# Traffic flow prediction model based on multi-scale pyramid space-time network

Yanlie Zheng[1,†], Hu Jiwei[2,†] , Xueying Li[1,*,†] Qingxia Shen[1,†] and Chunzhi Wang[3,†]

[1] *Fiberhome Telecommunication Technologies Co.,LTD, Wuhan 430068, China*
[2] *Wuhan Fiberhome Technical Services Co., Ltd., Wuhan 430205, China;*
[3] *Hubei University of Technology, Wuhan 430068, China*

**Abstract**

Traffic flow prediction is both one of the important components of intelligent transport systems (ITS) and a challenging task at the same time. Although the existing traffic flow prediction has achieved good results, the existing traffic flow prediction models only model dynamic spatio-temporal correlations on a single time or spatial scale, and have poor performance on long-distance prediction. Aiming at the above problems, this paper proposes a traffic flow prediction model based on multi-scale pyramid spatio-temporal network. Specifically, firstly, a local spatio-temporal grid is generated by combining traffic data features and adjacency matrix. Secondly, multiple convolutional layers are used to aggregate sequences with multiple resolutions, and at the same time, the spatio-temporal grids are merged into traffic event sequences based on the temporal and spatial dimensions. Next, the adaptive combination of pyramidal attention and multi-channel spatio-temporal convolution module is used to capture the spatio-temporal dependence of sequence dynamics and the global spatio-temporal features are obtained by optimal fusion using fully connected layers. Finally, the corresponding predicted values are output based on the global spatio-temporal features. Experimental results on two publicly available datasets show that the model largely improves the detection.

**Keywords**

Grid, Spatio-temporal,Attention, Channel

## 1. Introduction

In recent years, the sensor technology is developing rapidly, and the travel modes are rich and diverse, and the intelligent transportation system has become the key development object in many countries. In response to the problem of huge urban traffic flow and high speed area, traffic prediction has become a key research in intelligent transportation system, and traffic flow prediction methods are used to learn the highly nonlinear characteristics of traffic flow data in order to accurately predict the traffic flow of complex urban roads in the coming period. Stable and reliable traffic flow prediction algorithms can effectively alleviate traffic congestion and improve people's quality of life.

[1] *BAIT'2024: The 1st International Workshop on "Bioinformatics and applied information technologies", October 02-04, 2024, Zboriv, Ukraine*
* Corresponding author.
† These authors contributed equally.
zhengyanlie@ fiberhome.com (Y. Zheng); hujiwei@fiberhome.com (J. Hu); xli@ fiberhome.com (X. Li); qshen@ fiberhome.com (Q. Shen); chunzhiwang@hbut.edu.cn (C. Wang)
0000-0003-3412-1639 (Y. Zheng); 0000-0002-9622-744X (J. Hu); 0009-0008-2733-3565 (X. Li); 0009-0000-1600-4717 (Q. Shen); 0000-0002-6742-6644 (C. Wang)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3842/paper20.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Traffic data contains important transportation system features, such as traffic flow, speed, and time information recorded by road network sensors. Traditional statistical and machine learning methods [1,2] are used to analyze the complex spatio-temporal properties of these traffic data and thus predict traffic flow. However, these methods perform poorly in capturing high-dimensional spatio-temporal features. In recent years, researchers in the field of transportation have turned their attention to deep neural networks. Researchers have employed CNNs to capture spatial features of target nodes and nearby regional nodes [3].Li et al [4] modeled the diffusion process of directed graphs and combined diffusion convolution and GRU recurrent neural networks to fuse the temporal and spatial correlations of the traffic flow. Tang et al [5] proposed the spatio-temporal latent graph structure learning network STLGSL, which employs a multilayer perceptron and k-nearest neighbors to generate graph structure, and utilizes diffusion graph convolution and dilation causal convolution as well as gating mechanism to mine the spatio-temporal features of the generated graphs. Li et al [6] proposed Transformer-enhanced DetectorNet, which utilizes a multi-view temporal attention module to capture temporal correlation of distance and proximity, and combines graph convolution and a dynamic attention module to aggregate the spatial features of the generated dynamic graphs.

The aforementioned studies on traffic flow prediction have achieved impressive results. However, existing work prefers to capture the pairwise impacts of spatio-temporal traffic events and the spatio-temporal features of traffic data from a single temporal and spatial scope. However, this approach makes it difficult to learn the dependencies of distant locations in time and space, and does not comprehensively capture spatio-temporal dependencies at different scales.

In order to solve the above two problems, this paper proposes a new traffic flow prediction model based on multi-scale pyramidal hybrid spatio-temporal network, called MSLST. First, the original traffic data features (time and speed) and sensor distances are preprocessed, and a local spatio-temporal grid is constructed by selecting other nodes that are spatially correlated with the target node based on the generated adjacency matrices. Second, the temporal and spatial dimensions of the spatio-temporal grid are combined into a sequence of traffic events, and the sequence is processed using multiple convolutional layers with different Stride to obtain the feature information of the sequence at different resolutions. And stacking multiple pyramid attention is used to simulate the pairwise effects of traffic events under different spatio-temporal scales to obtain coarse and fine scale based spatio-temporal correlation features. Next, the spatio-temporal features at different scales are transformed into different channels using linear layers, and spatio-temporal convolution blocks are used for each layer separately to capture the spatio-temporal information of other nodes in the region near the target node in the local spatio-temporal space. Finally, the spatio-temporal features of each layer are merged into one channel, and the fully connected layer is used to transform the merged features to obtain the global spatio-temporal features. And finally, traffic flow prediction is performed based on the above output global spatio-temporal features.

To summarize the main contributions:

1. In this paper, the temporal and spatial dimensions of traffic data are merged into a single fluid, and pyramid attention is used to directly model the dynamic spatio-temporal associations between a target node in local spatio-temporal space and other

nodes at different moments, and to convey the spatio-temporal information of the nodes in different spatio-temporal ranges, which improves the model's ability to capture the highly nonlinear spatio-temporal features of the traffic flow.

2. In this paper, a multi-channel spatio-temporal convolutional block is proposed to perform gated aggregation of spatio-temporal information at various scales of the pyramid, flexibly mining the compact spatio-temporal feature representations of proximity and remoteness in the local spatio-temporal context, establishing one-to-many relationships between the target node and the other nodes in both time and space, and greatly improving the performance of long-distance multistep prediction.
3. MSLST is evaluated on two real-world public datasets to validate the effectiveness and sophistication of the model.

## 2. Model

This section describes a multi-scale local spatio-temporal network (MSLST) for modeling spatial and temporal information of traffic flow. As shown in Fig. 1, the MSLST is composed of three parts, which are local spatio-temporal grid construction, multi-scale spatio-temporal attention block, and multi-channel spatio-temporal convolution block. The local spatio-temporal mesh construction is responsible for stitching the spatio-temporal intervals between traffic events from target sensors and other sensors into a 3D mesh, the multi-scale spatio-temporal attention block is responsible for capturing the spatio-temporal correlation of traffic events of a single flow shape under multiple ranges of spatio-temporal contexts, and the multi-channel spatio-temporal convolution block is responsible for aggregating temporal and spatial information under multiple channels to obtain global spatio-temporal features. Finally, a dense fully connected layer is used to predict future traffic flow.

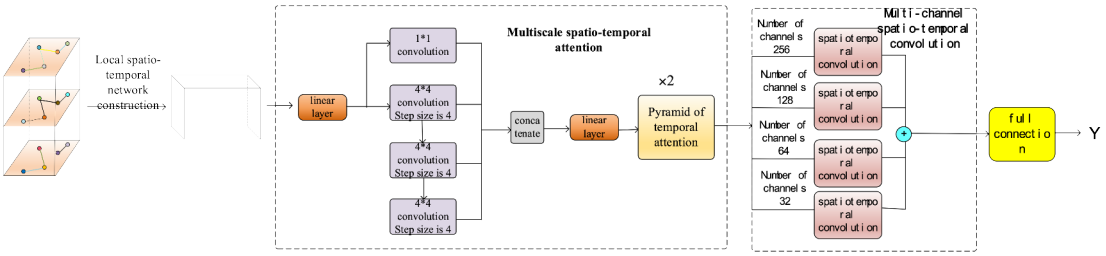


**Figure 1:** MSLST architecture diagram.

### 2.1. Local spatio-temporal grid construction

Inspired by the local spatio-temporal structure[7], we utilize a Gaussian kernel to transform the distance matrix of the sensors into a weight matrix to represent the connectivity of the sensors, where a larger weight represents a smaller distance of the sensors, and the formula is defined as follows:

$$A^{(\varepsilon)}(i,j)=\exp\left(-\frac{D^{(\varepsilon)}(i,j)^2}{\theta^2}\right) \quad (1)$$

where $\mathrm{A}^{(\varepsilon)}(i,j)\in[0,1]$ is the weight matrix at time step $\varepsilon$, $D^{(\varepsilon)}(i,j)$ is the distance matrix, $i$ and $j$ denote two nodes of the weight matrix $D^{(\varepsilon)}(i,j)$, $\varepsilon$ is some time step, and $\theta$ is a hyperparameter, which is usually set to the standard deviation of all $D^{(\varepsilon)}(i,j)$.

According to the weight size of $\mathrm{A}^{(\varepsilon)}(i,j)$ to extract a portion of nodes that are close to the target node, and combined into a node set $\mathrm{B}$, the relevant definition is as follows:

$$B=\left\{n_p\left|\max\left\{\mathrm{A}^{(\varepsilon)}(\mathrm{p},\mathrm{q}),\mathrm{A}^{(\varepsilon)}(\mathrm{q},\mathrm{p})\right|1\le\varepsilon\le T\right\}>\xi\right\} \quad (2)$$

where since the traffic graph is a bidirectional graph, the values of $\mathrm{A}^{(\varepsilon)}(\mathrm{p},\mathrm{q})$ and $\mathrm{A}^{(\varepsilon)}(q,p)$ are different unless $q=p$; $\xi$ is a weight threshold set in advance, which requires that the degree of connectivity of other nodes connected to the target node $n_p$ is greater than $\xi$; in addition to this, fix the size of the node set $\mathrm{B}$ to be $\rho$, and for the node set with the number of nodes less than $\rho$, the remaining nodes need to be filled with the features of 0 and have no connectivity to the target node $n_p$. And for the node number is greater than $\rho$, the $\rho$ nodes with the closest distance to the target node $n_p$ are selected. Finally, the nodes of node set $\mathrm{B}$ are arranged according to their weights from largest to smallest.

The information related to the traffic map structure at each time step $\varepsilon\in\{1,\ldots,P\}$ and the node traffic measurements are fused together to obtain the local spatio-temporal $LS_v$ of a target node $v$ as follows:

$$LS_v^{(\varepsilon)}(b_i)=concat\left(x_i^{(\varepsilon)},\mathrm{A}^{(\varepsilon)}(i,\mathrm{v})\right),b_i\in B \quad (3)$$

where $x_i^{(\varepsilon)}$ is the traffic measurement value recorded by node $b_i$ at time step $t$, and $concat(,)$ represents the stitching operation. The matrix $LS_v^{(\varepsilon)}$ not only contains the spatial structure relationship between the target node $v$ at time step $\varepsilon$ and node $b_i$ in the node set $\mathrm{B}$, but also encodes the traffic measurement information of node $b_i$. Therefore, the local spatio-temporal $LS_v$ is defined as

$$LS_v=\left(LS_v^{(1)},LS_v^{(2)},\ldots,LS_v^{(P)}\right)\in\mathbb{R}^{|B|\times(F+1)\times P} \quad (4)$$

It is the necessary information used to train the traffic prediction model to predict the traffic data of node $v$.

In summary, the traffic flow prediction problem in this paper is defined as:

$$\left(LS_v^{(1)}, LS_v^{(2)}, \ldots, LS_v^{(P)}\right) f_2\left[f_1(\square)\right] \left(y_v^{P+1}, \ldots, y_v^{P+Q}\right) \quad (5)$$

where $f_1(\square)$ is the mapping function learned by the multi-scale spatio-temporal attention block, $f_2(\square)$ is the training function learned by the multi-channel spatio-temporal convolution block, $P$ is the historical time step, $Q$ is the future time step, the nodes are $v \in \{v_1, v_2, \ldots, v_N\}$, and $N$ is the number of nodes in the traffic map.

## 2.2. Multi-scale spatio-temporal attention blocks

Inspired by Pyraformer [8], pyramid attention is introduced to describe the spatio-temporal dependencies of individual traffic events in multiple resolutions. FPN feature pyramids[66] utilize convolution to compute feature maps at different scales, preserving high-resolution fine-grained semantic features while incorporating low-resolution coarse-grained semantic features.

In this section, multiple convolutional layers with kernel size 4 and step size 4 are used for initialization in the dimension of time and space merging. Sequences of length $L/4^s$ are generated at scale $s$. The dimensionality of each node is then reduced by a fully connected layer, followed by connecting them to the output of the original sequence 1*1 convolution.

Assuming that $I$ and $O$ are the input and output of a single attention, first the input data $I$ is linearly transformed into three independent matrices $Query$, $Key$, and $Value$, where $Query = IW_q$, $Key = IW_k$, and $Value = IW_v$. The original attention mechanism can be represented as follows:

$$o_i = \sum_{j=1}^{L} \frac{\exp\left(q_i \cdot k_j^T\right)}{\sum_{j=1}^{L} \exp\left(\frac{q_i \cdot k_j^T}{\sqrt{d_k}}\right)} \cdot v_j \quad (6)$$

where $W_q \in \mathbb{R}^{d \times d}$, $W_k \in \mathbb{R}^{d \times d}$, $W_v \in \mathbb{R}^{d \times d}$ are the learnable weight matrices for transforming the traffic event features into the query, key, and value space, $q_i$ denotes the ith row of $Query$, $k_j$ denotes the jth row of $Key$, and $v_j$ denotes the jth row of $Value$.

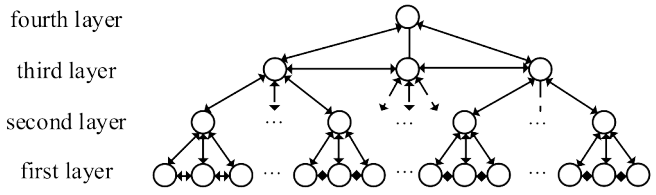


**Figure 2:** Pyramid diagram.

The structure of the pyramid is shown in Fig. 2, which is defined as a set of neighbor nodes $AS_l^{(s)}$ of the current node $n_l^{(s)}$, $n_l^{(s)}$ denotes the lth node in the sth layer and $AS_l^{(s)}$ contains the neighbor nodes of the current node in the same layer including its own node, $AS_l^{(s)}$, $C_l^{(s)}$ is the $C$ child node of the current node in the C-fork tree and $P_l^{(s)}$ is the parent node. They are specifically defined as:

$$AS_l^{(s)} = A_l^{(s)} \cup C_l^{(s)} \cup P_l^{(s)} \quad (7)$$

$$A_l^{(s)} = \left\{ n_j^{(s)} \middle| |j-l| \le \frac{N_{adj}-1}{2}, 1 \le j \le L/C^{s-1} \right\} \quad (8)$$

$$C_l^{(s)} = \left\{ n_j^{(s-1)} \middle| N_{child} \cdot (l-1) < j < N_{child} \cdot l \right\} \quad if \quad s \quad 2 \quad else \quad \varnothing \quad (9)$$

$$P_l^{(s)} = \left\{ n_j^{s+1} \middle| j = \left\lceil l/C \right\rceil \right\} \quad if \quad s \le S-1 \quad else \quad \varnothing \mathsf{S} \quad (10)$$

In this section, the number of fixed pyramid layers $S$ is 4, neighbor node $N_{adj}$ is 3 (including its own node), and child node $C$ is 4. Therefore, the pyramid attention of the current node $n_l^{(s)}$ can be expressed as:

$$o_i = \sum_{l \in AS_l^{(s)}} \frac{\exp\left(q_i \cdot k_l^T\right)}{\sum_{l \in AS_l^{(s)}} \exp\left(\frac{q_i \cdot k_l^T}{\sqrt{d_k}}\right)} \cdot v_j \quad (11)$$

## 2.3. Multi-channel spatio-temporal convolution block

The outputs of the four layers of the pyramid represent the spatio-temporal feature representations of traffic events at different scales. In this section, the features output from different layers of the last pyramid are first converted to different channel spaces, and then the spatio-temporal convolution block is used to integrate regional spatio-temporal features of traffic flows at different scales with different depths, and the regional spatio-temporal features at various scales are spliced together, and finally the fully-connected layer converts the spliced multiscale spatio-temporal features into predicted values of future traffic flows.

The spatio-temporal convolution block is composed of three convolutions that capture the effect of local spatio-temporal on the target traffic events, the temporal dependence of traffic events at different moments of the same node within the local spatio-temporal, and the spatial dependence between the node and its neighbors at the same moments within the local spatio-

temporal, the three convolution kernels are the spatio-temporal convolution kernel , the temporal convolution kernel $\eta_T \in \mathbb{R}^{f\times 1}$, and the spatial convolution kernel $\eta_S \in \mathbb{R}^{1\times f}$, and the output features of the sth layer of the pyramid, $CSTI^{(s)} \in \mathbb{R}^{d^{(s)}\times P\times |B|}$, serve as the input of the spatio-temporal convolution block, and the formula of the spatio-temporal convolution block is defined as follows:

$$H_{ST}^{(s)} = Leaky\,\mathrm{ReLU}\left(\eta_{ST} \otimes CSTI^{(s)}\right) \quad (12)$$

$$H_{T}^{(s)} = Leaky\,\mathrm{ReLU}\left(\eta_{T} \otimes CSTI^{(s)}\right) \quad (13)$$

$$H_{S}^{(s)} = Leaky\,\mathrm{ReLU}\left(\eta_{S} \otimes CSTI^{(s)}\right) \quad (14)$$

$$H^{(s)} = concat\left(H_{ST}^{(s)}; H_{S}^{(s)}; H_{T}^{(s)}\right) \quad (15)$$

$$CSTO^{(s)} = Leaky\,\mathrm{ReLU}\left(\eta_{o} \otimes H^{(s)}\right) \quad (16)$$

where $\otimes$ represents the convolution operation, $Leaky\,\mathrm{ReLU}(\cdot)$ denotes the leakage corrected linear unit function, $CSTO^{(s)} \in \mathbb{R}^{d\times P\times |B|}$ is the output of the spatio-temporal convolution block, the dimension of the output is the same as the dimension of the input as the convolution kernel $\eta_{ST}$ $\eta_S$ $\eta_T$ is set to $f = 3$ and the padding size to be 1. $\eta_o \in \mathbb{R}^{1\times 1}$ is a 1*1 convolution kernel for the aggregation of the three features of $H_{ST}^{(s)}$, $H_S^{(s)}$, and $H_T^{(s)}$ and a uniform number of channels to be $d$.

## 2.4. Prediction and Optimization

Next, prediction is performed by using two fully connected layers as prediction layers and using $CSTO^{(s)}$ as input to the prediction layers. The final prediction result $\hat{Y}$ is obtained.

$$\hat{Y} = \left(\mathrm{CSTO}^{(s)}\mathrm{W}_0 + \mathrm{b}_0\right)\mathrm{W}_1 + \mathrm{b}_1 \quad (17)$$

In order to learn the parameter settings of the model accurately, the final total loss function is expressed as follows.

$$\Lambda_{\mathrm{loss}} = \sqrt{\sum_{c=m}^{n}\left(\hat{Y}_{t,m} - \mathrm{Y}_{t,m}\right)^2} \quad (18)$$

where $\hat{Y}_{t,m}$ denotes the predicted value of the mth sensor at time t and $\mathrm{Y}_{t,m}$ denotes the true value of the mth sensor at time $t$. We optimize the model parameters using stochastic

gradient descent and backpropagation algorithms and further update the parameters using Adam optimizer.

## 3. Experimental comparison and analysis

### 3.1. Datasets

In this section, two publicly available real-world traffic prediction datasets are used to validate the effectiveness of the model, PeMS-Bay[9] and METR-LA dataset[9].The PeMS-Bay dataset contains measurements from 325 sensors for the period of January 1, 2017 to May 31, 2017 for the Bay area.The METR-LA dataset contains measurement data from 207 sensors for the period March 1, 2012 to June 30, 2012 in Bay. The speeds contained in both datasets are in miles per hour. Table 3.1 summarizes the statistics for both datasets.

**Table 1**
Statistics of the data set

| Datasets | Time Series | Nodes | Time Interval | Input Length | Output Length |
|---|---|---|---|---|---|
| META-LA | 34272 | 207 | 5min | 12 | 12 |
| PeMS-Bay | 52116 | 325 | 5min | 12 | 12 |

### 3.2. Evaluation Criteria

This section evaluates the performance of the model using three metrics that are more commonly used in traffic flow prediction, including MAE, RMSE, and MAPE, with the relevant formulas as follows:

$$MAE\left(y,\hat{y}\right)=\frac{1}{Q}\sum_{t=0}^{Q}\left|y_t-\hat{y}_t\right| \tag{19}$$

$$RMAE\left(y,\hat{y}\right)=\frac{1}{Q}\sum_{t=0}^{Q}\sqrt{\left(y_t-\hat{y}_t\right)^2} \tag{20}$$

where $y,\hat{y}\in\mathbb{R}^{Q}$, $Q$ denotes the length of the output sequence, $P$ denotes the length of the input sequence, $Q$ and $P$ are usually 12, $y$ denotes the true value of the output sequence of a node at a certain moment, and $\hat{y}$ denotes the predicted value of the input sequence of a node at a certain moment after the model.

### 3.3. Algorithm Comparison and Experimental Setup

The number of neighbor nodes constructed by the local spatio-temporal grid in the MSLST model architecture is 16. For the multi-scale spatio-temporal attention block, the number of

output channels is 128, the number of convolutional layers of the initialized pyramid graph is 4, including a 1*1 convolutional layer with stride 1 and three 4*4 convolutional layers with stride 4, the number of pyramidal spatio-temporal attentions of the stacked pyramids is 2, and the number of pyramidal layers of each pyramidal spatio-temporal The number of pyramidal layers of attention is 4. The number of channels of the four layers of spatio-temporal convolution of the multichannel spatio-temporal convolution block is 32, 64, 128, 256, and the size of convolution kernel of the spatio-temporal convolution is 3, respectively.

The dataset is set in chronological order with the first 70% as the training set, the middle 10% as the validation set, and the last 20% as the test set, and 0.2 70% is randomly selected as the final training set, and the length of the input sequence P and the length of the output sequence Q are 12. In terms of the model training, the batch size of the data is 80, and the number of iterations of the training is 50, and the optimizer for updating the parameters of the gradient descent is Adam, and the learning rate is 0.001, and the loss function is L1 loss value that minimizes the true and predicted values. dropout layer uses 0.3.

In order to verify the superiority of the model MSLST for the traffic prediction task, this section compares it with the following baseline methods:

1. ARIMA (Autoregressive Integrated Moving Average) [11], a well-known time series analysis method for predicting future values;
2. FC-LSTM (Fully Connected Short-Term Memory Network) [55], a sequence-to-sequence model with fully connected LSTM layers in both encoder and decoder;
3. STGCN (Spatio-Temporal Graph Convolutional Network) [60], a multi-scale traffic network that combines graph convolution and regular convolution;
4. DCRNN (Diffusion Convolutional Training Neural Network) [45], a diffusion convolution and GRU[?] based codec structure;
5. Graph WaveNet[46], an approach that combines diffusion factor convolution and graph convolution to model spatio-temporal dependencies;
6. DetectorNet (Transformer-enhanced spatio-temporal graph neural network) [63], a Transformer spatio-temporal network combining a multi-view temporal attention module and a dynamic attention module;

### 3.4. Experimental results and analysis

### 1) Comparative Analysis of Mainstream Advanced Algorithms

In this section, the performance of MSLST is evaluated on two real-world traffic prediction datasets and compared with state-of-the-art traffic flow prediction methods, as shown in Tables 2 and 3, and the following conclusions can be drawn from the comparison results:

1. Deep learning models have better prediction ability than traditional time series methods and machine learning models on both datasets, which indicates that deep learning methods are better able to model the nonlinear relationships between traffic data.
2. The traffic flow prediction methods STGCN, DCRNN, Graph Wavenet, DetectorNet, and MSLST combined with deep learning and graph structure generally perform

better than FC-LSTM, which demonstrates that information about the structure of the traffic road network is crucial for traffic prediction.

3. (c) DetectorNet and Graph WaveNet based on dynamic attention module have small detection errors at both time steps compared to STGCN and DCRNN with static road network structure, which reveals that the dynamic spatial correlation among modeled roads better reflects the dynamic changes of roads, and the learnable adaptive adjacency matrix can adapt to the uncertainty of traffic maps, both of which can retain valuable potential information.
4. (d) MSLST has a great improvement in prediction performance compared to DetectorNet in the short, medium, and long term, indicating that fusing temporal and spatial dimensions into a single spatio-temporal fluid to capture spatio-temporal features of traffic flow can directly model the dynamic spatio-temporal correlation among roads, which is a more complete characterization than that of aggregating independent temporal and spatial feature extraction modules. mSLST compared to Graph WaveNet at 60min has a decrease in MAE of 0.55, which implies that pyramidal attention modeling local spatio-temporal correlations at coarser scales enhances long-distance dependence, while multi-channel spatio-temporal convolution aggregating spatio-temporal features at each scale of pyramidal attention provides compact representations of proximity and long-distance spatio-temporal dependence.

**Table 2**
Comparison of overall performance in auroc, auprc, precision and recall

| **METR-LA** | **30min** | | | **60min** | | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAPE | MAE | RMSE | MAPE |
| **ARIMA** | 5.15 | 10.45 | 12.70% | 6.90 | 13.23 | 17.40% |
| **FC-LSTM** | 3.77 | 7.23 | 10.9% | 4.37 | 8.69 | 13.20% |
| **STGCN** | 3.47 | 7.24 | 9.57% | 4.59 | 9.40 | 12.70% |
| **DCRNN** | 3.15 | 6.45 | 8.80% | 3.60 | 7.60 | 10.50% |
| **GraphWavenet** | 3.07 | 6.22 | 8.37% | 3.53 | 7.37 | 10.01% |
| **DetectorNet** | 3.06 | 6.08 | 8.12% | 3.40 | 6.98 | 9.60% |
| **MSLST** | **2.50** | **5.19** | **6.78%** | **2.98** | **6.19** | **8.43%** |

**Table 3**
Error comparison between MSLST of this paper and baseline on PeMS-Bay dataset

| **METR-LA** | **30min** | **60min** |
|---|---|---|

| | MAE | RMSE | MAPE | MAE | RMSE | MAPE |
|---|---|---|---|---|---|---|
| **ARIMA** | 2.33 | 4.76 | 5.40% | 3.38 | 6.50 | 8.30% |
| **FC-LSTM** | 2.20 | 4.55 | 5.20% | 2.37 | 4.96 | 5.70% |
| **STGCN** | 1.81 | 4.27 | 4.17% | 2.49 | 5.69 | 5.79% |
| **DCRNN** | 1.74 | 3.97 | 3.90% | 2.07 | 4.74 | 4.90% |
| **GraphWavenet** | 1.63 | 3.70 | 3.67% | 1.95 | 4.52 | 4.63% |
| **DetectorNet** | 1.57 | 3.54 | 3.56% | 1.80 | 4.26 | 4.19% |
| **MSLST** | **1.47** | **3.22** | **3.29%** | **1.82** | **4.19** | **4.26%** |

## Component Analysis

In this section, component analysis is performed on two datasets, METR-LA and PeMS-Bay. Tables 4 and 5 show the statistical results of the three metrics for the component analysis of the two datasets, and the components are analyzed as follows:

1. w/o PM: eliminating the fusion of temporal and spatial dimensions into a single fluid, utilizing two independent self-attention for the temporal and spatial dimensions respectively for the capture of spatio-temporal correlation of traffic data, and then utilizing the gating mechanism to fuse the two features.
2. w/o R-PA: Replace pyramidal attention with ordinary self-attention.
3. w/o P-FPA: retain only the first layer output of the pyramid and extract regional spatio-temporal features using only a single-channel spatio-temporal convolution block.

**Table 4**
Component analysis of this paper's MSLST on the METR-LA dataset

| **METR-LA** | **30min** | | | **60min** | | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAPE | MAE | RMSE | MAPE |
| **MSLST** | **2.50** | **5.19** | **6.78%** | **2.98** | **6.19** | **8.43%** |
| **w/o PM** | 3.00 | 6.01 | 8.09% | 3.38 | 6.90 | 9.50% |
| **w/o R-PA** | 2.67 | 5.58 | 7.13% | 3.21 | 6.53 | 9.12% |

| | | | | | | |
|---|---|---|---|---|---|---|
| **w/o P-FPA** | 2.69 | 5.86 | 7.26% | 3.32 | 6.84 | 9.26% |

**Table 5**
Component analysis of MSLST in this paper on PeMS-Bay dataset

| **PeMS-Bay** | **30min** | | | **60min** | | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAPE | MAE | RMSE | MAPE |
| **MSLST** | **1.47** | **3.22** | **3.29%** | **1.82** | **4.19** | **4.26%** |
| **w/o PM** | 1.76 | 4.01 | 3.94% | 1.99 | 4.74 | 4.71% |
| **w/o R-PA** | 1.57 | 3.48 | 3.48% | 1.94 | 4.67 | 4.63% |
| **w/o P-FPA** | 1.61 | 3.51 | 3.53% | 1.95 | 4.70 | 4.67% |

The following conclusions can be drawn from the results in Table 4 and Table 4:

1. For both datasets, MSLST is much more effective than w/o PM, indicating that fusing the temporal and spatial dimensions of traffic flow data into a single fluid is more capable of modeling the spatio-temporal dynamics of traffic flow by using the attention mechanism that can unambiguously capture the spatio-temporal dependence of the target sensors at different moments.
2. The larger error of w/o R-PA than MSLST on both datasets, especially 60min, proves that the ability of ordinary self-attention in modeling a single spatio-temporal fluid with long-distance multi-step prediction is weaker compared to pyramidal attention, whose multi-resolution model effectively constructs the sensor's information transfer over long distances in time and space.
3. For METR-LA and PeMS-Bay, w/o R-PA works slightly better than w/o P-FPA, reflecting the multi-channel spatio-temporal convolutional block aggregation pyramid attention coarse and fine scale spatio-temporal features of the traffic flow are an integral part of the MSLST, and the finest scale spatio-temporal features are not able to simulate the long term dependence of traffic events.

### 3) Parametric Analysis

In order to verify how much the number of stacked pyramid spatio-temporal attention affects the model MSLST, 1~4 pyramid spatio-temporal attentions are stacked in this section, respectively. Figure 4.4 shows the results of the comparison of the average MAE for 12 time steps.

**Table 4**
Component analysis of this paper's MSLST on the METR-LA datase

| Stacks | METR-LA | | PeMS-Bay | |
|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE |
| | | 4.46 | | 2.54 |
| | | 4.18 | | 2.37 |
| 1 | 2.35 | 4.24 | 1.26 | 2.42 |
| | | 4.41 | | 2.49 |
| 2 | 2.22 | | 1.17 | |
| 3 | 2.27 | | 1.19 | |
| 4 | 2.31 | | 1.23 | |

From the experimental table, it can be seen that stacking 2 pyramids spatio-temporal attention is the best performance case for both METR-LA and PeMS-Bay datasets. As the number of stacked pyramids increases, the prediction error of the model does not decrease, but rather increases, the possible reason being that the model structure is too complex and there is overfitting. Apparently, 2 pyramids spatio-temporal attention is sufficient to deeply explore the complex spatio-temporal dependence of traffic events, which maintains the adequacy of the nonlinear structure of the model and does not lead to redundancy in the model structure.

## 4. Conclusion

In this chapter, a traffic flow prediction model based on a multi-scale local spatio-temporal network is proposed. First, the temporal and spatial dimensions are fused into a single fluid, the 3D local spatio-temporal grid is generated by combining the adjacency matrix and traffic flow features, multiple convolutional layers and pyramidal attention are introduced to learn the dynamic spatio-temporal dependence of the local spatio-temporal traffic events in different resolutions, and then the multi-channel spatio-temporal convolutional block is combined to merge and optimize spatial and temporal features among local spatio-temporal nodes, so as to get the global spatio-temporal features. Experiments on the traffic flow datasets PeMS and METR-LA show that the proposed model outperforms state-of-the-art methods.

## References


[1] Fang Zheng, Long Qingqing, Song Guojie, et al., Spatial-Temporal Graph ODE Networks for Traffic Flow Forecasting [C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp.364–373.

[2] Wang Xiaoyang, Ma Yao, Wang Yiqi, et al., Traffic Flow Prediction via Spatial Temporal Graph Neural Network[C]// Proceedings of The Web Conference 2020,pp. 1082–1092.

[3] Lu Bin, Gan Xiaoying, Jin Haiming, et al., Spatiotemporal Adaptive Gated Graph Convolution Network for Urban Traffic Flow Forecasting[C]// Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp.1025–1034.

[4] Li Y, Yu R, Shahabi C, et al. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting[C]//International Conference on Learning Representations 2018.
[5] Tang J, Qian T, Liu S, et al. Spatio-temporal latent graph structure learning for traffic forecasting[C]//2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022: 1-8.
[6] He Li, Shiyu Zhang, Xuejiao Li, et al., DetectorNet: Transformer-enhanced Spatial Temporal Graph Neural Network for Traffic Prediction [C]// Proceedings of the 29th International Conference on Advances in Geographic Information Systems,pp.133–136
[7] Yang S, Liu J, Zhao K. Space meets time: Local spacetime neural network for traffic flow forecasting[C]//2021 IEEE International Conference on Data Mining (ICDM). IEEE, 2021: 817-826.
[8] Liu S, Yu H, Liao C, et al. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting[C]//International conference on learning representations. 2021.
[9] Yu B, Yin H, Zhu Z. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018: 3634-3640.