

Large scale air pollution prediction with deep convolutional networks

Gao HUANG^{1†}, Chunjiang GE^{1†}, Tianyu XIONG², Shiji SONG^{1*},
Le YANG¹, Baoxian LIU^{3,4*}, Wenjun YIN⁵ & Cheng WU¹

¹Department of Automation, Tsinghua University, Beijing 100084, China;

²Department of Computer Science, University of Toronto, Toronto M5S1A1, Canada;

³School of Environment, Tsinghua University, Beijing 100084, China;

⁴Beijing Key Laboratory of Airborne Particulate Matter Monitoring Technology, Beijing Municipal Environmental Monitoring Center, Beijing 100084, China;

⁵Insights Value Technology CO., LTD., Beijing 100070, China

Received 17 February 2020/Accepted 15 April 2020/Published online 9 August 2021

Abstract Although considerable success has been achieved in urban air quality prediction (AQP) with machine learning techniques, accurate and long-term prediction is still challenging. One of the key issues for existing AQP approaches is that air quality monitoring stations are sparsely distributed, typically with around ten monitoring stations per city. As air quality may change abruptly in a local area, it is difficult to perform AQP accurately in areas that are far away from observation points. In addition, due to the large distance between every two monitoring stations, we cannot effectively leverage spatial relations among them to improve the AQP accuracy. In this paper, thanks to the development of low-cost air quality sensors, we are now able to collect a large-scale air quality dataset with 393 deployed air quality monitoring stations in a 120 km × 70 km region, which is more than ten times denser than existing AQP datasets. Further, we present a novel method to handle the data effectively. Specifically, we first convert the observed data from irregularly distributed monitoring stations into a regular image-like pollution map, which can then be processed with advanced deep convolutional networks. The experimental results show that the proposed approach can simultaneously model the temporal and spatial relations in our large-scale densely-observed dataset, leading to significantly improved AQP results.

Keywords air pollution prediction, artificial intelligence, convolutional neural network

Citation Huang G, Ge C J, Xiong T Y, et al. Large scale air pollution prediction with deep convolutional networks. *Sci China Inf Sci*, 2021, 64(9): 192107, <https://doi.org/10.1007/s11432-020-2951-1>

1 Introduction

Air quality control has become increasingly important in recent years, as air pollution's negative impacts have been extensively studied and spread to people [1]. In particular, the negative impacts of fine particulate matter (PM_{2.5}) pollution on human respiratory symptoms have been validated by a considerable number of researches [2–4]. Potential implications of PM_{2.5} air pollution include free radical peroxidation, imbalanced intracellular calcium homeostasis, inflammatory injuries, cardiovascular diseases, and lung cancer [5–7]. Hence, air quality prediction (AQP), a fundamental task for air quality control, has attracted great research interests in the artificial intelligence field since decades ago [8–10]. Among many AI-based AQP methods, machine learning has been demonstrated to be rather effective [11, 12] due to its strong modeling capability.

Although advanced machine learning algorithms have greatly improved the precision of AQP, they still have limitations in performing accurate and long-term predictions. It is well known that air quality data, like the PM_{2.5}¹⁾ observations, can be viewed as a typical spatiotemporal dataset, which ideally involves

* Corresponding author (email: shijis@mail.tsinghua.edu.cn, liubaixian28@163.com)

† Authors Huang G and Ge C J have the same contribution to this work.

1) PM_{2.5}, stands for atmospheric particulate matter that has a diameter of fewer than 2.5 μm, which is an essential metric for air quality.

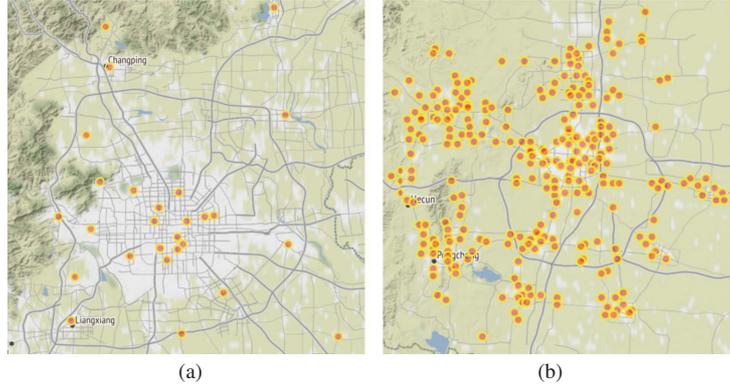


Figure 1 (Color online) A comparison of the number of monitoring stations in previous work [16] (a) and this paper (b). Both sub-figures show a region of approximately $70 \text{ km} \times 70 \text{ km}$. The density of observation points is about ten times higher than that in previous work.

both spatial modeling and temporal modeling. However, classical AQP models, such as auto-regressive moving average (ARMA) [13] and support vector regression (SVR) [14], treat each air quality monitoring station independently, totally ignoring the spatial relations among the air quality values of different stations. Therefore, they are not capable of using the data from neighboring stations to improve AQP accuracy. To address this issue, Zheng et al. [15] proposed the prediction aggregator (PA) to fuse the spatial information in AQP models. This approach is shown to outperform methods that are only using temporal information, but it is a complicated two-stage algorithm which is non-trivial to implement, and the improvement is also limited.

Overall, there are two challenges for existing AQP methods: (1) the lack of densely observed data that allow effective spatial modeling; and (2) the lack of integrated algorithms that can perform both spatial and temporal modeling simultaneously. The first challenge originates from the fact that high precision air quality sensors are expensive, and there are usually a limited number of monitoring stations in each city. Since air pollution has local property, it is difficult to predict the air quality in areas that are far away from monitoring stations. Moreover, we cannot adequately model the spatial dependency of the air quality at different stations as they are geographically far away from each other. The second challenge can be considered as a byproduct of the first issue. Because of the sparsely observed data, researchers have made little effort to develop practical algorithms to incorporate spatial information in temporal AQP.

In this paper, we aim to address the aforementioned two challenges, specifically. First, we collect a large-scale air quality dataset²⁾ with much denser monitoring stations, using low-cost sensors networks, which is shown in Figure 1 [16]. Compared to the existing air pollution datasets, e.g., those used in [14, 17, 18], which contain data observed from less than 40 monitoring stations in a single city, our dataset has 393 air pollution monitoring stations in a region of $120 \text{ km} \times 70 \text{ km}$. This large-scale dataset allows for accurate and long-term prediction of future air quality. Second, we propose a novel deep learning framework to handle this large-scale air quality dataset. This is achieved by transforming the dense air quality data to image-like pollution maps, as the monitoring points are dense enough to produce an image with decent resolution. Then image data can be effectively processed with advanced deep convolutional networks [19, 20]. It is well-known that convolutional networks can effectively explore spatial relations in image data; our model can perform spatial and temporal modeling simultaneously in an end-to-end manner, leading to a simple and accurate approach for AQP.

In summary, our contributions mainly include:

- A large-scale air pollution dataset is collected. The dataset contains the PM_{2.5} values from 393 monitoring stations in a region of $120 \text{ km} \times 70 \text{ km}$ over five months, with one observation per hour per station.
- An end-to-end deep learning framework for spatiotemporal AQP. With the collected densely observed data, we first transform them into 3-dimensional tensors, which contain both the temporal and spatial information. Then a deep convolutional network is deployed to predict the future air quality of the city.

²⁾ We focus on PM_{2.5} in this paper, although our proposed approach can be generally applied to other air quality metrics as well.

- Comprehensive experiments on the large-scale real dataset are conducted to demonstrate the effectiveness of the proposed approach.

2 Related work

2.1 Air quality prediction

The existing AQP methods are generally categorized into deterministic and statistical AQP models. The deterministic AQP models are designed to simulate pollutant discharge, transfer, diffusion, and removal processes with dynamic data of a limited number of monitoring points in a model-driven way [21, 22] based on theoretical meteorological and chemical models in environmental science [9, 22, 23]. However, deterministic AQP models have difficulty in modeling the generation and transformation of air pollution due to insufficient data and incomplete theoretical foundations, and thus frequently achieve low forecasting accuracy in urban AQP [24].

As the complexity in modeling the underlying factors causing air pollution, statistical methods in a data-driven manner have become more and more popular in AQP. The statistical AQP methods can be divided into linear AQP models, such as multi-linear regression [25] and ARMA [13], and nonlinear ones, such as long short-term memory (LSTM) [26, 27] and kernel SVR [14]. Furthermore, a stacked selective ensemble model is proposed for the PM_{2.5} forecast to boost the performance of a single model [28]. Although recent research has shown the high performance of statistical methods in AQP, these methods only predict air quality at monitoring stations and ignore the spatial correlations between the stations. As a strong correlation between these monitoring stations reflects air pollutant dispersion patterns, researchers propose several spatiotemporal AQP models to integrate spatial information during AQP. Zheng et al. [15] proposed a hybrid AQP model, which applies spatial aggregation in a temporal predictor based on a linear regression model. Moreover, the hybrid AQP model named deep distributed fusion network for AQP [29] also utilizes spatial aggregation in a deep network to accurately forecast future air pollution. Nevertheless, the spatial information can only be utilized by spatial aggregation in a non-trivial way because of the lack of densely observed data.

Different from these methods, our AQP model transforms the air quality data into image-like pollution maps due to the proposed densely observed data. Therefore, our network can conduct AQP with spatial and temporal information in an end-to-end manner. This non-trivial implementation further ensures the effectiveness of our network in AQP.

2.2 Deep learning for spatiotemporal prediction in urban computing

With the fast development speed of modern cities, urban computing [30, 31], especially spatiotemporal analysis [32, 33] and prediction [34], has become a popular research field in recent years. The latest research showed the ability of deep learning algorithms to solve spatiotemporal prediction. Lv et al. [35] made traffic prediction with a stacked auto-encoder model to learn generic traffic flow features with spatial and temporal correlations. ST-ResNet [36] is proposed to collectively forecast the inflow and outflow of crowds in each and every region of a city. Liang et al. [37] represented a deep multi-view network to predict taxi demand based on convolutional neural networks (CNN) and LSTM.

More recently, there has been a trend of applying deep learning methods to solve urban air quality challenges in the form of urban computing. U-air [16] is proposed to infer the real-time and fine-grained air quality throughout an entire city. Moreover, a general AQP model based on deep neural networks is proposed for spatiotemporal data [38].

However, these models are mainly built based on sparsely distributed monitoring stations. To effectively handle the proposed air quality dataset with densely deployed monitoring stations, we propose an AQP model based on deep architecture.

3 Large scale air pollution prediction

3.1 Datasets

In this subsection, we give a brief introduction of the proposed large-scale air pollution dataset. The dataset contains the PM_{2.5} concentrations sampled by 393 monitoring stations deployed in a region of

Table 1 Dataset description

Item	Description
Monitoring points	393
Collection time period	September/1st/2018–November/30th/2018 May/1st/2019–June/30th/2019
Resolution	1 hour per record
PM2.5 concentrations	$\mu\text{g}/\text{m}^3$
Longitude	Location information
Latitude	Location information

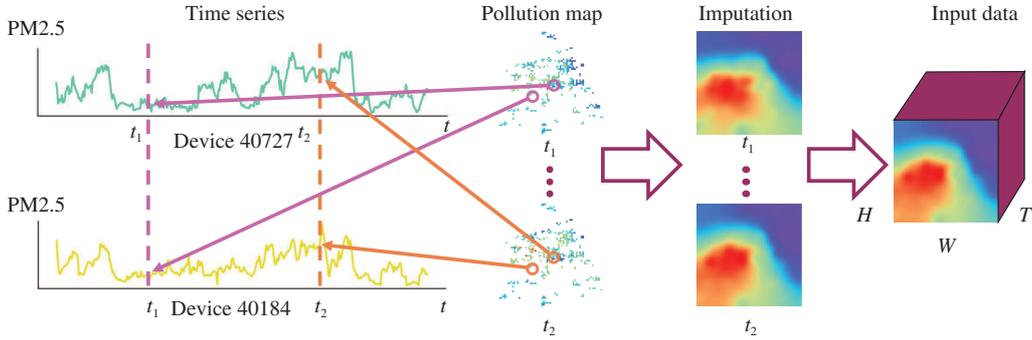


Figure 2 (Color online) An illustration of the transformation of our observed data. The observed PM2.5 concentrations of all monitoring points at time t_1 are firstly transformed into a pollution map according to the stations' geolocation. Then the pollution map is imputed into a regular image, and this transformation is conducted for all observed hours to consist of the input tensor to facilitates the use of a deep convolutional network.

120 km \times 70 km. The PM2.5 concentrations from each monitoring station are with the resolution of one hour or each record, and the datasets are comprised of two time periods: September/1st/2018–November/30th/2018 and May/1st/2019–June/30th/2019 (five months in total). See Table 1 for more details.

In general, our dataset contains two different kinds of information. (a) The geographic location of all monitoring stations. (b) The PM2.5 concentrations recorded as time sequence at each monitoring station, which are denoted by $\mathbf{y}^i = [y_1^i, y_2^i, \dots, y_N^i]^T \in \mathbb{R}^N$, where N and $i \in \{1, 2, \dots, M\}$ (M denotes the number of total monitoring points) are the number of total samples and the i th monitoring station, respectively.

3.2 Problem descriptions and data transformation

Before describing the problem, we first define a pollution map of a given region. The pollution map at time t is represented by matrix $\mathbf{X}_t \in \mathbb{R}^{H \times W}$, where H and W represent the relative height and width of the monitoring region. The value of each element x_t^{ij} represented the PM2.5 concentrations at time t at the corresponding location of the city, which can be further illustrated in Figure 2.

In this paper, we focus on using the pollution maps of T hours to predict the pollution maps of the next K hours. More specifically, we aim at predicting the pollution map $\hat{\mathbf{X}}_j$ at time t in next K hours, where $j \in \{t+1, \dots, t+K\}$, by a series of available pollution maps \mathbf{X}_i , where $i \in \{t-T+1, t-T+2, \dots, t-1, t\}$.

During the data transformation, we firstly equally divided the monitoring region into a $H \times W$ ($H \times W > M$) grid map based on the longitude and latitude, where the H and W can be selected to ensure that each grid contains less than one monitoring point. After defining the H and W , the pollution map \mathbf{X}_t is obtained by setting x_t^{ij} as the values of the PM2.5 concentrations of the corresponding monitoring points in time t . We repeat this procedure until we have the pollution maps of all valuable time.

Even the number of monitoring stations is much denser than in other researches. It is still too sparse as a neural network input. Therefore, an imputation procedure is further conducted, which is described as follows: for each point which does not have a corresponding monitor, we set an $m \times m$ grid window and the average value of the points in the grid window is considered as the value for the point [39]. We denote the

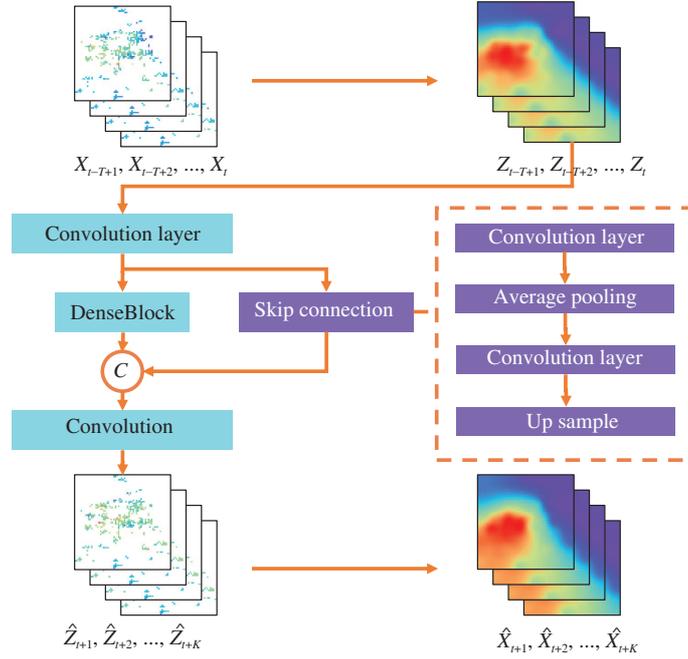


Figure 3 (Color online) The architecture of the proposed convolutional network.

imputed pollution maps of \mathbf{X}_i as \mathbf{Z}_i . Our task is then transformed to predict $\hat{\mathbf{Z}}_j$ ($j \in \{t+1, \dots, t+K\}$) at time t by \mathbf{Z}_i ($i \in \{t-T+1, \dots, t\}$) using a deep network. The predicted pollution map $\hat{\mathbf{X}}_j$ can then be extracted from the obtained $\hat{\mathbf{Z}}_j$.

3.3 The proposed network

In this subsection, we proposed our deep network, which is modified from the original DenseNet [20] for pixel-wise regression, and provides a possible way to predict the pollution map accurately. DenseNet [20] connects all layers with each other directly which facilitates the flow of information and reuse of features. Moreover, DenseNet [20] also requires fewer parameters than traditional convolutional neural networks. We design our proposed model based on DenseNet due to its computational efficiency and ability of feature extraction.

The proposed network is illustrated in Figure 3. It is comprised of three parts.

(1) A head layer. Given the pollution maps of T hours, the head layer is set to a 3×3 convolutional layer [40, 41]³⁾ with the number of input and output channels set to T and 24, respectively.

(2) A dense connection block (dense block). We apply a 4-layer (with bottleneck layer) dense block [20] in our network. The block applies the dense connection structure to explore the temporal variation of the pollution maps. The growth rate of the dense block is set to be 16. The dense block aims to learn the temporal trends of the air pollution maps and the correlation between monitoring stations from a global view.

(3) A skip connection block. Just like the skip connection in U-net [18] can refine the detailed information in pixel classification tasks, the skip connection in the proposed network can learn the regional variation trends of the pollution and thus can improve the performance of the proposed network. In the skip connection block, the features firstly pass through a 3×3 convolutional layer following by an average pooling layer. Then the obtained representations are fed to a 3×3 convolutional layer following by an up-sampling layer.

(4) A pixel level regression. After we concatenate the features from the dense block and the skip connection block, the obtained features are then fed into a 3×3 convolutional layer with the number of output channels as K to generate the final feature maps, where K is the number of the hours of the pollution maps that we want to predict. The K pollution maps are derived from selecting the values in

³⁾ In this paper, a convolutional layer sequentially contains a convolutional layer, a batch-normalization layer [40], and a ReLU layer [41].

the final feature maps. The objective function of the proposed network is set as

$$\text{Loss} = \frac{1}{KM} \sum_{i=1}^M \sum_{j=t}^{t+K-1} \lambda^{j-1} \|y_j^i - \hat{y}_j^i\|_1, \quad (1)$$

where y_j^i and \hat{y}_j^i are the observed and predicted PM2.5 concentrations at time t for i th monitoring points. Moreover, a decay factor $\lambda \in [0, 1]$ is introduced in the objective function for balancing the impact from different hours.

4 Experiments

In this section, three experiments are conducted to evaluate the benefits of AQP with a large-scale dataset. We further demonstrated the effectiveness of the proposed method in both one-hour AQP and multi-hour AQP.

4.1 Experiments setting

4.1.1 Data separation

In this subsection, we evaluate the AQP at Handan with the proposed large-scale dataset with different methods. The data from September/1st/2019 to November/30th/2019 are equally divided into five subsets, where the first three subsets are for training, and the rest two are used as the validation set and test set, respectively. Moreover, we further add the data from May/1st/2019 to June/30th/2019 as the training data. The input of our model is sliced from a time dimension with a length of $T = 4$ and a stride of 1.

4.1.2 Implementation details

The implementation details of our model are described as follows. Our optimizer is Adam [42] with a mini-batch size of 512. We train our model for a total of 70 epochs. The learning rate is set to 0.008 initially and is divided by 5 at the 30th and the 50th epoch, respectively. We use a weight decay of 5×10^{-4} and a momentum of 0.9. The decay factor λ is set as 0.95 in multi-hour AQP.

4.1.3 Evaluation metrics

To evaluate the performance of each AQP method, two metrics [26] for a predicted pollution map, including the mean absolute error (MAE) and the mean absolute percentage error (MAPE), are used in our experiments, which are respectively defined as follows:

$$\text{MAE} = \frac{1}{M} \sum_{i=1}^M \|\hat{y}^i - y^i\|_1, \quad (2)$$

$$\text{MAPE} = 1 - \frac{1}{M} \sum_{i=1}^M \frac{\|\hat{y}^i - y^i\|_1}{y^i}, \quad (3)$$

where y_i and \hat{y}_i are observed PM2.5 concentrations and predicted PM2.5 concentrations respectively, and M is the number of monitors.

4.1.4 Baseline models

We compare the proposed network with the following baselines AQP models:

- **SVR** [14]. SVR is an optimal regressor which minimizes a hinge loss. We use radial basis function (RBF) [43] as the kernel function in order to increase the ability of non-linearity fit.
- **GBDT** [44]. Gradient boosting decision tree is a regressor commonly used in the literature.
- **LSTM** [45]. Long short term memory network is commonly used to deal with sequence data.
- **1D CNN**. One-dimensional convolutional neural network with filters slide on the time dimension.
- **PA** [15]. Prediction aggregator combines a temporal model kernel SVR [14] which is mentioned before and the spatial predictor proposed by [15] making use of temporal and spatial information. Our implementation, which is an improved version of Zheng's model, increases the ability of non-linearity fit.

Table 2 One-hour AQP results

Type	Model	MAE	MAPE	Time (s)
Temporal	GBDT	23.81	79.61	0.0490
	SVR	20.96	81.47	0.0184
	LSTM	16.84	83.28	0.1190
	1D CNN	17.58	82.15	0.1231
Spatiotemporal	PA	17.89	82.36	0.0223
	Our model	13.26	85.04	0.0070

Table 3 Three-hour AQP results

Hour	Our method		PA	
	MAE	MAPE	MAE	MAPE
The 1st	15.38	83.37	17.89	82.36
The 2nd	19.36	77.52	26.91	72.60
The 3rd	25.24	71.17	34.66	64.29

4.2 One hour AQP

We first compare the proposed method with baseline AQP methods in the one-hour AQP setting. The experimental results are shown in Table 2.

From Table 2 we conclude that:

- The proposed method outperforms other non-neural network models (SVR and GDBT), which shows the effectiveness of the neural network-based method in AQP due to its strong ability in analyzing complex data.
- The experiments show that the spatiotemporal models are superior to the temporal-based models, from which we can conclude the importance of the spatial information in AQP. As the spatiotemporal AQP models disentangle the complex correlation between different monitoring stations, they can provide more precise PM2.5 concentrations predictions.
- Although the prediction aggregator and our model both make use of temporal and spatial information, our model achieves better results. Different from the prediction aggregator, which processes the spatial and temporal information separately, our method conducts the AQP in an end-to-end manner, which facilitates the model to learn more iteration information between two different kinds of information.

To further evaluate the effectiveness of our model, we plot the one-hour PM2.5 prediction results of two monitoring stations on the test set, which is illustrated in Figure 4. LSTM and PA are also implemented as comparisons. Moreover, we randomly select one predicted pollution map and visualize it in Figure 5. The low MAE shown in two figures verifies the effectiveness of our method. Table 2 also shows that our model achieves a faster inference speed with all the sites' data parallelly running on GPU by the implementation of deep learning models.

4.3 Multi-hour AQP

We further evaluate the proposed method in a multi-hour AQP setting, where we predict the pollution maps for the next three hours. We conduct the multi-hour AQP experiments on the proposed method and PA due to the ability of both methods to model the spatiotemporal information. The results are shown in Table 3.

From the experimental results, we conclude that:

- In general, our model outperforms the prediction aggregator for a stronger ability in the non-linear fitting of a deep network. The proposed method always achieves a lower MAE and higher MAPE, which means that our method accurately forecasts the future air quality in the next few hours. The results further indicate that our model has a longer-term prediction in AQP compared with the famous AQP model, prediction aggregator.
- The predicted error is becoming higher with the prediction in a longer term, which meets our intuition. Predicting the pollution in the long term can still be challenging for the uncertainty of complex factors causing air pollution.
- The performance of the 1st hour prediction in Table 2 is better than that in Table 3 because the model has to balance three hours' loss during the AQP procedure.

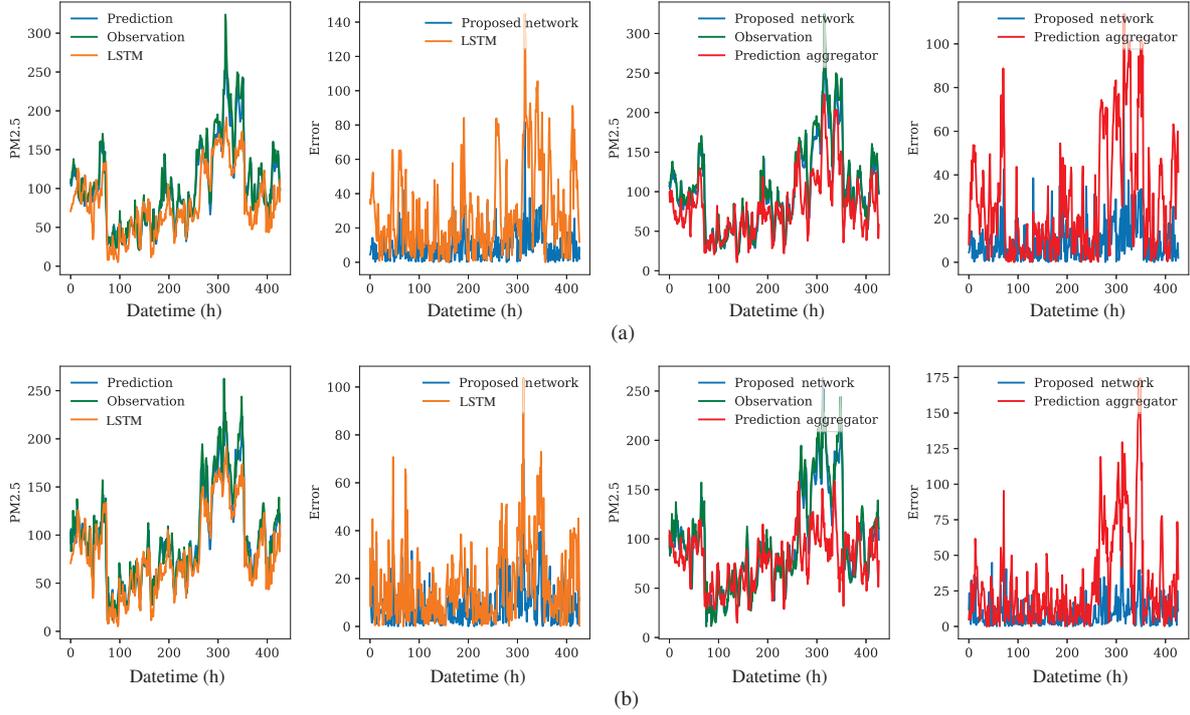


Figure 4 (Color online) A comparison of the proposed model, temporal model and existing spatiotemporal model for AQP at two randomly selected monitoring stations. (a) Monitor 10965's result; (b) monitor 16038's result.

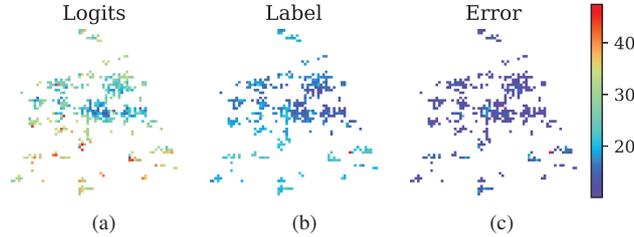


Figure 5 (Color online) The prediction of the proposed method (a), the real observation values (b), and mean absolute error (c) at a randomly selected time slot.

Table 4 Performance v.s. the number of monitors points

Monitors points	MAE	MAPE
20	21.54	72.21
30	20.54	72.98
50	14.12	78.68
393	13.26	85.04

4.4 Ablation study

4.4.1 Density of monitoring stations of the dataset

As the insufficiency of monitoring stations can limit the models to explore the spatial correlation between each monitoring point, we study the impact of the number of monitoring stations on the performance of AQP methods. As it is shown in Table 4, we demonstrate that denser monitoring points can provide a more accurate prediction in AQP, which further indicates the importance of the proposed dataset.

4.4.2 Spatial information of the dataset

In this subsection, we generate coordinates for each monitor randomly and transform the ordinal data into pollution maps to study the importance of spatial information in AQP. As such data does not have the correct geolocation information compared with the original data, the proposed method on this random

Table 5 Study of the effect of incorporating spatial information

Model	MAE	MAPE
Spatial information	13.26	85.04
Non-Spatial information	15.04	83.32

Table 6 Study on the effectiveness of the skip connection

Model	MAE	MAPE
Our method	13.26	85.04
Our method-nonSkip	14.74	84.29

Table 7 Study on the effectiveness of the dense connections

Model	MAE	MAPE
Our method	13.26	85.04
Without-dense-connections	15.00	83.36

pollution maps achieves a lower MAE and a higher MAPE in the experiments (Table 5), which shows the necessity for applying the spatial information in AQP.

4.4.3 Skip connection in the proposed network

We further conduct the experiments to show the effectiveness of the skip connection block in our network. The experiments on the proposed network with or without skip connection are shown in Table 6. Downsampling can enlarge the inception filed for making use of neighborhood points information. Moreover, similar to the skip connection in U-Net [18], the skip connection in our network can provide more regional information in AQP, which results in the higher performance in AQP.

4.4.4 Dense connections in the proposed network

We furthermore conduct the experiments to represent the effectiveness of the densely connected block in our proposed network. We replace our 4-layer dense block with 4-layer convolution layers which have the same number as channels of the dense block in order to keep the number of parameters unchanged. The experimental results are shown in Table 7. Dense connections can make better use of features and are of high efficiency [20]. Dense connections could preserve the information from shallow layers which could be of help to AQP.

5 Conclusion

To accurately predict future air quality (we focused on PM_{2.5} value in this paper) in an urban city, we collected a large-scale air quality dataset with 393 deployed air quality monitoring stations in a region of 120 km × 70 km and proposed a novel AQP approach based on deep convolutional networks. As the presented dataset is more than ten times denser than existing AQP datasets, we transform the data to image-like pollution maps, which can be effectively processed by convolutional networks. Different from the existing AQP methods, the proposed network simultaneously utilizes the temporal and spatial information to predict future air quality in an end-to-end manner and expertly explores the spatial correlation between each monitoring point, leading to a state-of-the-art performance in AQP. For future work, we will consider incorporating terrain information, wind data, and other factors to improve prediction accuracy further.

Acknowledgements This work was supported in part by National Science and Technology Major Project of the Ministry of Science and Technology of China (Grant No. 2018AAA0100701), National Natural Science Foundation of China (Grant Nos. 61906106, 62022048), and Institute for Guo Qiang of Tsinghua University and Beijing Academy of Artificial Intelligence.

References

- 1 Stern A C. Air Pollution: the Effects of Air Pollution. Amsterdam: Elsevier, 1977
- 2 Brunekreef B, Holgate S T. Air pollution and health. *Lancet*, 2002, 360: 1233–1242
- 3 Chow J C. Health effects of fine particulate air pollution: lines that connect. *J Air Waste Manage Assoc*, 2006, 56: 707–708

- 4 Dominici F, Peng R D, Bell M L, et al. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, 2006, 295: 1127–1134
- 5 Xing Y F, Xu Y H, Shi M H, et al. The impact of PM_{2.5} on the human respiratory system. *J Thoracic Dis*, 2016, 8: 69
- 6 Brook R D, Rajagopalan S, Pope C A, et al. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the american heart association. *Circulation*, 2010, 121: 2331–2378
- 7 Pope C A, Burnett R T, Thun M J, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*, 2002, 287: 1132–1141
- 8 James D E, Chambers J A, Kalma J D, et al. Air quality prediction in urban and semi-urban regions with generalised input-output analysis: the hunter region, australia. *Urban Ecol*, 1985, 9: 25–44
- 9 Bruckman L. Overview of the enhanced geocoded emissions modeling and projection (enhanced gemap) system. In: *Proceeding of the Air & Waste Management Association's Regional Photochemical Measurements and Modeling Studies Conference*, San Diego, 1993
- 10 Gu K, Qiao J F, Li X L. Highly efficient picture-based prediction of PM_{2.5} concentration. *IEEE Trans Ind Electron*, 2019, 66: 3176–3184
- 11 Corani G. Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. *Ecol Model*, 2005, 185: 513–529
- 12 Russo A, Raischel F, Lind P G. Air quality prediction using optimal neural networks with stochastic variables. *Atmos Environ*, 2013, 79: 822–830
- 13 Box G E P, Pierce D A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J Am Stat Assoc*, 1970, 65: 1509–1526
- 14 Liu B C, Binaykia A, Chang P C, et al. Urban air quality forecasting based on multi-dimensional collaborative support vector regression (SVR): a case study of Beijing-Tianjin-Shijiazhuang. *PLoS ONE*, 2017, 12: 0179763
- 15 Zheng Y, Yi X W, Li M, et al. Forecasting fine-grained air quality based on big data. In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015. 2267–2276
- 16 Zheng Y, Liu F R, Hsieh H P. U-air: when urban air quality inference meets big data. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013. 1436–1444
- 17 Kurt A, Oktay A B. Forecasting air pollutant indicator levels with geographic models 3days in advance using neural networks. *Expert Syst Appl*, 2010, 37: 7986–7992
- 18 Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015. 234–241
- 19 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 770–778
- 20 Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 4700–4708
- 21 Kim Y, Fu J S, Miller T L. Improving ozone modeling in complex terrain at a fine grid resolution: part I — examination of analysis nudging and all PBL schemes associated with LSMs in meteorological model. *Atmos Environ*, 2010, 44: 523–532
- 22 Baklanov A, Mestayer P G, Clappier A, et al. Towards improving the simulation of meteorological fields in urban areas through updated/advanced surface fluxes description. *Atmos Chem Phys*, 2008, 8: 523–543
- 23 Jeong J I, Park R J, Woo J H, et al. Source contributions to carbonaceous aerosol concentrations in Korea. *Atmos Environ*, 2011, 45: 1116–1125
- 24 Stern R, Builtjes P, Schaap M, et al. A model inter-comparison study focussing on episodes with elevated PM₁₀ concentrations. *Atmos Environ*, 2008, 42: 4567–4588
- 25 Li C, Hsu N C, Tsay S C. A study on the potential applications of satellite data in air quality monitoring and forecasting. *Atmos Environ*, 2011, 45: 3663–3675
- 26 Li X, Peng L, Yao X J, et al. Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. *Environ Pollution*, 2017, 231: 997–1004
- 27 Gu K, Qiao J F, Lin W S. Recurrent air quality predictor based on meteorology- and pollution-related factors. *IEEE Trans Ind Inf*, 2018, 14: 3946–3955
- 28 Gu K, Xia Z F, Qiao J F. Stacked selective ensemble for PM_{2.5} forecast. *IEEE Trans Instrum Meas*, 2020, 69: 660–671
- 29 Yi X W, Zhang J B, Wang Z Y, et al. Deep distributed fusion network for air quality prediction. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018. 965–973
- 30 Zheng Y, Capra L, Wolfson O, et al. Urban computing: concepts, methodologies, and applications. *ACM Trans Intell Syst Technol*, 2014, 5: 38
- 31 Xiong Z, Sheng H, Rong W G, et al. Intelligent transportation systems for smart cities: a progress review. *Sci China Inf Sci*, 2012, 55: 2908–2914
- 32 Deng M, Liu Q L, Wang J Q, et al. A general method of spatio-temporal clustering analysis. *Sci China Inf Sci*, 2013, 56: 102315
- 33 Wang C M, Hu X P, Yao L, et al. Spatio-temporal pattern analysis of single-trial EEG signals recorded during visual object recognition. *Sci China Inf Sci*, 2011, 54: 2499–2507
- 34 Wang W, Hu C B, Chen N C, et al. Spatio-temporal enabled urban decision-making process modeling and visualization under the cyber-physical environment. *Sci China Inf Sci*, 2015, 58: 100105

- 35 Lv Y S, Duan Y J, Kang W W, et al. Traffic flow prediction with big data: a deep learning approach. *IEEE Trans Intell Transp Syst*, 2015, 16: 865–873
- 36 Zhang J B, Zheng Y, Qi D K. Deep spatio-temporal residual networks for citywide crowd flows prediction. In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017
- 37 Liang Y X, Ke S Y, Zhang J B, et al. Geoman: multi-level attention networks for geo-sensory time series prediction. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018. 3428–3434
- 38 Zhang J B, Zheng Y, Qi D K, et al. Dnn-based prediction model for spatio-temporal data. In: *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2016
- 39 Parzen E. On estimation of a probability density function and mode. *Ann Math Stat*, 1962, 33: 1065–1076
- 40 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015. [ArXiv:1502.03167](https://arxiv.org/abs/1502.03167)
- 41 Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010. 807–814
- 42 Kingma D P, Ba J. Adam: a method for stochastic optimization. 2014. [ArXiv:1412.6980](https://arxiv.org/abs/1412.6980)
- 43 Broomhead D S, Lowe D. Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks. Royal Signals and Radar Establishment Malvern (United Kingdom) Technical Report, 1988
- 44 Friedman J H. Greedy function approximation: a gradient boosting machine. *Ann Stat*, 2001, 29: 1189–1232
- 45 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*, 1997, 9: 1735–1780