency ramp structure.
- We evaluated ISTNet on four real datasets and the experimental results showed that ISTNet consistently outperformed all baselines.

## 2 Related Work

### 2.1 Traffic Prediction

Traffic prediction is a classical spatial-temporal prediction problem that has been extensively studied in the past decades [22,23]. Compared with statistical methods VAR [12] and ARIMA [7], deep learning methods Recurrent Neural Networks (RNNs) [3], Long-Short-Term-Memory networks (LSTM) [19] break away from the assumption of smoothness and are widely used with the advantage of modeling serial data. Temporal convolutional networks(TCN) [2,15] is also a representative work, it considers a large receptive field through dilated convolutions, and thus enable the model process very long sequence with fewer time. Lately, more and more researchers have started to focus on spatial-temporal modeling studies, ConvLSTM [16] using CNN and RNN to extract local spatial patterns among variables and long-term patterns of time series, respectively. ST-ResNet [24] designed a deep residual network based on CNN for urban pedestrian flow prediction, these works can effectively extract spatial-temporal features, but the limitation is that the input must be standard spatial-temporal grid data, which cannot be applied in non-Euclidean spatial structures.

### 2.2 Graph Neural Network

Traffic data is collected by sensors deployed in traffic networks, and recently researchers have used GNN to model the spatial correlation between different sensors. For instance, DCRNN [11] models traffic flow as a spatially diffusive process and combines diffusion convolution and GRU to model spatial and temporal correlations. STGCN [23] integrates GCN and gated temporal convolution into one module to learn spatial-temporal dependence. Graph WaveNet [22] proposed an adaptive adjacency matrix and spatially fine-grained modeling of the output of the temporal module via GCN, for simultaneously capturing spatial-temporal correlations. STJGCN [25] performs GCN operations between adjacent time steps to capture local spatial-temporal correlations, and further proposes an dilated causal GCN layer to extract information on multiple spatial-temporal scales. Further to enhance the ability of the model to capture spatial-temporal correlations, ASTGCN [4], ASTGNN [5] added a complex attention mechanism as a complement to GCN. In addition, MTGNN [21] and DGCRN [9] proposed graph generation algorithms that fully consider real-world uncertainties and enhance the generalization ability of the models. Although these methods improve the traffic prediction accuracy to a large extent, they do not explicitly define the local and global relationships of the traffic network, which may lead to prediction errors.

## 3 Preliminary

The target of traffic prediction is to predict the traffic flow in the future period based on the previous observations of $N$ sensors in the traffic network. We construct a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ from the positions of the $N$ sensors, where $v$ is a set of $N = |\mathcal{V}|$ vertices, representing the sensors in the road network. $\mathcal{E}$ is the set of edges, $\mathcal{A} \in \mathbb{R}^{N \times N}$ is a weighted adjacency matrix representing the nodes proximity (e.g., road network distance of node pairs). We represent the traffic flow on the graph as a graphic signal $\mathcal{X} \in \mathbb{R}^{N \times C}$ (where $C$ is the number of signals, the signal can be traffic volume, traffic speed, etc.). Furthermore, we express the traffic prediction problem as follows: given a sequence of observations from $N$ sensors at $P$ time steps, predicting the traffic data at $Q$ future time steps by the function $\mathcal{F}$:

$$\{\mathcal{X}_{t-P+1}, \mathcal{X}_{t-P+2}, \cdots, \mathcal{X}_t\} = \mathcal{F}_\theta(\mathcal{X}_{t+1}, \mathcal{X}_{t+2}, \cdots, \mathcal{X}_{t+Q}; \mathcal{G}) \tag{1}$$

## 4 Methodology

Figure 1 shows our proposed ISTNet, which contains L ST-Blocks with residual connections and position encoding, and through a frequency ramp structure to control the ratio of local and global information of different blocks, lastly an attention mechanism generates multi-step prediction results at one time.

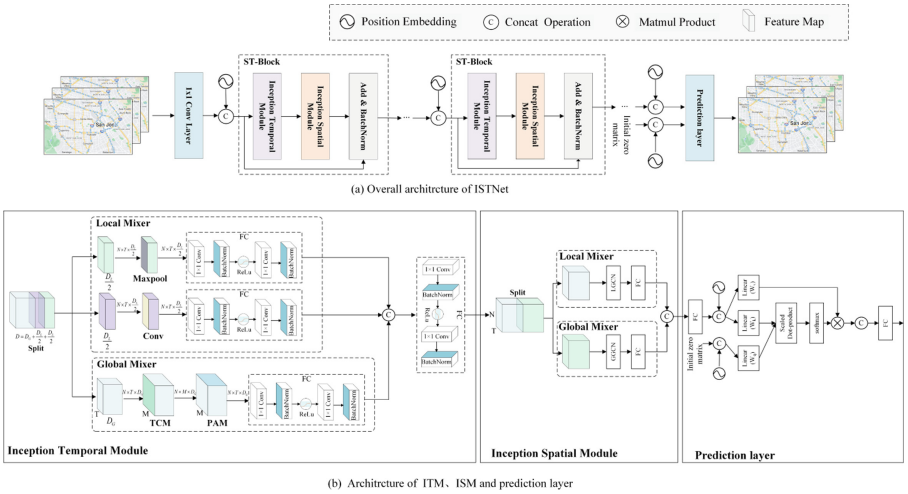



**Fig. 1.** The framework of Inception Spatial Temporal Trasnformer (ISTNet). (a) ISTNet consists of multiple ST-Blocks stacked on top of each other, each ST-Block is composed of inception temporal module and inception spatial module, and to synchronously capture local and global information in temporal or special dimensions. (b) Show details of ST-Block and prediction layer.

### 4.1 Inception Temporal Module

Inspired by the Inception module, we proposed an Inception Temporal Module (ITM) to model long-term and short-term patterns in the temporal dimension. As shown in Fig. 1, ITM first splits the input feature along the channel dimension, and then feeds the split components into local mixer and global mixer respectively. Here the local mixer consists of a max-pooling operation and a convolution operation, while the global mixer is implemented by a pyramidal attention.

Specifically, given the graph signal $\mathcal{X} \in \mathbb{R}^{N \times T \times D}$, it is factorized $\mathcal{X}$ into $\mathcal{X}^{local} \in \mathbb{R}^{N \times T \times D_L}$ and $\mathcal{X}^{global} \in \mathbb{R}^{N \times T \times D_G}$ along the channel dimension, where $D = D_L + D_G$. Then, $\mathcal{X}^{local}$ and $\mathcal{X}^{global}$ are assigned to local mixer and global mixer respectively.

**Local Mixer in Temporal Dimension.** Nodes normally present similar traffic conditions over a certain period of time, so it is necessary for us to focus on local contextual information. Considering the sharp sensitiveness of the maximum filter and the detail perception of convolution operation, we capture short-term temporal patterns in Local Mixer through a parallel structure. We divide the input $\mathcal{X}^{local} \in \mathbb{R}^{N \times T \times D_L}$ into $\mathcal{X}_M \in \mathbb{R}^{N \times T \times D_L/2}$ and $\mathcal{X}_C \in \mathbb{R}^{N \times T \times D_L/2}$ along the channel, then, $\mathcal{X}_M$ is embedded with a max-pooling and a fully-connected layer, and $\mathcal{X}_C$ is fed into a convolution layer and a fully-connected layer:

$$\mathcal{Y}_M^T = FC(MaxPool(\mathcal{X}_M)) \tag{2}$$

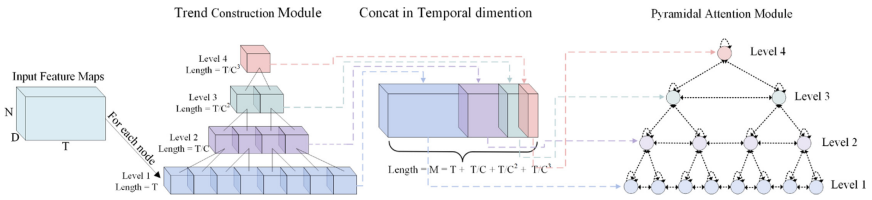



**Fig. 2.** Global Mixer in Temporal Dimension. We first obtain the trend representations of different levels by Trend Construction Module, then concatenate them and use them as initialized node representations of Pyramidal Attention Module.

$$\mathcal{Y}_C^T = FC(Conv(\mathcal{X}_C)) \tag{3}$$

where $\mathcal{Y}_M^T \in \mathbb{R}^{N \times T \times D_L/2}$ and $\mathcal{Y}_C^T \in \mathbb{R}^{N \times T \times D_L/2}$ are the output of local mixer.

**Global Mixer in Temporal Dimension.** There could be long-term effects of sudden traffic conditions at a node in a certain moment, and modeling long-term patterns in the temporal dimension is necessary. Inspired by [13], we propose a temporal attention mechanism with multiresolution structure, as shown in Fig. 2, consisting of two parts, Trend Construction Module (TCM) and Pyramidal Attention Module (PAM). TCM obtains local trend block and global trend block representations by successive convolution operations in the temporal dimension:

$$\begin{aligned} \mathcal{X}_{TCM} = concat(&\mathcal{X}^{global}, \\ &\sigma(\Phi_1 * \mathcal{X}^{global}), \\ &\sigma(\Phi_2 * \sigma(\Phi_1 * \mathcal{X}^{global})), \\ &\sigma(\Phi_3 * \sigma(\Phi_2 * \sigma(\Phi_1 * \mathcal{X}^{global})))) \end{aligned} \tag{4}$$

$\mathcal{X}^{global}$ is the component that is fed into the global mixer, where * denotes a standard convolution operation, $\Phi$ is the parameters of the temporal dimension convolution kernel, and $\sigma(\cdot)$ is the activation function. $\mathcal{X}_G = \sigma(\Phi_3 * \sigma(\Phi_2 * \sigma(\Phi_1 * \mathcal{X}^{global})))$ is a global trend block representation. After obtaining several trend blocks with different levels (the higher the level, the larger the time range represented, e.g. hourly, daily, weekly), we concatenate them as the output of the TCM $\mathcal{X}_{TCM} \in \mathbb{R}^{B \times (T+T/C+T/C^2+T/C^3) \times N \times D_G}$, it has a time length of $M$ and $C$ is the convolution kernel size and step size.

Further, we construct a pyramidal structure that allows the time step to enjoy a long-term horizon by adding a connection between the time step and the trend block. In detail: (1) initializes the node representation of the Pyramidal structure; (2) add connections to the Pyramidal structure with the following strategy: at the same level, nodes are connected to adjacent nodes, and at different levels, connections are added between nodes with their parents nodes and children nodes, with other pairs of nodes that are not connected subjected to the mask operation; (3) information interaction via attention mechanism in Pyramidal. Specifically expressed as:

$$Q^h = mask(\frac{(reshape(\mathcal{X}^{i,h}_{TCM} Z^h_Q))(reshape(\mathcal{X}^{i,h}_{TCM} Z^h_K))}{\sqrt{d_k}}) \tag{5}$$

$$\mathcal{Y}^{i,h} = reshape(softmax(Q^h)(\mathcal{X}^{i,h}_{TCM} Z^h_V)) \tag{6}$$

$$\mathcal{Y}^T_G = \sum_{i=1}^{N} concat(\mathcal{Y}^{i,1}, \cdots, \mathcal{Y}^{i,h}) \tag{7}$$

$\mathcal{X}^{i,h}_{TCM} \in \mathbb{R}^{M \times D_G}$ is the vector representation of node $v_i$ at the $h$th head, $Z^h_Q \in \mathbb{R}^{d_k \times d_k}$, $Z^h_K \in \mathbb{R}^{d_k \times d_k}$ and $Z^h_V \in \mathbb{R}^{d_k \times d_k}$ are learnable parameters of the $h$th header. $Q^h \in \mathbb{R}^{M \times M}$ is the mask matrix of the $h$th header, and $\mathcal{Y}^{i,h}$ is the vector representation of node $v_i$ after the update at the $h$th head. Lastly, we merge the output of $h$ headers and take the previous $T$ time steps as the output of PAM, and the output graph signal is $\mathcal{Y}^T_G \in \mathbb{R}^{N \times T \times D_G}$.

Finally, the outputs of local mixer and global mixer are concatenated along the channel dimension, and the output of the ITM is obtained by a fully-connected:

$$\mathcal{Y}_T = FC(concat(\mathcal{Y}^T_M, \mathcal{Y}^T_C, \mathcal{Y}^T_G)) \tag{8}$$

$\mathcal{Y}_T \in \mathbb{R}^{N \times T \times D}$ is the graph signal output from the ITM.

### 4.2 Inception Spatial Module

As with ITM, we use local mixer and global mixer in Inception Spatial Module to model local and global correlations in spacial synchronously. In the same ST-Block, ISM and ITM have the same channel split ratio. $\mathcal{X}^{local} \in \mathbb{R}^{N \times T \times D_L}$ and $\mathcal{X}^{global} \in \mathbb{R}^{N \times T \times D_G}$ are assigned to local and global mixer respectively.

**Local Mixer in Spatial Dimension.** It is well known that there tends to be a strong correlation between road pairs that are located closer together and a weaker correlation at further distance. This is a local correlation at the spatial level, and to measure the interaction between different roads, we use the thresholded Gaussian kernel function distance to measure the proximity between different pairs of roads:

$$\mathcal{A}_{i,j} = exp(-\frac{dist(v_i, v_j)}{\mu^2}) \tag{9}$$

$$\mathcal{Y}^S_L = FC(\mathcal{A}\mathcal{X}^{local} W_1 + b_1) \tag{10}$$

If $dist(v_i, v_j) \leq \varepsilon$ , $\mathcal{A}_{i,j} = 0$. Where $dist(v_i, v_j)$ represents the road network distance from sensor $v_i$ to $v_j$, $\mu$ is the standard deviation, and $\varepsilon$ is the threshold. $\mathcal{Y}^S_L \in \mathbb{R}^{N \times T \times D_G}$ is the output of local mixer.

**Global Mixer in Spatial Dimension.** In a traffic network, the traffic conditions of two nodes that are distant but have similar attributes (schools, apartments, etc.) are usually similar and can represent each other at a certain level. Distance-based adjacency matrix ignores this due to its local correlation limitation, therefore, we propose an adaptive adjacency matrix to learn this hidden relation. We start by initializing the spatial node embedding representation $E \in \mathbb{R}^{N \times N}$, and then construct the adjacency matrix by the dot product mechanism:

$$\mathcal{M} = softmax(EE^T/\sqrt{D}) \tag{11}$$

$$\begin{aligned} for \;\; & i = 1, 2, \cdots, N \\ & nodeId = argtopk(\mathcal{M}[i, :]) \\ & \mathcal{M}[i, -nodeId] = 0 \end{aligned} \tag{12}$$

$$\mathcal{Y}_G^S = FC(\mathcal{M}\mathcal{X}^{global}W_2 + b_2) \tag{13}$$

$\mathcal{M} \in \mathbb{R}^{N \times N}$ is an adaptive adjacency matrix, which has a set of learnable parameters. We reserve top-k closest nodes as its neighbors for each node by $argtopk(\cdot)$ to reduce the computing cost of GCN. $\mathcal{Y}_G^S \in \mathbb{R}^{N \times T \times D_G}$ is the output of global mixer.

Finally, the outputs of local mixer and global mixer are concatenated along the channel dimension, and get the output of ISM by a fully-connected:

$$\mathcal{Y}_S = FC(concat(\mathcal{Y}_L^S, \mathcal{Y}_G^S)) \tag{14}$$

$\mathcal{Y}_S \in \mathbb{R}^{N \times T \times D}$ is the output of ISM.

### 4.3 Prediction Layer

We direct interaction by adding position embedding to future data and historical data. The position embedding contains two parts, one is the spatial position embedding, and we take a set of learnable spatial embedding representations which share parameters with $E$ in Eq. (11). The other part is temporal embedding, by one-hot encoding of time-of-day, day-of-week and concatenation, and finally summing the two to get the spacial-temporal position embedding $U \in \mathbb{R}^{N \times T \times D}$. $U_H$ and $U_P$ denote the historical and future position embedding, respectively, and $U_H$ is taken as part of the input in each ST-Block.

We utilize attention mechanism to generate multi-step prediction results at one time. It is worth noting that the historical feature $F_H$ is obtained by concatenating the output of the ST-Block with $U_H$, and the future feature $F_P$ is obtained by concatenating the future observations (preset to zero) with $U_P$. We use $F_P$ as the query matrix and $F_H$ as the key and value matrices to obtain the final prediction:

$$R^h = \frac{(reshape(F_P^{i,h}M_Q^h))(reshape(F_H^{i,h}M_K^h))}{\sqrt{d_k}} \tag{15}$$

$$\mathcal{Y}^{i,h} = reshape(softmax(R^h)(F_H^{i,h}M_V^h)) \tag{16}$$

$$\mathcal{Y} = \sum_{i=1}^{N} concat(\mathcal{Y}^{i,1}, \cdots, \mathcal{Y}^{i,h}) \tag{17}$$

where $h$ is the number of attention head, $M_Q^h \in \mathbb{R}^{d_k \times d_k}$, $M_K^h \in \mathbb{R}^{d_k \times d_k}$ and $M_V^h \in \mathbb{R}^{d_k \times d_k}$ are learnable parameters, $R^h \in \mathbb{R}^{M \times M}$ is the temporal attention matrix and $\mathcal{Y}^{i,h}$ is the updated vector representation of node $v_i$, and $\mathcal{Y}$ is the output of the model.

### 4.4 Frequency Ramp Structure

Previous investigations [14] proved that bottom layers prefer local information, while top layers play a more significant role in capturing global information. With deeper layers, lower layers can capture short-term and local patterns in spatial and temporal dimension, and also gradually gather local information to achieve a global understanding of the input. For this reason, we have set up a frequency ramp structure in ISTNet, from the first layer of the ST-Block to the $l$th layer, we progressively decrease the channel dimension of the local mixer and increase the channel dimension of the global mixer. More specifically, for each ST-Block, we define a channel ratio to better balance the local and global components. Hence, with the flexible frequency ramp structure, ISTNet enables effective modeling of local and global spatial-temporal correlation to make precise predictions.

**Table 1.** Dataset description.

| Datasets | Time Range | Time Steps | Time Interval | Nodes |
|---|---|---|---|---|
| PEMS03 | 09/01/2018 - 11/30/2018 | 26202 | 5-min | 358 |
| PEMS04 | 01/01/2018 - 02/28/2018 | 16992 | 5-min | 307 |
| PEMS07 | 05/01/2017 - 08/31/2017 | 28224 | 5-min | 883 |
| PEMS08 | 07/01/2016 - 08/31/2016 | 17856 | 5-min | 170 |

## 5 Experiments

### 5.1 DataSets

We validated the performance of ISTNet on four public traffic datasets. The PEMS03, PEMS04, PEMS07, and PEMS08 datasets published by [1, 10], which are from four districts, respectively in California. The detailed information is shown in Table 1. We used Z-score normalization to standardize the input of data.

On the four traffic flow datasets, we divide them into training set, validation set, and test set according to the ratio of 6:2:2, respectively. We use one hour's historical data (12 steps) to predict the next hour's data.

Our experimental environment is with a 24G memory Nvidia GeForce RTX 3090 GPU. We train the model by Adam optimize and set the initial learning rate to 0.01, batchsize to 128, the maximum epoch is 80, the window size of the max-pooling layer and convolution layer is 3. The default settings for the TCM convolution kernel and step size are [2, 2, 3], frequency ramp structure sets the channel ratios of the local mixer and global mixer for different layers to [2, 1, 1/2].

### 5.2 Baseline Methods

(1) VAR [12] is a traditional time series model capable of capturing the pairwise relationship of time series. (2) SVR [17] utilizes a linear support vector machine to perform regression. (3) FC-LSTM [19] is an encoder-decoder framework using long short-term memory (LSTM) with peephole for multi-step time-series prediction. (4) DCRNN [11] employs an encoder-decoder architecture and combines GRU with a diffusion graph convolutional network to predict traffic speed. (5) STGCN [23] models spatial and temporal correlations using GCN and CNN, respectively. (6) ASTGCN [4] models spatial and temporal correlations via GCN and CNN. (7) Graph WaveNet [22] combines adaptive graph convolution and dilated casual convolution to capture spatial-temporal correlation. (8) STSGCN [18] constructs a local spatio-temporal graph and captures local spatio-temporal correlations by spatio-temporal synchronous graph convolution. (9) GMAN [26] captures spatio-temporal correlations through an attention mechanism and designs a transformation layer to reduce error propagation. (10) DSTAGNN [8] proposed a dynamic spatial-temporal aware graph and a gated convolution capable of capturing multiple ranges.

### 5.3 Experiment Results

**Prediction Performance Comparison.** Table 2 shows the average prediction results of ISTNet and baseline in the next hour. From Table 2 we can observe that: (1) VAR, SVR and LSTM only consider temporal correlation and as a result perform poorly in spatial-temporal data prediction. (2) The graph-based model takes into account spatial information, so the performance is further improved. (3) DCRNN and STGCN are highly dependent on predefined graph structure, CNN-based Graph WaveNet and STSGCN have difficulties in capturing long-term dependence, and attention-based GMAN and DSTAGNN loses local information, and it leads to their poor overall performance. (4) ISTNet achieves state-of-the-art prediction performance on four traffic flow datasets, ISTNet has three advantages compared to Baseline, First, ISTNet captures both local and global pattern in a parallel manner in both spatial and temporal dimensions. Second, ISTNet introduces trend blocks to model long-term temporal dependence. Third, ISTNet achieves optimal performance by controlling the proportion of local and global information at different layers.

**Table 2.** Prediction performance of different models on traffic flow datasets.

| Dataset | Metrics | VAR | SVR | FC-LSTM | DCRNN | STGCN | ASTGCN | Graph WaveNet | STSGCN | GMAN | DSTAGNN | ISTNet |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PEMS03 | MAE | 19.72 | 19.77 | 19.56 | 17.62 | 19.76 | 18.67 | 15.67 | 17.51 | 15.52 | 15.57 | **15.03±0.09** |
| | RMSE | 32.38 | 32.78 | 33.38 | 29.86 | 33.87 | 30.71 | 26.42 | 29.05 | 26.53 | 27.21 | **24.89±0.25** |
| | MAPE(%) | 20.50 | 23.04 | 19.56 | 16.83 | 17.33 | 19.85 | 15.72 | 16.92 | 15.19 | **14.68** | 15.24±0.19 |
| PEMS04 | MAE | 24.44 | 26.18 | 23.60 | 24.42 | 23.90 | 22.90 | 19.91 | 21.52 | 19.25 | 19.30 | **18.51±0.03** |
| | RMSE | 37.76 | 38.91 | 37.11 | 37.48 | 36.43 | 33.59 | 31.06 | 34.14 | 30.85 | 31.46 | **30.36±0.10** |
| | MAPE(%) | 17.27 | 22.84 | 16.17 | 16.86 | 13.67 | 16.75 | 13.62 | 14.50 | 13.00 | 12.70 | **12.36±0.16** |
| PEMS07 | MAE | 27.96 | 28.45 | 34.05 | 24.45 | 26.22 | 28.13 | 20.83 | 23.99 | 20.68 | 21.42 | **19.67±0.13** |
| | RMSE | 41.31 | 42.67 | 55.70 | 37.61 | 39.18 | 43.67 | 33.62 | 39.32 | 33.56 | 34.51 | **32.96±0.04** |
| | MAPE(%) | 12.11 | 14.00 | 15.31 | 10.67 | 10.74 | 13.31 | 9.10 | 10.10 | 9.31 | 9.01 | **8.57±0.20** |
| PEMS08 | MAE | 19.83 | 20.92 | 21.18 | 18.49 | 18.79 | 18.72 | 15.57 | 17.88 | 14.87 | 15.67 | **14.08±0.05** |
| | RMSE | 29.24 | 31.23 | 31.88 | 27.30 | 28.2 | 28.99 | 24.32 | 27.36 | 24.06 | 24.77 | **23.27±0.12** |
| | MAPE(%) | 13.08 | 14.24 | 13.72 | 11.69 | 10.55 | 12.53 | 10.32 | 11.71 | 9.77 | 9.94 | **9.34±0.09** |

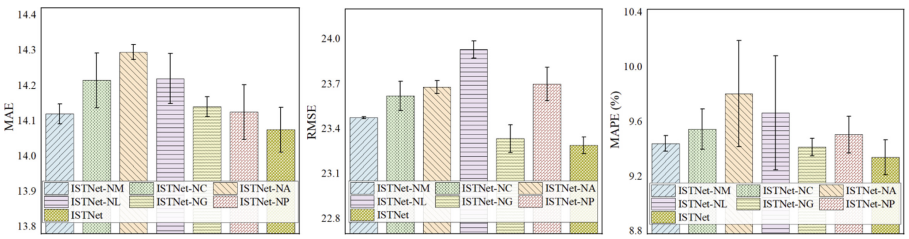



**Fig. 3.** The results of ablation study on the PEMS08 dataset.

**Ablation Study.** To further evaluate the effectiveness of each component in ISTNet, we conduct ablation studies on PEMS08 dataset. We named the variants of ISTNet as follows: (1) ISTNet-NM: We removed the maximum pooling layer from the ITM. (2) ISTNet-NC: We eliminate the convolutional layer from the ITM. (3) ISTNet-NA: We utilize a common temporal attention mechanism instead of PAM. (4) ISTNet-NL: We removed the local mixer from the ISM. (5) ISTNet-NG: We eliminate the global mixer from ISM. (6) ISTNet-NP: The prediction layer is replaced by a common multi-headed attention mechanism to investigate the impact of the prediction layer on model performance.

The experimental results are shown in Fig. 3. We observed that: (1) ISTNet-NM and ISTNet-NC show that removing the max-pooling layer and CNN layer in ITM degrades the performance of ISTNet due to their ability to efficiently extract local information from different perspectives. (2) ISTNet-NA demonstrates that PAM is critical to model performance because it can effectively model long-short-term temporal dependence and cope with temporal heterogeneity. (3) Distance-based graph and adaptive graph are highly effective, but for traffic prediction, the local information of the road is more significant than the global information. (4) The effectiveness of the prediction layer has been shown by the effect of ISTNet-NP, where we perform a direct interaction between future and historical observations, and this direct generation of multi-step prediction results facilitates the reduction of error propagation.

**Long Term Prediction Performance.** For evaluating the performance of ISTNet in long-term prediction, we predict the traffic data for the next 30, 60, 90, and 120 min on the PEMS04 dataset. The results are shown in Table 3, as the time step increases, ISTNet continues to outperform the most advanced baseline in the long-term, and the gap widens gradually, suggesting the effectiveness of ISTNet in long-term temporal modeling. Figure 4 shows the specific performance of the per prediction step.

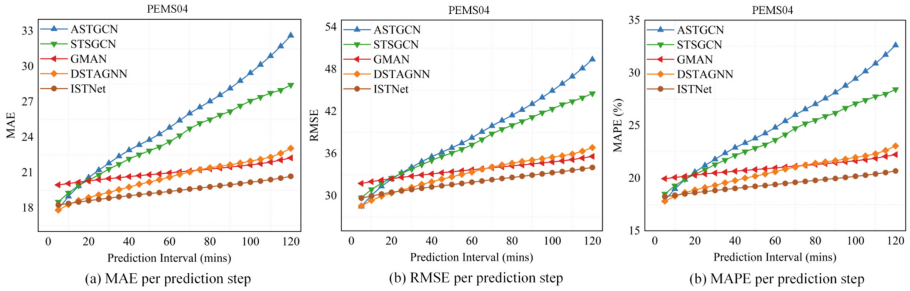



**Fig. 4.** Prediction performance comparison at each horizon on the PEMS04.

**Table 3.** Long-term prediction performance comparison of different models on PEMS04 dataset.

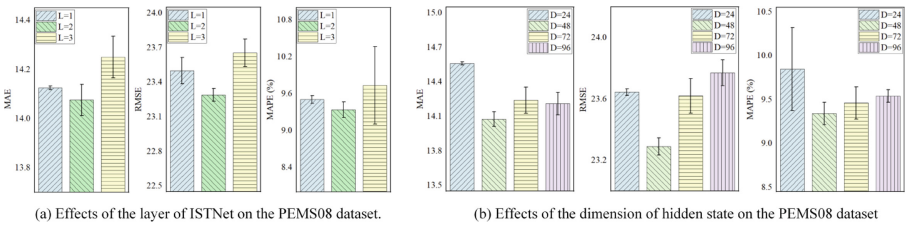
| Model | Metrics | 30 min | 60 min | 90 min | 120 min | Average |
|---|---|---|---|---|---|---|
| ASTGCN | MAE | 22.08±0.28 | 25.51±0.69 | 29.32±1.17 | 34.04±1.42 | 26.01±0.75 |
| | RMSE | 34.47±0.42 | 39.35±1.10 | 44.95±1.87 | 51.60±2.20 | 40.64±1.28 |
| | MAPE (%) | 14.70±0.10 | 16.84±0.19 | 19.28±0.28 | 22.49±0.31 | 17.22±0.19 |
| STSGCN | MAE | 21.66±0.36 | 24.04±0.41 | 26.70±0.52 | 29.07±0.64 | 24.35±0.47 |
| | RMSE | 34.56±0.75 | 37.98±0.72 | 41.91±0.75 | 45.45±0.90 | 38.46±0.79 |
| | MAPE (%) | 14.44±0.13 | 15.76±0.11 | 17.50±0.18 | 18.92±0.15 | 16.13±0.20 |
| GMAN | MAE | 20.50±0.01 | 21.02±0.04 | 21.55±0.08 | 22.29±0.05 | 21.08±0.05 |
| | RMSE | 33.21±0.42 | 34.18±0.48 | 35.09±0.56 | 36.13±0.54 | 34.24±0.49 |
| | MAPE (%) | 15.06±0.52 | 15.37±0.57 | 15.78±0.66 | 16.54±0.76 | 15.48±0.60 |
| DSTAGNN | MAE | 19.36±0.04 | 20.69±0.08 | 21.69±0.03 | 22.91±0.15 | 20.60±0.02 |
| | RMSE | 31.36±0.17 | 33.65±0.27 | 35.29±0.22 | 36.81±0.04 | 33.47±0.15 |
| | MAPE (%) | 12.88±0.02 | 13.54±0.03 | 14.22±0.01 | 15.04±0.05 | 13.58±0.02 |
| ISTNet | MAE | **18.93±0.10** | **19.51±0.11** | **20.11±0.15** | **20.85±0.17** | **19.58±0.13** |
| | RMSE | **30.96±0.15** | **32.01±0.08** | **32.97±0.03** | **34.00±0.08** | **32.05±0.06** |
| | MAPE (%) | **12.85±0.18** | **13.18±0.13** | **3.57±0.10** | **14.25±0.12** | **13.26±0.14** |

**Computation Cost.** We present the computation cost of ASTGCN, STSGCN, GMAN, DSTAGNN and ISTNet on the PEMS04 dataset in Table 4. We set the input and prediction step size to 24, batch-size to 16, and the rest of the parameters follow the best configuration in the text, and the five methods are tested uniformly in the Tesla V100 GPU environment.

**Table 4.** The computation cost on the PEMS04 dataset.

| Method | Computation Time | | Memory | |
|---|---|---|---|---|
| | Training (s/epoch) | Inference (s) | Parameter (M) | GPU Memory (G) |
| ASTGCN | 72.92 | 19.83 | 0.48 | 3.62 |
| STSGCN | 302.09 | 60.86 | 5.98 | 8.58 |
| GMAN | 574.2 | 30.6 | 0.51 | 28.67 |
| DSTAGNN | 379.22 | 67.19 | 0.44 | 9.78 |
| **ISTNet** | 157.9 | 7.4 | 0.13 | 5.85 |

From Table 4 we observe that in terms of computation time, based on the spatial-temporal attention mechanism GMAN and DSTAGNN run the slowest in the training stage, ASTGCN with the removal of the cycle component is the fastest, and ISTNet has a moderate training speed but the fastest inference speed. In respect of memory consumption, ISTNet has the smallest number of parameters, which can avoid overfitting to a certain degree, and the maximum GPU usage of ISTNet is only higher than ASTGCN, by further considering the prediction accuracy (refer to Table 1 and Table 2), ISTNet shows superior ability in balancing predictive performances and time consumption as well as memory consumption.




(a) Effects of the layer of ISTNet on the PEMS08 dataset.

(b) Effects of the dimension of hidden state on the PEMS08 dataset

**Fig. 5.** Parameter study.

**Parameter Study.** To further investigate the effect of hyperparameter settings on model performance, we conduct the study on the number of layers $L$, vector dimension $D$, and kernel size of TCM for ISTNet on the PEMS08 dataset. Except for the changed parameters, the other configurations remain the same.

We can observe from Fig. 5 that increasing the number of layers and the number of hidden units of ISTNet can improve the performance of the model, which is because increasing the number of layers expands the perceptual field of the nodes and the high-dimensional vectors can more fully express the hidden information. However, when the threshold is exceeded, the model performance gradually decreases, which means that an overfitting phenomenon occurs.

In addition, we have also investigated the convolutional kernel size of TCM. Here, we can choose to stack two or three CNN layers to build the global trend block. From Table 5 we observe that the model performs best when stacking three CNN layers and the kernel size is [2,2,3], which illustrates that a reasonable layer setting enables the representation range to be precise, and thus more conducive to modeling long-term dependence.

**Table 5.** Impact of kernel size in TCM on model performance.

| Metric | TCM kernel size | | | | | | |
|---|---|---|---|---|---|---|---|
| | [2, 6] | [3, 4] | [4, 3] | [6, 2] | [2, 3, 2] | [3, 2, 2] | [2, 2, 3] |
| MAE | 14.13±0.05 | 14.13±0.06 | 14.18±0.05 | 14.15±0.13 | 14.12±0.07 | 14.08±0.08 | **14.08±0.05** |
| RMSE | 23.63±0.13 | 23.40±0.06 | 23.55±0.08 | 23.49±0.10 | 23.54±0.10 | 23.55±0.12 | **23.37±0.12** |
| MAPE (%) | 9.48±0.11 | 9.48±0.13 | 9.58±0.14 | 9.41±0.09 | 9.53±0.11 | **9.33±0.01** | 9.34±0.09 |

**Visualization of PAM.** We randomly visualized pyramidal attention scores of three nodes on the PEMS04 dataset. In Fig. 6 we can see that: node 10 pays more attention to the features of neighboring time steps, node 51 and node 89 give more attention to local trend blocks, which indicates that most nodes require more trend features, it can contribute to modeling long-term dependence.

**Visualization of Prediction Results.** We visualized the traffic flow of the node for the next day and compared it with the predicted value. As shown in Fig. 7, where the boxes are highlighted, we observe that ISTNet's prediction curves coincide with the true values, and the performance at the peaks demonstrates ISTNet's ability to make accurate predictions for challenging situations, which further illustrates the effectiveness of ISTNet in modeling traffic data.

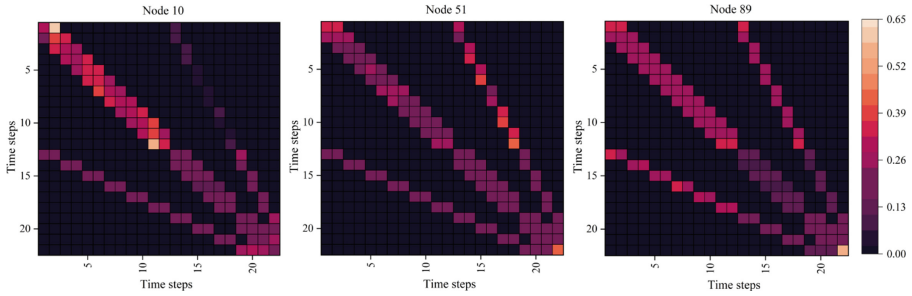



**Fig. 6.** Heatmap of pyramidal attention scores. The corresponding TCM kernel size is [2, 2, 3], and the total time step is 22.

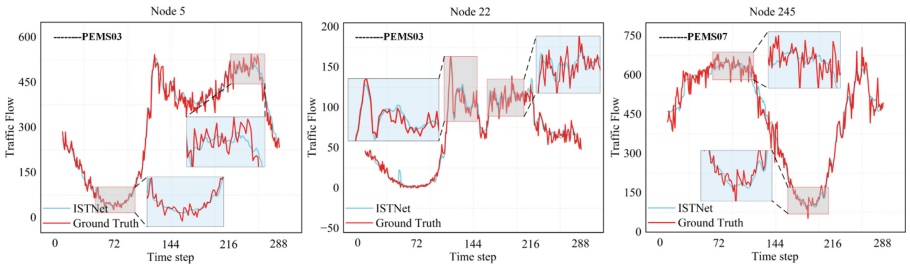



**Fig. 7.** Traffic prediction visualization on the PEMS03 and PEMS07 datasets.

## 6 Conclusion

In this paper, we propose a Inception Spatial Temporal Trasnformer (ISTNet) to perform the traffic prediction task. The core components of ISTNet include ITM and ISM, and ITM captures local and global correlations in temporal dimension with a parallel approach. In particular, we enhance the global sensing capability of the model by introducing trend blocks, which further improves the long-term prediction performance. ISM captures local and global correlations in traffic networks through distance-based graph and global dynamic graph. In addition, ISTNet controls the ratio of local and global information of ITM and ISM in different ST-Blocks by means of frequency ramp structure to achieve the best performance. Experiments on several traffic datasets show that ISTNet achieves the optimal performance, especially in long-term prediction, and ISTNet's performance is further improved. In the future, we will explore the application of ISTNet in the field of meteorological prediction and air quality prediction.

**Acknowledgment.** This work is supported in part by the National Natural Science Foundation of China (71961028), the Key Research and Development Program of Gansu (22YF7GA171), the Scientific Research Project of the Lanzhou Science and Technology Program (2018-01-58).

## References

1. Bai, L., Yao, L., Kanhere, S., Wang, X., Sheng, Q., et al.: STG2Seq: spatial-temporal graph to sequence model for multi-step passenger demand forecasting. arXiv preprint arXiv:1905.10069 (2019)
2. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
3. Connor, J.T., Martin, R.D., Atlas, L.E.: Recurrent neural networks and robust time series prediction. IEEE Trans. Neural Netw. **5**(2), 240–254 (1994)
4. Guo, S., Lin, Y., Feng, N., Song, C., Wan, H.: Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 922–929 (2019)

5. Guo, S., Lin, Y., Wan, H., Li, X., Cong, G.: Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. IEEE Trans. Knowl. Data Eng. **34**(11), 5415–5428 (2021)
6. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
7. Kumar, S.V., Vanajakshi, L.: Short-term traffic flow prediction using seasonal ARIMA model with limited input data. Eur. Transp. Res. Rev. **7**(3), 1–9 (2015)
8. Lan, S., Ma, Y., Huang, W., Wang, W., Yang, H., Li, P.: DSTAGNN: dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In: International Conference on Machine Learning, pp. 11906–11917. PMLR (2022)
9. Li, F., et al.: Dynamic graph convolutional recurrent network for traffic prediction: benchmark and solution. ACM Trans. Knowl. Discov. Data (TKDD) **17**, 1–21 (2021)
10. Li, M., Zhu, Z.: Spatial-temporal fusion graph neural networks for traffic flow forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 4189–4196 (2021)
11. Li, Y., Yu, R., Shahabi, C., Liu, Y.: Diffusion convolutional recurrent neural network: data-driven traffic forecasting. arXiv preprint arXiv:1707.01926 (2017)
12. Lippi, M., Bertini, M., Frasconi, P.: Short-term traffic flow forecasting: an experimental comparison of time-series analysis and supervised learning. IEEE Trans. Intell. Transp. Syst. **14**(2), 871–882 (2013)
13. Liu, S., et al.: Pyraformer: low-complexity pyramidal attention for long-range time series modeling and forecasting. In: International Conference on Learning Representations (2021)
14. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? In: Advances in Neural Information Processing Systems, vol. 34, pp. 12116–12128 (2021)
15. Sen, R., Yu, H.F., Dhillon, I.S.: Think globally, act locally: a deep neural network approach to high-dimensional time series forecasting. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
16. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
17. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Stat. Comput. **14**, 199–222 (2004)
18. Song, C., Lin, Y., Guo, S., Wan, H.: Spatial-temporal synchronous graph convolutional networks: a new framework for spatial-temporal network data forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 914–921 (2020)
19. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
20. Wang, X., et al.: Traffic flow prediction via spatial temporal graph neural network. In: Proceedings of the Web Conference 2020, pp. 1082–1092 (2020)
21. Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., Zhang, C.: Connecting the dots: multivariate time series forecasting with graph neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 753–763 (2020)
22. Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C.: Graph wavenet for deep spatial-temporal graph modeling. arXiv preprint arXiv:1906.00121 (2019)
23. Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875 (2017)

24. Zhang, J., Zheng, Y., Qi, D.: Deep spatio-temporal residual networks for citywide crowd flows prediction. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
25. Zheng, C., Fan, X., Pan, S., Wu, Z., Wang, C., Yu, P.S.: Spatio-temporal joint graph convolutional networks for traffic forecasting. arXiv preprint arXiv:2111.13684 (2021)
26. Zheng, C., Fan, X., Wang, C., Qi, J.: GMAN: a graph multi-attention network for traffic prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 1234–1241 (2020)