2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378156

MASTGN: Multi-Attention Spatio-Temporal Graph Networks for Air Pollution Prediction

Peijiang Zhao
Big Data Integration Research Center
National Institute of Information
and Communications Technology
Tokyo, Japan
dlzpj@nict.go.jp

Koji Zettsu
Big Data Integration Research Center
National Institute of Information
and Communications Technology
Tokyo, Japan
zettsu@nict.go.jp

***Abstract*—In recent years, the importance of air pollution issues has been increasingly discussed by the public and the government. Air pollution prediction has become a crucial reference to rely on when the government needs to formulate environmental policies. Considering the sparsity of atmospheric monitoring stations in the spatial distribution, in this study, we consider the atmospheric data as a spatio-temporal structure graph for better utilization of spatio-temporal information. Accordingly, we propose using multi-attention spatio-temporal graph networks (MASTGN) to exploit the graph structure atmospheric data for air pollution prediction tasks. The MASTGN model allows better mining the high-level spatial, temporal, and physical features corresponding to the atmospheric data through the multi-attention mechanism of the spatial, temporal, and channel attention. We conduct experiments on two datasets gathered in Japan and China to predict the concentration of $\mathrm{PM}_{2.5}$, O_3, and PM_{10}. The results indicate that the proposed MASTGN model outperforms the considered baseline approaches on prediction accuracy.**

***Index Terms*—Air pollution prediction, Graph networks, Multi-attention mechanism, spatio-temporal graphs.**

I. Introduction

In recent years, with an increase in public health awareness [1], the importance of air pollution issues has been increasingly acknowledged by the society and the government. At the same time, the demand for the high accuracy of air pollution prediction has also increased due to the fact that air pollution prediction is closely related to environmental policies issued by the government, raising an early warning of an air pollution level, and planning outdoor activities by citizens. Therefore, air pollution prediction has become a high-profile research direction.

Depending on the types of atmospheric datasets, the methods for air pollution prediction can be divided into the following groups: those based on the satellite data and those relying on the ground atmospheric monitoring station data. A number of atmospheric physics models [2] [3] based on the satellite data have been reported to succeed in air pollution prediction. These models usually can predict air pollution in a wide-scale range of thousands of kilometers. However, such atmospheric physics models are insufficient to ensure acceptable accuracy when predicting air pollution in an urban environment, as human activities and urban conditions are difficult to be modeled from the physical perspective. In contrast, the ground atmospheric monitoring station data contain the hidden information about the impact of human activities on an urban environment. Accordingly, the data-driven prediction models that can mine information effectively from the monitoring station data have become the key tool of urban atmospheric pollution prediction.

Recently, numerous data-driven models for air pollution prediction based on deep learning methods have been proposed. For example, [4] proposed using a convolutional neural network (CNN) for spatial feature extraction and a recurrent neural network (RNN) for time series modeling. In [5], a distributed network to identify the fusion heterogeneous features corresponding to the atmospheric data was introduced. Then, [6] suggested considering the urban traffic data to perform air quality predictions. The common approach incorporated in these studies focused on the spatial features of the atmospheric data was that they implied converting the sparse monitoring station data into the continuous spatial data to perform data preprocessing. The unknown atmospheric data between monitoring stations which cannot be obtained are considered as null values while using the spatial interpolation techniques, such as inverse distance weighting (IDW) [7]. However, due to the limitation on the number of atmospheric monitoring stations that can be deployed in the real world, the massive amounts of the unknown atmospheric data need to be interpolated, and therefore, the data preprocessing step related to spatial interpolation is associated with considerable computational costs and the high level of noise.

From the other perspective, a set of spatially sparse ground monitoring stations data can be considered as a typical graph structure data (see Fig 1 left). Each monitoring station can be regarded as the node of a graph, and spatial correlations - as the edges between nodes. Therefore, the atmospheric monitoring station data can be represented as a spatio-temporal structure graph (as shown in Fig 1 on the right).

The data that can be described using a graph structure widely exist in the real world, including social data [8], traffic data [9], human activity data [10], etc. To address the challenges associated with utilizing the spatio-temporal structure graph of the atmospheric data, we propose a multi-

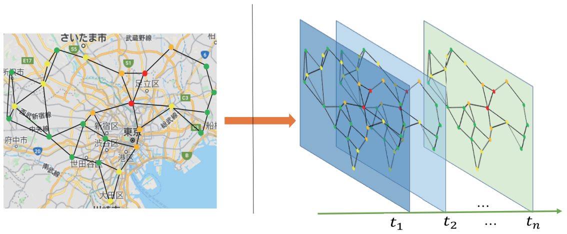


Fig. 1: (left) Undirected graph based on the atmospheric monitoring station data gathered in Tokyo. (right) The spatio-temporal structure graph G_t. Each graph represents the current atmospheric information at time step t.

attention spatio-temporal graph network (MASTGN) model for air pollution prediction. The MASTGN model is focused on multi-attention corresponding to the channel attention between all possible air pollutants. The spatio-temporal attention of the graph structure is considered to use the graph structure atmospheric data more effectively. By mining the high-level features in the atmospheric data in terms of multiple directions, such as the spatial, temporal, and physical ones, the MASTGN model can be used to achieve better air pollution prediction accuracy. The contributions of the present study can be summarized as follows:

- We consider the atmospheric data as a spatio-temporal structure graph aiming to suppress noise and to exploit the high-level spatio-temporal features in a more effective manner. Moreover, we propose a deep learning model based on multiple attention named MASTGN to perform air pollution prediction tasks on the graph structure data.
- The proposed MASTGN model can be used to construct three sub-networks: the channel attention, graph attention, and temporal attention networks that are based on a gated convolutional neural network to process the spatio-temporal graph data explicitly.
- The proposed MASTGN model can be applied to various real-world air pollution datasets with the purpose of predicting different types of air pollution. We evaluate the performance of the proposed model on two datasets gathered in Japan and China. The obtained results indicate that the MASTGN model outperforms the considered baseline approaches.

II. Proposed Method

A. Problem Definition

Graph Definition: We denote the atmospheric data registered by monitoring stations in a city as an undirected graph $g = (V, E)$ (Fig 1 left), where $V \in \mathbb{R}^{N \times C}$ is the node set; $E \in \mathbb{R}^{N \times N}$ is the adjacency matrix of a graph; N is the number of nodes; C is the number of input channels (in this case, it denotes the number of the atmospheric sensors). The adjacency matrix E is based on the coordinate of each monitoring stations. A time series graph with the time-step length t is denoted as $G_t = \{g_1, g_2, \ldots, g_t\}$(Fig 1 right).

Prediction Definition: Given a time series graph G_t over previous t time slices as the input, we can predict the future value $Y^c_{t+\tau} = \{y^{c,1}_{t+\tau}, y^{c,2}_{t+\tau}, \ldots, y^{c,N}_{t+\tau}\} \in \mathbb{R}^N$ of air pollutants c corresponding to all stations at time $t + \tau$, where $y^{c,n}_t \in \mathbb{R}$ is the value of the c-th air pollutant of station n at time t.

B. Framework

The framework underlying the proposed multi-attention spatio-temporal graph network (MASTGN) model, as presented in Fig 2, comprises four subnets: a channel attention network (CAT), a graph attention network (GAT), a gated 1-dimensional convolutional neural network (1D-GCNN), and a fully connected neural network (FC). Each subnet corresponds to one type of attention mechanisms: channel attention, spatial attention, and temporal attention. Input graph G_t is first processed applying CAT and GAT at each time step to derive channel features c_t and spatial features s_t. Then, features $[C_t, S_t] = [c_1+s_1, c_2+s_2, \ldots, c_t+s_t]$ corresponding to t time steps are fed into the 1D-GCNN to obtain temporal features R_t. Finally, all o extracted features $[C_t, S_t, R_t]$ are inputted into the FC connection layer to generate predictions.

C. Channel attention network

The channel attention mechanism [11] can be used to model interdependencies between channels. As atmosphere can be considered as a chaotic system [12], it is difficult to determine and item in graph G_t that can result in a positive response for prediction target $Y^c_{t+\tau}$. Here, we introduce a channel attention network inspired by [13] aiming to identify interdependencies between all atmospheric aspects. V

N
C □
C
N □

The input of original features constitutes the node set $V \in \mathbb{R}^{N \times C}$ of graph $g = (V, E)$. Then, the output $O \in \mathbb{R}^{C \times N}$ of the channel attention network can be calculated as follows:

$$X = softmax(V^T \cdot V) = \frac{exp(V^T \cdot V)}{\sum exp(V^T \cdot V)} \qquad (1)$$

$$O_j = \beta \sum_{i=0}^{C} x_{ji} V_i + V_j \qquad (2)$$

where V^T is the transpose matrix of V; $X \in \mathbb{R}^{C \times C}$ is the channel attention map in which each element x_{ji} of X reflects the impact of i^{th} channel on j^{th} channel; V_i, V_j are the elements of i^{th} channel and j^{th} channel of V; β is the learnable scale parameter that is initialized as 0. The output O is defined as the weighted sum of each channel over the original features. It is used to enhance the distinguishability of original features V.

x_{ji}
$j \leftarrow i$

D. Graph attention networks

The data from atmospheric monitoring stations as a dataset can be arranged into a graph structure dataset based on the geographic information. In the case of using a graph structure dataset, interpolation is not required outside atmospheric monitoring stations. At the same time, even in the same atmospheric geographic system, the influence of each leaf node to the root node may differ. Therefore, in the present study, we apply a

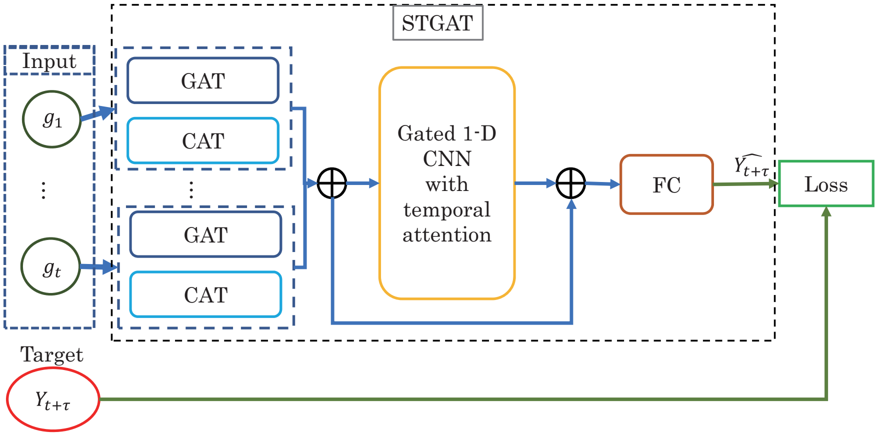


Fig. 2: The framework of the multi-attention spatio-temporal graph network (MASTGN) model. The MASTGN comprises four subnets: the channel attention networks (CAT), graph attention networks (GAT), a gated 1D convolutional neural network (1D-GCNN), and a fully connected one (FC). First, CAT and GAT are used to extract the channel and spatial features from the input g_t at each time step. Then, 1D-GCNN is applied to extract the temporal features corresponding to the channel and spatial features obtained as a result of t time steps. Finally, FC is utilized to generate the final predictions $\acute{Y}_{t+\tau}$ by integrating all extracted features.

graph attention neural network [8] to obtain the spatial features from the air pollution data.

The input into each graph attention layer is the node set $V = \{v_1, v_2, \ldots, v_N\}$, $v_i \in \mathbb{R}^C$ and the adjacency matrix $E \in \mathbb{R}^{N \times N}$. The normalized attention coefficients a_{ij} that indicate the impact of i^{th} node on j^{th} node can be calculated as follows:

$$a_{ij} = \frac{exp(\text{LeakyReLU}([Wv_i||Wv_j]))}{\sum_{k\in\mathcal{N}_i} exp(\text{LeakyReLU}([Wv_i||Wv_k]))} \tag{3}$$

where $W \in \mathbb{R}^{\overset{c\rightarrow c'}{C' \times C}}$ is a shared weight matrix; C' is the output channel size of each graph attention layer; $\mathcal{N}_i$ is a neighborhood node (including i) of node i (in the present study, $\mathcal{N}_i$ is determined according to adjacency matrix E; LeakyReLU is the nonlinearity function; $||$ denotes concatenation operation. The output of a single graph attention layer $V' = \{v'_1, v'_2, \ldots, v'_N\}, v'_i \in \mathbb{R}^{C'}$ can be expressed as follows:

$$v'_i = \sigma(\sum_{j\in\mathcal{N}_i} a_{ij} W v_j) \tag{4}$$

where, σ is the nonlinearity function. Finally, we concatenate the output of K graph attention layers to obtain the final output of the GAT as follows:

$$v'_i = ||_{k=1}^{K}(\sigma(\sum_{j\in\mathcal{N}_i} a_{ij}^k W^k v_j)) \tag{5}$$

The concatenation of multi-layers can be used to stabilize the learning process of the attention mechanism. Here, we concatenate 8 layers to obtain the final spatial feature.

E. The gated convolutional neural network with the temporal attention

The recurrent neural network (RNN)-based models, such as long short-term memory (LSTM) [14], GRU [15] have become the baseline method to solve time series tasks. However, the RNN-based models applied to the air pollution prediction tasks still exhibit the problem of the vanishing gradient of long term memory and time-consuming iterations. On the contrary, one-dimensional convolutional neural networks (1D-CNNs) [16] have the advantage of no constraint on the dependency between the previous step and the fast iteration in training so that 1D-CNNs are deemed applicable to the time series tasks. Inspired by [9], in this study, we apply a gated 1D-CNN based on the structure proposed by [17], and we add the temporal attention mechanism to this 1D-GCNN to obtain the final temporal feature.

The structure of the 1D-GCNN is as illustrated in Fig 3. It can be seen that the 1D-GCNN is a multi-layer structure in which each layer l has the same structure comprising multiple gated kernels to extract the temporal feature at different time steps t_n. The ratio of the number of time steps for each gated kernel grows exponentially. In this study, the time steps grow from 2^1 to 2^n, n is the number of the gated kernels in each layer l, and 2^n is the longest input time steps. Provided the input $I^l \in \mathbb{R}^{t_n \times C_i}$ with the time step t_n for each gated kernel n and the input channel size C_i of each node, the gated kernel first maps I^l to a single output element $[H^l, Q^l] \in \mathbb{R}^{(t_n - K_i)\times C_o}$ using the convolutional kernel $\Gamma \in \mathbb{R}^{K_i \times C_i \times 2C_o}$, where K_i is the width of kernel Γ; C_o is the output channel size of the convolutional network; $[H^l, Q^l]$ is split in half with the same size of C_o. Then, according to gated nonlinearity, an operation executed in the gated kernel can be described as follows:

$$\begin{aligned} H^l, Q^l &= \Gamma *_d I^l \\ h^l &= tanh(H) \odot \sigma(Q) \\ I^{l+1} &= W^l h^l + I^l \end{aligned} \tag{6}$$

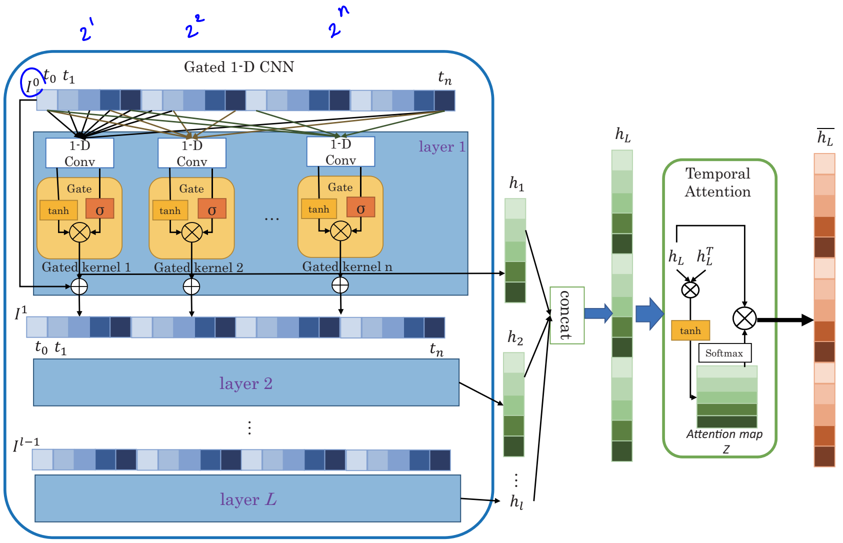


Fig. 3: The structure of the gated 1D convolutional neural network with the temporal attention. Each 1D-GCNN layer contains multiple gated kernels to extract the temporal feature h_l at different time steps. The ratio of the number of time steps for each gated kernel grows exponentially (2^n in this study). The input I_{l+1} of the next layer constitutes the element-wise sum of input I_l and output h^l of the current layer. We concatenate all features obtained at each layer (4 layers in this study) to obtain the feature h_L. Finally, we input h_L into the temporal attention unit to obtain final temporal feature $\overline{h_L}$.

where h^l is the temporal feature of layer l; I^{l+1} is the input feature into next layer $l+1$; W^l denote the convolution parameters; σ is the sigmoid function; $\odot$ corresponds to the element-wise production. The input into the first layer for all nodes $I_n^0 = transpose(O||V^{'}), I^0 \in \mathbb{R}^{N \times t \times (C+C^{'})}$ is the transposition of the concatenation of the outputs in the channel attention network and GAT. The tanh gate $tanh(H)$ is utilized to control which information can be learned in the current states. The sigmoid gate $\sigma(Q)$ is applied to control which compositional structure in a time series can be beneficial for the current states. The gated nonlinearity with multi-layers contributes to boosting the representative capability of an input feature based on the stacked temporal layers. Therefore, the gated kernels concatenates all h^l of each layer to obtain $h_L \in \mathbb{R}^{N \times t \times (LC_o)}$ that can be defined as follows:

$$h_L = ||_{l=1}^{L}(h^l) \tag{7}$$

The temporal attention unit is applied to emphasize the temporal relevance between each time step in the temporal feature h_L. Similarly as for the other subnets in this model, we use the attention mechanism to identify this temporal relevance:

$$\begin{aligned} Z &= tanh(h_L W_1)(h_L W_2)^T \\ z_{ij}^{'} &= \frac{exp(z_{ij})}{\sum_{j=1}^{t} exp(z_{ij})} \\ \overline{h_L} &= Z^{'}(h_L W_3) \end{aligned} \tag{8}$$

where $(W_1, W_2, W_3) \in \mathbb{R}^{C_o \times C_a}$ is the learnable parameters; C_a is the output channel size of the temporal attention unit; $Z = z_{00}, \ldots, z_{ij}, Z \in \mathbb{R}^{N \times T \times T}$ denote the attention coefficients in which each element z_{ij} is the strength of the relevance between time moments I and j; $Z^{'} = z_{00}^{'}, \ldots, z_{ij}^{'}, Z^{'} \in \mathbb{R}^{N \times T \times T}$ is the result of Z normalized by the softmax function. Here, $\overline{h_L} \in \mathbb{R}^{N \times T \times C_a}$ is the final temporal feature that is outputted by the 1D-GCNN unit.

F. Full-connection layer and loss function

After deriving the channel attention, the spatial feature, and the temporal feature, we finally employ a multi-layer perceptron (MLP) as the fully connected (FC) output layer. In detail, MLP includes two hidden layers to map all obtained features to derive the final prediction result. We denote all obtained features of G_t as $U_t = (O||V^{'}||\overline{h_L}^T), whereU_t \in \mathbb{R}^{T \times N \times (C+C^{'}+C_a)}$. Then, MLP is applied to perform the following procession:

$$\begin{aligned} D &= ReLu(W_M^1 U_t + b_1) \\ \acute{Y}_{t+\tau}^c &= \sigma(W_M^2 D + b_2) \end{aligned} \tag{9}$$

where W_M^1, W_M^2, b_1, b_2 are the learnable parameters; D_1 is the hidden feature; $\acute{Y}_{t+\tau}^c$ is the final prediction. We use the $ReLu$ activation function for the first hidden layer and σ is an optional activation function based on the prediction target. As the considered prediction target is the measurement of atmospheric pollutants, we apply the $sigmoid$ activation function to the output layer. Thereafter, we consider the mean absolute error (MAE) with $L2$ loss as the loss function to evaluate the performance of the MASTGN model. Therefore, the loss function can be written as follows:

$$\mathcal{L} = \frac{1}{N}\frac{1}{T}\sum_{n \in \mathcal{N}_{G_t}}\sum_{t}|\acute{Y}_{t+\tau}^c - Y_{t+\tau}^c| + \gamma\theta \tag{10}$$

where $\mathcal{N}_{G_t}$ denote all nodes of graph G_t; $\gamma = 0.0005$ is the weighted decay parameter; θ corresponds to all learnable parameter of the MASTGN model.

III. Evaluation

A. Dataset

To evaluate the performance of the proposed MASTGN model, we verify it on two types of the real-world atmospheric datasets: the urban atmospheric datasets gathered in Japan (J-datasets) that has been collected by the atmospheric environmental regional observation system (AEROS) [18], and the urban atmospheric datasets of China (C-datasets) collected by the Ministry of Ecology and Environment of China (MEMC) [19].

We obtained the data from 24 atmospheric monitoring stations in Tokyo and 101 monitoring stations in Chiba that were regarded as the J-datasets, as well as the data from 35 monitoring stations in Beijing denoted as C-datasets. The J-datasets comprised the information about ten types of air pollutants and four types of weather information. The C-datasets included the information about three types of air pollutants and four types of weather information. The detailed information about these two datasets can be found in Table I. Both considered types of datasets were composed of the atmospheric data recorded every hour. In the J-datasets, the data from 2016/1/1 to 2017/12/31 were used for training, and the data from 2018/1/1 to 2018/12/31 were employed for testing. Concerning the C-datasets, the data from 2017/1/1 to 2017/10/31 were utilized for training, and the data after 2017/11/1 were used for testing.

TABLE I: The details of the considered atmospheric datasets.

	J-datasets	**C-datasets**
Included Cities	Tokyo, Chiba	Beijing
Number of Stations	24,101	35
Air Pollutants	$\mathrm{PM}_{2.5}, \mathrm{SPM},$ $\mathrm{O}_3, \mathrm{NO}_2, \mathrm{NO},$ $\mathrm{NO}_X, \mathrm{SO}_2, \mathrm{NMHC},$ $\mathrm{CO}, \mathrm{CH}_4$	$\mathrm{PM}_{2.5}$
Meteorology	Temperature, Humidity, Wind Speed, Wind Direction	Temperature, Humidity, Wind Speed, Wind Direction
Records Interval	1 hour	1 hour

B. Baseline Models and Evaluation Metrics

We compare the proposed model with the following six baseline ones: MLP : Multi-layer perceptron is the basic neural network method for regression. Here, we adopt a network with the structure similar to the FC network in MASTGN as the baseline MLP. SVR : Support vector regression is a widely used regression method. CNN : Convolution neural network is a representational model for spatial data. We use a CNN comprising two convolution layers, two max-pooling layers, and one fully connected layer with the rectified linear unit (ReLU) activation function. Moreover, we apply IDW to insert a space value. $LSTM$: Long short-term memory [14] is a representational RNN model for time series prediction. $CRNN$: Convolution recurrent neural network [4] is a space-time series prediction method combining CNN and LSTM. Here, IDW is used to insert a space value. $MSSTN$: Multi-scale spatio-temporal graph convolutional networks for the space-time series prediction [20].

We employ the symmetric mean absolute percentage error (SMAPE) and mean absolute error (MAE) for evaluation. The evaluation metrics are defined as follows:

$$SMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \overline{Y}_t|}{(Y_t + \overline{Y}_t)/2}, \tag{11}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|Y_t - \overline{Y}_t|, \tag{12}$$

C. Results

Table II demonstrates the results of one-step prediction for all considered models applied to the J-datasets, concerning $\mathrm{PM}_{2.5}$, O_3, and PM_{10}. represents the prediction results for $\mathrm{PM}_{2.5}$ on the C-datasets. From Tables II, it can be seen that the proposed MASTGN model achieves the best performance for all datasets. Compared with the CNN with spatial interpolation, it can be observed that other models demonstrate superior performance, indicating that the usage of the spatial interpolation results in producing extra noise that deteriorates the prediction results. We can also observe that fitting the atmospheric monitoring station data through a graph network allows achieving better results in terms of accuracy, compared with the other considered models. By exploiting the concept of multiple attention, MASTGN can extract high-level features more effectively compared with others.

To evaluate the long-term prediction performance of MASTGN, we compare the corresponding results for the periods up to 48 hours. For convenience, we represent the prediction results corresponding to the dataset on $\mathrm{PM}_{25}, \mathrm{O}_3, \mathrm{PM}_{10}$ gathered in Tokyo, And compare MASTGN with LSTM, CRNN, and MSSTN models which get top 3 performance on single-step prediction of all baseline models. The prediction results presented in Fig 4a, 4b, 4c indicate that the MASTGN model can succeed in maintaining the advantages of high prediction accuracy also in the case of long-term prediction tasks including various prediction steps. Owing to the utilization of temporal attention, MASTGN can be used to identify positive information on multi-time scales, which has a beneficial effect on prediction targets.

The results presented in Tables II and Fig 4 can be interpreted as the evidence of the successful utilization of the multi-attention approach and the spatio-temporal graph structures. The change in air pollutant concentration is considerably influenced by the internal relations between air pollutants, regional pollutant accumulation, and region transportation, which is mainly captured by MASTGN.

TABLE II: Result of one step prediction on the J-datasets and C-datasets.

Methods	**J-datasets** ($\mathrm{PM}_{2.5}$/O_3/PM_{10})				**C-datasets**($\mathrm{PM}_{2.5}$)	
	Tokyo		**Chiba**		**Beijing**	
	SMAPE	MAE	SMAPE	MAE	SMAPE	MAE
MLP	0.272/0.210/0.321	3.094/0.558/0.348	0.445/0.193/0.477	3.356/0.459/0.396	0.364	12.482
SVR	0.249/0.382/0.530	2.923/0.384/0.345	0.619/0.231/0.459	2.824/0.371/0.364	0.495	9.631
CNN	0.542/0.626/0.505	7.034/0.772/0.567	0.623/0.462/0.562	4.357/0.624/0.462	0.521	14.251
LSTM	0.252/0.266/0.352	3.056/0.402/0.387	0.516/0.230/0.518	4.023/0.571/0.417	0.437	8.910
CRNN	0.242/0.233/0.299	2.832/0.353/0.274	0.465/0.193/0.464	2.765/0.368/0.323	0.273	6.624
MSSTN	0.214/0.184/**0.251**	2.454/0.308/0.169	0.432/0.170/0.429	2.458/0.304/0.312	0.250	**6.293**
MASTGN	**0.196/0.166**/0.261	**2.161/0.263/0.164**	**0.326/0.138/0.396**	**2.001/0.268/0.298**	**0.239**	6.317

IV. Conclusions

In the present paper, we introduced a multi-attention spatio-temporal graph network (MASTGN) model developed to perform air pollution prediction. We considered the atmospheric monitoring stations data obtained from the real-world urban atmospheric systems as a spatio-temporal structure graph that could be fitted by MASTGN. MASTGN was designed to rely on the three types of attention mechanisms: channel attention, spatial attention, and temporal attention, to mine the high-level features in terms of the physical, spatial, and temporal directions in a more effective way. MASTGN included the channel attention networks (CAT), graph attention networks (GAT), a gated 1D convolutional neural network (1D-GCNN), and a fully connected one (FC). CAT was implemented to identify the interdependencies between all atmospheric matters. GAT was used to extract spatial representations based on the mutual influence of each node. 1D-GCNN was realized to generate temporal features through multi-scale time superposition. Finally, FC was intended to combine all extracted features to generate resulting predictions. To confirm the applicability of the proposed model, we conducted the experiments on two datasets gathered in Japan and China to predict the concentration of $\mathrm{PM}_{2.5}$, O_3, and PM_{10}. The results indicated that the proposed MASTGN model surpassed the considered existing baseline approaches in terms of prediction accuracy.

For future research, we mainly focus on two aspects: (1) Clarify the processing capacity of MASTGN model for large-scale graph data. (2) Understand the prediction effect of other spatio-temporal structure graphs except atmospheric data.

References

[1] P.-U. Annette and C. Corvalan, "Preventing disease through healthy environments: towards an estimate of the environmental burden of disease," *Geneva: World Health Organization*, vol. 12, 2006.

[3] Z. Ma, X. Hu, A. M. Sayer, R. Levy, Q. Zhang, Y. Xue, and Y.Liu, "Satellite-based spatiotemporal trends in pm2.5 concentrations: China, 2004-2013," *Environmental Health Perspectives*, vol. 124, 2016.

[2] A. V. Donkelaar, R. V. Martin, M. Brauer, R. Kahn, R. Levy, C.Verduzco, and P. J.Villeneuve, "Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application." *Environmental Health Perspectives*, vol. 118, 2010.

[4] P. Zhao and K. Zettsu, "Convolution recurrent neural networks for short-term prediction of atmospheric sensing data," in *2018 IEEE International Conference on Internet of IEEE Smart Data (SmartData)*, 2018, pp. 815–821.

[5] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep distributed fusion network for air quality prediction." in *KDD 2018*, London, United Kingdom, Aug. 2018, pp. 965–973.

[6] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction." in *In Thirty-First AAAI Conference on Artificial Intelligence.*, California, USA, Feb. 2017, pp. 1655–1661.

[7] F.-W. Chen and C.-W. Liu, "Estimation of the spatial rainfall distribution using inverse distance weighting (idw) in the middle of taiwan," *Paddy and Water Environment*, vol. 10, no. 3, pp. 209–222, 2012.

[8] P. Velivkovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[9] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," *arXiv preprint arXiv:1709.04875*, 2017.

[10] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[11] A. A. Bastidas and H. Tang, "Channel attention networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[12] J. Shukla, "Predictability in the midst of chaos: A scientific basis for climate forecasting," *science*, vol. 282, no. 5389, pp. 728–731, 1998.

[13] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

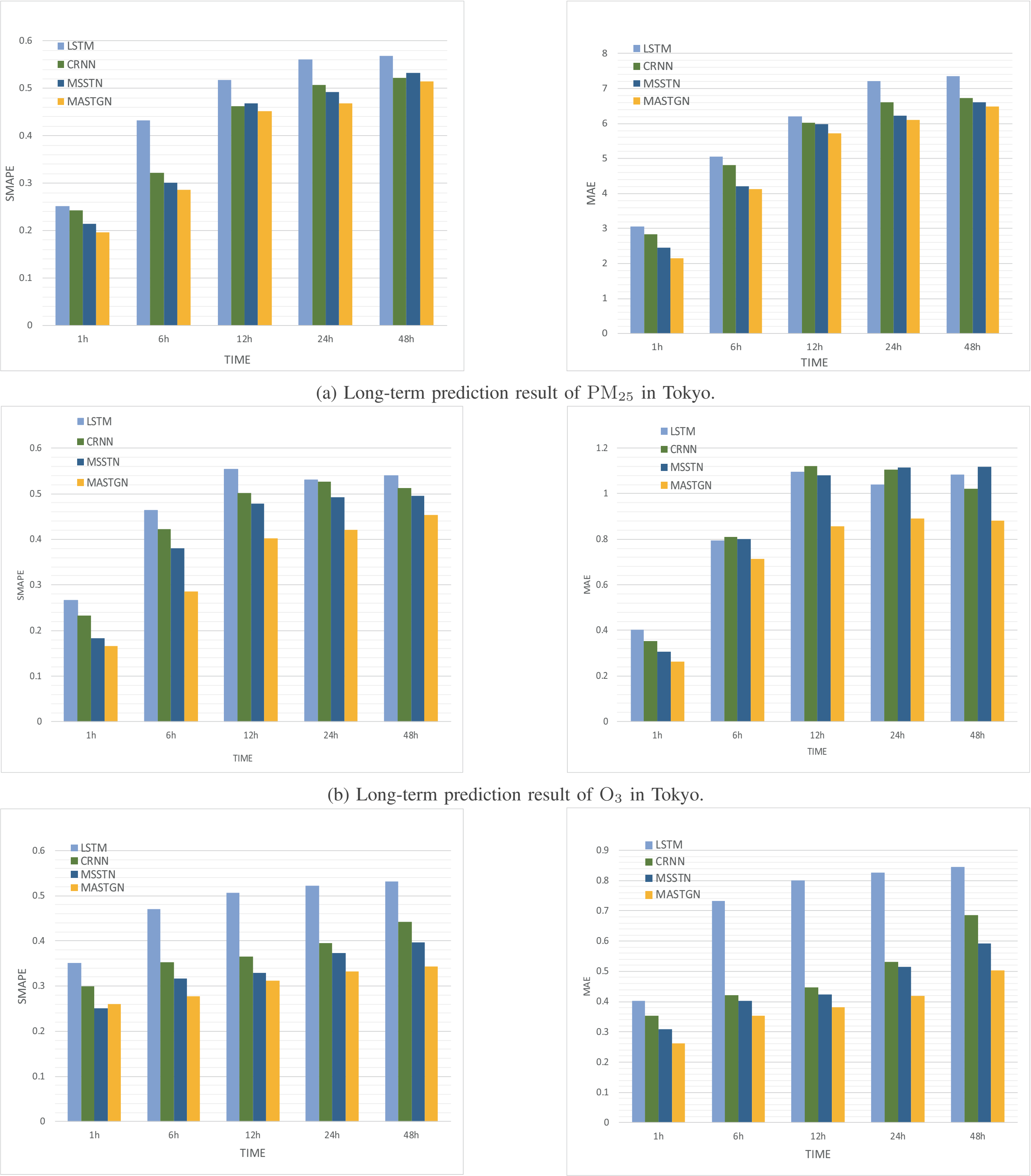
[16] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ecg classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2015.

[17] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[18] *Atmospheric Environmental Regional Observation System*, 2020. [Online]. Available: http://soramame.taiki.go.jp

[19] *Minstry of Ecology and Environment of China*, 2020. [Online]. Available: http://english.mee.gov.cn/

[20] Z. Wu, Y. Wang, and L. Zhang, "Msstn: Multi-scale spatial temporal network for air pollution prediction," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 1547–1556.



(a) Long-term prediction result of PM_{25} in Tokyo.

(b) Long-term prediction result of O_3 in Tokyo.

(c) Long-term prediction result of PM_{10} in Tokyo.

Fig. 4: Result of the long-term prediction of PM_{25}, O_3,PM_{10} in Tokyo. For each figure, left shows the SMAPE and right shows the MEA.