



Article Spatial-Temporal Diffusion Convolutional Network: A Novel Framework for Taxi Demand Forecasting

Aling Luo 🗅, Boyi Shangguan *, Can Yang 🕒, Fan Gao, Zhe Fang 🕒 and Dayu Yu

School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China; eeluo_2@whu.edu.cn (A.L.); yangcan@whu.edu.cn (C.Y.); fgao@whu.edu.cn (F.G.); fangzhe@whu.edu.cn (Z.F.); dayuyu@whu.edu.cn (D.Y.)

* Correspondence: sgby@whu.edu.cn

Abstract: Taxi demand forecasting plays an important role in ride-hailing services. Accurate taxi demand forecasting can assist taxi companies in pre-allocating taxis, improving vehicle utilization, reducing waiting time, and alleviating traffic congestion. It is a challenging task due to the highly nonlinear and complicated spatial-temporal patterns of the taxi data. Most of the existing taxi demand forecasting methods lack the ability to capture the dynamic spatial-temporal dependencies among regions. They either fail to consider the limitations of Graph Neural Networks or do not efficiently capture the long-term temporal dependencies. In this paper, we propose a Spatial-Temporal Diffusion Convolutional Network (ST-DCN) for taxi demand forecasting. The dynamic spatial dependencies are efficiently captured through a two-phase graph diffusion convolutional network where the attention mechanism is introduced. Moreover, a novel temporal convolution module is designed to learn various ranges of temporal dependencies, including recent, daily, and weekly periods. Inside the module, convolution layers are stacked to handle very long sequences. Experimental results on two large-scale real-world taxi datasets from New York City (NYC) and Chengdu demonstrate that our method significantly outperforms seven state-of-the-art baseline methods.

Keywords: demand forecasting; spatial-temporal dependencies; graph neural networks; attention mechanism

1. Introduction

The popularity of taxi requesting services nowadays has largely changed the travel behavior of people in the urban area. Taxi order forecasting plays a critical role in taxi requesting service as it could influence the preallocation of resources to fulfill the travel demand. Designing more accurate taxi order forecasting models could increase the efficiency of the taxi service and alleviate traffic congestion.

Benefiting from the wide deployment of GPS sensors in taxi vehicles, a large amount of taxi trip data have been collected, which brings opportunities to design more powerful data-driven models to improve the accuracy of taxi demand forecasting. However, taxi order data in real-life scenarios generally follow complex spatial-temporal patterns [1,2]. Figure 1a shows an example of the spatial distribution of one hour's taxi orders in New York City (NYC). It can be observed that the orders also tend to gather around hot spot areas in the city. The temporal distribution of the taxi orders is visualized in Figure 2 where the hourly demand is temporally correlated and contains both short-term and long-term periodicity. Another common pattern that is not shown here but has been observed in previous work is the correlations of demand in distant regions due to similar functionalities [3] or connections by public transportation system [1].

Taxi demand forecasting can be regarded as a special case of a more general spatialtemporal data forecasting problem. In addition to taxi order data [4–7], other types of spatial-temporal datasets have also been studied for prediction, including traffic volume [8–12], traffic flow [2,13–16], and bike-sharing demand [17]. Since taxi orders are



Citation: Luo, A.; Shangguan, B.; Yang, C.; Gao, F.; Fang, Z.; Yu, D. Spatial-Temporal Diffusion Convolutional Network: A Novel Framework for Taxi Demand Forecasting. *ISPRS Int. J. Geo-Inf.* 2022, *11*, 193. https://doi.org/ 10.3390/ijgi11030193

Academic Editor: Wolfgang Kainz

Received: 14 January 2022 Accepted: 10 March 2022 Published: 13 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/). continuously distributed in space, preprocessing is commonly performed to aggregate the data to grids [1,4,6], zones [18], or partitions created from the road network [2]. Consequently, the problem is transformed into predicting a matrix or graph where the challenges lie in modeling the complex and dynamic spatial-temporal dependencies in the demand data.



Figure 1. Spatial distribution of one hour's taxi order data in New York City (**a**) and virtual stations discovered by clustering (**b**). Each dot in (**a**) represents a taxi order origin and the underlying heatmap highlights the hot spots in the city.



Figure 2. Temporal distribution of hourly taxi order along one week period (**a**) and over day of week (**b**).

Conventional travel demand forecasting methods modeled the temporal correlation using time series analysis such as autoregressive integrated moving average (ARIMA) [19–21]. They could be weak in handling the complex spatial-temporal patterns in travel demand data. The recent advances in deep learning have largely promoted the usage of neural network models in travel demand forecasting. Zhang et al. [1] developed ST-ResNet where both local and global regional dependencies were captured by stacking multiple convolutional layers. The same approach was adopted in DMVST-Net [3], where semantic dependency was further considered by constructing a graph to represent the similarity between demand patterns among regions. Graph convolutional network (GCN) was also widely used to model the spatial dependencies in travel demand forecasting. Lin et al. [17] proposed GCN with graph filter for bike-sharing demand prediction where a graph filter encodes multiple features, including spatial distance, demand pattern, average trip duration, etc. Geng et al. [4] developed a multi-graph graph convolutional network to consider three types of adjacency graphs encoding spatial proximity, functional similarity, and transportation connectivity. Bai et al. [5] designed a hierarchical GCN that stacked multiple GCN layers to capture long-term spatial-temporal correlations. Sun et al. [2] fused the output of five GCN layers capturing different types of temporal views. Zhang et al. [18] performed clustering of taxi demand, then designed a multi-level recurrent neural network (MLRNN) to utilize inter-zone heterogeneity to improve the prediction. In order to capture the temporal dynamics, Ye et al. [22] developed a coupled layer-wise graph convolution mechanism where each GCN layer has a different adjacency matrix that is iteratively updated. Some studies further investigated the prediction of demand from an origin to destination (OD) region. Liu et al. [6] performed the convolution on the OD matrix to model the local spatial dependency. Wang et al. [7] developed a multi-task learning scheme with periodic-skip long short-term memory (LSTM) network for predicting the OD matrix and the inbound and outbound traffic flow of a grid.

Although many studies have been conducted to model the spatial-temporal dependencies in taxi demand data, they cannot capture the spatio-temporal dependencies effectively. On the one hand, the problem of the limitations of graph convolutional neural networks is not taken into account by any existing methods. On the other hand, although dilated causal convolution can learn longer-term temporal dependencies compared to Recurrent Neural Network (RNN) methods, it has the problem of gridding effects. Our proposed method, Spatial-Temporal Diffusion Convolutional Network (ST-DCN), effectively addresses these two challenges. The contributions of our work are summarized as follows:

- We design a two-phase graph diffusion convolutional network, which can effectively address the limitations of graph convolutional neural networks. During the diffusion process of the convolution, we use two types of adjacency matrices and introduce the attention mechanism to capture the dynamic spatial dependencies adaptively;
- 2. Hybrid Dilated Causal Convolution is used to capture the temporal dependencies, which can tackle the grid effect problem of conventional dilated convolution. We use a gating mechanism to efficiently control the information flow of nodes and further consider the periodicity of taxi demand data;
- 3. We evaluated our approach on two large-scale real-world datasets. The experimental results demonstrate that ST-DCN outperforms seven existing state-of-the-art base-line methods.

2. Preliminary

Virtual Station: Taxi order requests tend to gather in certain areas in a transportation mode like taxis. For example, at the entrance of a university or a residential area, which unconsciously forms a virtual station, there are usually more distinctive taxi demand characteristics [23]. The discovery of these virtual stations can help capture taxi demand characteristics and make the forecasting more accurate. It is worth mentioning that most existing works on transportation demand forecasting divide the city into grids and then consider each grid as a graph's node. Similar to CCRNN [22], we employ the Density Peak Clustering (DPC) [24] approach to partition regions into virtual stations and treat them as graphs' nodes. It more closely matches the structure of the road network in realistic scenarios and assists in achieving more accurate forecasting results.

Taxi demand forecasting: Given a graph G = (V, E, A), where V represents a set of nodes of the graph (|V| = N), which are virtual stations; E is a set of edges, which represent the connections between nodes. $A \in \mathbb{R}^{N \times N}$ is a weighted adjacency matrix of the graph, where each element A_{ij} stores a weight representing the strength of the connection between node i and j. At time step t, the graph G has a graph signal $X_t \in \mathbb{R}^{N \times C}$, C is the number of feature dimensions of input. Two features are considered, including the number of pick-up and drop-off of each node at time step t. Given a graph G and its history of H time step graph signals, the taxi demand forecasting problem is formulated as finding a mapping

function *f* that can predict its taxi pick-up for the next *P* time steps. The mapping relation can be defined as:

$$f(X_{(t-H+1):t}, G) \to X_{(t+1):(t+P)}$$
 (1)

where $X_{(t-H+1):t} \in \mathbb{R}^{H \times N \times C}$ and $X_{(t+1):(t+P)} \in \mathbb{R}^{P \times N \times C}$.

3. Methodology

In this section, we elaborate the proposed ST-DCN model with the technical details. As shown in Figure 3, the proposed ST-DCN network consists of (a) an input layer, (b) a temporal convolution module, (c) a spatial convolution module, and (d) an output layer. The temporal and spatial convolution modules consist of multiple T-blocks and S-blocks; each block correspondingly consists of stacked temporal and spatial convolution layers. Both temporal and spatial convolution layers are finally incorporated with residual connections to avoid the problem of gradient vanishing [25].



Figure 3. Illustrative architecture of the proposed spatial-temporal diffusion convolutional network (ST-DCN). ST-DCN consists of (**a**) an input layer, (**b**) a temporal convolution module, (**c**) a spatial convolution module, and (**d**) an output layer.

3.1. Spatial Dependency Modeling

The modelling of spatial dependencies is an important prerequisite study for achieving taxi demand forecasting. The rise of various graph neural networks in recent years has facilitated the task of dealing with data types that are graphical. Graph neural networks can be used to model intricate road networks when dealing with the problem of taxi demand forecasting. It addresses the limitations of Convolutional Neural Networks (CNNs) in coping with non-Euclidean data.

This paper applies diffusion convolution proposed by DCRNN [8] and employs the self-adaptive adjacency matrix designed in Graph WaveNet [11] for spatial dependency modeling. Specifically, we use \bar{A} to denote the stationary adjacency matrix where each value stores the distance between two nodes and \tilde{A} to denote the self-adaptive adjacency matrix with the following definition

$$\tilde{A} = softmax(ReLU(M_1, M_2^T))$$
(2)

$$h = \sum_{k=0}^{K} \mathcal{P}^k X W_{k1} + \tilde{A}^k X W_{k2}$$
(3)

where $M_1, M_2 \in \mathbb{R}^{N \times c}$ are source and target node embeddings, \mathcal{P} is the transition matrix, X denotes the input, and W denotes the model parameter matrix.

Equation (3) does not consider the different effects of the spatial dependencies represented by different adjacency matrices, which is important for effective learning spatial dependencies. Similarly, in the diffusion process of convolution, the different influences of each step should also be taken into consideration. Therefore, we adopt a diffusion process of convolution to control the flow of information on the nodes, consisting of two main phases: the information diffusion phase and the information control phase. The information diffusion phase is defined as follows:

$$X_k = \alpha X_{k-1} + (1-\alpha)\tilde{A}X_{k-1} \tag{4}$$

where α is a hyperparameter used to control the retaining rate of the original node's information. The same relation applies to stationary adjacency matrix by just replacing \tilde{A} with \bar{A} in the above equation.

The information diffusion phase will recursively diffuse the information of the nodes along with a given graph structure. One problem that needs to be overcome with graph convolutional networks is that the number of neighborhood nodes will grow exponentially when a multi-layer graph network is used. The problem of over-squashing will occur: a large amount of information about neighboring nodes has to be compressed into the feature vector of a single node [26]. As a result, information cannot be effectively propagated, and the model has poor performance. To solve this problem, we retain a certain percentage of the original information of the nodes during the information diffusion process, which can simultaneously retain the information of the original nodes and can effectively deepen the exploration of the neighboring nodes.

Graph convolutional networks also face the problem of over-smoothing [27,28]. After multiple graph convolutional layers, node features converge to the same or similar vectors, making them indistinguishable. The information control phase is adopted to address this problem effectively and can control the information generated by the nodes. Here, we use the attention mechanism [29] to control the information flow of nodes adaptively. The attention mechanism can concentrate limited attention on important information, thus saving computing resources and quickly acquiring the most helpful information. After combining the two phases of the diffusion process of convolution, Equation (3) will become the following Equation (6):

$$W_{i} = \frac{exp(Conv(X_{i}))}{\sum_{j=1}^{K} exp(Conv(X_{j}))}$$
(5)

$$h = \sum_{i=1}^{K} W_i X_i \tag{6}$$

where *K* is the depth of information diffusion, *X* is the output of the previous step of information diffusion, which is used as the input for the subsequent information diffusion, and *W* is the self-learned weights coefficient using the attention mechanism.

3.2. Temporal Dependence Modeling

In this section, we first discuss the importance of accounting for temporal periodicity when capturing temporal dependencies. Secondly, we describe the concept of conventional dilated causal convolution and its advantage over RNN to capture long-term temporal dependencies effectively. Then, Hybrid Dilated Convolution (HDC) is used to solve the gridding effect problem in conventional dilated convolution. Finally, to effectively control the information flow of nodes, a gating mechanism is used to improve the model's performance further. More specifically, the details of the temporal dependence model are presented as follows.

Temporal periodicity: Taxi demand data usually exhibit a strong daily or weekly periodic pattern. Figure 2 provides an example of one week's taxi demand data in New York. It can be observed that the demand curves from Monday to Friday are quite different from those on weekends.

Similar to ASTGCN [9] and ST-ResNet [1], this paper also considers taxi demand data's recent, daily, and weekly dependencies. Assuming that the current time is τ_0 , the historical time window size is T_H , the size of the time window to be predicted is T_P . The blue, red, and green parts in Figure 4 indicate the recent, daily, and weekly periods, respectively.



Figure 4. An example of constructing the input of time series segments (suppose the size of both historical and forecasting windows are 30 min).

It is necessary to note that in our model: $T_H \ge T_P$. Because the periodicity of taxi demand will have some fluctuation, it is not strictly periodic [13]. For example, the peak hours on weekdays usually fluctuate in the afternoon between 17:30 p.m. and 19:30 p.m.

Dilated Causal Convolution: The dilated causal convolution networks can exponentially increase the receptive field by stacking the depth of the network layers. Compared to RNN-based methods, dilated causal convolution networks can tackle long-term sequences in a non-recursive manner, enabling parallel computation and alleviating the gradient explosion issue [30]. Dilated causal convolutional networks keep the chronological causality sequence by padding zeros to the inputs. This way, it ensures that only historical information is used to predict without leaking any future information. More formally, for a one-dimensional sequence of inputs $X \in \mathbb{R}^T$ and the filter $f : \{0, ..., n - 1\}$, the dilation convolution operation F in the input sequence with element t can be defined as:

$$F(t) = \sum_{i=0}^{n-1} f(i) X_{t-d \cdot i}$$
(7)

where *d* is the dilation rate, *n* is the filter size, and $t - d \cdot i$ represents the past direction.

Hybrid Dilated Convolution: Wang et al. [31] points out that the conventional dilated convolution framework has the problem of gridding, i.e., dilated convolution inserts zero values between two sampled pixels of the convolution kernel. If the dilation rate becomes too large, the convolution will be too sparse and detrimental to learning because not all pixels are involved in the computation. This way, one will lose the consistency of information, which is fatal for pixel-level tasks (Figure 5a). Therefore, this paper use HDC to overcome the problems caused by the gridding effect. HDC uses a series of dilation rates, rather than a single one, to make the final receptive field fully cover the entire region with no holes or missing edges. At the same time, the receptive field of the network is also expanded to aggregate global information.

Hybrid dilated convolution is a simple solution proposed to overcome the gridding effect, which has the following three main features:

- 1. The dilation rate of a stacked dilated convolution should not have a common factor greater than 1. For example, [2, 4, 6] would not be a suitable three-layer convolution as it still has gridding effects;
- 2. The dilation rate is designed as a jagged structure, e.g., a cyclic structure like [1, 2, 5, 1, 2, 5];

3. The dilation rate needs to satisfy the equation:

$$M_{i} = max(M_{i+1} - 2d_{i}, M_{i+1} - 2(M_{i+1} - d_{i}), d_{i})$$
(8)

where the d_i is the dilation rate of the *i*-th layer, and M_i is the maximum dilation rate at the *i*-th layer. Assuming there are *n* layers and the default is $M_n = d_n$. If applied to a convolution kernel with size $k \times k$, the goal is to let $M_2 \leq k$.

As shown in Figure 5, increasing the dilation rates tend to change its focus from the local features to global ones. By only using a small number of dilated convolution layers, the receptive field can be significantly increased.



Figure 5. Illustration of the gridding problem. Dilated casual convolution with kernel size 2. (**a**) all convolutional layers have a dilation rate d = 2. (**b**) subsequent convolutional layers have dilation rates of d = 1, 2, 4, respectively.

Gated TCN: We adopt the Gated TCN designed by Graph WaveNet [11] to control the inflow of valid information and discard invalid information in the TCN. One temporal convolution is followed by a tangent hyperbolic activation function working as a filter. The other temporal convolution is followed by a sigmoid activation function that acts as a gate to control the amount of information passing out. Specifically, the Gated TCN takes the form:

$$Z = tanh(\Theta_1 * X + b_1) \odot \sigma(\Theta_2 * X + b_2)$$
(9)

where Θ_1 , Θ_2 , b_1 and b_2 are learnable parameters, \odot denotes the element-wise multiplication operator, $\sigma(\cdot)$ is a sigmoid function, * is the dilated convolution operation. Figure 6 illustrates the structure of Gated TCN.



Figure 6. Components of temporal block.

3.3. Extra Components

Skip Connection: As the depth of the network increases, it causes extra problems of gradient vanishing or explosion, which makes the training of deep learning models difficult. Meanwhile, Orhan and Pitkow [32] demonstrate that skip connection breaks the symmetry of the network forcibly and alleviates the degradation of the neural network. Therefore, we introduce skip connection to enhance the learning capability of the network, which can acquire activation from one network layer and then quickly give feedback to another layer or even deeper layers of the neural network.

Output Module: To achieve the goal of multi-step taxi demand forecasting, the output module of our ST-DCN network consists of a Multi-Layer Perceptron (MLP) and two 1×1 standard convolutional layers that convert the input dimensions into the desired output dimensions. ST-DCN treats the output $X_{(t+1):(t+P)}$ as a whole, which can effectively handle the dimensional inconsistency problem between training and testing. We can use the historical *H* consecutive time steps to predict the future *P* consecutive steps, just to set the temporal size of the expected output as *P*.

4. Experiments

4.1. Experimental Settings

Dataset Description: Experiments are conducted on two real-world datasets collected from NYC OpenData and Didi Chuxing.

- NYC Taxi (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page (accessed on 5 May 2021)): This dataset includes 91 days of 35 million taxi trip records of NYC in yellow taxis from 1 April 2016 to 30 June 2016.
- Didi Taxi (https://outreach.didichuxing.com/research/opendata/en/ (accessed on 7 July 2021)): The dataset contains taxi requests from 1 November 2016 to 30 November 2016 for the city of Chengdu with more than 7 million taxi trip records.

We only utilize the following information: pick-up and drop-off dates/times, pick-up and drop-off locations. In the experiments, we divide the training dataset, validation dataset, and testing dataset into the ratio of 7:1.5:1.5.

Preprocessing: We preprocess the data following the approach used in CCRNN [22]. The raw taxi records are aggregated into a 30 minute time window where missing values are replaced with zero and outliers are filtered out. We use a sliding window on training, validating, and testing data for sample generation. Z-score normalization is adopted to standardize the data inputs. The station-less NYC taxi orders are clustered into 248 virtual stations, as shown in Figure 1b. Chengdu Taxi orders are aggregated into 34 virtual stations;

Parameter setting: All experiments are conducted under the environment with one Intel(R) Xeon(R) Gold 6132 CPU @ 2.60 GHz and one NVIDIA Tesla P40 GPU card. The input data has dimension C of 2. We use the historical H = 12 continuous time steps to predict the taxi demand in the next $P \in \{3, 6, 12\}$ time intervals (i.e., short, mid, long-term) when testing the prediction result.

To cover the input sequence length, we use 9 layers Gated TCN with a sequence of dilation rates of [1, 2, 5, 1, 2, 5, 1, 2, 5]. We use Equation (6) as our graph convolution layer with a diffusion step K = 3. Our model is trained by the Adam optimizer [33] with an initial learning rate of 0.0015 and decays at a rate of 0.2 for every 5 epochs. Dropout is set as 0.3. The retain ratio from the information diffusion is set to 0.05. We also use the validation dataset with patience of 20 to early-stop our training algorithm for each model based on the best validation score.

We use three evaluation metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson Correlation Coefficient (PCC), to evaluate the performance of all methods. RMSE between the estimator and the ground truth is used as the loss function.

4.2. Baselines

This paper only compares our model with more recent deep learning models. We compare the performance of our proposed model (ST-DCN) with the following seven baselines:

- LSTM [34]: Long Short-Term Memory Network, a special RNN model for time series prediction;
- DCRNN [8]: Diffusion Convolutional Recurrent Neural Network, which combines diffusion graph convolutional networks with GRU in an encoder-decoder manner;
- STGCN [35]: A Spatial-Temporal Graph Convolutional Network uses ChebNet graph convolution and 1D convolutional networks to capture spatial dependencies and temporal correlations, respectively;
- GWNet [11]: A Spatial-Temporal Graph Convolutional Network integrates adaptive adjacency matrix into diffusion graph convolutions with 1D dilated casual convolutions;
- ASTGCN [9]: Attention Based Spatial-Temporal Graph Convolutional Networks, which introduces spatial attention and temporal attention mechanisms to model spatial and temporal dynamics, respectively. For fairness, we only take its recent components;
- MTGNN [36]: A Graph Neural Network designed for multivariate time series forecasting by adding a graph learning layer to capture the hidden relationships among time series data;
- CCRNN [22]: A Coupled Layer-wise Graph Convolution designed for transportation demand prediction.

4.3. Performance Comparison

Table 1 demonstrates the results of ST-DCN and baselines on the dataset NYC taxi. It shows that our ST-DCN outperforms other baseline models consistently and overwhelmingly in all metrics except the PCC reported from the short-term prediction experiment with p = 3. More specifically, our ST-DCN method achieves 4.31%, 4.04%, and -0.07% relative improvement when p = 3; 7.22%, 8.37%, and 0.44% relative improvement when p = 6; 9.89%, 9.16%, and 0.49% relative improvement when p = 12 over the best performance among baseline methods, respectively. Table 2 demonstrates the results of ST-DCN and baselines on the Chengdu Taxi dataset. It shows that our ST-DCN outperforms other baseline models consistently and overwhelmingly in all metrics. More specifically, our ST-DCN method achieves 8.78%, 10.91%, and 0.07% relative improvement when p = 3; 15.59%, 14.59%, and 0.09% relative improvement when p = 6; 15.74%, 11.40%, and 0.08% relative improvement when p = 12 over the best performance among baseline models improvement when p = 12 over the best performance among baseline.

 1		
 n=3	n = 6	n = 12

Models	<i>p</i> = 3			<i>p</i> = 6			<i>p</i> = 12		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
LSTM	22.2593	35.3812	0.0396	22.2777	35.3053	0.0846	22.3101	35.3657	0.0744
DCRNN	5.2734	8.7323	0.9691	5.3217	8.9063	0.9679	5.4931	9.1450	0.9665
ASTGCN	5.4692	9.4815	0.9650	5.3771	9.4569	0.9638	5.6197	9.9337	0.9608
MTGNN	5.4587	9.2379	0.9654	6.1552	10.4912	0.9554	7.3898	12.7436	0.9344
STGCN	6.2332	10.5332	0.9547	6.4520	10.8703	0.9517	6.6751	11.2684	0.9485
GWNet	5.2345	8.7947	0.9717	5.1035	8.8489	0.9690	5.3518	9.2376	0.9674
CCRNN	4.8576	8.2347	0.9754	5.2650	9.1107	0.9699	5.4746	9.5675	0.9672
ST-DCN	4.6481	7.9022	0.9747	4.7352	8.1085	0.9742	4.8226	8.3074	0.9721

Models		<i>p</i> = 3			<i>p</i> = 6		<i>p</i> = 12		
Models LSTM DCRNN ASTGCN MTGNN STGCN GWNet CCRNN ST-DCN	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
LSTM	186.5467	316.6753	0.0491	186.3385	316.9454	0.0271	185.8830	315.5982	0.0944
DCRNN	13.1659	25.6185	0.9967	13.3274	25.9840	0.9967	13.7034	25.9839	0.9966
ASTGCN	17.0968	34.7806	0.9940	17.2186	34.7206	0.9940	18.2122	37.4844	0.9930
MTGNN	14.2991	27.8156	0.9963	15.0434	30.1220	0.9956	16.2260	32.4549	0.9948
STGCN	16.5549	34.1537	0.9944	17.7299	37.1672	0.9932	19.9882	41.4199	0.9917
GWNet	12.0914	24.5351	0.9970	13.7161	27.4267	0.9963	14.1064	29.5471	0.9957
CCRNN	14.6755	28.3034	0.9962	15.2797	31.2508	0.9955	19.3367	40.3984	0.9933
ST-DCN	11.0293	21.8581	0.9977	11.2490	22.1927	0.9976	11.5460	23.0216	0.9974

Table 2. Performance comparison of ST-DCN and other baseline models on the Chengdu Taxi dataset.

The low performance of LSTM indicates the limitation of considering only temporal correlations and the necessity of utilizing the spatial dependencies of the spatial-temporal network. Methods like STGCN, DCRNN, and ASTGCN highly rely on a predefined graph, which may not capture crucial dependencies between nodes, therefore leading to worse performance. However, thanks to combining the encoder-decoder architecture for time series prediction with graph convolution, DCRNN has better performance. Benefiting from the self-learned adjacency matrix, MTGNN achieves competitive accuracy in short-term forecasting experiments. Although less competitive than our model, GWNt and CCRNN still report relatively high accuracies, which might be explained by adopting adaptive graphs in modelling relationships between nodes. It indicates that adaptive graph-based methods could effectively exploit valuable and latent spatial dependencies from historical taxi demand data.

Figure 7 shows the comparison of the forecasting results of various methods as the forecasting time increases. We exclude the results of LSTM since it performs poorly. Overall, as the forecasting time becomes longer, the forecasting becomes more difficult, and therefore the forecasting error becomes larger. As it is shown in the figure, MTGNN performs well compared to STGCN for short-term forecasting. However, when the forecasting time increases, its forecasting accuracy drops dramatically. The errors of the other approaches increase slowly when the forecasting time becomes longer, and their overall performance is relatively good. Our ST-DCN model achieves the best forecasting performance at all forecasting times, and its errors are the smallest and increase the slowest, indicating that our model is highly stable. All of these results suggest the effectiveness of our proposed method for spatiotemporal correlations modelling.



Figure 7. Comparison of the performance of different methods as the forecasting time increases on the NYC taxi dataset.

4.4. Component Analysis

To further evaluate the effect of different components of ST-DCN, we design six variants of the ST-DCN model. We compare these six variants with the ST-DCN model on the NYC Taxi dataset when p = 12. The difference between these seven models are described as below:

- 1. Basic: This model does not equip with hybrid dilated convolution, two-phase graph diffusion convolution, and temporal periodicity;
- 2. +HDC: This model uses hybrid dilated convolution to overcome the gridding effect;
- 3. Two-phase: This model uses two-phase graph diffusion convolution to address two limitations of graph convolution, but it does not employ hybrid dilated convolution;
- One T-block (1 day): This model considers the daily period in one T-block (only yesterday is included);
- 5. Multi T-block (1 day): This model considers the daily period in multi T-blocks (only yesterday is included);
- 6. One T-block (7 day): This model considers the daily and weekly period in one T-block;
- ST-DCN (multi T-block (7 day)): This model considers the daily and weekly period in multi T-blocks. It is the complete version of our proposed approach ST-DCN.

As shown in Table 3, we can observe that the complete version of ST-DCN outperforms other variants. The impact of HDC is significant in terms of MAE but less apparent in RMSE. The evident effect of two-phase graph diffusion convolution indicates the effectiveness of selecting useful information at each convolutional diffusion process. Compared with the model considering only daily periodicity, introducing the weekly periodicity into the model also improves its accuracy. In addition, the model outperforms its competitors after using multiple T-blocks instead of only one to process all the temporal dependencies. Hence, each designed sub-module has positive effects for forecasting performance improvement.

Models	MAE	RMSE	PCC
basic	5.4035	9.2762	0.9664
+HDC	5.3371	9.2543	0.9655
two-phase	5.3327	9.1792	0.9671
one T-block (1 day)	5.2181	9.0306	0.9679
multi T-block (1 day)	5.1653	8.8413	0.9685
one T-block (7 day)	5.0545	8.8018	0.9698
ST-DCN (multi T-block (7 day))	4.8226	8.3074	0.9721

 Table 3. Evaluation of different variants on the NYC taxi dataset.

5. Discussion

It is necessary to model the spatio-temporal information effectively to improve the taxi demand forecasting accuracy. Compared with HA, ARIMA, and LSTM, which only consider temporal information. ST-ResNet, STDN [13], and DMVST-Net, which combine spatio-temporal information, have improved forecasting accuracy, although these methods use CNN to obtain spatial information. The main idea of such methods is to consider traffic data like images and process their spatial correlation by CNN. However, in traffic forecasting tasks, the distribution properties of the data spatially are different from images, so there are limitations in the application of CNN-based methods to traffic problems. For example, in the taxi demand forecasting problem, there may be time-delayed correlations in the data of the origin and destination spatially. The origin-destination hotspot areas may cross all regions in the network. Data from regions with the same attributes are also spatially correlated, and their distributions are not restricted to fixed geometric regions.

Due to the ability of GCN to model complex road networks, scholars have used GCN-based methods for traffic forecasting in recent years. For example, STG2Seq [5], STS-GCN [14], STFGNN [16], and the baselines method chosen in this paper aim to improve the

adjacency matrix of GCN. However, they all missed the limitations of graph convolutional neural networks, which is one of the difficulties overcome in this paper.

In terms of temporal dependence, most deep learning-based models use RNN methods, such as ST-ResNet, STDN, DMVST-Net, DCRNN, CCRNN, etc. However, from the model optimization standpoint, RNNs cannot capture long-term dependencies well and suffer from gradient disappearance or explosion problems. There are also approaches using TCN, such as Graph WaveNet, STFGNN, and MTGNN, which cannot effectively improve the forecasting accuracy due to the grid effect problem of conventional dilated convolution. ST-DCN uses TCN to capture the long-term temporal dependence while using hybrid dilated convolution to overcome the grid effect problem, enabling ST-DCN to achieve high forecasting accuracy. Whether using NYC taxi's three-month dataset or Chengdu taxi's one-month dataset, the ST-DCN achieves state-of-the-art forecasting accuracy, which also proves the effectiveness of ST-DCN.

It should be mentioned that although ST-DCN can achieve high forecasting accuracy, it requires more memory and a longer training time compared to other methods. Although ST-DCN uses two types of adjacency matrix to capture spatial dependencies adaptively, it is essentially still a fixed graph structure, and the model's effectiveness may be further improved if dynamic graphs can be used to model spatial dependencies. The ST-DCN uses separate modules to capture temporal and spatial correlations, not simultaneously, which ignores the heterogeneity in spatio-temporal data.

6. Conclusions and Future Work

This paper proposes a novel spatial-temporal diffusion convolutional model called ST-DCN and successfully applies it to forecasting taxi demand. ST-DCN could capture spatial dependencies effectively in a two-phase graph diffusion convolutional network. Furthermore, our model considers the dynamic attribute in spatial correlation by using the attention mechanism. ST-DCN can learn long-term temporal dependencies through a hybrid dilated convolution, which stacks its convolutional layers exponentially to increase the receptive field. Moreover, we also consider the temporal periodicity to obtain more accurate prediction results. Experiments on two large-scale real-world taxi datasets demonstrate that our method can achieve state-of-the-art prediction performance, which illustrates the superiority of our model.

For future work, we will further optimize the network structure and parameter settings. Moreover, we plan to apply the proposed model to other spatial-temporal forecasting tasks. In addition, taxi demand is also affected by many external factors, such as weather and urgent events. In the future, we will take some external influences into account to further improve forecasting accuracy.

Author Contributions: Conceptualization, Aling Luo; Data curation, Aling Luo; Formal analysis, Aling Luo and Can Yang; Investigation, Aling Luo, Boyi Shangguan, Can Yang, Fan Gao, Zhe Fang and Dayu Yu; Methodology, Aling Luo; Resources, Boyi Shangguan; Supervision, Can Yang; Validation, Aling Luo; Visualization, Aling Luo and Can Yang; Writing—original draft, Aling Luo; Writing—review & editing, Aling Luo, Boyi Shangguan, Can Yang, Fan Gao, Zhe Fang and Dayu Yu. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Hubei Provincial Natural Science Foundation of China (Grant Number: 2020CFA001) and the National Key Research and Development Program of China (Grant Number: 2020YFC1512003)

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The New York City Taxi dataset that supports the findings of this study is available at https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page (accessed on 5 May 2021). The Didi Taxi dataset can be downloaded from https://outreach.didichuxing.com/research/opendata/en/ (accessed on 7 July 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

- 1. Zhang, J.; Zheng, Y.; Qi, D. Deep spatio-temporal residual networks for citywide crowd flows prediction. In Proceedings of the AAAI, San Francisco, CA, USA, 4–9 February 2017; pp. 1655–1661.
- 2. Sun, J.; Zhang, J.; Li, Q.; Yi, X.; Liang, Y.; Zheng, Y. Predicting Citywide Crowd Flows in Irregular Regions Using Multi-View Graph Convolutional Networks. *IEEE Trans. Knowl. Data Eng.* **2020**, *14*. [CrossRef]
- 3. Yao, H.; Wu, F.; Ke, J.; Tang, X.; Jia, Y.; Lu, S.; Gong, P.; Li, Z.; Ye, J.; Chuxing, D. Deep multi-view spatial-temporal network for taxi demand prediction. *arXiv* 2018, arXiv:1802.08714.
- 4. Geng, X.; Li, Y.; Wang, L.; Zhang, L.; Yang, Q.; Ye, J.; Liu, Y. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In Proceedings of the AAAI, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 3656–3663.
- Bai, L.; Yao, L.; Kanhere, S.S.; Wang, X.; Sheng, Q.Z. STG2seq: Spatial-temporal graph to sequence model for multi-step passenger demand forecasting. *IJCAI* 2019, 1981–1987. [CrossRef]
- 6. Liu, L.; Qiu, Z.; Li, G.; Wang, Q.; Ouyang, W.; Lin, L. Contextualized Spatialoral Network for Taxi Origin-Destination Demand Prediction. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 3875–3887. [CrossRef]
- Wang, Y.; Wo, T.; Yin, H.; Xu, J.; Chen, H.; Zheng, K. Origin-destination matrix prediction via graph convolution: A new perspective of passenger demand modeling. In Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1227–1235. [CrossRef]
- 8. Li, Y.; Yu, R.; Shahabi, C.; Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv* 2018, arXiv:1707.01926.
- 9. Guo, S.; Lin, Y.; Feng, N.; Song, C.; Wan, H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In Proceedings of the AAAI, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 922–929.
- 10. Zhao, L.; Song, Y.; Zhang, C.; Liu, Y.; Wang, P.; Lin, T.; Deng, M.; Li, H. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 3848–3858. [CrossRef]
- 11. Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Zhang, C. Graph wavenet for deep spatial-temporal graph modeling. *IJCAI* 2019, 1907–1913. [CrossRef]
- 12. Shi, X.; Qi, H.; Shen, Y.; Wu, G.; Yin, B. A spatial-temporal attention approach for traffic prediction. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 4909–4918. [CrossRef]
- 13. Yao, H.; Tang, X.; Wei, H.; Zheng, G.; Li, Z. Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. In Proceedings of the AAAI, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5668–5675.
- Song, C.; Lin, Y.; Guo, S.; Wan, H. Spatial-temporal synchronous graph convolutional networks: A new framework for spatialtemporal network data forecasting. In Proceedings of the AAAI, New York, NY, USA, 7–12 February 2020; pp. 914–921. [CrossRef]
- 15. Zhang, X.; Huang, C.; Xu, Y.; Xia, L.; Dai, P.; Bo, L.; Zhang, J.; Zheng, Y. Traffic Flow Forecasting with Spatial-Temporal Graph Diffusion Network. In Proceedings of the AAAI, Virtual, 2–9 February 2021; Volume 35, pp. 15008–15015.
- Li, M.; Zhu, Z. Spatial-Temporal Fusion Graph Neural Networks for Traffic Flow Forecasting. In Proceedings of the AAAI, Virtual, 2–9 February 2021; Volume 35, pp. 4189–4196.
- 17. Lin, L.; He, Z.; Peeta, S. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transp. Res. Part C Emerg. Technol.* **2018**, *97*, 258–276. [CrossRef]
- Zhang, C.; Zhu, F.; Lv, Y.; Ye, P.; Wang, F.Y. MLRNN: Taxi Demand Prediction Based on Multi-Level Deep Learning and Regional Heterogeneity Analysis. *IEEE Trans. Intell. Transp. Syst.* 2021, 1–11. [CrossRef]
- 19. Kaltenbrunner, A.; Meza, R.; Grivolla, J.; Codina, J.; Banchs, R. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive Mob. Comput.* **2010**, *6*, 455–466. [CrossRef]
- Moreira-Matias, L.; Gama, J.; Ferreira, M.; Mendes-Moreira, J.; Damas, L. Predicting taxi-passenger demand using streaming data. *IEEE Trans. Intell. Transp. Syst.* 2013, 14, 1393–1402. [CrossRef]
- 21. Williams, B.M.; Hoel, L.A. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *J. Transp. Eng.* 2003, 129, 664–672. [CrossRef]
- 22. Ye, J.; Sun, L.; Du, B.; Fu, Y.; Xiong, H. Coupled Layer-wise Graph Convolution for Transportation Demand Prediction. In Proceedings of the AAAI, Virtual, 2–9 February 2021; Volume 35, pp. 4617–4625.
- 23. Du, B.; Hu, X.; Sun, L.; Liu, J.; Qiao, Y.; Lv, W. Traffic demand prediction based on dynamic transition convolutional neural network. *IEEE Trans. Intell. Transp. Syst.* 2020, 22, 1237–1247. [CrossRef]
- 24. Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. Science 2014, 344, 1492–1496. [CrossRef] [PubMed]
- 25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NE, USA, 27–30 June 2016; pp. 770–778.
- Alon, U.; Yahav, E. On the Bottleneck of Graph Neural Networks and its Practical Implications. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
- Oono, K.; Suzuki, T. Graph Neural Networks Exponentially Lose Expressive Power for Node Classification. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.

- Li, Q.; Han, Z.; Wu, X.M. Deeper insights into graph convolutional networks for semi-supervised learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
- 29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 2017, 30, 5999–6009.
- 30. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271.
- Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding convolution for semantic segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1451–1460.
- 32. Orhan, E.; Pitkow, X. Skip Connections Eliminate Singularities. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
- 33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. arXiv 2014, arXiv:1412.6980.
- 34. Hochreiter, S.; Schmidhuber, J. Long short-term memory. Neural Comput. 1997, 9, 1735–1780. [CrossRef] [PubMed]
- 35. Yu, B.; Yin, H.; Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv* **2017**, arXiv:1709.04875.
- 36. Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Chang, X.; Zhang, C. Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks. *arXiv* 2020, 753–763. arXiv:2005.11650.