



Original Research

Surveillance-image-based outdoor air quality monitoring

Xiaochu Wang^{a, b, c}, Meizhen Wang^{a, b, c}, Xuejun Liu^{a, b, c, *}, Ying Mao^{a, b, c},
Yang Chen^{a, b, c}, Songsong Dai^{a, b, c}

^a School of Geography, Nanjing Normal University, Nanjing, 210023, China

^b Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Ministry of Education, Nanjing, 210023, China

^c Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing Normal University, Nanjing, 210023, China



ARTICLE INFO

Article history:

Received 23 February 2023

Received in revised form

13 September 2023

Accepted 14 September 2023

Keywords:

Outdoor air quality estimation

Hybrid deep learning model

Convolutional neural network

Long short-term memory

Image sequences

ABSTRACT

Air pollution threatens human health, necessitating effective and convenient air quality monitoring. Recently, there has been a growing interest in using camera images for air quality estimation. However, a major challenge has been nighttime detection due to the limited visibility of nighttime images. Here we present a hybrid deep learning model, capitalizing on the temporal continuity of air quality changes for estimating outdoor air quality from surveillance images. Our model, which integrates a convolutional neural network (CNN) and long short-term memory (LSTM), adeptly captures spatial-temporal image features, enabling air quality estimation at any time of day, including PM_{2.5} and PM₁₀ concentrations, as well as the air quality index (AQI). Compared to independent CNN networks that solely extract spatial features, our model demonstrates superior accuracy on self-constructed datasets with $R^2 = 0.94$ and RMSE = 5.11 $\mu\text{g m}^{-3}$ for PM_{2.5}, $R^2 = 0.92$ and RMSE = 7.30 $\mu\text{g m}^{-3}$ for PM₁₀, and $R^2 = 0.94$ and RMSE = 5.38 for AQI. Furthermore, our model excels in daytime air quality estimation and enhances nighttime predictions, elevating overall accuracy. Validation across diverse image datasets and comparative analyses underscore the applicability and superiority of our model, reaffirming its applicability and superiority for air quality monitoring.

© 2023 The Author(s). Published by Elsevier B.V. on behalf of Chinese Society for Environmental Sciences, Harbin Institute of Technology, Chinese Research Academy of Environmental Sciences. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Outdoor air quality monitoring is important to ensure public health and mainly relies on measurements from ground stations. Ground station-based monitoring is the most common measurement method, which has high precision and stability. However, the extensive deployment of these ground stations is impeded by the high cost of instruments, resulting in limited monitoring areas. Therefore, a low-cost and high-precision measurement method is urgently required for measuring ambient air pollution.

Air pollution can be roughly distinguished through visual observation; for example, air quality can be evaluated simply based on sky color (blueness) or the edges of distant buildings. This phenomenon is a consequence of the way light interacts with airborne particles, predominantly through atmospheric scattering,

influencing the process of visual perception [1]. The light reflected by the object's surface is attenuated and observed after being scattered by the atmospheric particles. Likewise, ambient light scatters upon encountering these particles, reaching the observer, resulting in the loss of object information and color deviation in imaging [2,3]. The degree of scattering varies with the size and concentration of the particles in the air, resulting in diverse visual effects [4]. Consequently, this observation has inspired researchers to use visual images captured by cameras to infer air quality levels around the range of the camera's view. Owing to its convenience and low cost of data acquisition, image-based air pollution detection has become an important development direction in recent years. This method can monitor air quality in areas lacking monitoring stations or during station malfunctions.

Accordingly, several studies have proposed efficient and accurate methods for estimating air quality based on visual images captured by portable cameras, smartphones, or surveillance cameras. These approaches can be roughly divided into three categories. (1) Physical model-based methods, where the atmospheric

* Corresponding author. School of Geography, Nanjing Normal University, Nanjing, 210023, China.

E-mail address: liuxuejun@njnu.edu.cn (X. Liu).

reflectance or medium transmission is calculated from images according to physical theories and models, such as the atmospheric scattering model [1] or dark channel prior [3], followed by the implementation of a simplified particulate matter model [5] or establishment of a linear relationship to determine $PM_{2.5}$ or PM_{10} concentrations [6–8]. This type of method is implemented with reliable theoretical support, but its estimation accuracy remains unsatisfactory due to the use of simplified models. (2) Traditional machine learning-based methods focus on the statistical relationship between visual image features and air quality data. According to atmospheric scattering theory, the size and distribution of atmospheric particulate matter affect the characteristics of the collected images [4]. Therefore, researchers have previously deduced air quality based on the differences in the image features under different air pollution conditions, such as image color [9,10], saturation [11,12], contrast [13,14], and edges and textures [15–17]. These studies mostly adopted linear regression [18,19], random forest [20], support vector regression (SVR) [13,16], and decision trees [10,17] to establish statistical models between image features and particulate matter indicators such as $PM_{2.5}$, PM_{10} , or the air quality index (AQI), which is a comprehensive measure of air cleanliness or pollution in a particular area and calculated by measuring the concentration of several harmful pollutants in the air, including O_3 , $PM_{2.5}$, PM_{10} , CO, SO_2 , and NO_2 . While these machine learning-centric methods are straightforward and effective, the choice of image attributes for modeling varies and is often subjective across different studies. (3) Deep learning-based methods. To avoid subjectivity in feature selection, deep learning has gained widespread acceptance for air pollution estimation from images due to its capacity to autonomously learn image features. Recently, researchers have used deep learning methods to accurately classify the air quality level (AQL) using images [21–24]. Additionally, these techniques have been employed to obtain accurate quantitative assessments of particulate matter (PM) concentrations or AQI by modifying the objective and activation functions of the deep neural network [25–28], even to estimate ultrafine particle pollution by combining the street-view images and satellite or audio data [29,30].

Image-based air quality monitoring has thus shown remarkable progress; however, certain shortcomings remain. The predominant focus of prevailing methods, whether traditional or deep learning methods, revolves around estimating air quality based on the spatial features of a single image taken at a certain moment, while only a few studies [31,32] considered the temporal dependency of the air pollution image changes. In addition, those studies simply focused on estimating daytime air quality. Some image features, such as transmission and color, can be effectively calculated and extracted from daytime images. However, they cannot be gleaned from nighttime images with low intensity due to environmental illumination, including sunlight and skylight, being minimal at night, and bright spots in the scene mainly comprise light sources such as street lamps and the windows of lit rooms [4], which estimates nighttime air quality remains a challenge. Kow et al. [32] attempted to estimate air quality at night, but daytime and nighttime air quality were predicted based on two separate models. Therefore, there is an urgent need to develop a comprehensive method to estimate air quality at any time during a day.

Indeed, air pollution is not a static process — existing temporal continuity in its revolution from one moment to the next, and the same is true for air quality images taken over time. Thus, image-based air quality estimation should be treated as a time series problem rather than a static image problem. Images of previous phases can provide useful information to help infer the air quality of the next phase. Therefore, we propose a hybrid deep learning model that takes image sequences as an input, applies a

convolutional neural network (CNN) to extract the spatial features of each image, and integrates long short-term memory (LSTM) to learn temporal information from sequential images, to improve air quality estimation at any time during a day.

The main contributions of this study cover the following:

- (1) Three time-series image datasets were constructed with $PM_{2.5}$, PM_{10} , and AQI labeled for air quality assessments, covering both daytime and nighttime images captured by surveillance cameras. The Shanghai dataset was used to test the model's effectiveness, and two other datasets were used to test applicability.
- (2) Considering the temporal correlation of air quality changes, a hybrid deep learning model was designed to improve the estimation accuracy of air quality at any time, especially at night, by incorporating CNN and LSTM to learn the spatio-temporal features of image sequences and build a regression relationship between features and air quality.

2. Materials and methods

2.1. Datasets

There are few publicly available time-series datasets for image-based air quality assessments, so three image datasets (I, II, and III) based on surveillance cameras were constructed, where each image corresponded to three air quality indicators ($PM_{2.5}$ concentration, PM_{10} concentration, and AQI). Dataset I contained a total of 8132 hourly images with the scene of Lujiazui extracted from the website of the Shanghai Municipal Bureau of Ecology and Environment (<https://sthj.sh.gov.cn/>) using web crawler technology, spanning between 00:00 on January 1 and 23:00 on December 31, 2021, including 4353 daytime (06:00–18:00) images and 3779 nighttime (19:00–05:00) images (the time division references Kow et al. [32]). Dataset II contained a total of 2691 hourly images with the scene of the Xianlin Campus of Nanjing Normal University captured by our surveillance camera hourly from 00:00 on July 26 to 23:00 on December 31, 2021, including 1413 daytime images and 1278 nighttime images. Dataset III contained a total of 6623 hourly images with the scene of a nearby community provided by the Department of Ecology and Environment of Jiangsu Province; images were taken from 00:00 on January 1 to 17:00 on November 21, 2021, including 3801 daytime images and 2822 nighttime images. Fig. 1 a, b, and c show the image scenes of the three datasets, respectively.

The corresponding air quality data for each image were collected from the historical hourly data of the air quality monitoring station closest to the photographing location published by the China National Environmental Monitoring Centre. The relative positions of each camera point to its nearest air quality monitoring stations are presented in Fig. 1 d, e, and f, respectively. The distance between the three photographing points and their monitoring stations was <4 km and within the general spatial representative radius of urban monitoring stations, which is defined by the technical regulation for selection of ambient air quality monitoring stations [33], demonstrating the rationality of taking the measurement data of these nearest monitoring stations as the labels of the images.

There was a small amount of missing data in the images and air quality data throughout the year. To avoid uncertain errors, we removed the hour records containing confidential data. Then, we classified the preprocessed data by daytime and nighttime and calculated the maximum value (Max), minimum value (Min), mean value (Mean), and standard deviation (Std) of each air quality

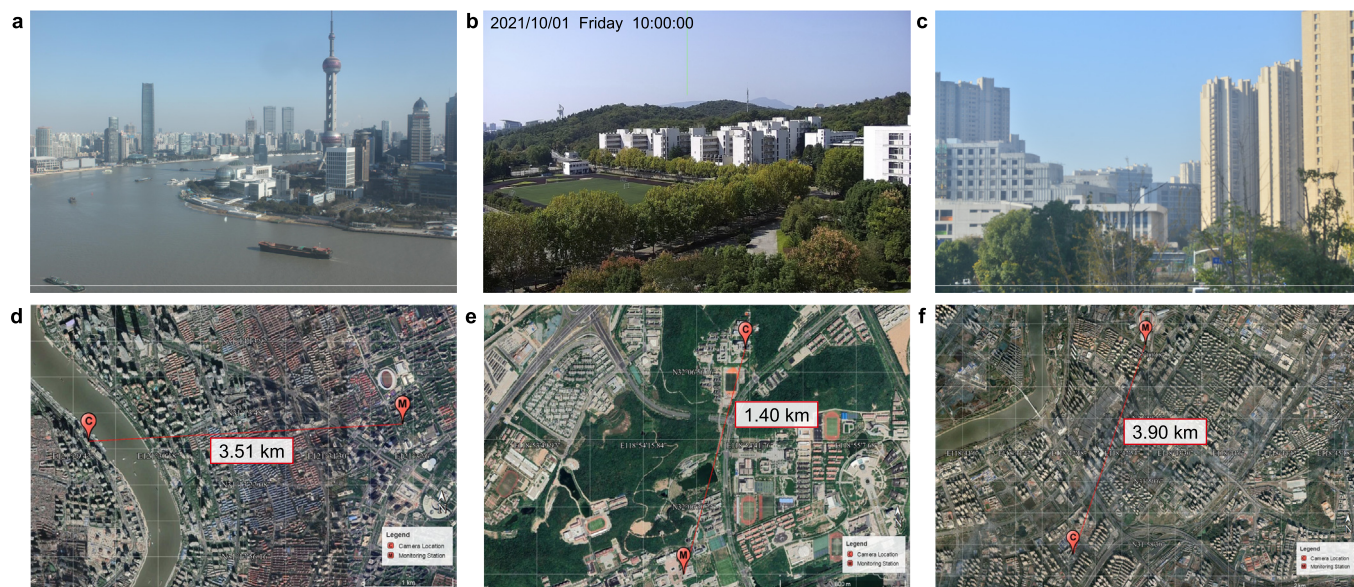


Fig. 1. Image scenes and data acquisition locations of the three image datasets. **a–c**, The image scenes of Datasets I (**a**), II (**b**), and III (**c**); **d–f**, The location maps of the image photographing points and their nearest air monitoring stations of Datasets I (**d**), II (**e**), and III (**f**).

indicator within the respective datasets. The calculation formulas of these metrics are shown in equations (1)–(4).

$$x_{\max} = \max(x_1, x_2, \dots, x_N) \quad (1)$$

$$x_{\min} = \min(x_1, x_2, \dots, x_N) \quad (2)$$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (3)$$

$$s = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (4)$$

In these equations, N is the sample size, x_i is the value of the i th sample ($i = 1, 2, \dots, N$), and x_{\max} , x_{\min} , \bar{x} , and s are the maximum value, minimum value, mean value, and standard deviation of the sample set $\{x_1, x_2, \dots, x_N\}$, respectively. Finally, the specific data statistics are presented in Table 1. The mean and standard deviation of the three datasets in Table 1 imply that the values of $PM_{2.5}$, PM_{10} , and AQI are mainly distributed in the low values. As also can be seen, Dataset I has the widest value ranges for the three air quality indicators among the three datasets. The value range of $PM_{2.5}$ in the

daytime is slightly smaller than that in the nighttime, whereas the value ranges of the other two indicators are larger in the daytime, especially that of PM_{10} . There is a small difference between the daytime and nighttime data ranges for each air quality indicator in Dataset II, while each air quality indicator in Dataset III has a larger value range in the daytime.

Then, we analyze the temporal correlations among the air quality time series according to autocorrelation functions [34]. As shown in Fig. 2, an obvious descending trend is observed with the lag time, reflecting that earlier status has a weaker influence on the current status. Additionally, the autocorrelation coefficients are higher than 0.5 when the time lag is less than 10 h, indicating a high temporal correlation. These findings can help select the appropriate sequence lengths for our estimation tasks. In this study, we employed these three datasets for different purposes. Dataset I was used to test the model performance, while Datasets II and III were employed to verify the applicability of the model's methodology.

2.2. CNN-LSTM model

We designed a hybrid deep learning model, namely CNN-LSTM, to construct robust regression relationships between images and multiple pollutants. This model facilitates the estimation of air quality at any time by integrating CNN and LSTM networks. This

Table 1
Data statistics of the three image datasets.

Indicator	Time	Dataset I				Dataset II				Dataset III			
		Max	Min	Mean	Std	Max	Min	Mean	Std	Max	Min	Mean	Std
$PM_{2.5}$ ($\mu g m^{-3}$)	Daytime	140	1	27.85	20.21	127	2	25.79	18.33	151	2	31.63	21.18
	Nighttime	159	1	28.02	20.71	136	2	27.06	20.07	133	1	33.87	22.02
	Whole day	159	1	27.93	20.45	136	2	26.40	19.19	151	1	32.58	21.57
PM_{10} ($\mu g m^{-3}$)	Daytime	834	1	44.06	37.48	180	2	53.22	28.72	551	2	66.55	45.04
	Nighttime	559	1	42.78	37.17	178	2	57.39	32.16	499	1	69.59	46.71
	Whole day	834	1	43.47	37.34	180	2	55.20	30.47	551	1	67.85	45.78
AQI	Daytime	500	11	49.36	30.49	168	11	53.50	23.34	451	10	61.09	30.64
	Nighttime	459	10	47.76	30.47	180	9	51.75	23.17	399	9	60.98	31.21
	Whole day	500	10	48.61	30.49	180	9	52.67	23.28	451	9	61.04	30.88

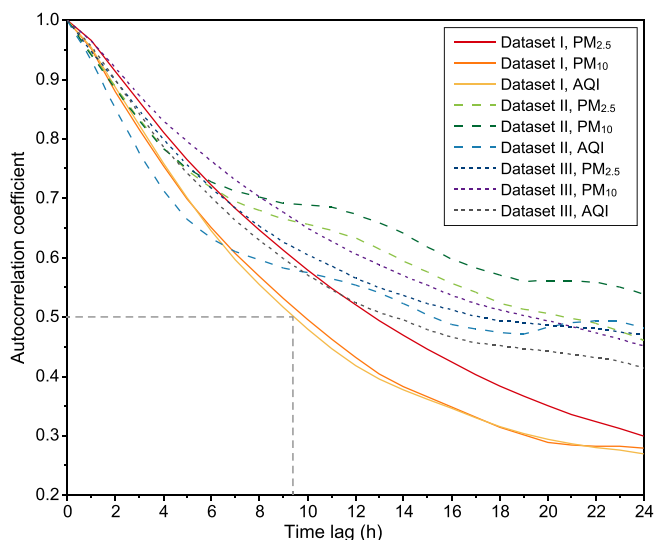


Fig. 2. The variations of the autocorrelation coefficient under different time lags.

integration enables the model to learn the spatiotemporal information of images. The proposed method is briefly described as follows.

CNN models have been identified as effective deep neural networks that can learn robust spatial features from images for object recognition and classification tasks. However, all images input into a CNN are considered independent of each other, disregarding any contextual relationships between them. Recurrent neural networks (RNN) are specifically designed to compensate for this inadequacy, and the LSTM network is the most popular RNN for learning long-term and short-term dependencies between sequential data. Owing to the strong temporal continuity of air pollution changes, CNN and LSTM can be combined to learn the differences in visual features and temporal correlation between consecutively acquired images to improve image-based air quality evaluations, especially the estimation of nighttime air quality.

The overall framework of the proposed CNN-LSTM model is presented in Fig. 3. First, the CNN architecture without fully connected layers and an output layer was used to extract the spatial features of images. We aimed to exploit LSTM to learn the contextual information between serial images; thus, in contrast to a general CNN that uses a single image with dimensions (H, W, C) as its input, an image sequence with dimensions (T, H, W, C) consisting of multiple images taken in succession served as the input for our CNN extractor, where $H, W,$ and C are the image height, width, and the number of color channels, respectively, and T is the length of an image sequence. The time gap of T determines the temporal resolution of the model. Given that the data we gathered in this investigation are recorded at hourly intervals, the temporal resolution of the predictions aligns with this hourly frequency.

Then, we seamlessly fused the CNN and LSTM networks; more specifically, the LSTM architecture was directly linked behind the processed CNN architecture and received the spatial features output from the CNN extractor as its input. These spatial features were stored and written to the memory cells of LSTM, and the temporal information was read and transferred via the interacting layers (known as “gates”) in the hidden state of each cell. Thereafter, two fully connected layers were added to synthesize and map the spatiotemporal features extracted by LSTM to the target vector. The activation function nested in the final fully connected layer was set as the sigmoid function to fit the nonlinear relationship between the spatiotemporal features and the target labels normalized

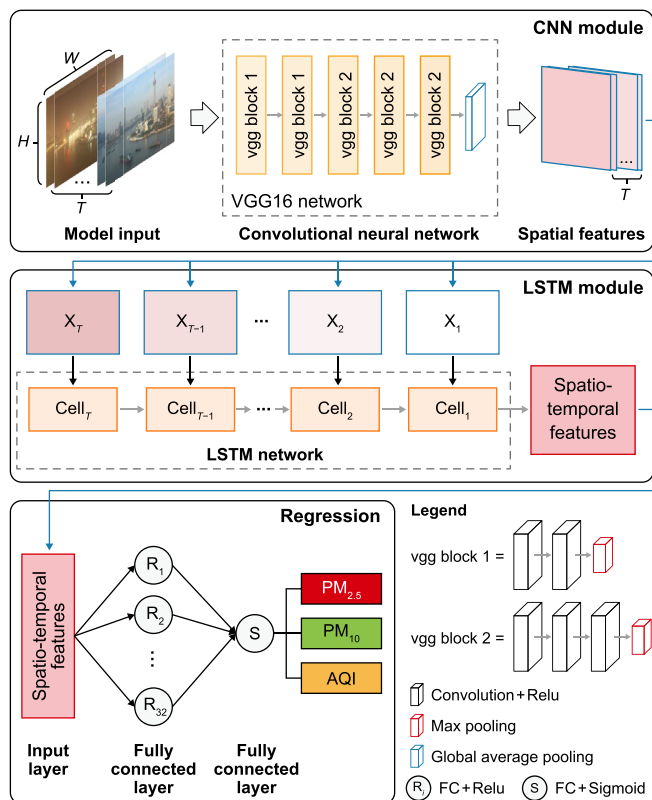


Fig. 3. The framework of the proposed CNN-LSTM model (VGG16 as CNN extractor for example).

to accelerate model convergence. The proposed CNN-LSTM network is a multivariate output regression model that yields three estimated values (PM_{2.5}, PM₁₀, and AQI) simultaneously because each target label contains three air quality indicators.

In practice, customizing a new CNN and optimally adjusting its parameters is a complex and difficult process. Additionally, a sufficiently large dataset is required to initialize the model weights. Thus, we utilized a pre-trained CNN based on the transfer learning method as the feature extractor to reduce the cost of building and training a new CNN architecture. This study tested two common pre-trained CNN schemes: VGG16 and ResNet50. These architectures had their model weights initialized using the ImageNet dataset, a large benchmark dataset for image classification and detection. Moreover, we added the same global average pooling (GAP) as the ResNet50 network at the end of the architecture when using the VGG16 network as the feature extractor, which could reduce the dimensions of the feature maps and the number of model parameters to avoid overfitting.

2.3. Model training

The models in our work are implemented using Python 3.7 and the TensorFlow deep learning framework, and the server configuration for training the model is Intel(R) Xeon(R) Silver 4216 @ 2.10 GHz CPU, NVIDIA GeForce GTX 2080Ti GPU, and Window 10 system. Some parameters need to be initialized before training the model. The loss function was set as the mean squared error (MSE) to measure the accuracy of the model training. The learning rate was set to 0.00001 through trial and error and was used to update the model weight. The batch size was set to 8, representing the number of samples utilized in one iteration and can consume

excessive computer memory if too large. Then, all images were resized to 224×224 pixels to agree with the input image size of the CNN. To accelerate the convergence of the model training, the pixel values of each image in our dataset were scaled from 0–255 to 0–1, and the $PM_{2.5}$, PM_{10} , and AQI values were normalized to 0–1 according to their respective upper limits (for example, the upper limit of AQI was 500).

Furthermore, we used two-fold cross-validation for model training and testing. More specifically, assuming a total of N image sequence samples of length T , $N/2$ samples were randomly selected as the training set, and the other $N/2$ samples were used as the testing set. Then, the two sets were switched, taking the second set as the training set and the first as the testing set. The average cross-validation result was regarded as the final prediction result of the model.

2.4. Evaluation metrics

Two common regression metrics were used to evaluate the prediction accuracy of the proposed CNN-LSTM model: the coefficient of determination (R^2) and root mean squared error (RMSE). R^2 , falling within the range of 0–1, reflects the closeness of fit between the ground truth data and the estimated values for air quality indicators; RMSE is used to calculate their errors. These metrics are defined as follows: n is the number of samples, y_i and y'_i are the i th ground truth and the corresponding estimated value ($i = 1, 2, \dots, n$), and \bar{y} is the mean of all ground truth data. A higher R^2 value and lower RMSE value indicate better model prediction performance.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - y'_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2} \quad (6)$$

3. Results and discussion

3.1. Overall results

3.1.1. Model accuracy

We applied the pre-trained VGG16 and ResNet50 as CNN extractors in the CNN-LSTM network to obtain two schemes, VGG16-LSTM and ResNet50-LSTM, and tested and compared these with the independent VGG16 and ResNet50 models on our constructed Shanghai dataset (Table 2). Independent CNNs demonstrated a moderate estimation accuracy, whereas the proposed hybrid model exhibited superior performance in comparison, with more than 0.12 higher at R^2 and $\sim 3.5 \mu\text{g m}^{-3}$ lower at RMSE for $PM_{2.5}$, more than 0.19 higher at R^2 and more than $8.3 \mu\text{g m}^{-3}$ lower at RMSE for

Table 2
Prediction accuracy (R^2 and RMSE) of different models on the constructed air quality image dataset.

Model	$PM_{2.5}$ ($\mu\text{g m}^{-3}$)		PM_{10} ($\mu\text{g m}^{-3}$)		AQI	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
VGG16	0.82	8.61	0.73	19.46	0.77	14.68
ResNet50	0.79	9.34	0.62	23.13	0.71	16.56
VGG16-LSTM	0.94	5.11	0.92	11.16	0.94	7.91
ResNet50-LSTM	0.92	5.87	0.88	13.18	0.92	8.73

PM_{10} , and more than 0.17 higher at R^2 and more than 6.7 lower at RMSE for AQI. These results demonstrate the feasibility and effectiveness of our proposed CNN-LSTM model based on surveillance images for air quality estimation, and the accession of the LSTM network significantly improved the accuracy of estimation.

All hourly estimates (7213 samples) of each air quality indicator derived from VGG16-LSTM and ResNet50-LSTM were close to those of ground-based measurements; most data samples were evenly scattered around the 1:1 line, with strong slopes (~ 0.89 – 0.93) and small intercepts (approximately 3.7–9.1), especially for the low-value range with the greatest data density (Fig. 4). However, the estimation accuracy of each model for $PM_{2.5}$ was higher than those for PM_{10} and AQI; this resulted from the larger range of PM_{10} and AQI values but the small sample size of high values, according to the data distribution in Section 2.1. It can be seen from the best-fit lines that the proposed CNN-LSTM model somewhat underestimated the air quality values on average, which was also caused by the high-value samples. The small number of samples for high-pollution cases hindered the ability to train the model, resulting in a serious underestimation of the high concentrations and increased estimation error.

Overall, the hybrid CNN-LSTM network has a strong predictive ability for $PM_{2.5}$, PM_{10} , and AQI, and the combination of CNN and LSTM efficiently improved the model's performance. VGG16-LSTM demonstrated the best results among the two CNN-LSTM schemes and served as the CNN-LSTM model in the following performance analyses.

3.1.2. Daytime and nighttime estimation

It is challenging to estimate nighttime air quality based on images owing to the low intensity of these images. Compared with the performance for daytime data, the accuracy of nighttime estimation decreased significantly, resulting in a low overall accuracy; this was confirmed by the separate assessment of daytime and nighttime estimates from VGG16 (Table 3). Table 3 also presents the estimation results of VGG16-LSTM using daytime and nighttime data for the three air quality indicators; the results demonstrate that VGG16-LSTM worked well regardless of the time of day, with R^2 values > 0.93 and RMSE values $< 5.2 \mu\text{g m}^{-3}$ for $PM_{2.5}$, R^2 values > 0.89 and RMSE values $< 12.5 \mu\text{g m}^{-3}$ for PM_{10} , as well as R^2 values > 0.92 and RMSE values < 8.5 for AQI. Furthermore, nighttime estimates for PM_{10} and AQI derived from VGG16-LSTM were more accurate than daytime estimates. This can be explained by the daytime and nighttime data distribution in Table 1, where it is shown that the cases with extremely high concentrations mainly occurred in the daytime for PM_{10} and AQI, but these high values had a smaller sample size, causing larger errors in their estimation.

We also depicted the dynamics of ground truths and estimated air quality values ($PM_{2.5}$, PM_{10} , and AQI) derived from VGG16-LSTM for 120 h over five consecutive days (March 26–30, 2021), as shown in Fig. 5. Close fits were observed between the ground truths and estimated values among the different air quality indicators, even during the transition from day to night and from night to day. Although high values were underestimated (see the right side of Fig. 5), the increases and decreases in air quality were perceived promptly and accurately.

The proposed CNN-LSTM model thus enhanced nighttime prediction accuracy and improved the overall model performance. This illustrates that temporal information is crucial in air quality estimations and should be carefully considered when introducing regression models for correlating images and air quality.

3.1.3. Comparison of different methods

We compared the performance of the proposed method with other traditional machine learning and deep learning methods

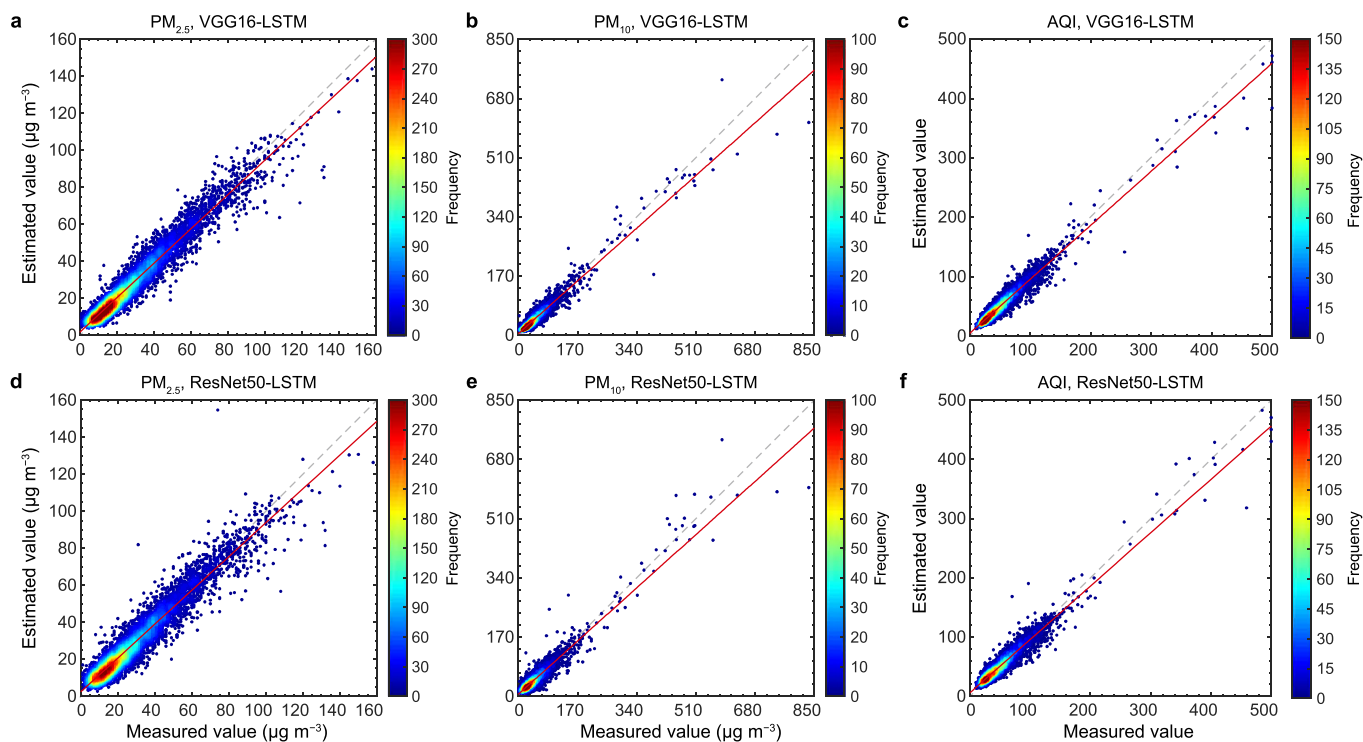


Fig. 4. Density scatterplots of the results for three air quality indicators estimated by two CNN-LSTM models. **a–c**, $PM_{2.5}$ (**a**), PM_{10} (**b**), and AQI (**c**) hourly estimates derived from VGG16-LSTM. **d–f**, $PM_{2.5}$ (**d**), PM_{10} (**e**), and AQI (**f**) hourly estimates derived from ResNet50-LSTM. Dashed lines denote 1:1 lines, and solid lines denote best-fit lines from the linear regression. The search radius for data density is two units.

Table 3

Statistical results of daytime and nighttime estimation derived from VGG16 and VGG16-LSTM.

Indicator	Time	VGG16		VGG16-LSTM	
		R^2	RMSE	R^2	RMSE
$PM_{2.5}$ ($\mu\text{g m}^{-3}$)	Daytime	0.829	8.35	0.937	5.07
	Nighttime	0.816	8.89	0.938	5.15
	Whole day	0.823	8.61	0.937	5.11
PM_{10} ($\mu\text{g m}^{-3}$)	Daytime	0.773	17.85	0.895	12.42
	Nighttime	0.675	21.16	0.940	9.46
	Whole day	0.729	19.46	0.916	11.16
AQI	Daytime	0.818	13.01	0.926	8.43
	Nighttime	0.711	16.39	0.945	7.26
	Whole day	0.768	14.68	0.935	7.91

[13,17,26,28]. Our experimentation employed an image dataset provided by Liu et al. [13], consisting of only 1954 daytime images (from 08:00 to 16:00) with the same scene as Dataset I. These images were solely annotated with $PM_{2.5}$ concentrations, ranging from 0 to $200 \mu\text{g m}^{-3}$. Due to the poor temporal continuity of the images in this dataset, we only generated 1032, 689, and 439 sequence samples with lengths 2, 3, and 4, respectively. Subsequently, we applied the VGG16-LSTM model to predict. Among these compared methods, Zhang et al. [28] and our method only depend on sequence images without using supplements of any other data. However, our combined model structure is generally more straightforward than the DCCN-ALSTM, which combines the DenseNet-121 architecture with a stacked module of three LSTM layers based on an attention mechanism. The comparative results on this daytime dataset are shown in Table 4. Our proposed method outperforms the conventional feature-based machine learning methods and even exceeds other hybrid models, such as the

combination of traditional machine learning and deep learning methods or the combination of different deep learning methods, thereby further demonstrating the superiority of the proposed method.

3.2. Performance analysis

3.2.1. Influence of image sequence length

In time-series models such as LSTM, the length of the input sequence is a key parameter. To this end, we analyzed the influence of image sequence length employed in the LSTM module of our proposed hybrid model on prediction accuracy. According to the temporal correlation of the air quality time series in Fig. 2, we set the length of the image sequence (T) to 2–9. Consequently, our dataset produced corresponding sequence samples amounting to 7957, 7791, 7633, 7486, 7346, 7213, 7082, and 6957. These sequence samples of different lengths were then separately input into the VGG16-LSTM model for training and testing, and the estimated results with different sequence lengths are shown in Fig. 6. To make a clearer comparison, the results of VGG16, which takes a single image as input, were depicted at $T = 1$ in the figure. The results show that even if the sequence length (T) was set to the minimum of 2, the prediction accuracies still exceed those of the independent CNNs whose input is a single image. This further supports the assertion that the deep learning model combining CNN and LSTM is extremely effective.

In addition, the estimated accuracy of VGG16-LSTM under different sequence lengths showed the same trend for the three air quality indicators. Specifically, as the sequence length increased, the performance initially exhibited a rapid increase, then reached a stable phase, and eventually demonstrated a slight decline. This can be attributed to the inadequate temporal information provided by excessively short sequence length, whereas the events with a long-

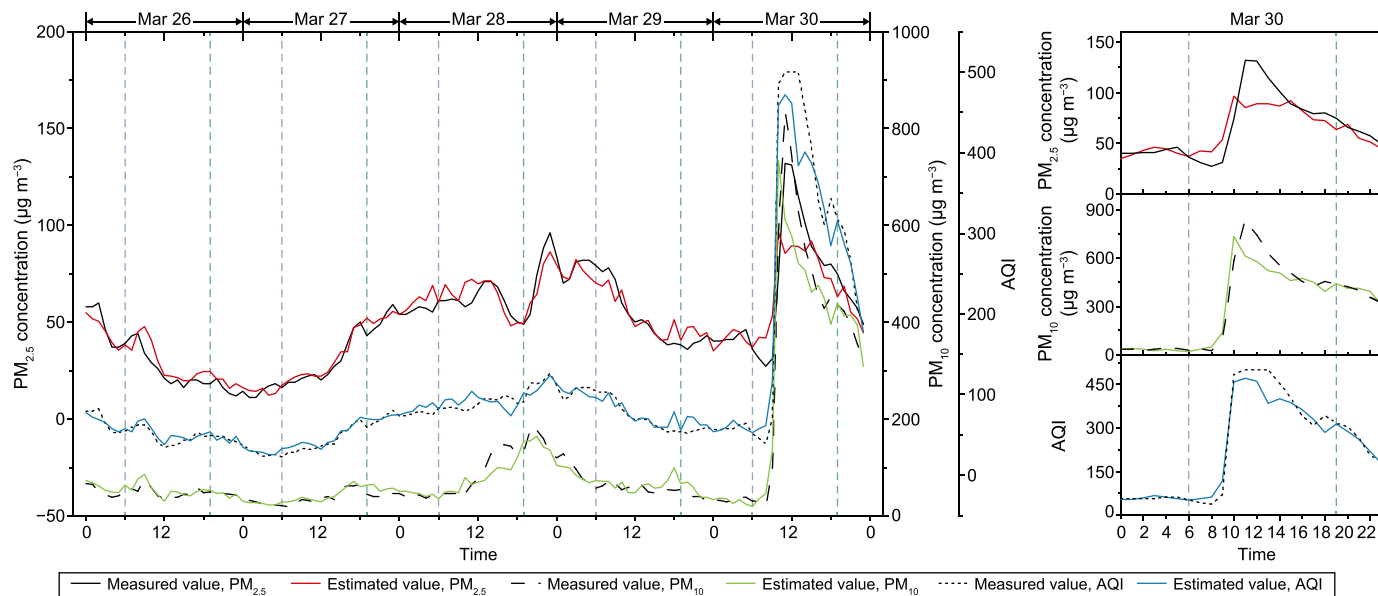


Fig. 5. Dynamics of ground truth data and estimated air quality values from VGG16-LSTM for five consecutive days. The two color-coded grid lines represent the time-dividing lines from night to day and day to night, respectively.

Table 4

Performance comparison of our method with other methods on the same PM_{2.5} daytime image dataset.

Method	Algorithm	Additional data	R ²	RMSE (µg m ⁻³)
Liu et al. [13]	SVR	Relative humidity and solar zenith angle	0.76	13.65
Luo et al. [26]	CNN-GBM	Weather conditions and photographing time	0.85	10.02
Wang et al. [17]	GBDT	Relative humidity, photographing month, and time	0.88	10.42
Zhang et al. [28]	DCCN-ALSTM (T = 4)	-	0.71	14.07
This study	VGG16	-	0.84	11.86
	VGG16-LSTM (T = 2)	-	0.92	7.37
	VGG16-LSTM (T = 3)	-	0.93	6.50
	VGG16-LSTM (T = 4)	-	0.90	7.53

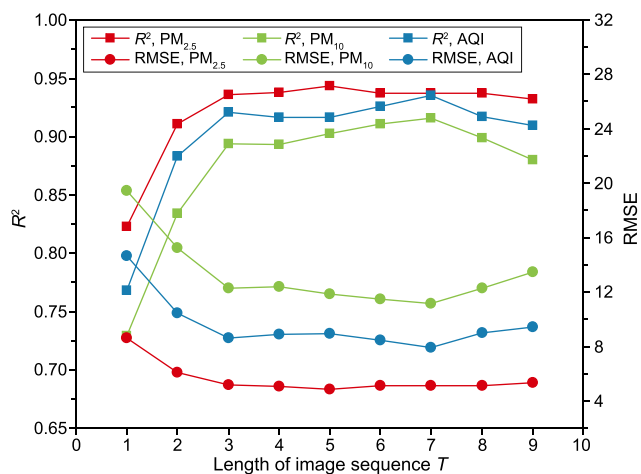


Fig. 6. Variations in the prediction performance by the image sequence length. The points at T = 1 are from the results of VGG16. The unit of RMSE of PM_{2.5} and PM₁₀ is µg m⁻³.

time lag have weak effects in a long sequence and may make the model unnecessarily complex; both extremes hinder the prediction accuracy. Accordingly, setting an appropriate sequence length for time-series models is important. By comparing the experimental

results in Fig. 6, the sequence lengths of 5–8 seem to be appropriate for our model, under which the estimated R² of all air quality indicators is higher than 0.9. Then, a sequence length of 7 was selected for this work because PM_{2.5}, PM₁₀, and AQI can obtain the highest average accuracy at T = 7. Unless otherwise specified, the experiments in this study were based on this configuration.

3.2.2. Influence of the ratio of daytime and nighttime images in a sequence

There are large differences in spatial features between daytime and nighttime images. It remains uncertain whether predictions can be affected by the presence of both daytime and nighttime images in a sequence, as well as whether the ratio of the two image types in the sequence affects the estimation accuracy. To answer these questions, we analyzed the prediction results of VGG16-LSTM with an image sequence length of T = 7 to explore the influence of the ratio of the number of daytime and nighttime images in the sequence and provided the results of VGG16 at the corresponding predicted time for comparison, as shown in Fig. 7.

The prediction accuracies of VGG16-LSTM under different ratios were clearly stable without large fluctuations and consistent with the overall accuracy shown in Section 3.1.1. This indicates that the ratio of daytime and nighttime images in the sequence had no obvious effect on the performance of VGG16-LSTM. Additionally, according to the results of VGG16 and VGG16-LSTM at multiple timings, VGG16-LSTM was more robust than VGG16; this result was

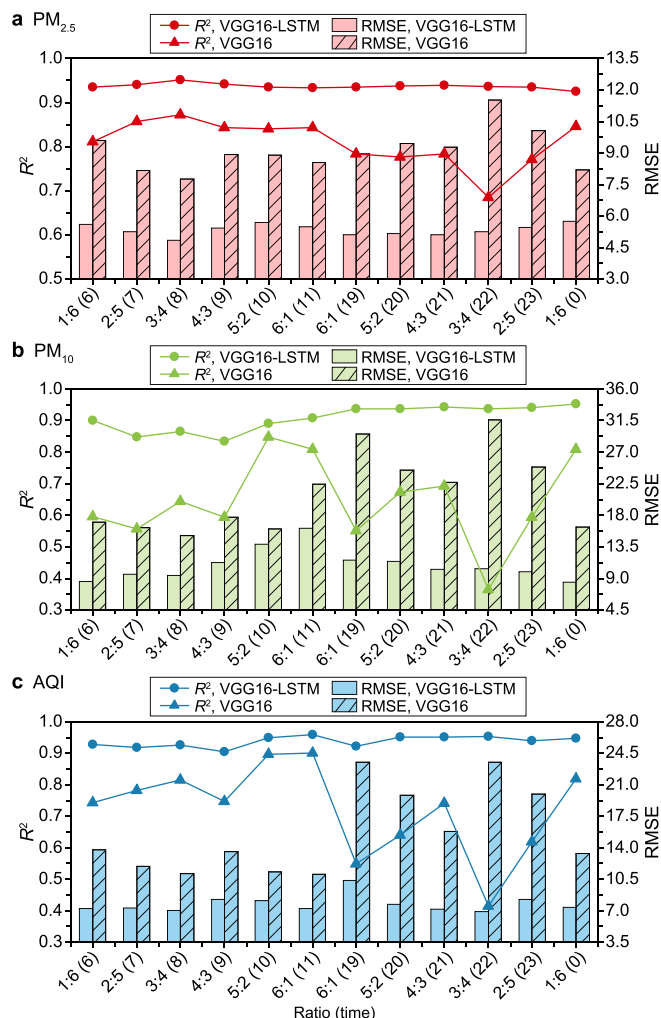


Fig. 7. Estimation accuracies of different ratios of daytime and nighttime images in a sequence: a. $PM_{2.5}$; b. PM_{10} ; c. AQI. The labels on the horizontal axis refer to the ratio of the daytime to nighttime images in the sequence and the corresponding predicted time. For example, "1:6 (6)" indicates that the ratio of daytime images to nighttime images in the sequence used for a prediction at 06:00 is 1:6.

observed as the prediction accuracy of VGG16 is easily affected by image changes caused by non-air pollution factors, as it is highly dependent on the spatial features of images. However, the CNN-LSTM model could perceive such changes and reduce their effects by learning the temporal information from the image sequence.

3.2.3. Key spatial features extracted from images

As mentioned in the Introduction, air quality variation can be evaluated by the sky color or building edge; thus, it is important to further explore the specific spatial features within the image scene that our model primarily relies on. We visualized the spatial features of images extracted by different vgg blocks in the CNN module. Table 5 shows the feature examples of the daytime and nighttime images taken at 13:00 on January 01, 2021, and 01:00 on September 02, 2021, respectively. A comparative analysis revealed that the features of the daytime images were indeed more obvious and clearer than those of the nighttime images, which is why the accuracy of the daytime estimation is higher than that of the nighttime estimation in the independent CNN results. The low-level features extracted by the initial vgg block have more obvious textures and edges, while the high-level features extracted

by the deeper vgg blocks are increasingly more abstract. However, both low-level and high-level features extracted by our model are mainly concentrated on the building regions in the image, regardless of whether the images are from daytime or nighttime. Thus, it can be concluded that building features are the key spatial features learned by the proposed model for capturing air quality variations. This observation can be further confirmed by relevant studies [17,23,26,32].

3.2.4. Applicability of the proposed method


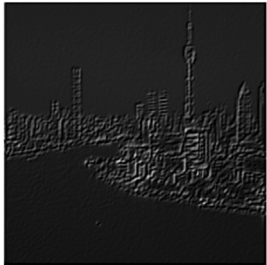
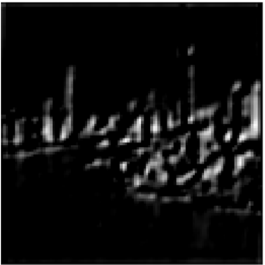
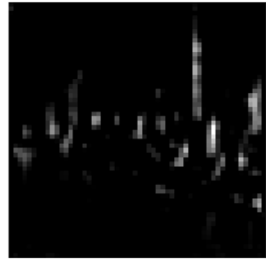
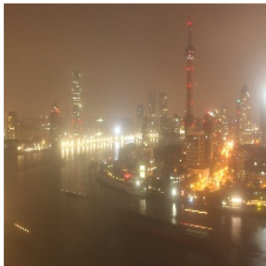

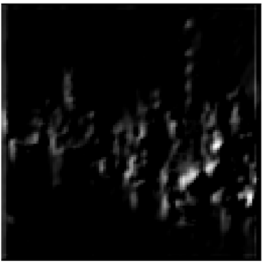
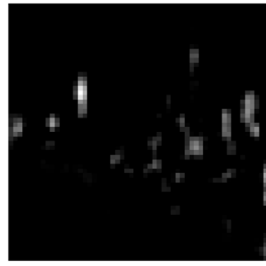
The viability of the suggested model necessitates validation across diverse datasets encompassing varied scenarios. Therefore, we incorporated Datasets II and III into our study. Initially, all images were resized to dimensions of 224×224 . This yielded 2256 image sequences from Dataset II and 4852 image sequences from Dataset III. However, air quality images and pollution trends exhibit regional disparities due to distinct regional climatic and atmospheric conditions [32]. Consequently, the model trained on Dataset I, collected at a fixed location, may not achieve the expected accuracy when directly applied to the datasets collected in other regions. Therefore, it needs to be retrained and adjusted with local data for other regions. To this end, we adopted two training strategies to test the applicability of the proposed model on Datasets II and III: one that employed the same strategies mentioned in Section 2.3 for both training and testing and another that utilized 10% of the new data to fine-tune the trained model on Dataset I and tested it with the remaining data as referenced in Luo et al. [26]. The validation results for Datasets II and III based on the two training strategies are shown in Fig. 8. Although the total sample sizes of Datasets II and III were much smaller than that of Dataset I, VGG16-LSTM nonetheless achieved adequate results for the two datasets with the same training strategy being applied, where the overall R^2 of the three air quality indicators on Dataset II was >0.84 and that on Dataset III was >0.89 . Decent results were also obtained based on the second training strategy, with R^2 values greater than 0.65 on Dataset II and greater than 0.68 on Dataset III. These results suggest that the proposed model applies to surveillance cameras with different scenes or regions. When applying our model in other regions, the model can be retrained from scratch if sufficient local data is available. Otherwise, it would be a good choice to fine-tune a trained model with a small amount of data, and then apply it for the air quality estimation.

4. Conclusion

The present study proposed a hybrid CNN-LSTM deep learning network for image-based air quality estimation, wherein LSTM was integrated with a CNN to learn the spatial and temporal features between image sequences to improve the estimation accuracy at any time of the day. Three air-quality image datasets with different surveillance scenes were compiled to evaluate the performance of the proposed method. The experimental results on these datasets show that our method enhances the estimation of nighttime air quality, improves the overall accuracy, and surpasses independent CNNs focused solely on extracting spatial image features, as well as other existing machine learning and deep learning methods. This confirms that integrating CNN and LSTM can effectively improve prediction accuracy and the temporal information is extraordinarily useful for air quality estimation.

In summary, based on the captured sequential images, the proposed CNN-LSTM network can simultaneously estimate high-precision $PM_{2.5}$, PM_{10} , and AQI data at any time, providing a promising solution for reliable and fast multi-pollutant estimations. Further investigations may involve extending this easily scalable measurement method from a single camera to multiple

Table 5
Feature examples of the daytime and nighttime images extracted by different layers of the proposed model.

	Original image (224 × 224)	The output of the first vgg block (224 × 224)	The output of the second vgg block (112 × 112)	The output of the third vgg block (56 × 56)
Daytime (January 01, 2021 13:00)				
Nighttime (September 02, 2021 01:00)				

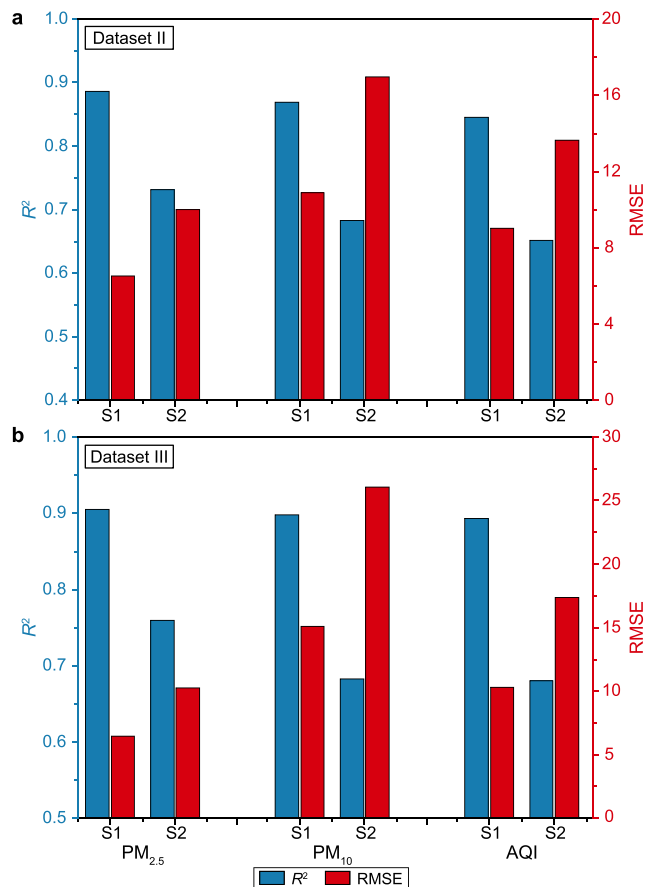


Fig. 8. Estimation results of the two other datasets based on two training strategies: **a**, Dataset II; **b**, Dataset III. S1 and S2 represent strategies 1 and 2. The unit of RMSE of PM_{2.5} and PM₁₀ is $\mu\text{g m}^{-3}$.

cameras to aid ground monitoring stations and enable regional air quality monitoring with a high spatio-temporal resolution.

CRediT author contribution statement

Xiaochu Wang: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing. **Meizhen Wang:** Validation, Formal Analysis, Supervision, Funding Acquisition. **Xuejun Liu:** Conceptualization, Investigation, Project Administration. **Ying Mao:** Validation, Visualization. **Yang Chen:** Validation, Software. **Songsong Dai:** Resources, Data Curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key Research and Development Program of China [2021YFE0112300]; the National Natural Science Foundation of China (NSFC) [41771420]; the State Scholarship Fund from the China Scholarship Council (CSC) [201906865016]; and the Postgraduate Research & Practice Innovation Program of Jiangsu Province [KYCX21_1341]. Thanks to the data providers: Images in Dataset I were published by the Shanghai Municipal Bureau of Ecology and Environment, and Images in Dataset III were provided by the Department of Ecology and Environment of Jiangsu Province.

References

[1] E.J. McCartney, *Optics of the Atmosphere: Scattering by Molecules and Particles*, John Wiley & Sons, New York, 1976, pp. 1–42.
[2] R.T. Tan, *Visibility in bad weather from a single image*, in: 26th IEEE

- Conference on Computer Vision and Pattern Recognition, 2008, pp. 2347–2354.
- [3] K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1956–1963.
- [4] S.G. Narasimhan, S.K. Nayar, Vision and the atmosphere, *Int. J. Comput. Vis.* 48 (3) (2002) 233–254.
- [5] H. Ozkaynak, A.D. Schatz, G.D. Thurston, R.G. Isaacs, R.B. Husar, Relationships between aerosol extinction coefficients derived from airport visual range observations and alternative measures of airborne particle mass, *J. Air Pollut. Control Assoc.* 35 (11) (1985) 1176–1185.
- [6] C.J. Wong, M.Z. MatJafri, K. Abdullah, H.S. Lim, K.L. Low, Using image processing technique for the studies on temporal development of air quality, computer graphics, imaging and visualisation, *N. Adv.* (2007) 287–291.
- [7] H. Wang, X. Yuan, X. Wang, Y. Zhang, Q. Dai, Real-time air quality estimation based on color image processing, in: 2014 IEEE Visual Communications and Image Processing Conference, 2014, pp. 326–329.
- [8] B. Yang, Q. Chen, PM_{2.5} Concentration estimation based on image quality assessment, in: Proceedings 2017 4th Asian Conference on Pattern Recognition (ACPR), 2017, pp. 676–681.
- [9] B. Pudasaini, M. Kanaparthi, J. Scrimgeour, N. Banerjee, S. Mondal, J. Skufca, S. Dhaniyala, Estimating PM_{2.5} from photographs, *Atmos. Environ.: X* 5 (2020).
- [10] L. Feng, T. Yang, Z. Wang, Performance evaluation of photographic measurement in the machine-learning prediction of ground PM_{2.5} concentration, *Atmos. Environ.* 262 (2021).
- [11] K. Gu, J. Qiao, X. Li, Highly efficient picture-based prediction of PM_{2.5} concentration, *IEEE Trans. Ind. Electron.* 66 (4) (2019) 3176–3184.
- [12] G. Yue, K. Gu, J. Qiao, Effective and efficient photo-based PM_{2.5} concentration estimation, *IEEE Trans. Instrum. Meas.* 68 (10) (2019) 3962–3971.
- [13] C. Liu, F. Tsow, Y. Zou, N. Tao, Particle pollution estimation based on image analysis, *PLoS One* 11 (2) (2016).
- [14] Z. Zhang, H. Ma, H. Fu, L. Liu, C. Zhang, Outdoor air quality level inference via surveillance cameras, *Mobile Inf. Syst.* 2016 (2016).
- [15] M.M. Samsami, N. Shojaee, S. Savar, M. Yazdi, Classification of the air quality level based on analysis of the sky images, in: 2019 27th Iranian Conference on Electrical Engineering, 2019, pp. 1492–1497.
- [16] J.J. Liaw, K.Y. Chen, Using high-frequency information and RH to estimate AQI based on SVR, *Sensors* 21 (11) (2021).
- [17] X. Wang, M. Wang, X. Liu, X. Zhang, R. Li, A PM_{2.5} concentration estimation method based on multi-feature combination of image patches, *Environ. Res.* 211 (2022) 113051.
- [18] J.J. Liaw, Y.F. Huang, C.H. Hsieh, D.C. Lin, C.H. Luo, PM_{2.5} concentration estimation based on image processing schemes and simple linear regression, *Sensors* 20 (8) (2020).
- [19] W.M. Yang, X. Chen, Q.M. Liao, Air quality evaluation based on local normalized image contrast, *Appl. Mech. Mater.* 511–512 (2014) 413–416.
- [20] C. Feng, Y. Tian, X. Gong, X. Que, W. Wang, MCS-RF: mobile crowdsensing-based air quality estimation with random forest, *Int. J. Distributed Sens. Netw.* 14 (10) (2018).
- [21] A. Chakma, B. Vizena, T. Cao, J. Lin, J. Zhang, Image-based air quality analysis using deep convolutional neural network, in: 2017 24th IEEE International Conference on Image Processing (ICIP), 2017, pp. 3949–3952.
- [22] J. Ma, K. Li, Y. Han, J. Yang, Image-based air pollution estimation using hybrid convolutional neural network, in: 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 471–476.
- [23] Z. Wang, W. Zheng, C. Song, Z. Zhang, J. Lian, S. Yue, S. Ji, Air quality measurement based on double-channel convolutional neural network ensemble learning, *IEEE Access* 7 (2019) 145067–145081.
- [24] C. Zhang, J. Yan, C. Li, H. Wu, R. Bie, End-to-end learning for image-based air quality level estimation, *Mach. Vis. Appl.* 29 (4) (2018) 601–615.
- [25] Q. Bo, W. Yang, N. Rijal, Y. Xie, J. Feng, J. Zhang, Particle pollution estimation from images using convolutional neural network and weather features, in: 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 3433–3437.
- [26] Z. Luo, F. Huang, H. Liu, PM_{2.5} concentration estimation using convolutional neural network and gradient boosting machine, *J. Environ. Sci.* 98 (2020) 85–93.
- [27] Q. Zhang, F. Fu, R. Tian, A deep learning and image-based model for air quality estimation, *Sci. Total Environ.* 724 (2020) 138178.
- [28] B. Zhang, Z. Geng, H. Zhang, J. Pan, Densely connected convolutional networks with attention long short-term memory for estimating PM_{2.5} values from images, *J. Clean. Prod.* 333 (2022).
- [29] K.Y. Hong, P.O. Pinheiro, S. Weichenthal, Predicting outdoor ultrafine particle number concentrations, particle size, and noise using street-level images and audio data, *Environ. Int.* 144 (2020).
- [30] M. Lloyd, E. Carter, F.G. Diaz, K.T. Magara-Gomez, K.Y. Hong, J. Baumgartner, V.M. Herrera G, S. Weichenthal, Predicting within-city spatial variations in outdoor ultrafine particle and black carbon concentrations in Bucaramanga, Colombia: a hybrid approach using open-source geographic data and digital images, *Environ. Sci. Technol.* 55 (18) (2021) 12483–12492.
- [31] S. Song, J.C.K. Lam, Y. Han, V.O.K. Li, ResNet-LSTM for real-time PM_{2.5} and PM₁₀ estimation using sequential smartphone images, *IEEE Access* 8 (2020) 220069–220082.
- [32] P.Y. Kow, I.W. Hsia, L.C. Chang, F.J. Chang, Real-time image-based air quality estimation by deep learning neural networks, *J. Environ. Manag.* 307 (2022).
- [33] M. o. E. P. o. t. P. s. R. o. China, Technical Regulation for Selection of Ambient Air Quality Monitoring Stations (On Trial), Vol. HJ 664-2013, 2013, pp. 1–14.
- [34] M. Geurts, G.E.P. Box, G.M. Jenkins, Time series analysis: forecasting and control, *J. Market. Res.* 14 (2) (1977), 269.