Hindawi
Mobile Information Systems
Volume 2021, Article ID 6630944, 12 pages
https://doi.org/10.1155/2021/6630944

Research Article

Air Quality Prediction Based on a Spatiotemporal Attention Mechanism

Xiangyu Zou,[1,2] **Jinjin Zhao**,[1,2] **Duan Zhao**,[1,2] **Bin Sun**,[1,2] **Yongxin He**,[1,2] **and Stelios Fuentes**[3]

[1]*National and Local Joint Engineering Laboratory of Internet Application Technology on Mine, China University of Mining and Technology, Xuzhou 221116, China*
[2]*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China*
[3]*Leicester University, Leicester, UK*

Correspondence should be addressed to Jinjin Zhao; 295318654@qq.com

Received 25 November 2020; Revised 24 January 2021; Accepted 8 February 2021; Published 19 February 2021

Academic Editor: Xiaoxian Yang

With the rapid development of the Internet of Things and Big Data, smart cities have received increasing attention. Predicting air quality accurately and efficiently is an important part of building a smart city. However, air quality prediction is very challenging because it is affected by many complex factors, such as dynamic spatial correlation between air quality detection sensors, dynamic temporal correlation, and external factors (such as road networks and points of interest). Therefore, this paper proposes a long short-term memory (LSTM) air quality prediction model based on a spatiotemporal attention mechanism (STA-LSTM). The model uses an encoder-decoder structure to model spatiotemporal features. A spatial attention mechanism is introduced in the encoder to capture the relative influence of surrounding sites on the prediction area. A temporal attention mechanism is introduced in the decoder to capture the time dependence of air quality. In addition, for spatial data such as point of interest (POI) and road networks, this paper uses the LINE graph embedding method to obtain a low-dimensional vector representation of spatial data to obtain abundant spatial features. This paper evaluates STA-LSTM on the Beijing dataset, and the root mean square error (RMSE) and *R*-squared (R^2) indicators are used to compare with six benchmarks. The experimental results show that the model proposed in this paper can achieve better performance than the performances of other benchmarks.

1. Introduction

The rapid development of next-generation information technologies such as the Internet of Things and Big Data has promoted the concept of "smart cities." Smart cities use information and communication technology (ICT) to make city services and monitoring highly perceptual, interactive, and efficient, thereby promoting city harmony and sustainable development [1]. Among these technologies, the construction of smart environments is an important part of smart cities because air pollution is one of the most important factors that seriously threaten people's health [2]. A large number of diversified air quality monitoring systems are currently deployed in cities. For example, an air quality monitoring station is set up at a specific location in the city to monitor the conventional pollution factors (PM2.5, PM10, SO2, etc.) and meteorological parameters (temperature, humidity, etc.) at all hours [3]. In addition, Yang [4] designed a UAV-based mobile sensing system to effectively capture meter-level air quality index (AQI) changes while also analyzing the corresponding fine-grained distribution. However, monitoring the air quality alone is not enough to meet the needs of smart city construction. Analyzing and mining dynamic city data is an inevitable step in building a smart city [5]. The prediction of air quality can provide early warnings to the public and the government before serious air pollution occurs, enabling them to take corresponding emergency measures as soon as possible [6]. Therefore, the air quality analysis and prediction of the acquired big data are essential parts of constructing smart cities.

The AQI is calculated from six major pollutants, including SO_2, NO_2, PM10, PM2.5, CO, and O_3, to evaluate

daily air quality. However, the prediction of the AQI requires the consideration of more influencing factors. Figure 1(a) shows a true description of the physical world at different moments. Figure 1(b) shows the mathematical model and models the physical world in Figure 1(a), where the nodes represent the area where the air monitoring station is located at different times. It shows the factors influencing air quality prediction, including time, space, and nonsequential information. Zhang et al. [7] pointed out that the geosensory time series, similar to an air quality sequence, usually follows a periodic pattern, which changes with time. In addition, the air quality is also affected by complex spatial factors. For example, if the environment around the predicted area is good, then its air quality will also be good and will change nonlinearly with time. In addition, nonsequential information such as the POI and road network [8, 9] also affects the prediction of air quality. For example, the air quality near a park is much better than the air quality near a factory. The road network has a strong correlation with the mode of traffic. Traffic flow is one of the main factors contributing to air pollution [10], so it also reflects the air quality to a certain extent. In other words, air quality prediction is affected by many factors in time and space, and this is also a major challenge.

Recently, there have been many studies on the prediction of air quality. Qin et al. [11] only took the meteorological conditions and pollutant concentrations in the past few hours as the input of their prediction model. Huang and Kuo [3] combined a convolutional neural network (CNN) and LSTM [12] for air quality prediction. The model achieved good prediction results for time-series data (meteorological data, traffic flows, factories, etc.). However, the proposed model could not handle nonsequential information related to spatial features such as POIs and road networks. Zhao et al. [13] proposed that the use of processing times and non-time-series information separately can better capture the impact of temporal and spatial characteristics on air quality prediction than using both together, and it also considered the impact of adjacent areas on the measured area. The modeling method is more conducive to the prediction of air quality than other methods. However, different neighboring areas have different effects on the target area. If we treat the spatial impact of each region equally, the prediction effect may have climbing space. In other words, the existing works may have the following defects: (1) the time factors are not considered comprehensively; (2) the nonsequential information is not handled well; and (3) existing methods fail to fully consider spatial factors, for example, the correlation between the surrounding area and the predicted area is different due to distance, POI, etc.

Therefore, to solve the defects of the existing works, this paper proposes an LSTM prediction model based on a spatiotemporal attention mechanism (STA-LSTM), whose structure is in the form of an encoder-decoder, and its purpose is to predict the air quality index in the next few hours. First, this paper considers various complex factors that affect air quality prediction, including information data related to temporal characteristics and spatial characteristics. The temporal information mainly includes the AQI, meteorological data (temperature, humidity, wind speed, wind direction, etc.), traffic flows, and factory emissions in the past few hours, and the nonsequential information includes POIs and road networks. Then, the paper uses an LSTM network that is good at handling long-term sequences for analysis and processing according to the characteristics of time-series information. For non-time-series information that cannot be directly processed by deep learning models, this paper considers using the LINE method [14] of graph embedding to transform the information into a vector and then use that vector as the input of the model. Finally, to model the dynamic temporal and spatial dependence, we incorporate a spatiotemporal attention mechanism into the model [15, 16]. In the encoder, spatial attention is introduced to capture the different influences of the surrounding areas at different distances from the target area. In the decoder, we introduce temporal attention to select relatively important historical time information. Compared with the method of giving equal weights to different regions in [11], the model proposed in this paper can obtain more accurate prediction results and higher performance.

The contributions of this paper are as follows:

(1) This paper proposes an STA-LSTM model based on spatiotemporal attention that not only considers time-series information (such as historical AQIs and meteorological data) but also uses non-time-series information (POIs and road networks) as auxiliary predictors. The model adopts the LSTM network and LINE graph embedding method to extract features.

(2) The model proposed in this paper uses an encoder-decoder structure and introduces a spatiotemporal attention mechanism, and it can automatically capture the relative dependence of time and space.

(3) The deep learning model proposed in this paper can jointly grasp and predict air quality locally and globally. Compared with other benchmark air quality prediction models, the accuracy and performance of the model in this paper are greatly improved.

The rest of this paper is arranged as follows. Section 2 summarizes the related works. In Section 3, we introduce the details of the model presented in this paper. Section 4 presents the experiment conducted in the paper. We present a summary and conclusion in Section 5.

2. Related Works

With the rapid development of science and technology, many fields have involved forecasting technologies, such as personnel trajectory forecasting, traffic forecasting, air quality forecasting, and other daily fields. In addition, optimization problems [17, 18], quality-of-service prediction [19], and user recommendations [20, 21] also involve prediction technology. In this paper, we mainly study the prediction of air quality because it is an important part of building a smart city and it is closely related to people's lives and health. At present, there are many studies on air quality

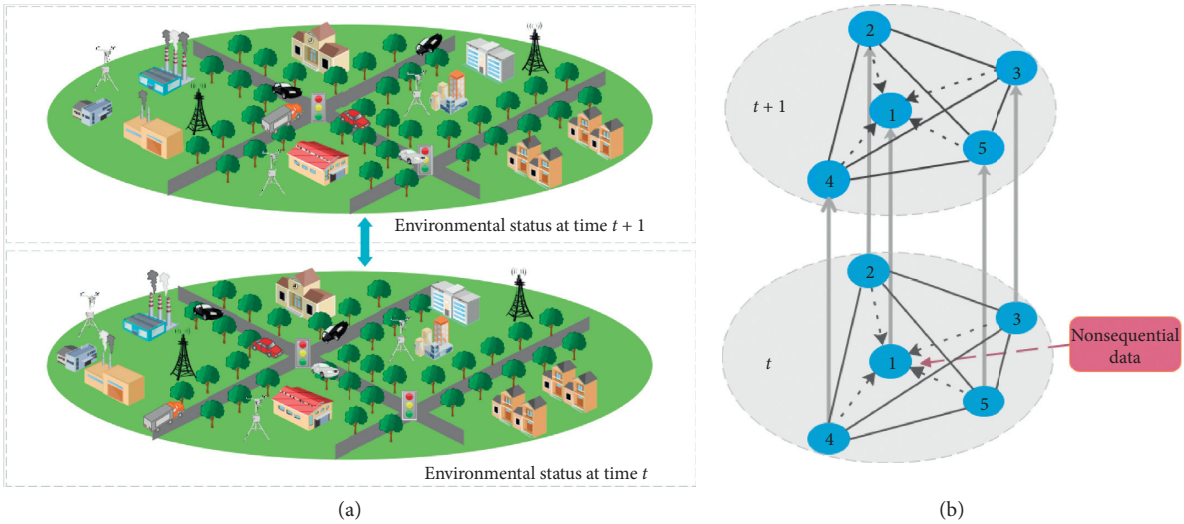

FIGURE 1: Influencing factors of air quality. (a) Physical world. (b) Mathematical model.

prediction, which can be roughly divided into prediction methods based on physical models, traditional linear statistical models, machine learning techniques, and deep learning. Among them, the prediction method based on a physical model uses a physical model to simulate the formation, diffusion, and transfer of various pollutants in the air to predict the concentration of air pollutants. But most of the prediction methods based on physical models require many empirical parameters and assumptions, which may be true for a specific environment but not for all urban environments [22]. Therefore, to obtain more accurate prediction results than those obtained by these methods, an increasing number of researchers have proposed data-driven methods to predict air quality, including traditional linear statistical models, machine learning techniques, and deep learning methods.

The method based on the traditional linear statistical model is used to describe the linear relationship between air quality and related impact characteristics. Jian et al. [23] used the autoregressive integrated moving average (ARIMA) to predict the effects of meteorological factors on the concentration of submicron particles. In [24], Genc et al. used a multiple linear regression model to predict Ankara's air pollution index. Moisan et al. [25] proposed a method based on dynamic multivariate linear equations to predict PM2.5 pollution concentrations at different monitoring stations. The abovementioned studies are all based on linear model prediction methods. However, the relationships between air quality and its related factors are mostly nonlinear. The linear models mentioned above do not represent their complex interrelationships well.

Therefore, machine learning technology has received increasing attention for air quality prediction. This prediction method takes the nonlinearity between air quality and its influencing factors into account and is more suitable for describing problems with complex relationships. For example, Niu et al. [26] proposed an integrated empirical mode decomposition and least-squares support vector machine (LSSVM) method based on phase space reconstruction for PM2.5 concentration prediction. However, for complex problems with high-dimensional nonlinear long-term time series, machine learning methods still seem to be incapable of solving them [27].

With the rapid growth in data volume, the advantages of deep learning methods in responding to forecasting problems are slowly being revealed. Lipton et al. [28] found that the recurrent neural network (RNN) model showed very good performance when modeling a time structure. Zhao et al. [29] proposed a model based on LSTM and the firework algorithm to predict the air quality of Wuhan. The RNN and LSTM networks mentioned in the above studies are deep learning methods that can model a time structure very well. However, the RNN is very sensitive to short sequence data, and once the data are very long, the problems of gradient disappearance and gradient explosion appear. LSTM is better at processing longer time-series data, so it is more suitable for the air quality prediction problem in this paper.

There have been many studies on applying deep learning to air quality prediction. Zhang et al. [30] combined a CNN and an LSTM network to forecast air quality. The model achieved good prediction results for time-series data (meteorological data, traffic flows, factory air pollutant emissions, etc.). Ge et al. [31] regarded time-series and non-time-series information as influencing factors in air quality prediction. Qi et al. [32] proposed a mixed model called GC-LSTM, in which graph convolutional networks were used to extract the spatial correlation between different sites, and LSTM was used to capture the temporal correlation between different time observations. The fully connected neural network based on spatial combination was used to capture the correlation between the target area and its five neighboring sites in [13]. However, different surrounding areas may have different effects on the target area due to the distance between them or differences in their POI types. Therefore, this paper proposes to introduce

a spatiotemporal attention mechanism into the model to capture the relative importance of different surrounding areas.

The essence of the attention mechanism comes from human visual attention. For example, when observing a scene, people pay attention to a specific part of the scene according to their own needs, and they ignore irrelevant information [33]. The attention mechanism was originally used in machine translation [34], but it is now an important part of neural network structures, and it is also widely used in image processing, speech recognition, and computer-related fields [35]. In the recent literature, Li et al. [36] proposed to use the attention mechanism to capture the most important part of the past state, but ignored the relative importance of neighboring sites. In addition, non-time series (road network and POI) also affects the prediction of the target area. Therefore, in response to the above problems, this paper introduces a spatiotemporal attention mechanism to capture temporal and spatial correlations in air quality prediction.

3. Problem Definition and Model Framework

This section first defines the air quality prediction problem, then proposes the overall model for prediction, and finally introduces the various components of the model in detail.

3.1. Problem Definition. Assuming that there are n regions with air quality monitoring stations, the characteristics of the time series for prediction can be obtained. The time series of the area i to be predicted is expressed as $\mathbf{X}^i = (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n)^{\mathrm{T}} \in \mathbf{R}^{n\times T}$, where T is the length of the set time window, n indicates the number of time series (including the AQI index, meteorological data, traffic flows, and factory pollution emissions), and the row vectors $\mathbf{x}$ represent the time series of each feature considered in this paper. At the same time, X^ican also be expressed as $\mathbf{X}^i = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) \in \mathbf{R}^{n\times T}$, where $\mathbf{x}_t = (x_t^{1,i}, x_t^{2,i}, \ldots, x_t^{n,i}) \in \mathbf{R}^n$ represents the monitoring value of each feature in region i at time t. In addition to the influence of the feature values in the target area on the predicted air quality, the environmental conditions in the surrounding area also have different degrees of influence on the predicted results. Therefore, the paper expresses the global features as $(X^1, X^2, \ldots, X^N)$.

According to the temporal data, spatial data, and global characteristics of the target area, the STA-LSTM model is used to predict the air quality of area i at a future time T'. The result is expressed as $\widehat{y}^i = (\widehat{y}_{T+1}, \widehat{y}_{T+2}, \ldots, \widehat{y}_{T+T'})$, where $\widehat{y}_{T+t'}$ represents the predicted AQI value at time $T + t'$ in the future.

3.2. Overall Framework. To predict air quality, this paper proposes an STA-LSTM model based on a spatiotemporal attention mechanism and uses an encoder-decoder architecture. As shown in Figure 2, the model is mainly composed of three parts: (1) A spatial attention mechanism is used to capture the dynamic spatial correlation between sensors. In the encoder, we design a spatial attention mechanism to automatically capture the relative influence of different regions on the target region and assign different weights to different regions, namely, $(\alpha_t^1, \alpha_t^2, \ldots, \alpha_t^N)$, where α_t^j represents the degree of influence of area j on the target area at time t. Furthermore, the weight of each area is determined jointly by the historical information of each monitoring station, the hidden state $\mathbf{h}_{t-1}$, and the cell state $\mathbf{c}_{t-1}$ of the LSTM of the encoder. (2) Feature extraction of nonsequential information for auxiliary prediction is performed. Nonsequential data similar to those of POIs and road networks cannot be directly used as the input of the LSTM. Therefore, the solution is to preprocess the spatial data and use its output $\mathbf{ns}_{t'}$ as the input of the LSTM of the decoder, where t' is the future time. (3) A temporal attention mechanism is used to capture the dynamic temporal correlation. In the decoder, the model uses a temporal attention mechanism to automatically select the relevant hidden state of the output of the LSTM of the encoder to obtain the temporal context vector $\mathbf{y}_{t'}$, which is connected with the auxiliary vector $\mathbf{ns}_{t'}$ and the prediction result obtained at the previous time. Then, it is used as the input information for the LSTM of the decoder to predict the air quality at time t'. The weight $\beta_{t'}^t$ of the attention mechanism is calculated according to the hidden state $\mathbf{h}'_{t'-1}$ and cell state $\mathbf{c}'_{t'-1}$ in the LSTM of the decoder at time $t' - 1$.

3.2.1. Encoder with a Spatial Attention Mechanism. The purpose of this paper is to predict the air quality at time T' in the future. In previous studies, some methods [13] only considered the relevant influencing factors of the target area. Even though some methods [11] considered the influence of surrounding areas, they simply gave the same weight to different areas. In fact, different regions play different roles, and their impact on the target region also changes with time. For example, the data in the area closest to the target area have a relatively important reference value. Similarly, if strong winds are blown from a certain area, the impact on the air quality of the target area is greater than if the area has no wind. In addition, Liang [16] pointed out that there may be sequences with little correlation or relevance in other regions. If the temporal data of all regions are directly used as the input of the encoder to capture the influence of other regions, the result is a high computational cost and a reduction in performance [16].

Therefore, we propose a spatial attention mechanism to automatically capture and utilize the relative importance of different regions, thereby grasping the spatial influencing factors of each region in the overall situation and enhancing the traditional LSTM that is good at solving time-dependent problems. The specific process is as follows. Given the hidden state $\mathbf{h}_{t-1}$ and cell state $\mathbf{c}_{t-1}$ of the LSTM of the encoder at time $t - 1$, we can calculate the attention weight of the surrounding area l in terms of its influence on the target area i according to the following formula:

$$s_t^l = \mathbf{V}_s^{\mathrm{T}} \tanh\left(\mathbf{W}_s\left[\mathbf{h}_{t-1}; \mathbf{c}_{t-1}\right] + \mathbf{U}_s \mathbf{X}^l \mathbf{Z}_s + \mathbf{b}_s\right), \quad (1)$$

where $\mathbf{X}^l \in \mathbf{R}^{N\times T}$ represents all historical time-series data at time T (in the past) for region l, and $\mathbf{V}_s \in \mathbf{R}^T$, $\mathbf{W}_s \in \mathbf{R}^{T\times 2M}$,

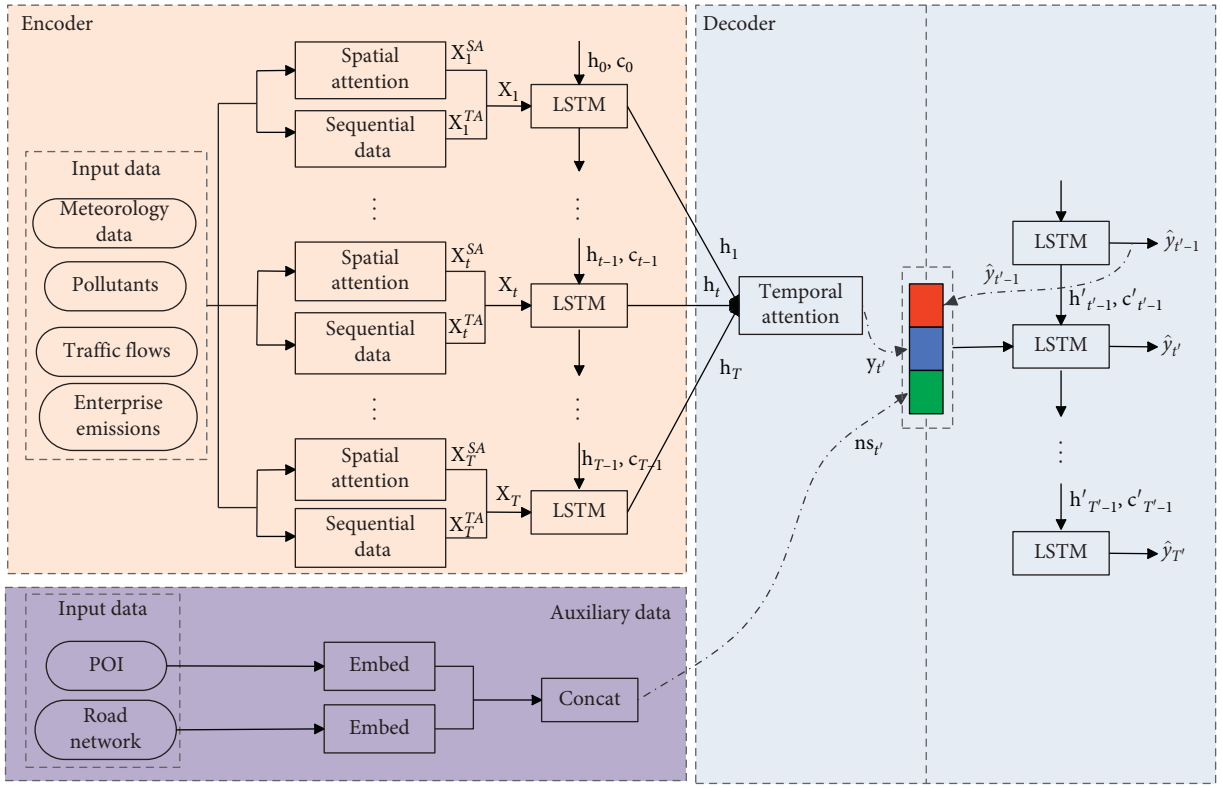

Figure 2: STA-LSTM model.

$\mathbf{U}_s \in \mathbf{R}^{T\times N}$, $\mathbf{Z}_s \in \mathbf{R}^T$, and $\mathbf{b}_s \in \mathbf{R}^T$ are the parameters of the attention model, which can be obtained through learning. The weight s_t^l obtained by each set of time-series data for area l represents the influence of the area on the target area.

In addition, the geographical distance between the two regions also affects the degree of correlation, that is, the closer the distance, the stronger the correlation. Therefore, the model uses the distance correlation matrix $\mathbf{D} \in \mathbf{R}^{N\times N}$ to represent the correlation between each region and the target region i, where $d_{i,l}$ is the reciprocal of the distance between regions i and l, and $\mathbf{D}$ is a diagonal matrix. Finally, we use the softmax function to normalize all the spatial attention weights to [0, 1] and ensure that the sum is 1. The formula for this calculation is as follows:

$$\alpha_t^l = \frac{\exp\left(\lambda s_t^l + \lambda' d_{i,l}\right)}{\sum_{j=1}^{N} \exp\left(\lambda s_t^j + \lambda' d_{i,j}\right)}. \tag{2}$$

Therefore, α_t^l comprehensively considers the importance of area l to the target area. In other words, it controls the amount of information in area l input into the LSTM of the encoder. Among the terms in the formula, $\lambda + \lambda' = 1$, and λ is an adjustable hyperparameter that determines the proportions of s_t^l and $d_{i,j}$ when calculating the weight. According to the above process, the attention weight of each area at time t can be obtained in turn, namely,

$$\mathbf{e}_t = \left(\alpha_t^1, \alpha_t^2, \ldots, \alpha_t^l, \ldots, \alpha_t^N\right)^{\mathrm{T}}. \tag{3}$$

Then, the vector output through the spatial attention mechanism at time t is as follows:

$$\mathrm{X}_t^{SA} = \left(\alpha_t^1 x_t^{1,1}, \alpha_t^2 x_t^{1,2}, \ldots, \alpha_t^l x_t^{1,l}, \ldots, \alpha_t^N x_t^{1,N}\right)^{\mathrm{T}}, \tag{4}$$

where $x_t^{1,l}$ represents the AQI value of area l at time t.

The spatial influence factor X_t^{SA} at time t is connected with the temporal data $\mathrm{X}_t^{TA} = (x_t^1, x_t^2, \ldots, x_t^n)^{\mathrm{T}}$ of the target area (where x_t^i is the i-th temporal data at time t, such as AQI, temperature, wind speed, etc.) to obtain the input of the LSTM of the encoder, namely, $\mathrm{X}_t = [\mathrm{X}_t^{SA}; \mathrm{X}_t^{TA}]$, $\mathrm{X}_t \in \mathbf{R}^{N+n}$. Then, we use $\mathbf{h}_{t-1}$, $\mathbf{c}_{t-1}$, and X_t at the previous time t to update the hidden state $\mathbf{h}_t$ [12]. The calculation process is as follows:

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f\left[\mathbf{h}_{t-1}, \mathbf{X}_t\right] + \mathbf{b}_f\right), \tag{5}$$

$$\mathbf{i}_t = \sigma\left(\mathbf{W}_i\left[\mathbf{h}_{t-1}, \mathrm{X}_t\right] + \mathbf{b}_i\right), \tag{6}$$

$$\mathbf{o}_t = \sigma\left(\mathbf{W}_o\left[\mathbf{h}_{t-1}, \mathrm{X}_t\right] + \mathbf{b}_o\right), \tag{7}$$

$$\widehat{c}_t = \tanh\left(\mathbf{W}_c\left[\mathbf{h}_{t-1}, \mathrm{X}_t\right] + \mathbf{b}_c\right), \tag{8}$$

$$\mathbf{c}_t = \mathbf{f}_t * \mathbf{c}_{t-1} + \mathbf{i}_t * \widehat{c}_t, \tag{9}$$

$$\mathbf{h}_t = \mathbf{o}_t * \tanh\left(\mathbf{c}_t\right), \tag{10}$$

where $\mathbf{f}$, $\mathbf{i}$, and $\mathbf{o}$ represent the forget gate, input gate, and output gate, respectively, $\widehat{c}_t$ is the candidate cell information,

W is the weight parameter, **b** is the bias term, and σ represents the sigmoid activation function.

3.2.2. Feature Extraction of Nonsequential Information for Auxiliary Prediction. The spatial data, similar to POIs and road networks, directly or indirectly affect air quality, so the model uses these spatial data as auxiliary information for air quality prediction. However, these data cannot be directly input into the LSTM. Therefore, this paper proposes using the LINE method to embed the information network composed of the coordinates, POIs, and road networks of the prediction area into a low-dimensional vector to improve the prediction effect for air quality. The following figure is an information network diagram composed of spatial information such as coordinates, POIs, and road networks.

As shown in Figure 3, the network graph $G_{aa} = (A \cup A, \beta_{aa})$ between prediction regions represents the distance relationship of each region, where A represents the region to be predicted, β_{aa} represents the set of edges e_{ij} between any two regions, and the weight w_{ij} represents the distance between the two areas. On the right side of Figure 4, the network graph $G_{ap} = (A \cup P, \beta_{ap})$ between the area and the POIs represents the distribution of POIs in the prediction area, where P represents the collection of POI categories, and the categories $p_1 \sim p_{10}$ are, respectively, expressed as transportation spots, factories, parks, stores, eating and drinking establishments, stadiums, schools, real estate, entertainment establishments, and other establishments [9]. β_{ap} represents the set of edges e_{ij} between the region and the POI category, and its weight w_{ij} represents the number of POIs containing category p_j in the prediction area i. The network graph $G_{ar} = (A \cup R, \beta_{ar})$ between the area and the road network in the left part of the figure represents the distribution of road segments in the prediction area, where R represents the set of road segment categories, β_{ar} represents the set of edges e_{ij} between the region and the road segment categories, and its weight w_{ij} represents the total length of the roads of category r_j included in the prediction area i.

According to the network graph defined above, this paper uses the LINE method to learn the low-dimensional vector representation of the spatial data in the prediction area. The objective functions are shown in the following formulas:

$$L(G_{aa}) = -\sum_{e_{ij} \in \beta_{aa}} w_{ij} \log p(v_j \mid v_i), \tag{11}$$

$$L(G_{ap}) = -\sum_{e_{ij} \in \beta_{ap}} w_{ij} \log p(v_j \mid v_i), \tag{12}$$

$$L(G_{ar}) = -\sum_{e_{ij} \in \beta_{ar}} w_{ij} \log p(v_j \mid v_i), \tag{13}$$

$$L(G) = L(G_{aa}) + L(G_{ap}) + L(G_{ar}). \tag{14}$$

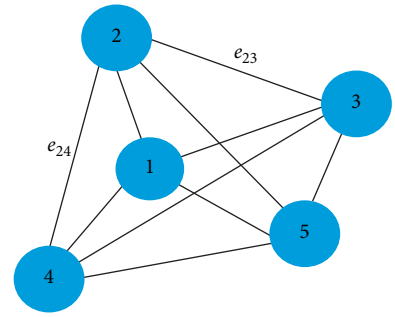

Figure 3: Network diagram with connections between regions.

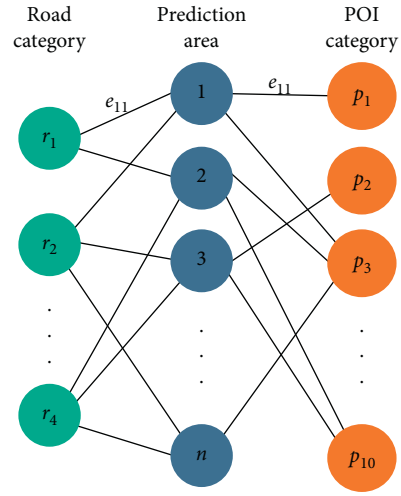

Figure 4: Network diagram between the prediction area, POIs, and road networks.

By optimizing the objective function $L(G)$, a low-dimensional vector representation of the spatial information of each region d_i can be obtained, that is, $\mathbf{ns}^i \in \mathbf{R}^\phi$, and $\mathbf{R}^\phi$ represents a ϕ-dimensional vector space.

3.2.3. Decoder for Air Quality Prediction. When the traditional encoder-decoder model performs air quality prediction, the hidden states $(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T)$ obtained by the LSTM of the encoder are directly input into the decoder to obtain a fixed-length target sequence. However, Cho [37] found that the performance of the model decreases rapidly as the input length of the decoder increases. Therefore, this paper introduces a temporal attention mechanism in the decoder to assign different temporal weights to the hidden states of the encoder output. At the same time, all hidden states are weighted and summed, and the result is used as the input of the LSTM of the decoder at the future time t' to capture the dynamic time correlation between the future and the historical times [38]. The specific process is as follows.

Given the hidden state $\mathbf{h}'_{t'-1}$ and the cell state $\mathbf{c}'_{t'-1}$ of the LSTM of the decoder at time $t' - 1$, we can use the following formula to calculate the attention weight of the hidden state $\mathbf{h}_t$ output by the encoder at time t':

$$u_{t'}^{t} = \mathbf{v}_d^{\mathrm{T}} \tanh\left(\mathbf{W}_d\left[\mathbf{h}'_{t'-1}; \mathbf{c}'_{t'-1}\right] + \mathbf{U}_d \mathbf{h}_t + \mathbf{b}_d\right), \tag{15}$$

where $\mathbf{v}_d, \mathbf{b}_d \in \mathbf{R}^L$, $\mathbf{W}_d \in \mathbf{R}^{L\times 2P}$, and $\mathbf{U}_d \in \mathbf{R}^{L\times M}$.

Similar to the weight for the spatial attention mechanism in the previous section, the weight of the hidden state at the historical time is normalized to [0, 1], as shown in the following formula:

$$\beta_{t'}^{t} = \frac{\exp\left(u_{t'}^{t}\right)}{\sum_{j=1}^{T} \exp\left(u_{t'}^{j}\right)}. \tag{16}$$

According to the above equation, the weights of all the historical hidden states output by the encoder can be calculated, and then, the hidden state $\mathbf{h}_t$ is weighted and summed to obtain the time context vector y, that is,

$$\mathbf{y}_{t'} = \sum_{t=1}^{T} \beta_{t'}^{t} \mathbf{h}_t. \tag{17}$$

We connect $\mathbf{y}_{t'}$ with the nonsequential auxiliary information $\mathbf{ns}_{t'}$, and the output result $\widehat{y}_{t'-1}$ at time $t'-1$ is used as the input for the LSTM of the decoder at time t', and it is used to update the hidden state $\mathbf{h}'_{t'}$. This process is similar to the calculation procedure for the LSTM of the encoder, and it is briefly expressed as

$$\mathbf{h}'_{t'} = \mathrm{LSTM}\left(\mathbf{h}'_{t'-1}, \left[\widehat{y}_{t'-1}; \mathbf{ns}_{t'}; \mathbf{y}_{t'-1}\right]\right). \tag{18}$$

Then, the model uses the updated hidden state $\mathbf{h}'_{t'}$ and the context vector $\mathbf{y}_{t'}$ to jointly calculate the AQI prediction result $\widehat{y}_{t'}$ of the target area. The calculation process is as follows:

$$\widehat{y}_{t'} = \mathbf{v}_y^{\mathrm{T}}\left(\mathbf{W}_y\left[\mathbf{h}'_{t'}; \mathbf{y}_{t'}\right] + \mathbf{b}_y\right) + b_y, \tag{19}$$

where $\mathbf{W}_y \in R^{Q\times(P+M)}$ and $\mathbf{v}_y, \mathbf{b}_y \in R^Q$.

Finally, during model training, we choose the Adam optimization algorithm [39] to minimize the mean squared error function between the predicted value $(\widehat{y}_{T+1}, \widehat{y}_{T+2}, \ldots, \widehat{y}_{T+T'})$ and the true value $(y_{T+1}, y_{T+2}, \ldots, y_{T+T'})$. The formula for the calculation is as follows:

$$\mathrm{Loss}(\Theta) = \|\widehat{y} - \mathbf{y}\|^2, \tag{20}$$

where Θ represents all the parameters learned by the STA-LSTM model.

4. Experiments

For the prediction model proposed in this paper, the effectiveness of the model is verified in this section by several sets of comparative experiments. First, we introduce the experimental datasets and their evaluation criteria. Second, we use other air quality prediction methods as benchmarks for comparison with the STA-LSTM model proposed in the paper. Finally, we verify the effectiveness of different input features. In addition, we evaluate the impact of the spatial and temporal attention mechanism modules on air quality prediction in turn.

4.1. Experimental Settings

4.1.1. Datasets and Settings. In this experiment, we use the monitoring data from the Beijing area with a total of 36 monitoring stations, some of which are shown in Figure 5. The time span is from January 1, 2018, to December 31, 2018, with an interval of 1 hour. Figure 5 shows the distribution of some of the monitoring stations.

(1) Historical meteorological data: meteorological data mainly include temperature, humidity, wind speed, and wind direction data. We mainly obtain them through the Chinese weather website, and the time granularity is an hour.

(2) Historical air quality data: air quality data mainly include historical AQI, PM2.5, PM10, CO, NO_2, O_3, and SO_2 data. These data are mainly obtained through the PM2.5 historical data website, with a time granularity of an hour.

(3) Historical factory pollutant emission data: the factory pollutant emissions record the concentration of air pollutants emitted by the factory, which is obtained through the company's self-monitoring information disclosure platform.

(4) Historical traffic flow data: the obtained traffic flow data contain the traffic index, that is, the traffic congestion index, which is obtained through the platform of the Beijing Transportation Development Research Institute.

(5) POI data and road network data: these data are extracted by downloading OpenStreetMap data.

In the experiment, the above datasets are randomly divided into a training set, verification set, and test set according to a ratio of 6 : 2 : 2. In the training phase, we set the batch size to 512, the learning rate to 0.001, the time window T to {6, 12, 24, 36, 48}, and the predicted future time length to 24 h. The model was trained on a server with Tesla K40m GPU and Intel Xeon E5 CPU.

4.1.2. Metrics. This experiment uses two common regression evaluation indicators to evaluate the performance of the prediction model proposed in this paper, namely, RMSE and R^2.

The RMSE is used to measure the deviation between the predicted value and the true value of a variable, namely,

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \widehat{y}_i\right)^2}, \tag{21}$$

where $\mathbf{y}_i$ is the true value, $\widehat{y}_i$ is the predicted value, and m represents the number of all predicted values. When the RMSE value is large, the error between the predicted value and the true value is also large.

R^2 usually indicates the quality of fit of the model, and its definition is as follows:

Figure 5: Distribution of some monitoring stations in Beijing.

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{m} (\overline{y}_i - y_i)^2}, \tag{22}$$

where $\overline{y}$ represents the mean value of y, and its value range is usually [0, 1], but sometimes it is also a negative number. Generally, if the result of R^2 is 0, it means that the model fitting effect is very bad; if it is 1, it means that the model predicts the result without error.

4.1.3. Compared Methods. We use the following methods as benchmarks and compare them with the model proposed in the paper:

(1) ARIMA: this is a method based on traditional linear statistical models, which can be used to predict temporal data.

(2) MFSVR: this is a predictive model with machine learning technology based on SVR. To improve the prediction accuracy, the model uses a feature fusion method based on partial least squares to extract the original features and reduce the dimensions of the input variables of the SVR model [40].

(3) DeepST: this is a prediction model for spatiotemporal data based on deep learning [30].

(4) LSTM: this method uses LSTM to automatically extract useful features from historical data, and it takes the spatial and temporal correlation of influencing factors into account [41].

(5) GC-LSTM: this method is a hybrid model based on deep learning methods. It integrates a graph convolution network and an LSTM network to predict the spatiotemporal changes in PM2.5 concentrations [32].

(6) ADAIN: this method combines feedforward and recurrent neural networks while adding an attention-based pooling layer to learn the functional weights of different monitoring stations [42].

5. Results

First, we compare the prediction model with the six benchmarks mentioned above. Then, we evaluate the effectiveness of each module of the model.

5.1. Model Comparison. To verify the feasibility and effectiveness of our model, we compare the STA-LSTM model proposed in this paper with six other AQI prediction methods, including ARIMA, MFSVR, DeepST, LSTM, GC-LSTM, and ADAIN. We use the same datasets and appropriate parameters to train these models to obtain prediction results at different scales and use the RMSE evaluation criterion to evaluate the performance of these models. The results are shown in Table 1. Obviously, as the prediction time becomes longer, the performances of all models show downward trends. The reasons for this result may be because of the following: (1) the temporal information that affects the prediction of air quality sometimes fluctuates greatly with time, and the prediction effect will be reduced under long-term prediction in the future; (2) when predicting the AQI at a certain time in the future, the prediction result from the previous time will also be introduced, resulting in the continuous accumulation of prediction errors; and (3) as time passes, the correlation between the predicted value and the input data decreases, which leads to poor prediction performance.

It can be seen from the table that the proposed STA-LSTM model has a lower RMSE value for air quality prediction than other methods. The reason may be that the STA-LSTM model considers the interaction between direct and indirect factors when modeling. In addition, the model also uses data information between neighboring stations as an influencing factor in predicting the target area. For example, the GC-LSTM and ADAIN models in Table 1 also consider the effects of spatial factors. It can be found that their results are significantly better than those of other methods, so spatial information is important for air quality prediction. Compared with the GC-LSTM model, STA-LSTM has a better prediction effect. The reason may be that

Table 1: RMSE comparison of the STA-LSTM model with other benchmarks.

Models	1 h	2 h	3–6 h	7–12 h	13–24 h
ARIMA	20.97	25.41	33.53	42.21	54.42
MFSVR	18.43	23.63	29.94	35.68	48.31
DeepST	15.89	17.28	25.41	27.67	34.46
LSTM	17.64	18.49	23.86	29.78	35.43
GC-LSTM	13.41	17.95	24.54	28.27	34.56
ADAIN	12.78	15.76	20.56	24.62	29.06
STA-LSTM	12.23	15.58	20.52	23.59	28.71

Table 2: RMSE and R^2 obtained for different combinations of input features.

models	RMSE	R^2
$F^a + F^m$	26.72	0.74
$F^a + F^c + F^t$	28.47	0.71
$F^a + F^s$	31.25	0.65
$F^a + F^m + F^c + F^t + F^s$	22.63	0.78

this paper introduces a spatial attention mechanism, which transforms the equal treatment of data information from surrounding sites into weighted data by considering the importance of the differences between regions. From the results, STA-LSTM is superior to the ADAIN model, which also introduces the attention mechanism. This may be because the model proposed in this paper introduces a temporal attention mechanism in the decoder, which can be used to learn the dynamic correlation between future and historical time data. Therefore, it is more targeted for important historical time data.

5.2. STA-LSTM Evaluation. To verify the effectiveness of the different input features of the STA-LSTM model proposed in this paper, we can limit the input of some features when conducting experiments while keeping other modules the same. As shown in Table 2, F^a, F^m, F^c, F^t, and F^s represent the characteristics of the AQI, meteorological data, factory air pollutant emissions, traffic flows, and spatial data (POIs and road networks), respectively. The following table shows the RMSE values obtained by combining different input features. It is not difficult to see that the group of experiments that combines all the features obtains the lowest RMSE value. We can also observe from the figure that, compared with the data with spatial characteristics such as POIs and road networks, the experiments with meteorological data, factory pollutant emissions, and traffic flows as input characteristics have better prediction effects. The reason for this may be that continuous temporal data such as wind speed, enterprise emissions, and vehicle exhaust are highly correlated with air prediction. However, it can also be seen from the figure below that effectively capturing the potential relationship between spatial data is very helpful for prediction. Therefore, when making air quality predictions, we need to consider more relevant factors to achieve better prediction results.

To determine the effect of the spatial attention mechanism, we compare it with that of the GC-LSTM model mentioned in the previous section, and the conclusion is that the spatial attention mechanism of STA-LSTM is more conducive to the prediction of air quality. The GC-LSTM model mainly has the following shortcomings: (1) because GC-LSTM inputs the data from the surrounding monitoring stations equally, it cannot accurately capture their spatial dependence, and (2) the performance of GC-LSTM may gradually decrease as the number of nearby monitoring stations increases. Therefore, this paper chooses to introduce a spatial attention mechanism to capture the different effects of data information from different sites on the target area, thereby improving the prediction accuracy.

Next, we verify the effectiveness of the temporal attention mechanism in the decoder. The temporal attention mechanism is used to adaptively select the relevant hidden state of the encoder, so we can use different encoding lengths to evaluate its prediction effect. We manually delete different modules and obtain three variants of STA-LSTM, STA-ns, STA-ne, and STA-nt, and compare them with the STA-LSTM model. Among the variants, STA-ns removes the spatial attention mechanism in the encoder; STA-ne deletes the spatial information of the POIs and road networks used for auxiliary prediction; and the STA-nt variant removes the temporal attention mechanism in the decoder. Figure 6(a) shows the RMSE values obtained by the STA-nt and STA-LSTM models. It is not difficult to see that the model proposed in this paper is much better than STA-nt because the temporal attention mechanism improves the long-term prediction performance for air quality. Figure 6(b) shows the prediction results of various models with different coding lengths. We can clearly observe that the error of each model is at its minimum value at $T = 12$, possibly because air quality does not exhibit any long-term time dependence.

The above experimental results show that the STA-LSTM model proposed in this paper has a better prediction effect compared to the other six benchmarks. And it also discussed the effectiveness of each module of the STA-LSTM model. Next, Figure 7 shows an optimal prediction result. When the number is less than 25, it is obvious that the fitting

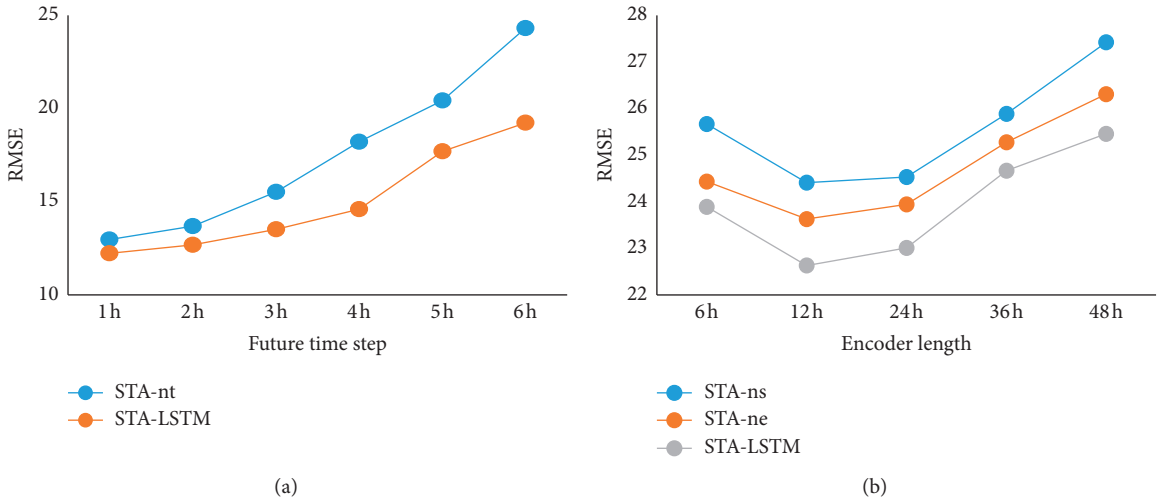

Figure 6: (a) Comparison of the RMSE values obtained by the STA-nt and STA-LSTM models. (b) Prediction results with different coding lengths.

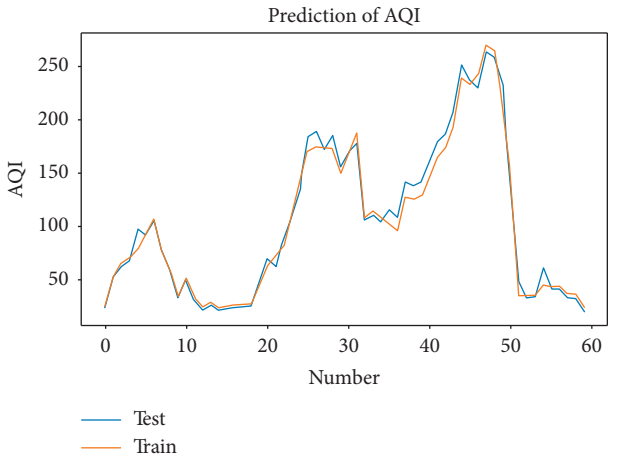

Figure 7: Prediction results of AQI.

result is very good. When it is greater than 25, there is a certain deviation. The predicted effect is consistent with the results discussed above.

6. Conclusion

In this paper, we propose an air quality prediction model based on a spatial and temporal attention mechanism, namely, the STA-LSTM model. The model adopts an encoder-decoder architecture. First, a spatial attention mechanism is introduced into the encoder to capture the relative importance of adjacent monitoring sites to the target area. Second, a temporal attention mechanism is added to the decoder to capture the dynamic correlation between future and historical times. In addition, the model uses the spatial data of the target area as auxiliary information for prediction to improve the prediction accuracy. We use real datasets to evaluate the effectiveness of the model proposed in this paper. The experiments show that our model shows the best performance when compared to 6 benchmarks. In addition, we also verify the effectiveness of modules with different features and spatiotemporal attention mechanisms. The best results are obtained by combining all the features proposed in this paper.

Data Availability

The data used to support the findings of this study are available through a public website http://zx.bjmemc.com.cn/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (2017YFC0804402).

References

[1] J. Bélissent, "Getting clever about smart cities: new opportunities require new business models," *Cambridge, Massachusetts, USA*, vol. 193, pp. 244–277, 2010.

[2] T. Zaree and A. R. Honarvar, "Improvement of air pollution prediction in a smart city and its correlation with weather conditions using metrological big data," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 26, pp. 1302–1313, 2018.

[3] C.-J. Huang and P.-H. Kuo, "A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities," *Sensors*, vol. 18, no. 7, p. 2220, 2018.

[4] Y. Yang, "Real-time profiling of fine-grained air quality index distribution using UAV sensing," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 186–198, 2017.

[5] G. Pan, G. Qi, W. Zhang, S. Li, Z. Wu, and L. Yang, "Trace analysis and mining for smart cities: issues, methods, and applications," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 120–126, 2013.

[6] L. Bai, J. Wang, X. Ma, and H. Lu, "Air pollution forecasts: an overview," *International Journal of Environmental Research and Public Health*, vol. 15, no. 4, p. 780, 2018.

[7] Yu Zheng, F. Liu, and H.-P. Hsieh, "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*, Chicago, IL, USA, 2013.

[8] J. Yuan, Yu Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing China, 2012.

[9] L. Chen, Y. Ding, D. Lyu, X. Liu, and H. Long, "Deep multitask learning based urban air quality index modelling," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 1, pp. 1–17, 2019.

[10] A. Rakowska, K. C. Wong, T. Townsend et al., "Impact of traffic volume and composition on the air quality and pedestrian exposure in urban street canyon," *Atmospheric Environment*, vol. 98, pp. 260–270, 2014.

[11] D. Qin, J. Yu, G. Zou, R. Yong, Q. Zhao, and B. Zhang, "A novel combined prediction scheme based on CNN and LSTM for urban PM2.5 concentration," *IEEE Access*, vol. 7, pp. 20050–20059, 2019.

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

[13] J. Zhao, F. Deng, Y. Cai, and J. Chen, "Long short-term memory—fully connected (LSTM-FC) neural network for PM2.5 concentration prediction," *Chemosphere*, vol. 220, pp. 486–492, 2019.

[14] J. Tang, "Line: large-scale information network embedding," in *Proceedings of the 24th International Conference on World Wide Web*, Florence Italy, 2015.

[15] A. Vaswani, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017.

[16] Y. Liang, "Geoman: multi-level attention networks for geosensory time series prediction," in *Proceedings of the IJCAI*, Stockholm, Sweden, 2018.

[17] H. Gao, "V2VR: reliable hybrid-network-oriented V2V data transmission and routing considering RSUs and connectivity probability," *IEEE Transactions on Intelligent Transportation Systems*, vol. 99, pp. 1–14, 2020.

[18] S. Deng, Z. Xiang, P. Zhao et al., "Dynamical resource allocation in edge for trustable internet-of-things systems: a reinforcement learning method," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 9, pp. 6103–6113, 2020.

[19] Y. Yin, "QoS prediction for service recommendation with features learning in mobile edge computing environment," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, 2020.

[20] H. Gao, "Mining consuming behaviors with temporal evolution for personalized recommendation in mobile marketing apps," *Mobile Networks and Applications*, vol. 25, 2020.

[21] X. Yang, S. Zhou, and M. Cao, "An approach to alleviate the sparsity problem of hybrid collaborative filtering based recommendations: the product-attribute perspective from user reviews," *Mobile Networks and Applications*, vol. 25, pp. 1–15, 2019.

[22] S. Vardoulakis, B. E. A. Fisher, K. Pericleous, and N. Gonzalez-Flesca, "Modelling air quality in street canyons: a review," *Atmospheric Environment*, vol. 37, no. 2, pp. 155–182, 2003.

[23] L. Jian, Y. Zhao, Y.-P. Zhu, M.-B. Zhang, and D. Bertolatti, "An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China," *Science of the Total Environment*, vol. 426, pp. 336–345, 2012.

[24] D. D. Genc, C. Yesilyurt, and G. Tuncel, "Air pollution forecasting in Ankara, Turkey using air pollution index and its relation to assimilative capacity of the atmosphere," *Environmental Monitoring and Assessment*, vol. 166, no. 1–4, pp. 11–27, 2010.

[25] S. Moisan, R. Herrera, and A. Clements, "A dynamic multiple equation approach for forecasting PM2.5 pollution in Santiago, Chile," *International Journal of Forecasting*, vol. 34, no. 4, pp. 566–581, 2018.

[26] M. Niu, K. Gan, S. Sun, and F. Li, "Application of decomposition-ensemble learning paradigm with phase space reconstruction for day-ahead PM 2.5 concentration forecasting," *Journal of Environmental Management*, vol. 196, pp. 110–118, 2017.

[27] J. Ma, Z. Li, J. C. P. Cheng, Y. Ding, C. Lin, and Z. Xu, "Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network," *Science of The Total Environment*, vol. 705, Article ID 135771, 2020.

[28] Z. C. Lipton, J. Berkowitz, and E.. Charles, "A critical review of recurrent neural networks for sequence learning," 2015, https://arxiv.org/abs/1506.00019.

[29] J. Zhao, T. Dong, and B. Cai, "AQI prediction based on long short-term memory model with spatial-temporal optimizations and fireworks algorithm," *Journal of Wuhan University*, vol. 65, no. 3, pp. 250–262, 2019.

[30] J. Zhang, Y. Zheng, D. Qi 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*, pp. 1–4, Atlanta, GA, USA, 2016.

[31] L. Ge, A. Zhou, H. Li et al., "Spatially fine-grained air quality prediction based on DBU-LSTM," in *Proceedings of the 16th ACM International Conference on Computing Frontiers*, pp. 202–205, Ischia, Italy, 2019.

[32] Y. Qi, Q. Li, H. Karimian, and D. Liu, "A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory," *Science of the Total Environment*, vol. 664, pp. 1–10, 2019.

[33] K. Xu, "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of the International Conference on Machine Learning*, Lille, France, 2015.

[34] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, https://arxiv.org/abs/1409.0473.

[35] S. Chaudhari, "An attentive survey of attention models," 2019, https://arxiv.org/abs/1904.02874.

[36] S. Li, G. Xie, J. Ren, L. Guo, Y. Yang, and X. Xu, "Urban PM2.5 concentration prediction via attention-based CNN-LSTM," *Applied Sciences*, vol. 10, no. 6, p. 1953, 2020.

[37] K. Cho, "On the properties of neural machine translation: encoder-decoder approaches," 2014, https://arxiv.org/abs/1409.1259.

[38] 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, 2018.

[39] D. P. Kingma and Ba. Jimmy, "Adam: a method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[40] Y. Li and Y. Tao, "Daily PM10 concentration forecasting based on multiscale fusion support vector regression," *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 6, pp. 3833–3844, 2018.

[41] X. Li, L. Peng, X. Yao et al., "Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation," *Environmental Pollution*, vol. 231, pp. 997–1004, 2017.

[42] W. Cheng, "A neural attention model for urban air quality inference: learning the weights of monitoring stations," in *Proceedings of the AAAI*, Riverside, CA, USA, 2018.