IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 8, 2021, accepted February 1, 2021, date of publication February 3, 2021, date of current version February 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3056926

Modeling Dynamic Spatio-Temporal Correlations for Urban Traffic Flows Prediction

NABEELA AWAN[1], (Member, IEEE), AHMAD ALI[2], FAZLULLAH KHAN[3], MUHAMMAD ZAKARYA[3], RYAN ALTURKI[4], MAHWISH KUNDI[3], MOHAMMAD DAHMAN ALSHEHRI[5], AND MUHAMMAD HALEEM[6]

[1]Department of Computer Science, Government Girls Degree College at Mardan, Mardan 23200, Pakistan
[2]School of Electronics Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[3]Department of Computer Science, Abdul Wali Khan University, Mardan 23200, Pakistan
[4]Department of Information Sciences, College of Computer and Information Systems, Umm Al-Qura University, Makkah 24372, Saudi Arabia
[5]Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
[6]Department of Computer Science, Faculty of Engineering and Technology, Kardan University, Kabul 1003, Afghanistan

Corresponding author: Fazlullah Khan (fazlullah@awkum.edu.pk)

This work was supported by the Taif University, Taif, Saudi Arabia, through the Taif University Researchers Supporting Project, under Grant TURSP-2020/126.

ABSTRACT Prediction of traffic crowd movement is one of the most important component in many applications' domains ranging from urban management to transportation schedule. The key challenge of citywide crowd flows prediction is how to model spatial and dynamic temporal correlation. However, in recent years several studies have been done, but they lack the ability to effectively and simultaneously model spatial and temporal dependencies among traffic crowd flows. To address this issue, in this article a novel spatio-temporal deep hybrid neural network proposed termed STD-Net to forecast citywide crowd traffic flows. More specifically, STD-Net contains four major branches, i.e., closeness, period volume, weekly volume, and external branches, respectively. We design a residual neural network unit for each property to depict the spatio-temporal features of traffic flows. For various branches, STD-Net provides distinct weights and then combines the outputs of four branches together. Extensive experiments on two large-scale datasets from New York bike and Beijing taxi have demonstrated that STD-Net achieves competitive performances the existing state-of-the-art prediction baselines.

INDEX TERMS Deep learning, urban crowd flows, neural networks, long short-term memory, convolutional neural network.

I. INTRODUCTION

In urban computing environments, the future of the urban traffic crowd flows prediction, i.e., the amount of inflow and outflow of taxi, buses, and pedestrians is a critical research problem. This problem is very critical in the context of public-safety and traffic management in city-wide planning [1]. Based on these works, a city is partitioned into a grid map based on the longitude and latitude, as shown in figure 1, respectively. A large number of traffic conditions occur because of many complex factors, i.e., temporal changes, road networks, and accidents. Predicting traffic crowd flows is an important part of intelligent transportation systems (ITS), particularly on highways with quick driving speed and a huge amount of traffic crowd flows. Since the highway is largely closed, it will significantly impact the traffic flow efficiency if congestion occurs. Traffic crowd flows is an important factor that actively reflecting the highway state. According to this, if it can forecast correctly in advance since the authorities of traffic management would be able to guide the flow of crowds more appropriately, so as, to boost the performance of the highway network, as shown in Figure 1.

The two major limitations of previous studies by using deep learning approaches for urban crowd flow prediction. Firstly, the spatial dependency between the regions relies on the similarity of historic traffic data [1], [2] as well as the model learn spatial dependency statically. However, between the regions, dependencies change over time, for example during the early morning, the dependency between business and residential area is very strong while in the late evening

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Liu.

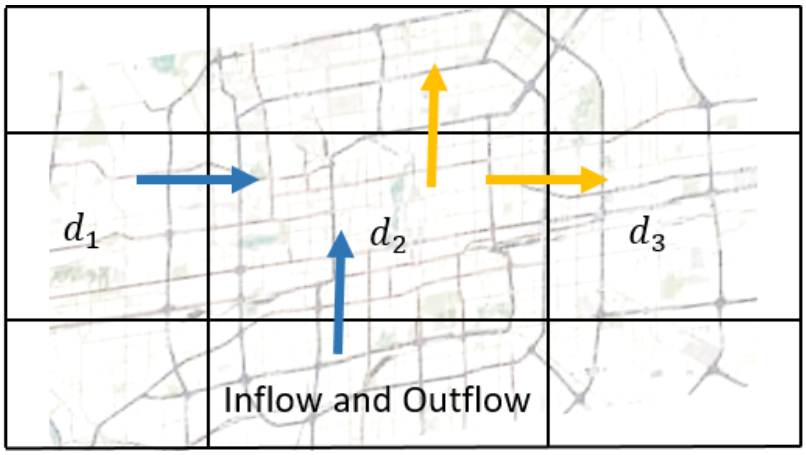

FIGURE 1. Traffic flows of urban crowd.

the dependency between the working area and restaurant might be strong. Secondly, the existing approaches avoid the shifting of long-term periodic dependency. One issue is that urban crowd flows data not firmly periodic, for example, the peak hours on weekdays might vary from 4 PM to 6 PM on different days. According to [1], [2], the authors described periodic crowd flows traffic data, but they fail in sequential dependency and temporal data.

To tackle the above challenges, a novel spatio-temporal dynamic neural networks proposed, namely STD-Net that dynamically predicts urban traffic crowd flows. To the best of our knowledge, the proposed STD-Net is the first work that combines a Convolutional neural network (CNN) and long short-term memory (LSTM) for solving the urban traffic prediction problem. The key limitation of ST-ResNet model is too lacking the ability to model temporal correlation in a timely manner. The ST-RseNet model only focus on spatial correlation, but ignores the temporal correlation. ST-ResNet used a simple combination of spatial features along with various cycles and did not understand the trends of traffic scenarios of crowd flows. However, they did not well performed to depict the features when included a temporal extra dimension. The key novelty of our proposed framework is to learn dynamically the spatio-temporal correlation of urban traffic flows prediction. The proposed STD-Net model captures long-term spatio-temporal correlations via ConvLSTM network. In addition, our approach uses the LSTM model, in a hierarchical way, in order to manage sequential dependency. Briefly, this article has the following main contributions:

- we propose a novel dynamic deep hybrid Spatio-temporal prediction model that explore the spatial dependencies as well as temporal dependencies at the same time in any two regions of a city
- STD-Net summarize the temporal dependencies into three sections, i.e., temporal closeness, daily-periodic, and weekly-periodic
- the proposed model combines the results of three components respectively and assigned weights according to them
- extensive experiments on two real-world datasets show the advantage of STD-Net over several state-of-the-art existing baselines.

We organize the rest of our paper as follows. We offer an overview of the related work in Section II. Section III describes some preliminary work along with the problem formulation for the city-wide crowd flows. In Section IV, we discuss the proposed STD-Net model for traffic prediction. In Section V, we evaluate the performance of the proposed model based on two real-world datasets. Finally, Section VI concludes the paper along with several directions for future research.

II. RELATED WORK

The prediction of citywide crowd flows problem has gained large attention from the last decades. The main aim of urban traffic crowd flows is to predict the crowd traffic value for a location at a timestamp by using historical data. In this section, we described in more detail the related work of urban traffic flows. In [1], the authors proposed a deep learning novel residual model for the prediction of crowd inflow and outflow in a city. In this research work [3], they proposed a novel architecture to estimate traffic volume at citywide using GPA Beijing taxis trajectories data. The same research work has been conducted for the utilization of current development in the machine learning area for the prediction of crowd flows in the urban computing environment. There are some existing, published, research work related to crowd flow prediction available that predict each moment based on their location information as illustrated in [4], [5]. Some other researchers [6]–[8] have proposed to forecast speed of travel volume on the road. The majority of them focus on multiple or single segments of the road instead of citywide segments.

TABLE 1. Summary of the related works.

Baselines	Spatial		Temporal			Work in this paper
	intra	inter	closeness	period volume	weekly volume	
HA [9]	×	×	×	implicit	implicit	×
SARIMA [10]	×	×	explicit	×	×	×
MST3D [11]	✓	✓	explicit	explicit	explicit	✓
DHSTNet [12]	✓	✓	explicit	explicit	explicit	✓
STD-Net	✓	✓	explicit	explicit	explicit	✓

Currently, research scholars started working on citywide crowd flows prediction task [2], [13]. But, their research work is different from our proposed work because they just focus on an individual area, not the overall city as well as they do not divide the city in grid-based. Most of the research scholars applied CNN for different kinds of problems specifically in the computer vision area [14]. According to [15] presented residual learning that allows those networks which have a more deep structure. According to [16], they used a recurrent neural network (RNN) for sequence learning tasks as well as researchers used long short term memory (LSTM) that enables CNN to understand long term temporal dependency. But both methods just capture spatial or temporal dependencies.

Currently, scholars combined both of them methods called the ConvLSTM network [17] to simultaneously capture both spatial and temporal dependencies. For urban crowd flows prediction the machine learning approaches i.e. k-nearest neighbours (KNN) [18] and support vector regression (SVR) [19]. However, the output of SVR and KNN manually extracts the features from traffic data. With the enhancement of machine learning approaches i.e. CNN for image recognition and LSTM for video tracing, there is some research work published on urban traffic flow prediction. According to [3], [20], [21], the researchers described the CNN-LSTM model to evaluate mitotic events in patch sequences of variable length. The simulation results show outstanding the gradational graph model-based method with a large margin using two datasets i.e. C3H10 and C2C12. Mostly they forecast even billions of individual mobility traces instead of aggregated flows of the crowd in a region. Similarly [12], they proposed an approach namely DHSTNet to forecast citywide traffic crowd flows prediction using ConvLSTM network. In addition, some studies first depict the states of the traffic into images and then 2D CNNs applied for citywide crowd flows prediction [22]. To deal with the constraints and to properly capture the spatial-temporal the combinations of RNNs and CNNs are easily considered to be associated [17], [23]–[25]. Those kinds of jobs required such a huge number of computational resources as well as not always necessary for the public safety application scenario.

In the human visual system, visual attention is a central point of view. It refers to the mechanism by which individuals concentrate the statistical assets of their brains on specific areas of the visual field when observing the external environment. In neural networks, an attention-based mechanism were successfully applied to handle various tasks, such as the answering of vision questions [26], natural language processing, image caption [27], [28], machine translation [29] and speech recognition [30], [31]. The main objective of the attention-based mechanism is to identify truly challenging information from all inputs to the current task. In the image caption task [27], they proposed two attention-based models and employ a visualization method to show the effect of attention-based mechanisms. To predict a time series [32], they proposed a multi-level attention model to adjust the correlations between multiple sensor time series. Similarly, for speech recognition [30], they proposed an RNN bi-directional and an attention mechanism for textual and visual question answering. However, these methods are time-consuming since for each time series a good model is required to be trained. Therefore, we proposed a deep hybrid neural network model to dynamically learn spatio-temporal correlation throughout a city. The summary of the comparison between our proposed technique "STD-Net" and other closely related works is given in Table 1. We believe, the information in Table 1 would also help our readers to quickly identify gaps for further research.

III. BACKGROUND

In this section, we present the basic notations of traffic flows and then illustrates the problem of city-wide crowd flows. Moreover, we mathematically formulate the problem scenario.

Definition 1 (Region Partition): In existing research works [12], [22], [33], we divided the whole city into map grid of ($M \times N$) based on latitude and longitude and grid represents a region.

Definition 2 (Traffic Flows): From the previous knowledge in [23], we assume at time t, let P be a set of trajectories. The traffic flows inflow and outflow at a time interval of t are illustrated using the following Eq. 1 and Eq. 2.

$$y_t^{in,m,n} = \sum_{T_r \in P} |\{t > 1 | g_{t-1} \notin (m,n) \wedge g_t \in (m,n)\}| \tag{1}$$

$$y_t^{out,m,n} = \sum_{T_r \in P} |\{t \geq 1 | g_t \in (m,n) \wedge g_{t+1} \notin (m,n)\}| \tag{2}$$

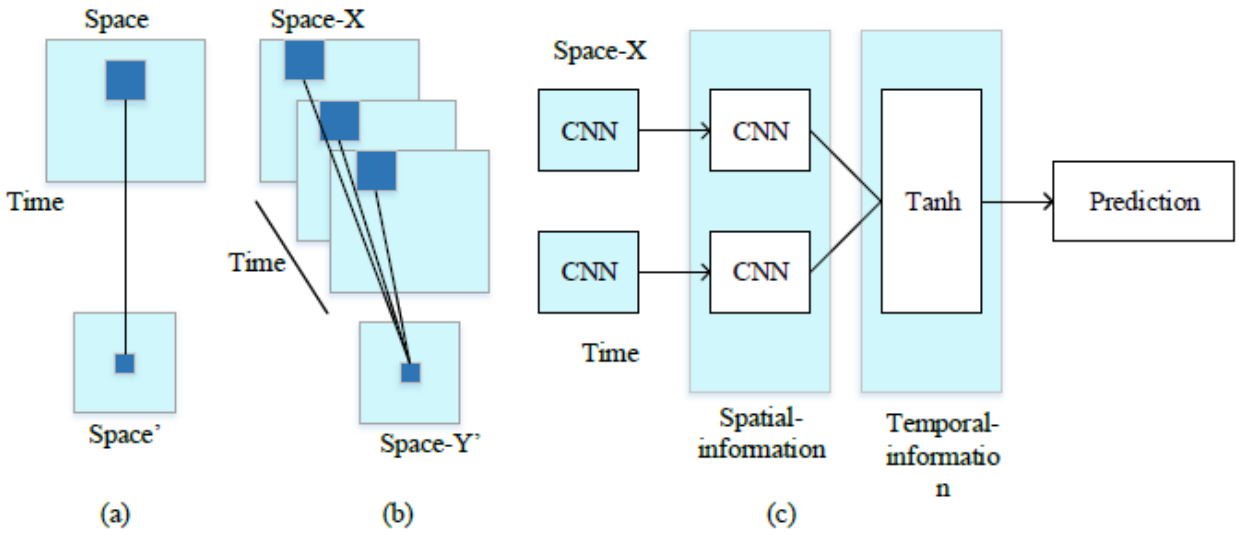

FIGURE 2. 2D-CNN methods limitations (a) Modelling 2D image Spatial and temporal dependencies, (b) Construct time dimension through utilizing RGB channels, (c) 2D-CNN utilizing for image slice with a time dimension and integrate them through *tanh* Function.

where, T_r:$g_1 \rightarrow g_2 \rightarrow, \ldots, \rightarrow g_{|Tr|}$ is a trajectory in P, g_t represent the geospatial coordinate; $g_t \in (m, n)$ means that the point g_t lies within the grid (m, n), and vice versa.

A. 2D-CNN METHODS LIMITATIONS FOR URBAN TRAFFIC FLOWS PREDICTION

We aim to jointly study the Spatio-temporal dependencies in traffic flows prediction. We observed that our proposed approach STD-Net is more appropriate for spatial and temporal correlation features compared with ST-ResNet and other 2D-CNN models. For 2D-CNN, features of 2D dimensions can be trained with the operations of Conv2D and pooling respectively. CNN is well suited in learning the images features that only contain 2 dimensions i.e. longitude and dimensionality. Spatio-temporal fully learning features is important for the urban crowd flows prediction problem. However, they did not well performed to depict the features when included a temporal extra dimension. Many previous research works have made their endeavours to utilize 2D-CNN to capture jointly both spatial and temporal correlations features. These research studies can be divided into three catalogues: (1) an image of the 2D dimension can be treated as time and another as space as illustrated in Figure 2 (a); (2) another approach is to adopt 2D-CNN to extract spatial and temporal features is to change the channels of RGB (three-colour images i.e. red, green, and blue) with time sequence. Since, after each convolution operation, it is losing temporal information of the input signal; and (3) some researchers focused on 2D-CNN to extract spatial features for image slice in the time dimension domain and combined them with the activation (*tanh*) function. However, still, the temporal dependency cannot learn easily of the low-level spatial features.

IV. PROPOSED STD-NET FRAMEWORK

In this section, we describe the framework of our proposed STD-Net model in more detail. Figure 3 demonstrates the framework of our approach, which is composed of four major components, including temporal closeness, period volume, weekly volume, and external factors, respectively. To jointly strengthen the urban crowd traffic flows prediction, i.e., inflow and outflow, we predict and integrate them. In this part of our model, we can consider a hybrid spatial-temporal network (HSTN). To accurately predict the target region with low correlations hurts the performance. To solve the issue, we prefer to use CNN to model the spatial dependency near the regions. We applied the LSTM network to model the sequential temporal dependency, which solves the problem of traditional recurrent neural network and vanishing gradient problem.

The primary components of our network are convolutional-LSTM (ConvLSTM) which replace the kernel size with convolutional. Our network focus on hidden features to extract spatio-temporal dependencies of urban crowd flows. This approach can handle both spatio-temporal information within the data. Our hybrid neural network focus on hidden features to captures both spatial and temporal information of urban crowd flows. After loading several layers of the deep hybrid model, we keep the last time interval in the result for the last layer of the model. The deep hybrid neural network proposes by us, first, we divide the DNN into two sub-DNN. The first one captures the spatial features through CNN, while the

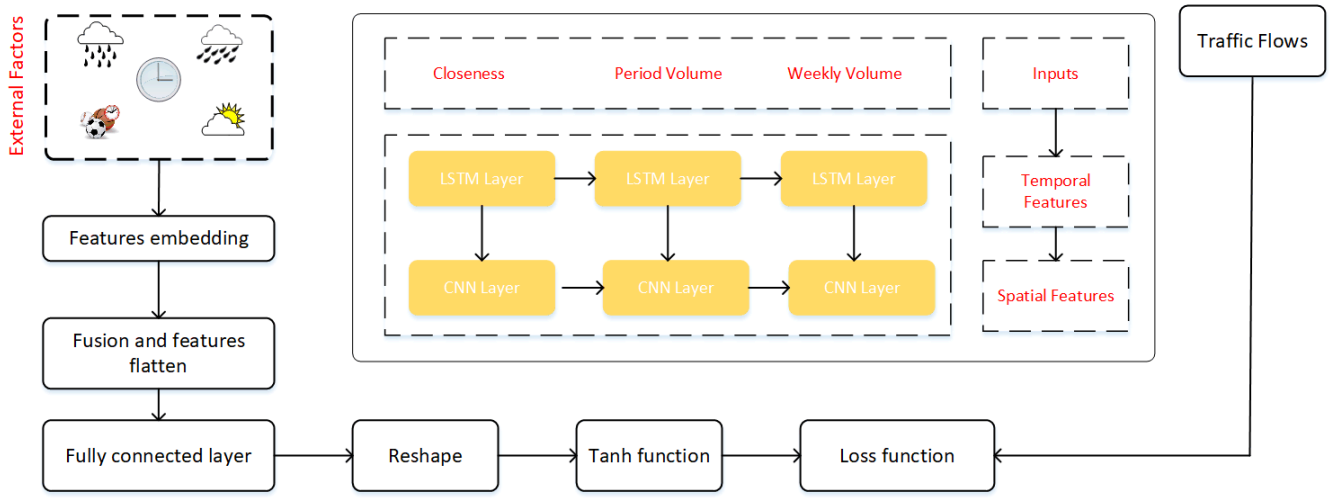

FIGURE 3. The architecture of the proposed STD-Net model.

second one learns temporal features over some time through LSTM. Further, the first three properties outputs are fused as $\mathbf{Z_{fusion}}$. Besides, the $\mathbf{Z_{fusion}}$ is combined with the external branch as $\mathbf{X_{ext}}$ output. We then fused them both external and spatio-temporal features.

A. CONVOLUTIONAL NEURAL NETWORK AND LONG SHORT TERM MEMORY

The deep neural network (DNN) is extensively used in different kinds of applications. CNN is one of the most powerful DNNs for videos and image processing. For the urban crowd flows predictions problem, firstly use the CNN model to extract the spatial data from trajectories GPS data. The CNN is very good in spatial data extraction while GPS trajectories data different from manually images because it includes temporal information which directly affects the prediction accuracy. For accurately explore the temporal information, we consider the recurrent neural network (RNN) to simultaneously capture the temporal information proposed by Sun [34]. However, the RNN faces the problem of vanishing gradient in various kinds of applications. Currently, the LSTM deals with standard RNN to solve the vanishing gradient problem as well as perform to explore the time series features for some time. We used three different components i.e. closeness, period, and trend, the LSTM can be calculated as follows. The various inputs for the closeness, period and trend are $[Y_{t-l_r}, Y_t - (l_{r-1}), \ldots, Y_{t-1}]$, $[Y_{t-l_p \times p}, Y_{t-(l_{p-1})\times p}, \ldots, Y_{t-1}]$, and the final one is $[Y_{t-l_w \times t}, Y_{t-(l_{t-1})\times w}, \ldots, Y_{t-1}]$, respectively. The mathematical formulation is given by Eq. 3:

$$\begin{aligned} i_t &= \sigma\left(W_{yi} * \mathbf{Y}_{r,t} + W_{di} * \mathbf{Y}_{r,t-1} + W_{ri} \circ \mathcal{C}_{t-1} + b_i\right) \\ f_t &= \sigma\left(W_{yf} * \mathbf{Y}_{r,t} + W_{df} * \mathbf{Y}_{c,t-1} + W_{rf} \circ \mathcal{C}_{t-1} + b_f\right) \\ \mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ \tanh\left(W_{yr} * \mathbf{Y}_{r,t} + W_{dr} * \mathbf{Y}_{r,t-1} + b_c\right) \\ o_t &= \sigma\left(W_{yo} * \mathbf{Y}_{r,t} + W_{ho} * \mathbf{Y}_{r,t-1} + W_{ro} \circ \mathcal{C}_t + b_o\right) \\ \mathbf{Y}_{r,t} &= o_t \circ \tanh\left(\mathcal{C}_t\right) \end{aligned} \tag{3}$$

where $*$ denotes the convolutional operation and $\circ$ represents the Hadamard product, whereas W_{yi}, W_{yf}, W_{rf}, W_{df}, W_{yr}, W_{dr}, b_i, b_f, b_c, b_o represent the learnable parameters. The variables i_t, f_t, C_t, and o_t are represented as input gate, forget and output gate, respectively.

Similarly, we also structure the period volume and weekly volume branch by using the same process as mentioned above. Suppose that, l_p is a time intervals of daily segment and d denotes daily. Thus, the daily dependent sequence is $[Y_{t-l_p \times p}, Y_{t-(l_{p-1})\times p}, \ldots, Y_{t-1}]$. The output of daily branch is $\mathbf{Y}_{p,t}^{(L+2)}$. The dependent sequence of trend component is $[Y_{t-l_w \times w}, Y_{t-(l_{w-1})\times w}, \ldots, Y_{t-1}]$ and weekly component output is $\mathbf{Y}_{w,t}^{(L+2)}$, where l_w represent the dependent sequence length of trend branch and w denotes the trend. In our implementation process, p is equal to 1-day that denotes the periodicity of day-level, and w is equal to one-week that describes the trend patterns.

B. CONVOLUTION AND RESIDUAL UNIT

Normally the city contains a huge size with several regions as well as distinct distances. In the nearby area the crowd flows affect each other, we can handle this issue through the ConvLSTM model to explore the Spatio-temporal information. As we know highways and subway connect two regions with a far distance so, in order to extract the Spatio-temporal dependency of any area, we need to design ConvLSTM architecture. It is a fact that deep convolutional neural network (DNN) concession effectiveness of training according to well-known activation function i.e. ReLU as well as some regularization mechanisms already used in [14], [35]. Similarly, we need to use DNN to explore big citywide dependencies. We assume the kernel size is 3×3 and input size is 32×32 for crowd flows of data. We deployed a residual learning scheme in our model because training the super deep neural network more than 1000 layers effectively. In our model, we stack L residual units with both CNN-LSTM.

TABLE 2. Details and characteristics of the evaluated datasets [12].

Dataset	**TaxiBJ**	**BikeNYC**
Data type	Taxi GPS	Rent bike
Location	Beijing	New York
Time span	3/9/2013-30/08/2013 3/5/2014-01/06/2014 3/5/2015-01/06/2015 3/11/2015-10/06/2016	4/07/2014-01/08/2014 - - -
Interval of sampling	60 minutes	30 minutes
# Available time period	23,560	5,394
# Taxis/Bikes	32,000+	7,900+
Size of map	32×32	16×8
Average sampling rate (s)	$\sim$ 55 sec	-
External data (holidays)	40	18
Weather situations	15 types i.e.,, rainy, sunny etc.	-
Temperature/C°	[-25.7, 39.0]	-

We also focus on the Batch Normalization method in our research work used by [6]. We added this method before residual units (ReLU) as shown in our architecture Figure 3. We added Convolutions and ConvLSTM layers on top of the residual unit. The output of closeness component with ConvLSTM, convolutions and L residual units is $Y_{Ext(l+2)}$.

C. FUSION

In this section, we need to know the concept of fusion as four components shown in Figure 3. First, we fuse the 3 components i.e. closeness, period volume, and weekly volume with parametric based fusion method and then integrated with external features. Through the parametric fusion method, we fuse the three components as follows, given by Eq. 4:

$$\mathbf{Z_{fusion}} = \mathbf{W_r} \circ \mathbf{Y}_{r,t} + \mathbf{W_p} \circ \mathbf{Y}_{p,t} + \mathbf{W_w} \circ \mathbf{Y}_{w,t} \quad (4)$$

Here $\odot$ represents the Hadamard product, while W_w, W_p, and W_r, indicate learnable parameters that modify the degrees which are affected through weekly volume, period volume, and temporal closeness. We directly combine the result of the first three components with an external component. At time interval t the predicted value indicates as $Y_{r,t}$. The $Y_{r,t}$, $Y_{p,t}$, and $Y_{w,t}$ represent closeness, period, and weekly components respectively.

Our STD-Net model can be trained to predict z_t from four properties, i.e.,, weekly, period, closeness and external components respectively by reducing the value of Mean Squared Error (MSE) between the ground truth and predicted flows at time interval t. The training model is given by the following Eq. 5.

$$L_(\phi) = \|z_t - \hat{z}_t\|^2 \quad (5)$$

where ϕ includes W_r, W_p, W_w, and other learnable parameters.

D. EXTERNAL COMPONENT FUSION

In this step, we combined directly the first three properties output with an external component. The output of fused components $\hat{Z}_t$ and external components is defined in Eq. 6.

$$\hat{Z}_t = (\mathbf{Z_{fusion}} + \mathbf{X_{ext}}) \quad (6)$$

V. EXPERIMENTAL SETUP

In this section, we use two kinds of datasets, i.e., BikeNYC and TaxiBJ GPS trajectories data both datasets consist of trajectories and weather information. We implemented all our experiments in python, through the Keras library and the Theano in the back-end, which is the most important and famous open-source framework. Table 2 shows the details of two datasets, while Table 3 discuss the comparative results with our proposed STD-Net.

A. COMPARING BASELINES

We compare our model with the following competitive spatio-temporal baselines.

- HA: This method is very easy in implementation but poorly perform under unexpected traffic situations. This model cannot handle temporal dependencies.

TABLE 3. Comparison of TaxiBJ dataset with previous baselines [minimum values are better than the others].

Models	Runtime (epoch)	RMSE	MAPE
HA	-	58.59	39.20
SARIMA	-	25.86	27.20
ARIMA	-	23.87	29.30
ST-ANN	-	18.75	21.35
VAR	-	21.79	23.20
DeepST	2435s	17.89	20.45
DeepST-CPTM	2023s	17.43	19.56
ST-ResNet	4994s	16.69	17.64
STDN	4566s	16.59	18.87
MST3D	5902s	15.99	16.98
STD-Net (Ours)	3967s	14.36	16.04

TABLE 4. Comparison of BikeNYC dataset with previous baselines [minimum values are better than the others].

Models	RMSE	MAPE
HA	15.59	38.32
SARIMA	13.68	28.31
ARIMA	12.41	27.42
ST-ANN	11.57	26.53
VAR	11.12	24.31
DeepST	10.51	22.31
DeepST-CPTM	9.32	21.32
ST-ResNet	6.33	21.23
STDN	6.2	20.98
MST3D	5.81	20.68
STD-Net (Ours)	5.34	17.45

- ARIMA: The objective of this model is to fit the linear model and for time series data forecasting.
- SARIMA: enhanced of ARIMA known as seasonal ARIMA. A regular pattern of changes in a time series over s periods.
- VAR: The Vector Auto-Regressive (VAR) more powerful Spatio-temporal data because it explores pairwise correlations between flows of data and the computational cost is too high because of the large number of parameters.
- ST-ANN [36]: According to this Spatio-temporal Artificial Neural Network first it captures spatial 8 regions nearby values and temporal (previous 8-time intervals) and finally fed them in ANN.
- DeepST [2]: This model is a kind of deep neural network used for Spatio-temporal data prediction and provides state-of-art results for prediction of crowd flows. It consists of 4 different kinds of variants, i.e., DeepST-CP, DeepST-C, DeepST-CPT, and DeepST-CPTM that concentrate on external features as well as temporal dependencies.
- STDN [37]: This model applied an attention layer with deep neural networks to model spatio-temporal dependencies.
- MST3D [11]: This model applied 3D CNN to exploit citywide traffic flows prediction.
- ST-ResNet [1]: This model is used for crowd flows prediction using a convolutional neural network (CNN) with residual learning [2].

B. PREPROCESSING

In the final output of our model, we use the activation function tanh whose range in between -1 to 1 as see in Equation (1). We also focus on the method of Min-Max normalization to balance data between $[-1, 1]$. For external features, we transform metadata using the concept of one-hot coding schemes such as weather, weekday, holidays, and weekend as well use the same concept of Min-Max normalization to balance the temperature and wind and within [0, 1].

C. HYPER-PARAMETER SETTINGS

To validate our proposed STD-Net model, we used PyTorch with the Keras (2.0.1) and Tensor flow (2.1.6). The three branches' lengths (closeness, daily-periodic, and weekly-periodic) are set to (4, 4, and 4) for BikeNYC. For (TaxiBJ), We set the lengths of three dependent sequences that are set to (6, 4, and 4), respectively. We apply two convolutional layers for (BikeNYC) dataset with a small size, i.e., $(4 \times 8 \times 16)$ in all branches. To reduce the issue of over-fitting, we set a dropout rate to 0.25.

D. MEASUREMENT METRIC

Finally, we evaluate our proposed method by using Mean Average Percentage Error (MAPE) given in Equation (7) and Root Mean Square Error (RMSE) given in Equation (8).

$$\text{MAPE} = \frac{1}{t}\sum_{i=1}^{t}\frac{|\hat{x}_i - x_i|}{x_i} \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(\hat{x}_i - x_i)^2} \tag{8}$$

where $\hat{x}_i$ and x_i denote the prediction and ground truth value, respectively; at time interval t, and t is the total number of samples.

E. RESULTS ANALYSIS

In this section, we compare our STD-Net model with state-of-the-art competing models for BikeNYC and TaxiBJ datasets as shown in Table 2 and Table 3, respectively. It is obliviously demonstrated that, STD-Net outperforms than existing methods by achieving lowest RMSE and MAPE values. For TaxiBJ, the value of RMSE and MAPE are

14.36 and 16.04 respectively. Likewise, for BikeNYC, the values of RMSE and MAPE are 5.34 and 17.45, respectively. For TaxiBJ dataset, the proposed model 12% better than the existing baselines in terms of RMSE, while 8% better than in terms of MAPE. Similarly, for BikeNYC dataset, the STD-Net 14.8% better than the existing baselines using RMSE, while 15% better than in terms of MAPE.

In contrast, the prediction methods of traditional time-series i.e., (HA and ARIMA) cannot obtain better results as they dependent on historical records to overlook spatial and other related external features and predict the future values. Our proposed STD-Net outperforms significantly above all the existing methods. The results shows the effectiveness of our schemes adopting jointly traffic crowd flows. Our proposed model also achieves better performance than DeepST-CPTM and ST-ResNet. One of main reasons is that DeepST-CPTM cannot model temporal sequential dependency and explicitly model spatial dependency. Also, ST-ResNet just focus on CNN to extract spatial correlation without considering the sequential temporal dependency. STD-Net also outperforms the category of methods (e.g., MST3D and STDN) which use LSTM and CNNs together for traffic crowd flows prediction. This category of baselines only captures the temporal dependencies for the high-level spatial features, but not considers the temporal correlations with low-level spatial features.

F. DIFFERENT COMPONENTS EFFECTS

We compared the L12-e model, as we see that the results of L2-E, L4-E, and L12-E show that an increase in the residual units and a decrease in the RMSE. Deeper the network results will be more accurate through using deep learning residual networks.

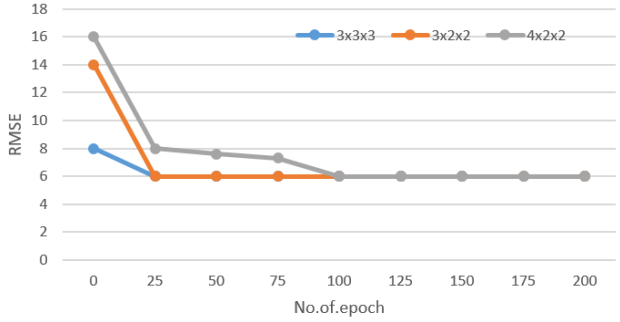

FIGURE 4. Combinations of different kernel size.

We have collected various combinations of training datasets to get better kernel size as shown in Figure 4. The output demonstrates that kernel size 3 ×3 × 3 is the optimal solution for urban traffic transportation information as mentioned in the previous papers in some other behaviour's detection tasks. We compile each model 10 times as well as focus on bike New York and taxi Beijing data for external and without external as shown in Figure 5.

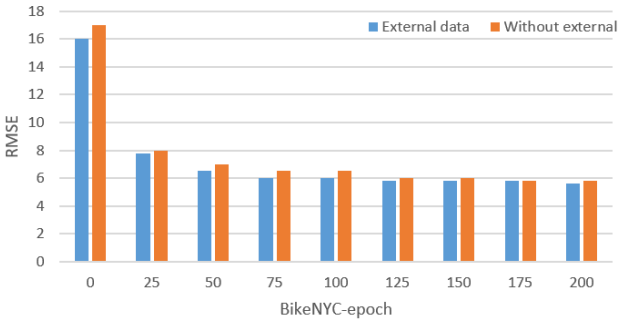

FIGURE 5. Results of BikeNYC with external and without external data.

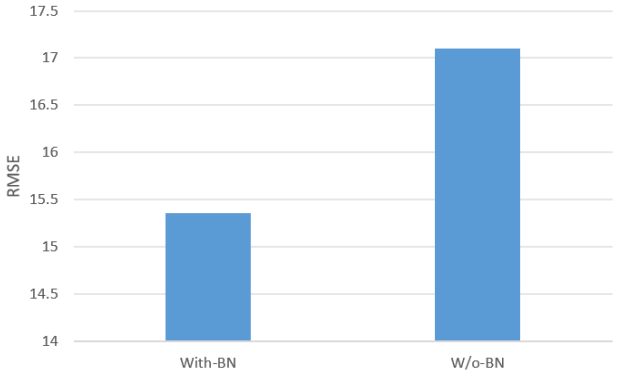

FIGURE 6. Impact of batch normalization.

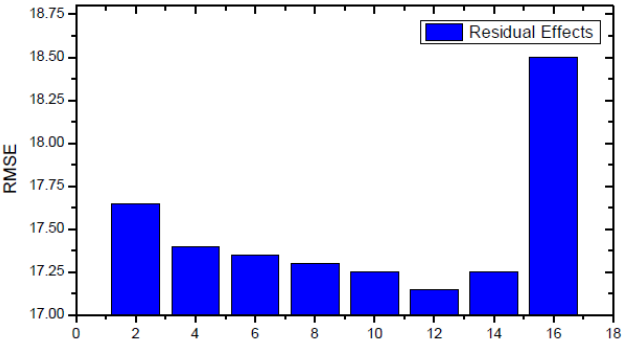

FIGURE 7. Impact of different residual units.

G. BATCH NORMALIZATION EFFECT

We utilize to integrate BN into each residual unit, discovering that the RMSE is improved slightly in urban traffic flows prediction as shown in Figure 6.

H. IMPACT OF DIFFERENT NETWORKS

We discuss briefly the impact of depth network in detail as shown in Figure 6, and Figure 7. As we can see that the network is going deeper, i.e., increase residual number units, reduce RMSE of the model, then increases. It is clear that it can get a better result if the network is deeper, so it not only captures the near spatial dependency, but can also catch one that is distant. The network, however, goes deeper and deeper, i.e. the number of residual units greater than or equal

to 14, as the training phase becomes more complicated and the problem of over-fitting is likely to occur.

VI. CONCLUSION AND FUTURE WORK

In this article, we analyse the citywide traffic crowd flows prediction problems. We propose the STD-Net model to simultaneously learn Spatio-temporal correlations among different types of traffic crowd flows. We conducted simulations on two real-world datasets, which demonstrate that our proposed method outperforms in prediction accuracy than existing baselines significantly as well confirmed that more applicable for urban crowd flow prediction. Our evaluation suggests that, for various datasets, the proposed STD-Net model is approximately 14.67% better than the existing baselines in terms of RMSE, while approximately 11.23% better than in terms of MAPE.

In the future, we aim to develop cloud-based systems know as Urban Flow that can check the real-time crowd flow prediction. If we implement the training module on the cloud and then the prediction model is implemented on the edge computing or small data centre, then the benefit is that each vehicle should quickly predict and take appropriate decisions for re-routing to avoid crowed or traffic congestion [38]–[43]. For that, we need to deploy fog devices on local regions such as shopping malls or mobile base stations, etc. Because prediction and decision denote real-time applications. Hence, the process is lengthy if the data goes to the cloud, latency that could be improved through deploying fog devices in the local area (near the vehicles). So, we need to analyse the related datasets (which are very rare) and check various constraints to formulate the optimization problem [44].

REFERENCES

[1] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI*, 2017, pp. 1655–1661.
[2] M. X. Hoang, Y. Zheng, and A. K. Singh, "Forecasting citywide crowd flows based on big data," in *Proc. ACM SIGSPATIAL*, 2016, pp. 1–10.
[3] J. Zhao, Q. Qu, F. Zhang, C. Xu, and S. Liu, "Spatio-temporal analysis of passenger travel patterns in massive smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3135–3146, Nov. 2017.
[4] Z. Fan, X. Song, R. Shibasaki, and R. Adachi, "Citymomentum: An online approach for crowd behavior prediction at a citywide level," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput.*, 2015, pp. 559–569.
[5] X. Song, Q. Zhang, Y. Sekimoto, and R. Shibasaki, "Prediction of human emergency behavior and their mobility following large-scale disaster," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 5–14.
[6] R. Silva, S. M. Kang, and E. M. Airoldi, "Predicting traffic volumes and estimating the effects of shocks in massive transportation systems," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 18, pp. 5643–5648, 2015.
[7] A. Abadi, T. Rajabioun, and P. A. Ioannou, "Traffic flow prediction for road transportation networks with limited traffic data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 653–662, Apr. 2015.
[8] Y. Xu, Q.-J. Kong, R. Klette, and Y. Liu, "Accurate and interpretable Bayesian MARS for traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 6, pp. 2457–2469, Dec. 2014.
[9] E. Zivot and J. Wang, "Vector autoregressive models for multivariate time series," in *Modeling Financial Time Series With S-Plus*. Springer, 2006, pp. 385–429. [Online]. Available: https://link.springer.com/book/10.1007/978-0-387-32348-0#authorsandaffiliationsbook
[10] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results," *J. Transp. Eng.*, vol. 129, no. 6, pp. 664–672, Nov. 2003.
[11] C. Chen, K. Li, S. G. Teo, X. Zou, K. Li, and Z. Zeng, "Citywide traffic flow prediction based on multiple gated spatio-temporal convolutional neural networks," *ACM Trans. Knowl. Discovery Data*, vol. 14, no. 4, pp. 1–23, Jul. 2020.
[12] A. Ali, Y. Zhu, Q. Chen, J. Yu, and H. Cai, "Leveraging spatio-temporal patterns for predicting citywide traffic crowd flows using deep hybrid neural networks," in *Proc. IEEE 25th Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Dec. 2019, pp. 125–132.
[13] Y. Li, Y. Zheng, H. Zhang, and L. Chen, "Traffic prediction in a bike-sharing system," in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2015, p. 33.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[16] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.
[17] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.
[18] S. Clark, "Traffic prediction using multivariate nonparametric regression," *J. Transp. Eng.*, vol. 129, no. 2, pp. 161–168, Mar. 2003.
[19] C.-H. Wu, J.-M. Ho, and D. T. Lee, "Travel-time prediction with support vector regression," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 276–281, Dec. 2004.
[20] F. Altche and A. de La Fortelle, "An LSTM network for highway trajectory prediction," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 353–359.
[21] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, "Deep learning: A generic approach for extreme condition traffic forecasting," in *Proc. SIAM Int. Conf. Data Mining*, 2017, pp. 777–785.
[22] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction," *Sensors*, vol. 17, no. 4, p. 818, Apr. 2017.
[23] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, "Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction," 2018, *arXiv:1803.01254*. [Online]. Available: http://arxiv.org/abs/1803.01254
[24] J. Ke, H. Zheng, H. Yang, and X. Chen, "Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach," *Transp. Res. C, Emerg. Technol.*, vol. 85, pp. 591–608, Dec. 2017.
[25] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, "Deep multi-view spatial-temporal network for taxi demand prediction," 2018, *arXiv:1802.08714*. [Online]. Available: http://arxiv.org/abs/1802.08714
[26] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2397–2406.
[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[28] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2361–2369.
[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473
[30] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4945–4949.
[31] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.

[32] Y. Liang, S. Ke, J. Zhang, X. Yi, and Y. Zheng, "Geoman: Multi-level attention networks for geo-sensory time series prediction," in *Proc. IJCAI*, 2018, pp. 3428–3434.

[33] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, "Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction," 2018, *arXiv:1803.01254*. [Online]. Available: http://arxiv.org/abs/1803.01254

[34] C. Sun, "Fundamental Q-learning algorithm in finding optimal policy," in *Proc. Int. Conf. Smart Grid Electr. Automat. (ICSGEA)*, May 2017, pp. 243–246.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[36] Z. He, C.-Y. Chow, and J.-D. Zhang, "STANN: A spatio–temporal attentive neural network for traffic prediction," *IEEE Access*, vol. 7, pp. 4795–4806, 2019.

[37] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, "Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5668–5675.

[38] A. A. Khan, M. Zakarya, R. Buyya, R. Khan, M. Khan, and O. Rana, "An energy and performance aware consolidation technique for containerized datacenters," *IEEE Trans. Cloud Comput.*, early access, Jun. 5, 2019, doi: 10.1109/TCC.2019.2920914.

[39] M. Zakarya, "An extended energy-aware cost recovery approach for virtual machine migration," *IEEE Syst. J.*, vol. 13, no. 2, pp. 1466–1477, Jun. 2019.

[40] A. A. Khan, M. Zakarya, and R. Khan, "Energy-aware dynamic resource management in elastic cloud datacenters," *Simul. Model. Pract. Theory*, vol. 92, pp. 82–99, Apr. 2019.

[41] M. Zakarya and L. Gillam, "Managing energy, performance and cost in large scale heterogeneous datacenters using migrations," *Future Gener. Comput. Syst.*, vol. 93, pp. 529–547, Apr. 2019.

[42] L. Gillam, K. Katsaros, M. Dianati, and A. Mouzakitis, "Exploring edges for connected and autonomous driving," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 148–153.

[43] A. Ali, L. Lu, Y. Zhu, and J. Yu, "An energy efficient algorithm for virtual machine allocation in cloud datacenters," in *Proc. Conf. Adv. Comput. Archit.* Singapore: Springer, 2016, pp. 61–72.

[44] M. R. Jabbarpour, H. Zarrabi, R. H. Khokhar, S. Shamshirband, and K.-K.-R. Choo, "Applications of computational intelligence in vehicle traffic congestion problem: A survey," *Soft Comput.*, vol. 22, no. 7, pp. 2299–2320, Apr. 2018.