Deep Spatio-temporal Adaptive 3D Convolutional Neural Networks for Traffic Flow Prediction

HE LI, XUEJIAO LI, LIANGCAI SU, DUO JIN, and JIANBIN HUANG, Xidian University
DESHUANG HUANG, Tongji University

Traffic flow prediction is the upstream problem of path planning, intelligent transportation system, and other tasks. Many studies have been carried out on the traffic flow prediction of the spatio-temporal network, but the effects of spatio-temporal flexibility (historical data of the same type of time intervals in the same location will change flexibly) and spatio-temporal correlation (different road conditions have different effects at different times) have not been considered at the same time. We propose the Deep Spatio-temporal Adaptive 3D Convolution Neural Network (ST-A3DNet), which is a new scheme to solve both spatio-temporal correlation and flexibility, and consider spatio-temporal complexity (complex external factors, such as weather and holidays). Different from other traffic forecasting models, ST-A3DNet captures the spatio-temporal relationship at the same time through the Adaptive 3D convolution module, assigns different weights flexibly according to the influence of historical data, and obtains the impact of external factors on the flow through the ex-mask module. Considering the holidays and weather conditions, we train our model for experiments in Xi'an and Chengdu. We evaluate the ST-A3DNet and the results show that we have better results than the other 11 baselines.

CCS Concepts: • **Information systems** → **Data mining**; • **Applied computing** → **Transportation;**

Additional Key Words and Phrases: Convolutional neural networks, spatial-temporal information, traffic prediction

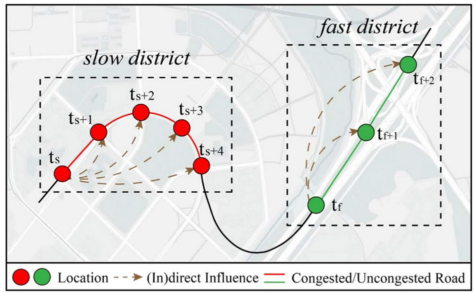
ACM Reference format:
He Li, Xuejiao Li, Liangcai Su, Duo Jin, Jianbin Huang, and DeShuang Huang. 2022. Deep Spatio-temporal Adaptive 3D Convolutional Neural Networks for Traffic Flow Prediction. *ACM Trans. Intell. Syst. Technol.* 13, 2, Article 19 (January 2022), 21 pages.
https://doi.org/10.1145/3510829

1 INTRODUCTION

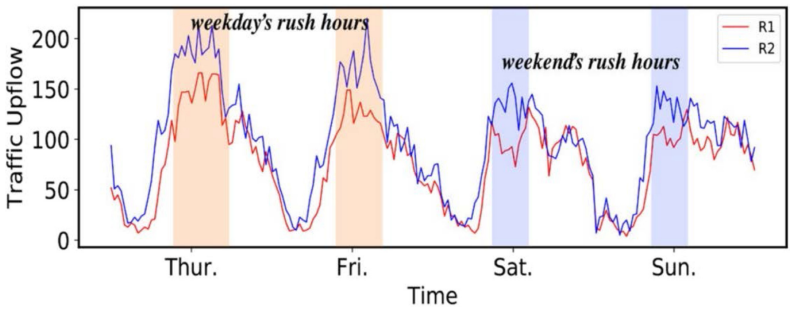
In the process of urbanization, with the continuous expansion of urban boundaries, the number of cars is also in excess growth. Traffic congestion has become a major challenge to urban construction and management in various countries. The analysis and prediction of dynamic traffic

This work was supported by the National Natural Science Foundation of China (No. 61602354, 61876138) and Natural Science Foundation of Shaanxi Province (No. 2019JM-227).
Authors' addresses: H. Li (corresponding author), X. Li, L. Su, D. Jin, and J. Huang (corresponding author), School of Computer Science and Technology, Xidian University, No. 2 South Taibai Road, Xi'an, Shaanxi 710071, China; emails: heli@xidian.edu.cn, {xjli_521, suliangcai, djin}@stu.xidian.edu.cn, jbhuang@xidian.edu.com; D. Huang, College of Electronic and Information Engineering, Tongji University, No. 1239 Siping Road, Shanghai 200092, China; and Guangxi Academy of Science, No. 98 Dalings Road, Nanning, Guangxi Zhuang Autonomous Region 530007, China; email: dshuang@tongji.edu.cn.

2157-6904/2022/01-ART19 $15.00
https://doi.org/10.1145/3510829



(a) Spatiotemporal Correlations



(b) Spatiotemporal Flexibility

Fig. 1. Spatio-temporal correlations and flexibility: (a) The trajectory of a car passing through congested and non-congested roads; (b) Traffic up-flow in R1 and R2 for four days.

situation is of great significance for solving the problem of congestion and road construction and traffic management of smart cities in the new era, and it has also played a role in promoting the rapid development of self-driving technology in recent years.

The goal of traffic flow prediction is to predict future traffic information by analyzing historical traffic GPS data and to help people make better travel decisions. In the real world, accurate traffic forecasting is a huge challenge, which is affected by three complex factors:

— `Spatio-temporal correlation`: In the field of transportation, the influence of one region and other regions is closely related to time and space. For example, in Figure 1(a), the figure shows the transportation tracks of motor vehicles at different timestamps, and the dots indicate the area in which they are located. At the timestamp t, each region has a direct effect on the region at the next time and an indirect effect on the region at a subsequent time. Specifically, for the expressway with good traffic conditions (such as urban expressway), the influence range of the current area is large and the influence time is short; on the contrary, for the slow speed area with poor traffic conditions (such as speed limit section), the current area has a greater impact on the traffic in the nearby area, while the future traffic flow in the farther area has less influence and longer influence time. This shows that the relationship between time and space is inseparable.

— `Spatio-temporal flexibility`: Spatio-temporal flexibility means that the reference value and credibility of historical data with fixed time will change flexibly, and inputs can not be treated equally. We can see from Figure 1(b) that the peak period varies from day to day or from district to district. During holidays and weekends, the traffic flow will change greatly. In other words, if we attach equal importance to historical data, it is not easy to pick out the most valuable data to predict future traffic volume.

— `Spatio-temporal complexity`: Spatio-temporal complexity refers to the sudden change of traffic flow state with time. Traffic flow will be affected by traffic accidents, weather, large-scale activities, and other complex factors, and show a different state than usual.

In fact, in the past few decades, many attempts have been made to make accurate traffic flow forecasts. Traffic forecasting methods are mainly divided into two categories: traditional methods and deep learning methods. The traditional learning methods are mainly divided into classical statistical methods and machine learning methods. The classical statistical method is to establish a statistical model based on the data to predict and analyze the data. The most representative and typical are the **historical average** (**HA**) [19] and **autoregressive integral moving average** (**ARIMA**) [23]. However, this kind of method requires that the data satisfy certain assumptions (e.g., ARIMA requires that the time series data must be smooth), while the spatio-temporal traffic data are too complex to satisfy these assumptions. Also, these methods are usually only suitable for small datasets, so their performance is usually poor in practical applications. Then, aiming at the problem of traffic flow forecasting, a variety of machine-learning methods represented by **support vector regression** (**SVR**) [26] are proposed. These methods can deal with high-dimensional data and capture complex relationships, but most of these methods rely on hand-designed features, which can not adequately describe the attributes of the data. The feature engineering of machine learning also requires a lot of domain expertise [10]. Deep learning models usually consist of multiple layers. They usually combine simpler models and pass data from one layer to another to build more complex models. The model is automatically derived by training a large amount of data, without the need for manual feature extraction [11].

At present, the better way is based on deep learning, which is divided into three categories: CNN [16], RNN [12], and GCN [6]. Generally, the convolution neural network is used to mine the spatial characteristics of traffic data, and the recurrent neural network is used to process time-series information. That is, they all try to capture the spatio-temporal correlation of the data by splitting the data from spatial correlation and nonlinear temporal correlation. However, spatial and temporal dependence that is not suitable for disassembly is a set of common features. Therefore, GNN models such as ASTGCN and GAAN [31] have been proposed to capture spatial and temporal information at the same time, where ASTGCN extracts temporal and spatial dependencies separately. These models usually combine the recurrent neural network and graph neural network to model the spatio-temporal information in traffic data. The performance of this kind of model has been significantly improved, but there are still some limitations, that is, the neglect of spatio-temporal flexibility.

To meet the above challenges, we propose a spatio-temporal traffic flow prediction network ST-A3DNet, based on deep learning to predict future traffic data. Our research has three main contributions:

— ST-A3DNet designed a deep neural network, which contains an Adaptive 3D Convolution block for capturing spatio-temporal information and an Ex-mask block for dealing with external factors, which are fused together by weighted fusion module. Adaptive 3D Convolution block includes a 3DConvSE component for capturing spatio-temporal correlation and different effects of different times, an Adaptive Transformation component for capturing spatio-temporal flexibility. In order to capture the different spatio-temporal information in the input closeness flow, daily period flow, and weekly period flow, a different number of Adaptive 3D Convolution blocks are designed in each part.
— In the 3DConvSE component, in order to mine both temporal and spatial features to preserve the spatio-temporal correlation, we use 3D convolution to reconstruct the SE-Resnet structure. In the Adaptive Transformation component, in order to capture spatio-temporal

flexibility, we flexibly allocate different weights for different 3D convolution mining by our proposed Selection mechanism. In the Ex-mask block, in order to effectively deal with the information of external factors, we use the mask matrix.
— We used Didi Taxi Dataset in Chengdu and Xi'an to strictly evaluate our method. The results show that our ST-A3DNet has advantages over ten baselines.

2 PROBLEM FORMULATION

Definition 1 (Regions of a City). We regard a city as the composition of $M \times N$ regions based on the longitude and latitude, denoted by $V = \{r_{1\times1}, r_{1\times2}, \ldots, r_{i\times j}\}$, $i \leq M$, $j \leq N$, each of which represents a spatial grid. The region can be defined as a pair (i, j), where it is located in row i and column j of the grid map.

Definition 2 (Traffic Inflow and Outflow). The inflow and outflow of a region are the total traffic volume of crowds entering the region and the total traffic volume of crowds leaving the area, respectively. The inflow and outflow of the grid in the ith row and jth column at time t are expressed respectively as

$$X_t(0, i, j) = \left|(s, e) \in \mathbb{P} : (x_e, y_e) \in r_{ij} \wedge \tau_e \in t\right|, \tag{1}$$

$$X_t(1, i, j) = \left|(s, e) \in \mathbb{P} : (x_s, y_s) \in r_{ij} \wedge \tau_s \in t\right|, \tag{2}$$

where $X_t(0, :, :)$ and $X_t(1, :, :)$ represent inflow and outflow, respectively. Let (τ, x, y) be a temporal geospatial coordinate, of which τ denotes timestamp, and (x, y) denotes region. The movement of an object can be recorded as a time-ordered spatial trajectory, among which the start point and endpoint are denoted by $s = (\tau_s, x_s, y_s)$ and $e = (\tau_e, x_e, y_e)$ represent the source and destination, respectively. Let $\mathbb{P}$ be all start-end (s, e) pairs. $(x, y) \in r_{i,j}$ represents the point (x, y) is within the node $r_{i,j}$, $\tau \in t$ represents the timestamp τ is in the time interval t.

Definition 3 (Historical Traffic State). The entire period of historical traffic states can be divided into non-overlapping intervals.

Problem 1 (Traffic Flow Prediction). *The problem of traffic flow prediction is to predict the inflow and outflow of each region of the city in the next time interval $T + 1$ based on the continuous historical flow observation data up to time T $\{X_t | t = t_1, \ldots, t_T\}$ and some external factors at T time intervals.*

3 METHODOLOGY

To solve the problem of traffic flow prediction, we propose an end-to-end deep learning model ST-A3DNet. Figure 2 shows its overall architecture, which is mainly composed of two parts. The first part describes the spatio-temporal features of traffic data, and the second part is used to describe the external factors that affect traffic flow data.

In the first part, we classify the temporal attributes of traffic data into three types, namely, Closeness, Daily, and Weekly period. Closeness mainly refers to the local time pattern in the traffic data, of which the current flow data is closely related to the recent historical data. The current traffic data in the daily period is related to the historical data within a few days, which can be understood at the same time in these days. Weekly period mainly refers to the weekly time pattern in traffic data, that is, the current traffic data is related to this day a week ago, which can be understood as that human behavior is periodic, and people always do some specific activities on a certain day in a particular week, such as fitness or watching movies. The difference between these three parts is in each part, the number of Adaptive 3D convolution blocks is not same. In the second part of the external factors, we from many influencing factors, such as sudden traffic accidents, meteorology,

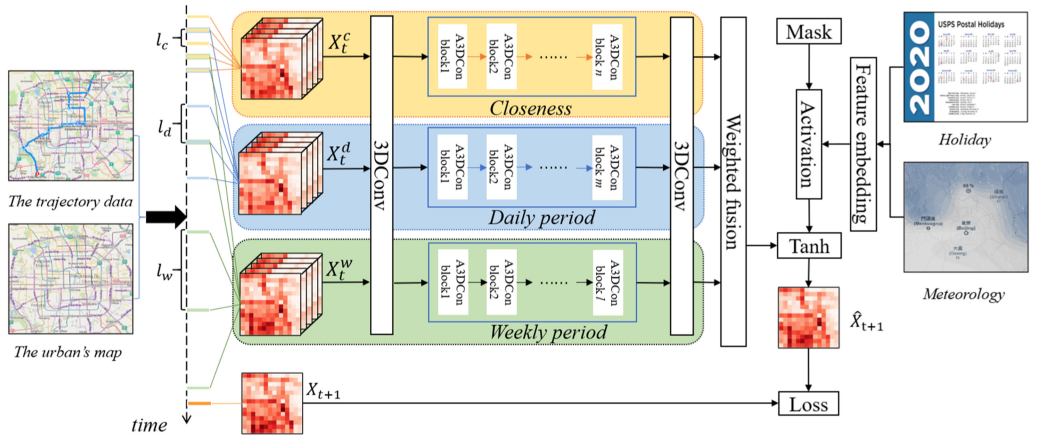


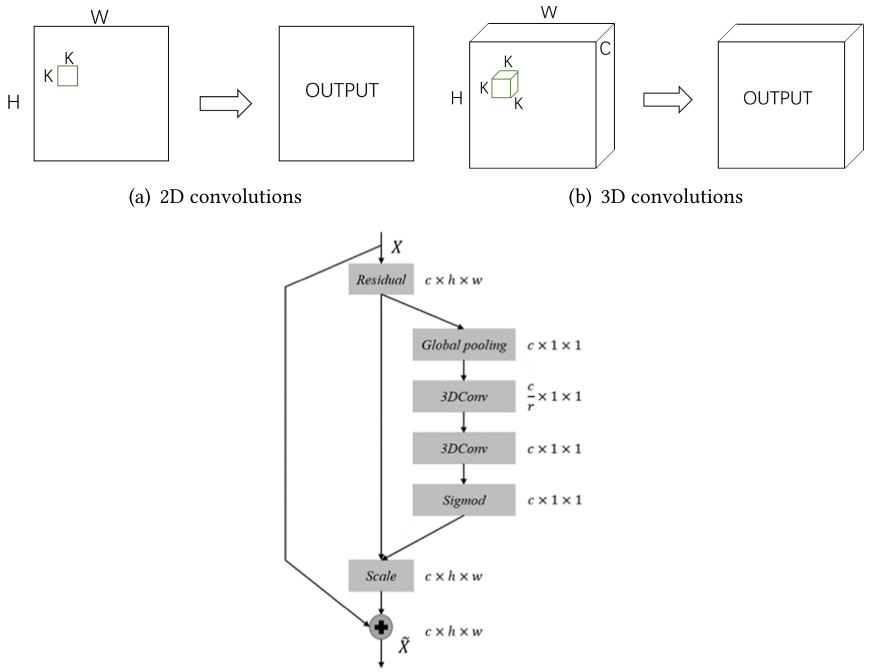
Fig. 2. ST-A3DNet architecture. A3DConv block: Adaptive 3D Convolution block; $\hat{X}_{t+1}$: Prediction result at time $t+1$; 3DConv: 3D convolution.

holidays, road reform, and other events pick out the most typical and most likely to cause regular changes in traffic flow, the meteorology, and holidays for research. After we finish the feature embedding work of weather and holiday through Dense, we deal with it through mask matrix, then use our proposed weighted fusion module to fuse the input data processed by Adaptive 3D Convolution block and external factor module, and finally complete the model training through loss function.

3.1 Adaptive 3D Convolution Block

The main function of the Adaptive 3D convolution block is to capture the spatio-temporal features of traffic data. In order to capture the spatio-temporal flexibility and spatio-temporal correlation we proposed, we designed two components, namely, the 3DConvSE component and the Adaptive transformation component. We propose two combination methods based on these two modules. Figure 4 depicts its structure. In the following, we will introduce the two components (take the Figure 4(b) structure for example). The reason we propose to choose 3D convolution is to simultaneously capture spatio-temporal correlation. But 3D convolution will generate a lot of parameters and calculations. In order to overcome this shortcoming and make the model have more nonlinearity to better fit the complex correlation between channels, we proposed a new 3DConvSE module by combining the SE module with the 3D convolution method. We proposed a new adaptive transformation component to automatically generate and match corresponding weights to input variables to capture the spatio-temporal flexibility of traffic data. Due to this simple and effective attention mechanism, our method can well simulate the dynamic temporal and spatial dependence of traffic flow.

Adaptive 3D convolution block mainly consists of three channels. However, the main difference is that the number of blocks is inconsistent, so in the following, we mainly use the closeness channel as a specific description. The channel is designed to capture spatio-temporal features based on the latest historical data. Therefore, its input is a subsequence of spatio-temporal raster data in the latest period. Let the subsequence be $X_t^c = \{X_{t-l_c}, \ldots, X_{t-1}\}$, where l_c is the length of the part of dependent sequences. As the same, the subsequences of the daily period and weekly period are $X_t^d = \{X_{t-l_d}, \ldots, X_{t-1}\}$ and $X_t^w = \{X_{t-l_w}, \ldots, X_{t-1}\}$, d and w are daily period and weekly period span, where l_d, l_w are the lengths of these parts of dependent sequences.



(a) 2D convolutions (b) 3D convolutions

(c) 3DConvSE architecture

Fig. 3. 2D/3D convolution method and 3DConvSE structure.

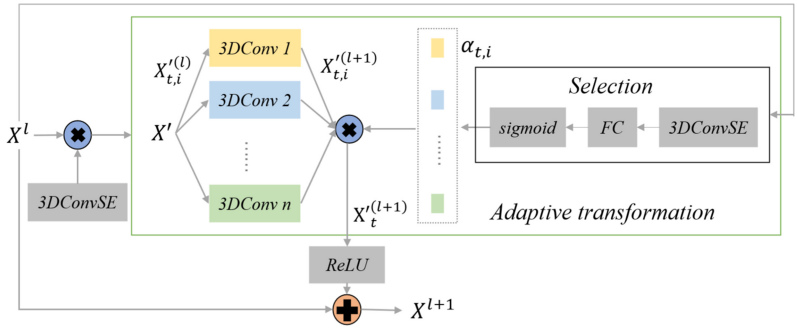
3.1.1 3DConvSE Component. How to effectively learn spatio-temporal information without relying on handcrafted features is a key issue. Ref. [22] found that compared with 2D convolution, 3D convolution captures spatio-temporal features more effectively. Figure 3(a) and (b) shows a comparison of 2D and 3D convolutions. In Figure 3(a), the 2D convolution performed in the convolutional layer can only extract features in the spatial dimension from the local neighborhood of the previous layer and result in a feature map. The 3D convolution operation is to convolve a 3D filter on a cube generated by superimposing multiple consecutive frames, as shown in Figure 3(b). The 3D convolution that convolves the 3D filter on a cube composed of multiple consecutive frames can generate a cube-style feature map, which not only preserves the spatial information but also contains temporal information.

As mentioned earlier, for predicting traffic flow, the importance of different time channel information is high or low, depending on its information and external context at each time slot. For example, the reference value of the traffic flow when the weather conditions are the same is often higher than the weather when conditions vary widely. Therefore, we cannot treat the input of all time slices equally, so we need to enhance or suppress the traffic flow information. Figure 3(c) depicts its structure. For the Input X, first obtain the global average pooling value for each timestamp, which represents the proportion of the global traffic flow at each time, denoted as z, and the calculation process is expressed as follows:

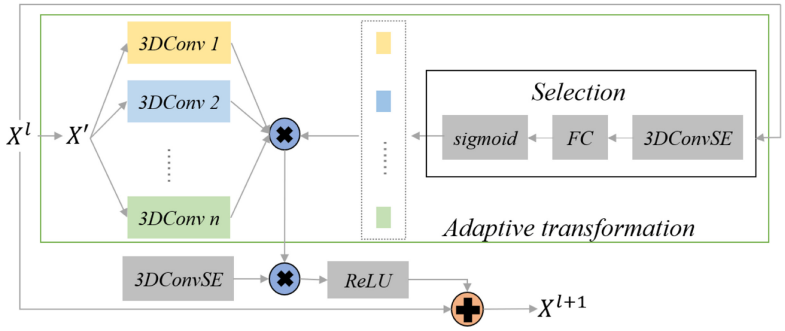
$$F_{sq}(X_t) = \frac{1}{W \times H \times C} \sum_{i=1}^{C} \sum_{j=1}^{W} \sum_{n=1}^{H} X_t(n, i, j), \tag{3}$$

$$z_t = F_{sq}(X_t), \tag{4}$$

where $F_{sq}(\cdot)$ is the squeeze function.



(a) Adaptive 3D Convolution block(A)



(b) Adaptive 3D Convolution block(B)

Fig. 4. Two structures of adaptive 3D convolution block.

Furthermore, activate all the traffic flow and capture the non-linear dynamic relationship between space and time. We formulate the activation process as

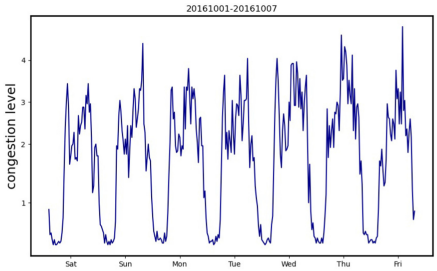
$$s = F_{ex}(z, W) = Sigmoid(W_2 * (Relu(W_1 * z + b_1)) + b_2), \tag{5}$$

where $*$ denotes the operation of 3D convolution, $F_{ex}(\cdot)$ is the extract function; $b_1, b_2, W_1 \in \mathbb{R}^{\frac{c}{r} \times r}$, and $W_2 \in \mathbb{R}^{\frac{c}{r} \times r}$ are learnable hyperparameters. Here, the structure containing two 3D Convolution Layers are adopted, where the first 3D Convolution layer plays a role in dimensionality reduction, and the dimensionality reduction coefficient r is a hyperparameter. The final 3D Convolution layer restores the original dimensions. Scale-down correction of historical traffic flow map information based on activation value s and express it as

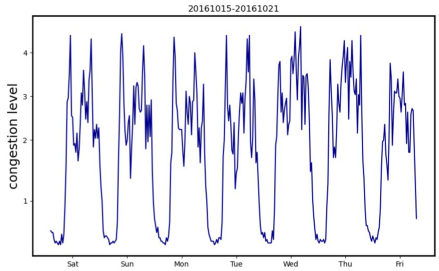
$$\tilde{X_t} = F_{scale}(X_t, s) = X_t \circ s, \tag{6}$$

where $F_{scale}(\cdot)$ refers to the scale function, the operator $\circ$ refers to the corresponding multiplication on the channel by the scalar s.

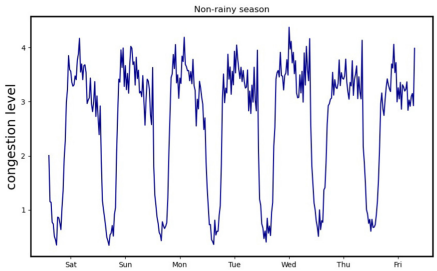
3.1.2 Adaptive Transformation Component. In order to capture the impact of the sudden change and the transition between holidays and ordinary time, and to consider that different times and different regions have different effects on the traffic flow at the predicted time, we propose the Adaptive Transformation component. This component is mainly composed of two parts, namely, the 3D convolution module of feature extraction, the selection module of matching weights for convolution training. Inspired by the popular residual network and Dynamic convolution, we design the block composed of **Batch Normalization** (**BN**), **Rectified Linear Unit** (**ReLU**), and 3D



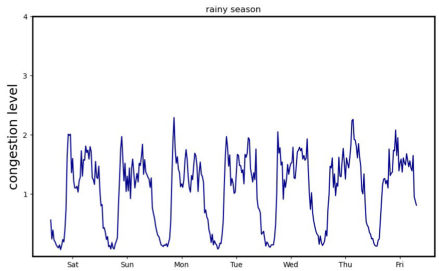
(a) The inflow of a region during the seven-day National Day holiday in China



(b) The inflow of a region during a normal week



(c) Non rainy season



(d) Rainy season

Fig. 5. Weekly periodicity of a region in Xi'an.

convolutions [11]. Adaptive Transformation component can be formalized as

$$X'^{(l+1)}_{t,i} = f\left(W^{(l+1)}_{t,i} * X'^{(l)}_{t,i} + b^{(l+1)}_{t,i}\right), l = 1, \ldots, L, \tag{7}$$

where $*$ denotes the operation of 3D convolution. $W^{(l+1)}$ and $b^{(l+1)}$ are parameters and f is an activation function.

$$\alpha_{t,i} = Sigmoid\left(FC\left(3DConvSE\left(X^{(l)}_{t,i}\right)\right)\right), \tag{8}$$

where $\alpha_{t,i}$ is the weight of time t and $i-th$ 3D convolution channel obtained after the selection mechanism. As shown in formula 8, the Selection module is composed of 3DConvSE and FC layers, and the sigmoid is selected as the activation function. This module can automatically generate a matching weight for each input variable, which can more accurately measure the complex impact of each variable on other variables.

$$\mathrm{X}^{(l+1)}_t = \sum_{i=1}^{n} \alpha_{t,i} \mathrm{X}^{(l+1)}_{t,i}, \alpha_{t,i} \in \mathbb{R}^{C\times 1\times 1}, \tag{9}$$

$$X^{(l+1)}_t = \mathrm{ReLU}\left(\mathrm{X}^{(l+1)}_t\right) + X^{(l)}_t, \tag{10}$$

$X^{(l)}_t \in \mathbb{R}^{C\times 1\times 1}$ is the input of the $l-th$ 3D convolution layer, and C_l is the number of channels.

3.2 Ex-maxk Block

Traffic flow will be affected by weather, holidays, and so on. We have constructed four figures to illustrate the situation. As shown in Figure 5, the above two figures compare the impact of holidays on traffic flow. Figure 5(a) shows the traffic flow state of seven days of China's National Day and Figure 5(b) shows the normal traffic flow state of the week. It can be seen from the figure

that during the holiday period, the traffic flow in each morning is significantly lower than that in the normal week, which is consistent with common sense that people will rest at home during the holidays. The following two figures compare the difference between the rainy season and the non-rainy season. The average traffic flow in the rainy season is lower than that in the non-rainy season, which is in line with common sense that people do not like to go out on rainy days. We embed the features of the weather and holiday data at first. Then through the further processing of the mask matrix, we get the matrix of the influence of external factors on the traffic flow data $\varepsilon_T \in \mathbb{R}^{M\times N}$, which represents the external characteristics at time T. Below we will explain these two parts, respectively.

3.2.1 Feature Embedding. The holiday factor is expressed as the time of the day, the day of the week, the week of the month, the month of the year, and whether it is a holiday. Meteorology factors are expressed as the temperature and Meteorology type of the current weather, such as strong wind, light rain, and heavy rain. Let the matrix of the Meteorology factors at T time be expressed as $M_T \in \mathbb{R}^{C\times M\times N}$, the matrix of the holiday factors at T time be expressed as $S_T \in \mathbb{R}^{C\times M\times N}$, the combination of the two is the matrix of external factors $E_T \in \mathbb{R}^{C\times M\times N}$. The Dense layer is used to embed its features, and the model is as follows:

$$E'_T = Sigmoid(E_T \bullet W_3 + b_3), \tag{11}$$

where $E'_T \in \mathbb{R}^{C\times M\times N}$ is the external feature matrix after feature embedding, and W_3 and b_3 are learnable parameters. Then we hand over the feature matrix to the mask matrix.

3.2.2 Mask. Holidays and meteorology factors will only affect the traffic flow in some regions, for example, there will be obvious differences between entertainment regions and work regions during holidays, but there will be no obvious differences in traffic flow in residential regions. A rainstorm will only affect traffic flow in wet areas. This external factor is like putting a mask on the urban area; if it is covered, the flow will change greatly. Based on this view, we develop an eigenvector transformation matrix based on the mask principle. At the time T, the corresponding external features can be obtained in the ST network, which can be expressed as $E'_T \in \mathbb{R}^{C\times M\times N}$, where Eigenvector of a specific region is denoted by $E'_T(:, i, j) \in \mathbb{R}^C$. Formally expressed as

$$\varepsilon_T(i, j) = \sigma(W_e(:, i, j) \bullet E'_T(:, i, j) + b_e(i, j)), 1 \le i \le M, 1 \le j \le N, \tag{12}$$

where W_e and b_e are learnable parameters. $\varepsilon_T \in \mathbb{R}^{M\times N}$ is the data of mask, where $\varepsilon_T(i, j)$ is expressed as the eigenvector corresponding to the region $r_{i,j}$ in the spatio-temporal graph. $\sigma(\cdot)$ is the sigmoid function, and $\bullet$ is dot product (inner product) of two vectors.

3.3 Weighted Fusion

To break free from the shackles and integrate features more flexibly, we propose a new weighted fusion module. First of all, it fuses the spatio-temporal features of the traffic data output from the three channels. Formally, the three outputs are fused as follows:

$$\mathrm{X}_{t_{T+1}} = W_1 \odot X^c_{t_{T+1}} + W_2 \odot X^d_{t_{T+1}} + W_3 \odot X^w_{t_{T+1}}, \tag{13}$$

where $\odot$ is the Hadamard product, $W_1 \in \mathbb{R}^{C\times M\times N}$, $W_2 \in \mathbb{R}^{C\times M\times N}$, and $W_3 \in \mathbb{R}^{C\times M\times N}$ are learnable parameters that reflect the degrees of the closeness influence, the daily period influence and the weekly period influence on the predicted target.

External factors, such as holiday and meteorology that can affect the flows. Then we fuse external factors as follows:

$$X_{t_{T+1}} = \tanh\left(\mathrm{X}_{t_{T+1}} \odot \varepsilon_T\right), \tag{14}$$

where $\varepsilon_T \in \mathbb{R}^{M\times N}$ denotes external features at time T.

ALGORITHM 1: ST-A3DNet Training Algorithmic

Input: The continuous historical flow observations $\{X_t|t = t_1, \ldots, t_T\}$; the lengths of closeness, daily period and weekly period sequences: l_c, l_d, l_w; the time interval between the last observed time interval; the predicted time interval $\triangle t$ and external factors $\{\varepsilon_t|t = t_1, \ldots, t_T\}$

Output: Learned ST-A3DNet model

// construct train dataset

1: $\mathcal{D} \leftarrow 0$

2: **for** all available time interval $t(0 \leq t \leq |T|)$ **do**

3: $X_t^c = \left[X_{t-l_c}, X_{t-(l_c-1)}, \ldots, X_{t-1}\right]$

4: $X_t^d = \left[X_{t-l_d}, X_{t-(l_d-1)}, \ldots, X_{t-d}\right]$

5: $X_t^w = \left[X_{t-l_w}, X_{t-(l_w-1)}, \ldots, X_{t-w}\right]$

//$X_{t+\triangle t}$ is the target at time interval $t + \triangle t$

6: Put $(X_t^c, X_t^d, X_t^w, \varepsilon_t, X_{t+\triangle t})$ into $\mathcal{D}$

7: **end for**

// training model

8: initialize all learnable parameters θ in ST-A3DNet

9: **repeat**

10: randomly select a batch of instances $\mathcal{D}_{batch}$ from $\mathcal{D}$

11: find θ by minimizing loss function

12: **until** stopping criteria is met

13: output the learned ST-A3DNet model

3.4 Optimizer

ST-A3DNet is trained by minimizing the **mean squared error (MSE)** loss function, which is defined as the MSE between the real value and the predicted value

$$L_\theta = \left\| X_{at_{T+1}} - X_{t_{T+1}} \right\|_2^2, \tag{15}$$

where θ is all the learnable parameters in this network. $X_{at_{T+1}}$ is the real flow value of the next time and $X_{t_{T+1}}$ is the predicted flow value of the next time.

3.5 Training Algorithm

Algorithm 1 outlines the ST-A3DNet training process. We first construct the training dataset (rows 1–9) from the original dataset. During each iteration, we optimize the object (15) on a selected batch of training examples $\mathcal{D}_{batch}$ (lines 10–11).

4 EXPERIMENTS

In this part, we conducted experiments on two datasets on traffic tracks to answer the following questions:

— Q1: How does ST-A3DNet compare with the state-of-the-art traffic flow forecasting methods in our mission?
— Q2: How do the components in ST-A3DNet affect performance?
— Q3: How do different designs, such as the number of Adaptive 3D convolution block in ST-A3DNet, affect the performance of ST-A3DNet?

Table 1. Statistics of the Traffic Dataset

Dataset	Chengdu	Xi'an
Location	In Chengdu, within the second Ring Road	In Xi'an, within the second Ring Road
Time interval	30 minutes	30 minutes
Raster size	(14,14)	(14,14)
Number of available time intervals	2592	2592
Data volume	10.6G	17.9G
Time span	1st Oct. 2016–30th Nov 2016	1st Oct. 2016–30th Nov 2016
Latitude and longitude range	[30.727818, 104.043333], [30.726490, 104.129076] [30.655191, 104.129591], [30.652828, 104.042102]	[108.92309, 34.279936], [109.008833, 34.278608] [109.009348, 34.207309], [108.921859, 34.204946]

— Q4: How does ST-A3DNet predict the performance of traffic data flow at different time intervals? How does ST-A3DNet perform on the test dataset?

4.1 Dataset

We have carried out experiments on the following two real traffic trajectory datasets,[1] namely, Didi-Chengdu-2016 and Didi-Xian-2016, which are called Chengdu and Xi'an for simplicity. Chengdu is a non-tourist city, while Xi'an is a tourist city. When it is on holidays, the change in traffic flow felt by tourist cities will be greater than that felt by non-tourist cities. Table 1 lists the statistics for the dataset.

Didi-Chengdu-2016: This dataset is provided by Didi and is used for traffic research, including flow forecasting, path planning, and so on. From the dataset, we can obtain the time, longitude, and latitude of the order driver track data of the Didi Express Taxi platform in the local area of the second Ring Road of Chengdu from October to November 2016, and the collection interval of the track points is 2–4 s. The track points have been processed by binding, which ensures that the data can correspond to the actual road information. We preprocess this dataset for traffic prediction. That is to say, 14-14 grids are divided by about 1 km $\times$ 1 km, and then the inflow and outflow of each region at each time are obtained according to the longitude and latitude coordinates and time stamp. We superimposed the preprocessed data at 30-minute intervals and got a total of 2,592 snapshots of traffic flow.

Didi-Xian-2016: This dataset is also provided by Didi. From the dataset, we can obtain the order driver track data of the Didi Express Taxi platform in the local area of the second Ring Road of Xi'an City from October to November 2016. The data information and preprocessing methods are consistent with the Chengdu dataset.

4.2 Experimental Setup

We consider two types of traffic forecasting tasks. The first is the full-time flow forecast, which means that we need to forecast the full-time traffic flow. The second is the traffic flow forecast during the peak hour, which means that we will make the traffic flow forecast during the period that we set as the peak hour. A total 80% of the data is used for training and 20% of the data is used for testing. The training was conducted from 1 October 2016 to 19:00 on 18 November 2016 and tested from 19:00 on 18 November 2016 to 30 November 2016. The peak periods of our design are 8–9 in the morning, 12–1 in the afternoon, and 6–7 in the evening. By default, the number of layers of ST-A3DNet is set to 2, the learning rate is set to 0.005, the number of iterations is 4,000 and the type of optimizer is ADAM. We choose MAE, RMSE, and MAPE as the criteria to evaluate the performance of our model.

[1] https://gaia.didichuxing.com/.

Table 2. Performance Comparison of Different Models

Dataset	Model	Test(Full)			Test(Peak)		
		MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
Chengdu	HA	39.21	79.08	20.86	42.63	97.04	25.67
	ARIMA	39.03	78.99	20.65	42.42	96.83	25.46
	GRU	35.57	55.31	32.37	38.91	54.59	33.82
	LSTM	36.29	47.47	37.49	38.79	55.18	40.97
	ConvLSTM	27.02	65.31	25.11	40.41	84.71	29.89
	STResNet	29.75	42.97	36.75	31.87	46.47	36.94
	ASTGCN	32.43	52.81	38.75	33.19	54.15	30.75
	ST-3DNet	28.67	46.20	33.67	30.54	47.32	32.89
	STSGCN	28.57	56.40	24.96	29.74	49.40	21.28
	STFGNN	29.85	61.69	22.54	31.24	50.23	20.44
	AGCRN	28.47	49.92	26.71	44.46	66.27	28.03
	ST-A3DNet(A)	**19.95**	**31.56**	**19.32**	**19.74**	**31.03**	**20.17**
Xi'an	HA	19.01	29.55	21.28	19.57	32.25	26.48
	ARIMA	18.89	29.42	21.09	19.48	32.16	26.37
	GRU	19.63	29.01	25.61	16.63	24.46	28.21
	LSTM	21.48	32.05	34.79	20.77	29.13	39.25
	ConvLSTM	21.22	32.99	26.06	19.51	30.54	30.08
	STResNet	18.12	26.65	40.41	15.65	23.49	46.93
	ASTGCN	18.11	27.65	36.51	18.01	28.21	27.16
	ST-3DNet	17.22	26.20	35.01	15.67	24.00	39.26
	STSGCN	16.04	26.45	30.07	19.62	27.54	25.41
	STFGNN	16.00	25.78	25.45	19.37	27.24	25.93
	AGCRN	17.58	27.34	37.85	25.94	36.75	32.84
	ST-A3DNet(A)	**15.42**	**23.30**	**20.87**	**13.53**	**21.04**	**24.42**

4.3 Baselines

We compare our model with the following seven models and Table 2 lists the baselines compared.

- **HA:** Prediction is the average of recent historical traffic data.
- **ARIMA:** Autoregressive integral moving average is the classical method of predicting time series data in traditional statistical methods.
- **GRU:** A kind of RNN network proposed to solve the problems of long-term memory and gradient in backpropagation. [5]
- **LSTM:** To solve the problem of gradient disappearance and gradient explosion in long sequence training, proposed a special RNN.
- **ConvLSTM:** A variant of LSTM. The change point is mainly to change the weight calculation to convolution calculation, which can extract the features of the image. [15]
- **STResNet:** A spatio-temporal data prediction model with advanced performance based on deep learning. [34]
- **ASTGCN:** A spatio-temporal graph convolution network with advanced performance based on attention mechanism, which can be used for spatio-temporal data prediction.
- **ST-3DNet:** ST-3DNet introduces 3D convolutions to automatically capture the correlations of traffic data in both spatial and temporal dimensions. [8]
- **STSGCN:** A spatio-temporal Synchronous graph convolutional neural network that uses local spatio-temporal subgraphs to synchronously model spatio-temporal features. [20]

Table 3. Performance Comparison of Ablation Study

Dataset	Model	Test(Full)			Test(Peak)		
		MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
Chengdu	ST-A3DNet(A)	19.95	**31.56**	19.32	19.74	31.03	20.17
	ST-A3DNet(B)	28.88	42.85	22.32	28.76	43.78	26.01
	ST-A2DNet	27.67	33.12	36.53	30.19	42.59	35.26
	ST-A3DNet(static)	29.69	40.65	33.19	31.19	42.77	35.81
	ST-A3DNet using Regular SE	28.01	44.54	23.92	33.73	49.37	32.15
	ST-A3DNet + L2	23.85	39.22	25.66	24.79	43.52	36.49
	ST-A3DNet-Ext	**19.78**	32.17	**16.59**	**18.77**	**30.07**	**17.99**
Xi'an	ST-A3DNet(A)	15.42	23.30	**20.87**	**13.53**	21.04	24.42
	ST-A3DNet(B)	19.83	29.85	30.16	19.51	29.72	39.05
	ST-A2DNet	17.94	26.52	39.32	15.41	23.17	43.68
	ST-A3DNet(static)	18.41	26.62	30.22	15.38	23.81	34.41
	ST-A3DNet using Regular SE	17.65	26.12	23.45	15.73	23.36	26.99
	ST-A3DNet + L2	15.17	23.40	22.32	15.10	24.68	33.23
	ST-A3DNet-Ext	**15.02**	**22.91**	21.68	13.62	**20.01**	**24.07**

— **STFGNN:** Spatial-Temporal Fusion Graph Neural Networks can effectively learn the hidden spatio-temporal relationship by processing the fusion operation of multiple spatial and temporal graphs in parallel at different times. [17]
— **AGCRN:** An adaptive graph convolutional recurrent network based on capturing specific node patterns and automatically inferring the interdependence between different traffic time series, and combining a recurrent network to automatically capture the temporal and spatial correlation in the traffic flow sequence. [2]

4.4 Performance Comparison(Q1)

Table 3 summarizes the comparison results. Below, we discuss the results of two tasks, namely, full-time traffic prediction and peak traffic prediction.

4.4.1 Result of Full-time Traffic Prediction. We have the following observations on the results of the full-time traffic prediction task:

— Our proposed ST-A3DNet (A) performs much better than all other baselines on the two datasets, which verifies the effectiveness of our model in solving full-time prediction tasks. Compared with the best baseline results on the Chengdu dataset, ST-A3DNet (A) shows an increase of about 35.4%, 36.1%, and 6.8% in terms of MAE, RMSE, and MAPE, respectively. On the Xi'an dataset, it has been improved by about 3.7%, 10.6%, and 11.1%, respectively. The performance difference between the two datasets also shows the impact of external factors on traffic prediction, because ST-A3DNet (A) has not taken into account external factors, but Chengdu, as a tourist city, is more affected by external factors, such as holidays, which is reflected in the experimental results that the performance of the model on Xi'an dataset is not better than that on the Chengdu dataset.
— On the Xi'an dataset, the performance of the ASTGCN model is slightly better than that of the STResNet model, indicating that the model has good robustness. But at the same time, in Chengdu dataset, the performance of graph-based ASTGCN model is not as good as that of STResNet model based on raster data, which shows that not all graph-based methods can beat the traditional methods based on raster data. How to use the spatio-temporal information in the way of the graph is the key to the graph-based method, not the graph form itself.

This also proves that ST-A3DNet (A) can make more effective use of the spatio-temporal information in the graph. STFGNN is an improved version of STSGCN, but it can be seen that STFGNN performs slightly better than STSGCN on the Xi'an dataset, while the situation on the Chengdu dataset is the opposite. This is because STSGCN proposes a localized spatio-temporal subgraph to synchronously capture the local correlation, but only designs the local and ignores the global information. STFGNN integrates some graphs into spatio-temporal fusion graphs to obtain hidden spatio-temporal correlations. In order to break the balance between local and global correlations, a dilation convolution module is proposed, in which a larger dilation rate can capture long-term correlations. Therefore, STFGNN is more likely to be affected by additional factors, and its performance in Chengdu, which is a tourist city and greatly affected by the weather, is not as good as Xi'an.

4.4.2 Result of Peak Period Traffic Prediction. We have the following observations on the results of the traffic prediction task during peak hours:

— Our proposed ST-A3DNet (A) model still outperforms all other baselines on the two data sets, which verify the effectiveness of our model in solving peak-hour prediction tasks. Compared with the best baseline results on the Chengdu data set, ST-A3DNet (A) shows an improvement of about 50.6%, 49.7%, and 1.3% in terms of MAE, RMSE, and MAPE, respectively. On the Xi'an dataset, it has been improved by about 15.6%, 11.6%, and 4.1%, respectively. It can be seen that, compared with the full-time prediction results, the performance of our model has been greatly improved on the Chengdu dataset, indicating the superiority of our model in solving peak-time prediction tasks.
— On both datasets, STResNet outperforms ASTGCN in terms of MAE and RMSE. And on the Xi'an dataset, STResNet outperforms all GCN models in terms of MAE and RMSE. It shows that in the peak time prediction task, the performance of the grid-based model is not worse than that of the graph-based model. This further illustrates the effectiveness of ST-A3DNet (A) in solving peak-hour forecasting tasks. Our model performs better than ST-3DNet on both datasets, because the recalibration block it used is inflexible and it does not consider the difference in time-related contributions. The adaptive attention mechanism we proposed and the 3D convolution method combined with the SE mechanism can more accurately quantify the difference in spatio-temporal correlation contributions.

4.5 Ablation Study(Q2)

In this section, we examine how components in ST-A3DNet, namely, convolution methods (2D convolution and 3D convolution), convolution kernel, component location (where 3DConvSE components are located in Adaptive 3D Convolution block), types of SE, external factors, and L2 regularization affect performance, which is also Q2's answer. Table 4 summarizes the comparison results.

— **Component location:** In Adaptive 3D Convolution block, we have designed two combination modes, one is the mode in front of the 3DconvSE component, that is, ST- A3DNet (B); and the other is the mode after the 3DConvSE component, that is, ST-A3DNet (A). It can be seen that the performance of ST-A3DNet (B) is far worse than that of ST-A3DNet (A). This is because the 3DConvSE component is designed to overcome the space-time flexibility, to enhance or suppress the traffic flow information as needed. The Adaptive Transformation component is designed to capture spatio-temporal correlation. When the 3DConvSE component is in front, it first enhances or suppresses the traffic data, which will partially destroy the spatio-temporal correlation and affect the performance of the Adaptive Transformation

Table 4. Performance Comparison of ST-A3DNet with Different Number of Propagation Layers

Dataset	Model	Test(Full)			Test(Peak)		
		MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
Chengdu	ST-A3DNet-1	21.43	32.41	21.28	21.75	33.58	25.53
	ST-A3DNet-2	**19.95**	**31.56**	19.32	**19.74**	**31.03**	20.17
	ST-A3DNet-3	20.49	32.65	**17.14**	20.15	**31.03**	**18.34**
	ST-A3DNet-4	20.58	32.64	17.94	19.94	31.83	18.88
Xi'an	ST-A3DNet-1	16.22	24.58	25.52	15.13	23.59	32.03
	ST-A3DNet-2	15.42	23.30	**20.87**	**13.53**	**21.04**	**24.42**
	ST-A3DNet-3	**15.15**	**22.65**	25.69	14.89	22.79	27.57
	ST-A3DNet-4	15.16	23.18	24.19	14.61	23.36	33.76

component. The Adaptive Transformation component in the previous, although have the convolution and other operations, but did not essentially change the characteristics of the original data, does not affect the performance of the 3DConvSE component. Therefore, the performance of ST-A3DNet (B) will be worse than that of ST-A3DNet (A).

— **Convolution methods:** The convolution method chosen in this article is 3D convolution, and the selected control experiment is 2D convolution, which is the combination of STResNet and ConvSE. It can be seen that, regardless of any task on any data set, the performance of our proposed model is better than that of 2D convolution, and the improvement effect is obvious, especially on MAPE. Specifically, on the Chengdu dataset, the performance of ST-A3DNet (A) is 89% and 74.8% higher than that of ST-A2DNet, while on the Xi'an dataset, it is 88.4% and 78.8% higher, respectively. This is because 3D convolution can capture the spatio-temporal characteristics of inseparable traffic data, and the spatio-temporal correlation of spatio-temporal features is more complete. However, 2D convolution can only capture the spatial characteristics of traffic data.

— **Convolution kernel:** As can be seen from the table, the effect of static convolution is far inferior to our model. This is because our model borrows the idea of dynamic convolution to increase the complexity of the model without increasing the depth or width of the network to better capture spatio-temporal features.

— **Types of SE:** The control experiment we selected in this article is RegularSE. It can be seen from the table that the performance of 3DConvSE selected in this article is much better than that of RegularSE because 3DConvSE has a process of expansion and contraction, which can effectively assign different importance weights to channels of different importance, while RegularSE provides a priori probability to make the optimization process tend to the desired goal, which can not effectively assign different weights to different channels according to their importance and can not effectively capture the spatio-temporal flexibility.

— **External factors:** As can be seen from Table 3, the optimization effect of the external factor module on the Xi'an dataset is greater than that of the Chengdu dataset. Specifically, the external factor module improved 2.6% on MAE on full-time tasks on Xi'an datasets and only 0.8% on Chengdu datasets. This is because Xi'an is a tourist city, which is greatly affected by external factors, so after adding the module of external factors, the improvement effect is greater than that of Chengdu.

— **L2 regularization:** It can be seen that the optimization effect of the L2 regularization task on the model during the peak period is better than that of the full-time task, this is because the training data of the task during the peak period is less, and L2 regularization can solve

the problem of over-fitting, one of the reasons for over-fitting is the lack of training data. The optimization effect of Chengdu dataset is better than that of Xi'an dataset because the data of Chengdu dataset is less likely to be disturbed by external factors.

4.6 Design Choices of ST-A3DNet(Q3)

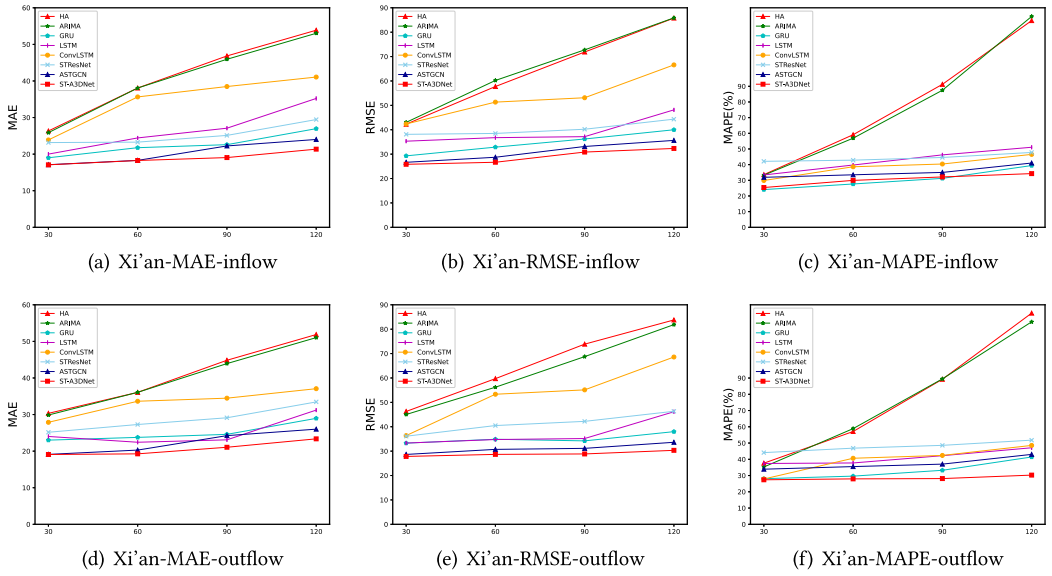
In this part, we study the effect of the number of layers on performance. To explore how the number of propagation layers affects performance, we changed the number of model layers. In particular, we have carried out experiments in the range of the number of layers of 1, 2, 3, 4. Table 4 summarizes the experimental results, where ST-A3DNet-X represents the model with X layer. Table 4 summarizes the experimental results, where ST-A3DNet-X represents the model with X layer. From the results, we have the following observations:

— Too many layers (more than three layers) will not bring additional improvement. From the overall results, the performance of ST-A3DNet-2 is slightly better than that of ST-A3DNet-3 but much better than that of ST-A3DNet-4. Specifically, in the full-time prediction task, ST-A3DNet-2 is 2.7% and 3.4% higher than ST-A3DNet-3 in terms of MAE and RMSE, respectively, in Chengdu dataset, while ST-A3DNet-3 is 0.4% lower in MAE and only 0.03% higher in RMSE than ST-A3DNet-4. It can be said that ST-A3DNet-2 is the best choice when dealing with full-time tasks on Chengdu dataset. Too many layers can lead to redundancy, which can affect performance.
— Broadcast ST-A3DNet-1 is worse than ST-A3DNet-2, 3, and 4, indicating that only one propagation layer is not enough to achieve excellent performance. This is reasonable because the single-layer 3DCNN can only capture the spatial information of the nearest neighbor, while the multi-layer can capture the spatial information of the distant neighbor. The transfer of traffic flow is not only related to the adjacent position, but also the long-distance position. Therefore, the single-layer network does not make explicit use of near-distance spatial correlation and long-distance correlation, resulting in poor performance. Therefore, it is necessary to overlay at least two layers of the network.

4.7 Results with Different Time Intervals(Q4)

In this part, we study the performance of traffic prediction of the model at different time intervals and the performance on the test dataset. Figure 6 is the performance evaluation result of traffic prediction of ST-A3DNet and its baselines at different time intervals on Xi'an data set. As shown in the figure, we divide the interval into 30 minutes, 60 minutes, 90 minutes, and 120 minutes. Through the results, we have the following observation:

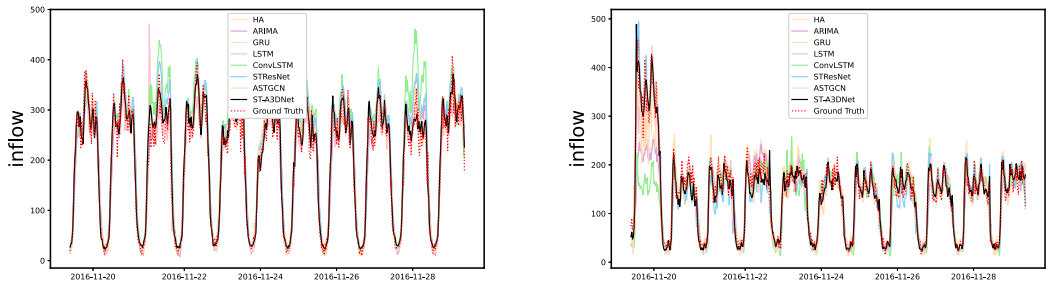
— Whether in terms of RMSE, MAE, or MAPE, our proposed ST-A3DNet outperforms other networks in 120-minute long interval prediction tasks. In the 60-minute and 90-minute mid-interval traffic prediction task, ST-A3DNet is better than other networks in terms of RMSE and MAE, only in the inflow traffic prediction task of MAPE, ST-A3DNet is slightly inferior to GRU. In the 30-minute short-term traffic prediction task, the performance of the model is slightly better than that of the best ASTGCN. It shows the effectiveness of our model in long-term and short-term prediction, especially in the task of long-term traffic prediction, which is more difficult to predict and has stronger spatio-temporal uncertainty.
— From the experimental results, we can see that LSTM, GRU, and ASTGCN can effectively solve the long-term prediction task. In terms of RMSE and MAE, the performance of ConvLSTM is weaker than that of LSTM, especially in medium-and long-term prediction tasks, because ConvLSTM has a stronger ability to capture spatial features than LSTM, but the ability to capture time features is weaker than LSTM.



(a) Xi'an-MAE-inflow (b) Xi'an-RMSE-inflow (c) Xi'an-MAPE-inflow

(d) Xi'an-MAE-outflow (e) Xi'an-RMSE-outflow (f) Xi'an-MAPE-outflow

Fig. 6. Performance evaluation results of the inflow and outflow flow prediction of the models at different time intervals on the Xi'an dataset.



(a) Xi'an (b) Chengdu

Fig. 7. Performance of the inflow prediction of the models on the test dataset.

Figure 7 shows the performance of ST-A3DNet and its baselines on the test set. Through these results, we have the following observation: as shown in Figure 7(b), when there is an anomaly, that is, the real data of 2016-11-20 in Chengdu is obviously higher than that of other dates, the performance of ConvLSTM and ARIMA is obviously worse than that of other models, and has obvious delayed processing ability in predicting the actual results. It can be seen from the diagram that the spatio-temporal flexibility has a certain influence on the prediction results, and because our model captures this characteristic, the gap between the predicted results and the actual results is small, and the actual effect is better than other models.

5 RELATED WORK

Urban Traffic Flow Prediction. Traffic flow prediction is one of the important research issues in spatio-temporal data and urban computing. In the early days, many studies focused on the movement of people by studying trajectories [21], and there were also many studies on local area traffic (e.g., vehicles per unit time, average vehicle speed.) [18]. In recent years, many scholars have

begun to study city-level traffic flow prediction. In the past, some linear models focused only on temporal relationships. For example, ARIMA was widely used [23], and it also produced a variety of deformations [25]. Later, some studies used machine learning [30], including deep learning, for traffic prediction, which will be described below.

Deep learning for spatio-temporal Prediction. There are many excellent studies published on the application of deep learning in spatio-temporal prediction. Early, DNN is used to predict the flow of the crowd [33]. Besides, convolutional networks are often used for spatial dependence in grid-maps [24, 32, 35] and have also achieved success. Recently, some authors used 3D convolution to capture both spatial and temporal relevance, and we were also inspired by this. ST-3DNet introduces 3D convolution for the first time to automatically capture the correlation of traffic data in space and time dimensions. A new recalibration block is proposed to quantify the difference in spatial correlation contribution. Mainly consider two-time characteristics of traffic data, namely, short-term and long-term [8]. However, this recalibration block is not flexible and does not consider the difference in temporal correlation contribution, so we introduce an adaptive attention mechanism to more accurately quantify the difference in spatial and temporal correlation contribution. And we also divide time characteristics into three categories. TF-3DNet also uses 3D convolution to obtain spatial and temporal features at the same time, but its main contribution is the missing value completion method [29]. Some works take advantage of RNNs to process time-series data and capture the temporal dependence of spatio-temporal data [27]. M-B-LSTM model proposes a deep bidirectional long and short-term memory network for short-term traffic flow prediction, which reduces the uncertainty problem through the approximation process of the context before and after the randomness reduction layer. Its disadvantage is that due to the limitation of LSTM itself, it can only predict the short-term flow of observation points and cannot capture spatial correlation [37].

In recent years, the use of graph neural networks to predict traffic flow has also been a hot topic [1, 7]. TGCN introduced GCN, combined with GRU to extract the spatial and temporal characteristics of traffic data, and achieved good results [36]. STGCN uses pure convolution to extract temporal features for the first time and proposes a new neural network composed of spatio-temporal blocks. Since this architecture is a pure convolution operation, it is 10 times faster than RNN-based models and requires fewer parameters[28]. ASTGCN considers the influence of different time periods, and also uses GCN to extract spatial features, and ordinary convolution to extract temporal features, but on this basis, an attention mechanism is added [7]. MRA-BGCN considers the relationship of the edges and extracts the correlation of the edges by constructing a graph on the edges [3]. GSTNet considers non-local (global) spatial dependencies and proposes a new network structure, including a time module that can extract both long-term and short-term dependencies, and a spatial module that can extract both local and global relationships [4]. Some scholars have previously proposed the heterogeneity of Spatio-temporal predictions [13], but the research is not in-depth for the time being. STSGCN proposes a localized spatio-temporal subgraph to synchronously capture the local correlation, but only designed the local information while ignoring the global information. When there is missing data, the effect will be worse because only local noise is learned [20]. On the basis of STSGCN, in order to simultaneously capture the local and global complex spatio-temporal correlations, a new framework based on CNN, STFGNN, is proposed. In order to break the balance between local and global correlations, a dilation convolution module is proposed, in which a larger dilation rate can capture long-term correlations, but the long-term correlations can be captured by stacked layers. It is susceptible to external factors [17].

Attention mechanism in deep learning. In deep learning, the attention mechanism is a means to effectively improve model performance. In traffic prediction problems, it is a common

method to add attention mechanism to LSTM. Besides, attention mechanisms can also be subdivided into spatial attention mechanisms, channel attention mechanisms [9], and mixed attention mechanisms. There have been a number of notable ideas in the mixed attention mechanism in recent years. GMAN proposed a spatio-temporal attention mechanism based on gated fusion to model complex spatio-temporal correlations. The experimental results on the sensor experiment set show that it improves the performance of long-term traffic flow prediction [38]. MRA-BGCN proposes a multi-level attention mechanism, which can aggregate information in different neighborhoods and learn their importance [3]. ATFM calculates the attention weight of the spatial region in each time interval and combines two ConvLSTM units to dynamically learn the spatio-temporal representation [14]. Inspired by these works, our ST-A3DNet flexibly captures spatio-temporal correlation through the Adaptive 3D convolution module combined with an adaptive attention mechanism. Due to this simple and effective attention mechanism, our method can well simulate the dynamic temporal and spatial dependence of traffic flow.

6 CONCLUSIONS

We propose a new deep learning model ST-A3DNet, to simultaneously predict input/output flows in spatio-temporal networks. ST-A3DNet can capture spatio-temporal correlation (the combination of close/long-distance in space and closeness/daily period/weekly period in time), spatio-temporal flexibility (with different weights in different combinations) and external factors (such as holidays and weather). We evaluated our ST-A3DNet on two real datasets in Chengdu and Xi'an and achieved significantly better performance than seven baselines.

REFERENCES

[1] Lei Bai, Lina Yao, Salil Kanhere, Xianzhi Wang, Quan Sheng, et al. 2019. Stg2seq: Spatial-temporal graph to sequence model for multi-step passenger demand forecasting. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI'19)*. International Joint Conferences on Artificial Intelligence Organization, 1981–1987. https://doi.org/10.24963/ijcai.2019/274

[2] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 17804–17815. https://proceedings.neurips.cc/paper/2020/file/ce1aad92b939420fc17005e5461e6f48-Paper.pdf.

[3] Weiqi Chen, Ling Chen, Yu Xie, Wei Cao, Yusong Gao, and Xiaojie Feng. 2020. Multi-range attentive bicomponent graph convolutional network for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 3529–3536.

[4] Shen Fang, Qi Zhang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. 2019. GSTNet: Global spatial-temporal network for traffic flow prediction.. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 2286–2293.

[5] Rui Fu, Zuo Zhang, and Li Li. 2016. Using LSTM and GRU neural network methods for traffic flow prediction. In *Proceedings of the 2016 31st Youth Academic Annual Conference of Chinese Association of Automation*. IEEE, 324–328.

[6] Xu Geng, Xiyu Wu, Lingyu Zhang, Qiang Yang, Yan Liu, and Jieping Ye. 2019. Multi-modal graph interaction for multi-graph convolution network in urban spatiotemporal forecasting. arXiv:1905.11395. Retrieved from https://arxiv.org/abs/1905.11395.

[7] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 922–929.

[8] Shengnan Guo, Youfang Lin, Shijie Li, Zhaoming Chen, and Huaiyu Wan. 2019. Deep spatial–temporal 3D convolutional neural networks for traffic data forecasting. *IEEE Transactions on Intelligent Transportation Systems* 20, 10 (2019), 3913–3926.

[9] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7132–7141.

[10] Michael I Jordan and Tom M Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349, 6245 (2015), 255–260.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[12] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SJiHXGWAZ.

[13] Youru Li, Zhenfeng Zhu, Deqiang Kong, Meixiang Xu, and Yao Zhao. 2019. Learning heterogeneous spatial-temporal representation for bike-sharing demand prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1004–1011.

[14] Lingbo Liu, Jiajie Zhen, Guanbin Li, Geng Zhan, Zhaocheng He, Bowen Du, and Liang Lin. 2020. Dynamic spatial-temporal representation learning for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems* 22, 11 (2020), 7169–7183.

[15] Yipeng Liu, Haifeng Zheng, Xinxin Feng, and Zhonghui Chen. 2017. Short-term traffic flow prediction with Conv-LSTM. In *Proceedings of the 2017 9th International Conference on Wireless Communications and Signal Processing*. IEEE, 1–6.

[16] Xiaolei Ma, Zhuang Dai, Zhengbing He, Jihui Ma, Yong Wang, and Yunpeng Wang. 2017. Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. *Sensors* 17, 4 (2017), 818.

[17] Mengzhang Li and Zhanxing Zhu. 2021. Spatial-temporal fusion graph neural networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 5 (May 2021), 4189–4196. https://ojs.aaai.org/index.php/AAAI/article/view/16542.

[18] Ricardo Silva, Soong Moon Kang, and Edoardo M Airoldi. 2015. Predicting traffic volumes and estimating the effects of shocks in massive transportation systems. *Proceedings of the National Academy of Sciences* 112, 18 (2015), 5643–5648.

[19] Brian L Smith and Michael J Demetsky. 1997. Traffic flow forecasting: Comparison of modeling approaches. *Journal of Transportation Engineering* 123, 4 (1997), 261–266.

[20] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. 2020. 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. 914–921.

[21] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. 2014. Prediction of human emergency behavior and their mobility following large-scale disaster. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 5–14.

[22] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4489–4497.

[23] Mascha Van Der Voort, Mark Dougherty, and Susan Watson. 1996. Combining Kohonen maps with ARIMA time series models to forecast traffic flow. *Transportation Research Part C: Emerging Technologies* 4, 5 (1996), 307–318.

[24] Leye Wang, Xu Geng, Xiaojuan Ma, Feng Liu, and Qiang Yang. 2018. Cross-city transfer learning for deep spatio-temporal prediction. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China (IJCAI'19)*. AAAI Press, 1893–1899.

[25] Billy M Williams and Lester A Hoel. 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of Transportation Engineering* 129, 6 (2003), 664–672.

[26] Chun-Hsin Wu, Jan-Ming Ho, and Der-Tsai Lee. 2004. Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems* 5, 4 (2004), 276–281.

[27] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Proceedings of the Advances in Neural Information Processing Systems*. 802–810.

[28] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence Stockholm, Sweden (IJCAI'18)*. AAAI Press, 3634–3640.

[29] Feng Yu, Dan Wei, Shuting Zhang, and Yanli Shao. 2019. 3D CNN-based accurate prediction for large-scale traffic flow. In *Proceedings of the 2019 4th International Conference on Intelligent Transportation Engineering*. IEEE, 99–103.

[30] Hongyuan Zhan, Gabriel Gomes, Xiaoye S Li, Kamesh Madduri, Alex Sim, and Kesheng Wu. 2018. Consensus ensemble system for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems* 19, 12 (2018), 3903–3914.

[31] Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King, and Dit-Yan Yeung. 2018. Gaan: Gated attention networks for learning on large and spatiotemporal graphs. *Uncertainty in Artificial Intelligence* (2018), 339–349.

[32] Junbo Zhang, Yu Zheng, and Dekang Qi. 2016. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA (AAAI'17)*. AAAI Press, 1655–1661.

[33] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. 2016. DNN-based prediction model for spatio-temporal data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 1–4.

[34] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, Xiuwen Yi, and Tianrui Li. 2018. Predicting citywide crowd flows using deep spatio-temporal residual networks. *Artificial Intelligence* 259, 0004–3702 (2018), 147–166.

[35] Junbo Zhang, Yu Zheng, Junkai Sun, and Dekang Qi. 2019. Flow prediction in spatio-temporal networks based on multitask deep learning. *IEEE Transactions on Knowledge and Data Engineering* 32, 3 (2019), 468–478.

[36] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. 2019. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems* 21, 9 (2019), 3848–3858.

[37] Qu Zhaowei, Li Haitao, Li Zhihui, and Zhong Tao. 2020. Short-term traffic flow forecasting method with MB-LSTM hybrid network. *IEEE Transactions on Intelligent Transportation Systems* (2020), 1–11.

[38] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. 2020. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1234–1241.

Received November 2020; revised March 2021; accepted May 2021