IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 30, 2020, accepted November 24, 2020, date of publication December 3, 2020,
date of current version December 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3042278

ResNet-LSTM for Real-Time $PM_{2.5}$ and PM_{10} Estimation Using Sequential Smartphone Images

SHIGUANG SONG[1], JACQUELINE C. K. LAM[1], (Member, IEEE), YANG HAN[1], AND VICTOR O. K. LI[1,2], (Life Fellow, IEEE)

[1]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong
[2]Clare Hall, University of Cambridge, Cambridge CB3 9AL, U.K.

Corresponding authors: Shiguang Song (sgsong@eee.hku.hk) and Jacqueline C. K. Lam (jcklam@eee.hku.hk)

This work was supported by the Theme-based Research Scheme of the Research Grants Council of Hong Kong SAR Government under Grant T41-709/17-N.

ABSTRACT Attempts have been made to estimate $PM_{2.5}$ and PM_{10} values from smartphone images, given that deploying highly accurate air pollution monitors throughout a city is a highly expensive undertaking. Departing from previous machine learning studies which primarily focus on pollutant estimation based on single day-time images, our proposed deep learning model integrates Residual Network (ResNet) with Long Short-Term Memory (LSTM), extracting spatial-temporal features of sequential images taken from smartphones instead for estimating $PM_{2.5}$ and PM_{10} values of a particular location at a particular time. Our methodology is as follows: First, we calibrated two small portable air quality sensors using the reference instruments placed in the official air quality monitoring station, located at Central, Hong Kong (HK). Second, we verified experimentally that any $PM_{2.5}$ and PM_{10} values obtained via our calibrated sensors remain constant within a radius of 500 meters. Third, 3024 outdoor day-time and night-time images of the same building were taken and labelled with corresponding $PM_{2.5}$ and PM_{10} ground truth values obtained via the calibrated sensors. Fourth, the proposed ResNet-LSTM was constructed and extended by incorporating meteorological information and one short path. Results have shown that, as compared to the best baselines, ResNet-LSTM has achieved 6.56% and 6.74% reduction in MAE and SMAPE for $PM_{2.5}$ estimation, and 13.25% and 11.03% reduction in MAE and SMAPE for PM_{10} estimation, respectively. Further, after incorporating domain-specific meteorological features and one short path, Met-ResNet-LSTM-SP has achieved the best performance, with 24.25% and 20.17% reduction in MAE and SMAPE for $PM_{2.5}$ estimation, and 28.06% and 24.57% reduction in MAE and SMAPE for PM_{10} estimation, respectively. In future, our deep-learning image-based air pollution estimation study will incorporate sequential images obtained from 24-hr operating traffic surveillance cameras distributed across all parts of the city in HK, to provide full-day and more fine-grained image-based air pollution estimation for the city.

INDEX TERMS Deep learning, $PM_{2.5}$ and PM_{10} estimation, ResNet-LSTM, Met-ResNet-LSTM, ResNet-LSTM-SP, Met-ResNet-LSTM-SP, smartphone-taken images, sequential images, domain-specific knowledge, meteorological features.

I. INTRODUCTION

$PM_{2.5}$ and PM_{10} have presented great public health challenges given their devastating health impacts, especially for those who are constantly living under high levels of air pollution, such as China and India. Citizens in these countries are consistently exposed to high levels of air pollution due to rapid industrialization and urbanization. In response to this air pollution related health challenge, governments all over the world have set up regular stationary air quality monitoring systems. Stringent air quality regulatory standards are put forward and air pollution reports are provided to inform the public the level of air pollution on a regular basis. For example, Pollutant Standards Index (PSI) has been introduced in Singapore and Daily Air Quality Index (DAQI) has been deployed in the United Kingdom (UK).

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

In Hong Kong (HK), the Air Quality Health Index (AQHI) is used by the Environmental Protection Department (EPD) to report the level of air pollution and provide hourly health advice to HK citizens. AQHI is calculated based on the average of the last 3-hr consecutive readings covering five types of air pollutants, including, ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and two types of particulate matters (PM), with health risks of all air pollutants being taken into account [1]. Air pollution, especially $PM_{2.5}$ and PM_{10}, are detrimental to the environment [2] and presents tremendous health risks to the public, especially those suffering from respiratory diseases, the elderly and young children [3]. $PM_{2.5}$ is defined as an atmospheric particulate having a diameter of less than 2.5 μm, whereas PM_{10} is one having a diameter of less than 10 μm. $PM_{2.5}$ and PM_{10} can be easily inhaled into lungs and absorbed by the alveoli directly. It can affect the respiratory and cardiovascular system and trigger asthma attacks, respiratory inflammations, and even cancers [4]–[6].

Given its harmful health consequences, it is important to make $PM_{2.5}$ and PM_{10} readings throughout the city publicly available so any citizens can plan their outdoor activities accordingly. However, due to the high cost of building and operating government-run AQMS, only a limited number of AQMSs can be provided throughout a city to provide regular air pollution readings. In HK, only 18 AQMSs across an area of 1100 sq. km are available [1]. People residing in areas without an AQMS can only rely on measurements obtained from nearest AQMSs, which may not necessarily reflect the actual pollution readings of their own locations.

Further, the lack of real-time pollutant measurements presents a real challenge to public health. Air pollutant levels in HK are updated on an hourly basis [1]. If any unexpected events such as traffic jams or fires occur, AQMS cannot provide timely alerts to HK citizens and inform them of the sudden increase in pollution level until the next hourly update.

To overcome such challenge, previous estimation studies designed and used feature-based machine learning models [7]–[11], such as Support Vector Regression (SVR), to estimate $PM_{2.5}$ values, capitalizing on features manually selected from images (such as contrast and saturation). However, as these feature-based models are highly dependent on how features are constructed, their performance can be easily distorted by any change in environmental conditions. For example, night-time images normally have a lower intensity when compared to day-time images. Since image features are manually selected for air pollution estimation, conventional machine learning methods can be less robust when compared to deep learning-based methods, which can extract image features automatically.

In recent years, some studies [12]–[15] have used deep learning models to improve image estimation accuracy and performance. Although deep learning models have been able to extract image features automatically, they have relied only on day-time images as inputs. Night-time image estimation remains a challenge, given that the low image intensity during night-time has often resulted in poor estimation accuracy. Furthermore, these studies have only considered spatial features extracted from single images via CNN models, while temporal correlations of sequential images are ignored. This is because CNN models can only learn features from single images, but not how images are changed from one temporal instance to the next, and recurrent patterns such as day-time vs. night-time images. To improve the quality of image-based PM estimation using night-time images, an end-to-end ResNet-LSTM model is proposed in this study. Our proposed model can estimate both $PM_{2.5}$ and PM_{10} values directly from pictures taken by smartphones. It achieves lower estimation errors and possess a higher pollutant estimation capability, after incorporating both day-time and night-time images, when compared to other baselines such as CNN-based models.

Main novelties of this image-based pollutant estimation study cover the following:

- A temporally fine-grained image set is constructed with corresponding $PM_{2.5}$ and PM_{10} values labelled. 3024 images have been taken consecutively in a time-sequential order at the Cyberport Waterfront Park in HK, covering both day-time and night-time images.
- An end-to-end ResNet-LSTM model using sequential images as inputs is constructed and achieves the best $PM_{2.5}$ and PM_{10} estimation, when compared to state-of-the-art baselines.
- Apart from taking single images as inputs, ResNet-LSTM also incorporates sequential images taken once every minute for estimating $PM_{2.5}$ and PM_{10} values of a specific location to enhance robustness.
- First deep learning model used to estimate $PM_{2.5}$ and PM_{10} values based on night-time images.
- A novel Met-ResNet-LSTM model is developed based on the newly developed ResNet-LSTM model, taking into account six meteorological features, in addition to images taken from smartphones as inputs, which gives even better estimation performance when compared with the ResNet-LSTM model.
- After incorporating deep supervising techniques, ResNet-LSTM-SP and Met-ResNet-LSTM-SP are constructed to improve estimation performance by comparing results of each with that of ResNet-LSTM and Met-ResNet-LSTM.

The rest of this paper is organized as follows. Related studies are reviewed in Section II. The dataset and the methodology are described in Section III. Section IV presents both our experimental results and analyses, and Section V concludes and highlights directions for future study.

II. RELATED WORK

This section reviews deep learning techniques and methods used for $PM_{2.5}$ estimation, based on images.

A. DEEP LEARNING IN GENERAL

Krizhevsky *et al.* [16] first constructed a deep convolutional neural network (CNN), AlexNet, to classify images into different categories in 2012. Its accuracy is higher when compared to manually selected feature methods. Simonyan [17] further increased the depth of CNN by adding very small convolution kernels. This further improves the image classification performance based on deep learning. The deep residual network, ResNet [18], also a CNN, was developed to tackle the issue of degradation that occurred when the number of neural network layers was increased in 2016. CNN is good at extracting key information from images. The technique has been widely implemented in the field of computer vision, such as facial recognition [19], image classification [16]–[18] and visual tracking [20]. Zheng *et al.* [21] and Hong *et al.* [22] used CNN models to analyze satellite images and estimate ground-level $PM_{2.5}$ values. Apart from CNN, Hochreiter and Schmidhuber [23] proposed an LSTM model to extract features from sequential data for neural machine translation [24], [25]. Using these techniques, efforts have been made on combining CNN with LSTM for extracting spatial-temporal features. Chen *et al.* [26] used CNN-LSTM models to forecast typhoon formation and hourly air pollution across the city [27], [28]. These models have shown that pollution estimation performance can be improved by combining temporal features with spatial features.

B. MACHINE LEARNING FOR $PM_{2.5}$ ESTIMATION

$PM_{2.5}$ can affect the light scattering coefficient [29] when a picture is taken, as it obscures the scene and blurs the sky, which eventually degrades the visibility [30]. Estimating pollution level based on smartphone-taken images is handy, as this allows one to easily capture any change in air pollution level.

Conventional machine learning approach maps ambient light scattering coefficients with $PM_{2.5}$ values. The haze-image model [31] was widely utilized to estimate scattering coefficients from single images. This model learned the formation of observed images (haze-images) from pure scenes that carry light-scattering effects. To estimate $PM_{2.5}$ values, some studies combined the dark channel model [32] with the haze-image model to calculate light coefficients directly from single images [33], and Yang and Chen [34] made good use of the relative humidity to improve pollution estimation. As compared to [33], [34], Liu *et al.* [7] and Zhang *et al.* [8] extracted image features such as image entropy, contrast, and saturation for further improvement. Capitalizing on the haze-image model and extracted image features, Liu *et al.* [7] deployed Support Vector Regression (SVR) to estimate $PM_{2.5}$ concentrations. Zhang *et al.* [8] made good use of multi-kernel learning to estimate air quality. Liu *et al.* [9] adopted similar features and used a linear least square regression to estimate $PM_{2.5}$ values via smartphone-taken images. Instead of using basic image features, Gu *et al.* [10] constructed a picture-based predictor. The entropy information from the image saturation map was extracted and non-linear mapping was used to estimate $PM_{2.5}$ values based on the overall likelihood of naturalness. Yue *et al.* [11] combined the color information loss with the structural information loss and applied a five-parameter logistic function to estimate $PM_{2.5}$ values, which achieved a high estimation performance.

After all, the performance of pollution estimation based on the conventional machine learning approaches can be easily distorted due to changes in meteorological conditions and discrepancies of light intensity between the day and the night.

C. DEEP LEARNING FOR $PM_{2.5}$ AND PM_{10} ESTIMATION

Capitalizing on strengths of the previous deep learning techniques, Li *et al.* [35] combined the dark channel and the haze-image models to estimate the scattering ability of a medium, also referred to as the transmission layer in the haze-image model. Following this, the depth of an image was further estimated by Deep Convolutional Neural Fields (DCNF) [36], and a non-linear mapping was designed to estimate $PM_{2.5}$ values based on experimental results. Since CNN can extract important spatial features from images, efforts have been made to construct CNN models to analyze images directly, without selecting image features manually. CNN models were used to analyze images and estimated $PM_{2.5}$ [12]–[14] and PM_{10} levels [12], [14]. To further improve classification accuracy, Ma *et al.* [37] and Wang *et al.* [38] developed two parallel CNN models to analyze single images. Ma *et al.* [37] used two parallel CNNs to analyze both the original image and the transmission layer, while outputs from these two CNNs were used to estimate $PM_{2.5}$. Images collected by Wang *et al.* [38] covered both skies and buildings, using the same weights, similar to the method outlined in [13]. Wang *et al.* [38] split images into two parts, the sky and buildings, and constructed a Double-Channel CNN to estimate air quality. Instead of providing a rough estimation, Liu *et al.* [15] used Long Short-Term Memory (LSTM) [23] to analyze the meteorological data, and used a CNN model to process images. Results from these two were combined to estimate $PM_{2.5}$ concentrations. However, the models did not take sequential images into account.

The pollutant estimation models above usually took $PM_{2.5}$ values from government AQMSs close to the place where images were taken, rather than the exact measurements of the place where images were taken, as ground truths [39]. Further, these models had yet considered how environmental variations will affect image estimation performance. Image-based estimation performance could be seriously degraded given that images taken and $PM_{2.5}$ measurements were not co-located, and that $PM_{2.5}$ concentrations could change when environmental conditions, such as the street canyon effect, urban morphology and traffic conditions, change overtime. To tackle the current research gaps, we used high accuracy calibrated portable pollutant sensors to provide accurate $PM_{2.5}$ and PM_{10} measurements of the exact locations where pollutant pictures have been taken. Furthermore,

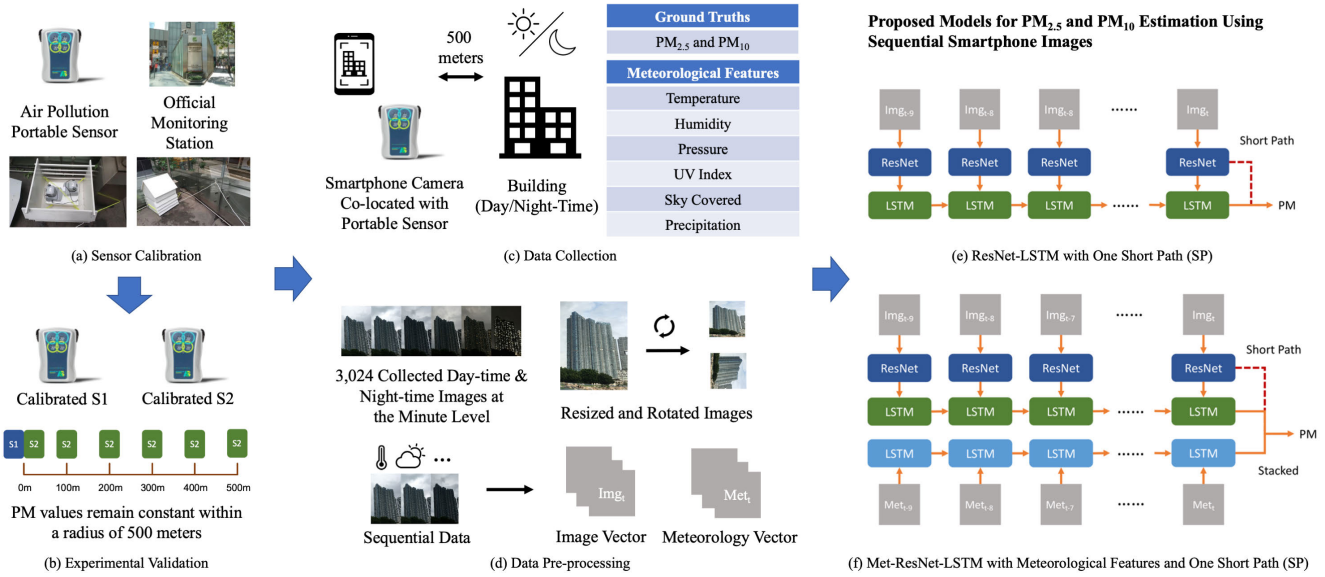


FIGURE 1. An overarching framework highlighting the novelties of ResNet-LSTM for $PM_{2.5}$ and PM_{10} estimation based on sequential smartphone images.

we incorporated six most relevant features reflecting meteorological conditions of locations that these pictures were taken into our deep learning model.

Furthermore, previous studies extracted image features, such as entropy, saturation, or high-level features layers, from the depth and the transmission map. However, these features could be easily distorted by any change in environmental conditions and light intensity, especially features that are extracted from low-intensity images, such as night-time images. In our study, we combined ResNet and LSTM to extract spatial-temporal domain features. In contrast to [15], in our proposed model, an LSTM model is added to the ResNet to extract the temporal features from images. As single night-time images are difficult to be analyzed directly, earlier efforts were only based on single day-time images as inputs. Instead of simply processing single images, our model considers the sequential day-time and night-time images in estimating $PM_{2.5}$ and PM_{10} values of a specific location.

III. METHODOLOGY

Our overarching methodology for estimating $PM_{2.5}$ and PM_{10} values via sequential smartphone taken images consists of four stages (see Figure 1). First, two portable air quality sensors were calibrated using reference instruments of the official air quality monitoring station located in Central, HK. Second, an experimental study was conducted to verify that $PM_{2.5}$ and PM_{10} values obtained via high-accuracy calibrated portable sensors remain constant within a radius of 500 meters. Third, 3024 outdoor day-time and night-time images of the same building (within a radius of 500 meters) were taken by a smartphone camera labelled with corresponding PM values using a co-located calibrated high-accuracy sensor. The collected data were pre-processed. Fourth, our proposed ResNet-LSTM was further refined by incorporating meteorological features and one short path to exploit the fullest potential of ResNet-LSTM for $PM_{2.5}$ and PM_{10} pollutant estimation.

A. CALIBRATION OF PORTABLE SENSOR

Before data collection, calibration was performed to reduce biases between portable sensors and a certified equipment. As shown in Figure 1(a), two Atmospheric Sensor 520 (AS520) sensors, using the Alphasense OPC-N2 sensor for measuring PM via the light scattering approach, were co-located with a certified device (TEOM 1405DF). The certified device is installed by EPD, HKSAR Government and located in the AQMS in Central, HK. Humidity, temperature, $PM_{2.5}$, and PM_{10} values were measured by the two portable sensors every minute, for twelve days. The $PM_{2.5}$ and PM_{10} ground truth values were recorded via a certified air pollution measuring instrument, based on an oscillating microbalance approach. This device is able to accurately measure the ambient PM concentrations, by detecting changes in oscillation frequencies when air particulates are flowing through the device and depositing onto an internal filter.

Calibration models, including Polynomial Regression (PR) and Support Vector Regression (SVR), were trained on the ground truths provided by EPD, taking AS520 sensor measurements as model inputs. During training, cross-validation was conducted to select the best parameters for each calibration model. The last 10% sequential data was used as the test set and mean absolute error (MAE) was used as the evaluation metric. Finally, the model that achieved the lowest MAE was selected for calibration.

B. $PM_{2.5}$ AND PM_{10} MEASUREMENTS OVER DIFFERENT DISTANCES

To develop a sequential image set labelled with corresponding $PM_{2.5}$ and PM_{10} values, a picture focusing on one single building located in Cyberport, HK, within a distance of 500m

FIGURE 2. Sequential images taken by a smartphone across different time spans at the same location.

was taken once every minute, continuously and sequentially, with $PM_{2.5}$ and PM_{10} values obtained from the portable sensor within the same radius of 500 meters serving as ground truths of the building. Our calibration study verified that the variation in $PM_{2.5}$ or PM_{10} measurements over a distance of up to 500 meters is negligible. To verify this, we placed two calibrated AS520 sensors (S1 and S2) across different locations of the Cyberport Waterfront Park, HK. Sensor 1, S1 was set as a stationary sensor, and Sensor 2, S2, was a mobile sensor placed at a distance of 0m, 100m, 200m, 300m, 400m, and 500m from S1, as shown in Figure 1(b).

$PM_{2.5}$ and PM_{10} values were obtained via calibrated S1 and S2 sensors consecutively once every minute across a 30-min time span, at each of the six S2 positions. Both units were placed 1.6m above the ground level, as $PM_{2.5}$ and PM_{10} values at such height resembles the level where humans are exposed to air pollution. Mean absolute difference between the two measurements of different distances were compared, with results shown in Section IV.

C. DATA COLLECTION

The calibrated AS520 sensor was used to measure $PM_{2.5}$ and PM_{10} concentrations once every minute. As shown in Figure 1(c), a smartphone co-located with the portable sensor was used to capture images of a building when $PM_{2.5}$ and PM_{10} were being measured. As validated by our empirical experiment (see Section III.B), we used PM values measured at the location of the camera (where the portable sensor was deployed and images of the building were taken), to represent the actual outdoor PM values of the building. An image of the same shot was taken once every minute from 14:30 until 19:30 daily across a ten-day period (which was taken after the twelve-day calibration period), with the aim to cover both day-time and night-time images at higher frequencies. Compared to prior studies that primarily focused on daily/hourly images and pollution values, the use of a high temporal frequency dataset can enable a more representative analysis of sequential images taken. Finally, an image set consisting of 3024 day-time and night-time images taken within a distance of 500 meters, capturing images of the same building in Cyberport, HK, was developed. Images taken after 18:00 after the sunset, at the HK time, were considered as night-time images. A set of images taken at different time spans at the same location is shown in Figure 2.

Given that temperature, humidity and weather conditions are important features affecting the performance of $PM_{2.5}$ and PM_{10} estimation, hourly meteorological data were collected from Dark Sky, a provider of meteorological forecasting and visualization. The dataset covers four features, including precipitation intensity, percentage of sky covered by cloud, UV index and pressure. Moreover, temperature and humidity measured directly by AS520 sensor were collected.

D. IMAGE DATA AUGMENTATION

To reduce overfitting and improve estimation performance, smartphone images had been augmented before being used to train the model parameters (Figure 1(d)). The details are as follows. First, these images were resized to (224, 224) to couple with the input dimensions of ResNet (see Figure 3(b)). Next, they were rotated by 0, 90, 180 or 270 degrees randomly with an equal probability (see Figure 3(b-e)). Finally, they were flipped horizontally for randomization (see the flipped results in Figure 3(f-i)). Augmentation would not only increase the size of the dataset but also improve estimation stability, as these images should give the same PM values even when being rotated or flipped during training.

E. NOVEL RESNET-LSTM, MET-RESNET-LSTM, RESNET-LSTM-SP, AND MET-RESNET-LSTM-SP

1) RESIDUAL NETWORK

Deep Convolutional Neural Network (CNN) can better extract different levels of spatial features from images, by stacking more network layers. However, previous study found it difficult to train deep CNN models, based on the observation that as network depth increases, training accuracy gets increasingly saturated and starts to degrade rapidly afterwards [18]. To overcome degradation, a deep residual network was proposed to facilitate the training (optimization) of deep CNN models [18]. Compared to other deep CNN models such as AlexNet and VGGs, the deep residual network addresses the issue of degradation by adding an identity mapping, as represented by the curved line in Figure 4. Let x represent the inputs to the first layer and $H(x)$ represent the desired mapping function from x to the

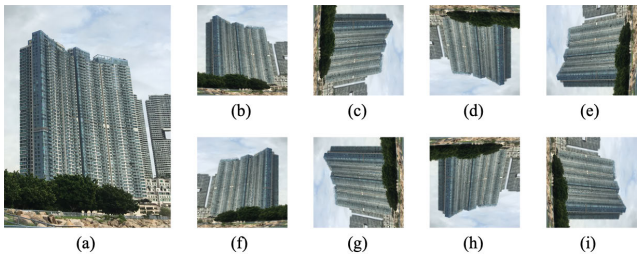


FIGURE 3. Image augmentation configuration: (a) original image (b) resized image (c-e) image rotated by 90, 180, 270 degree respectively (f-i) image flipped horizontally.

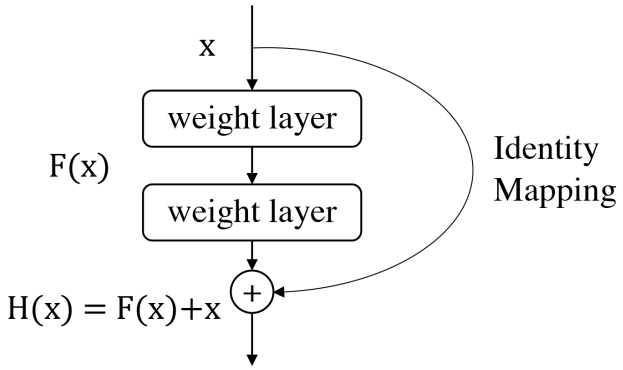


FIGURE 4. A residual block.

output of stacked layers. During model training, the problem of learning the desired mapping function $H(x)$ can be reformulated as learning the residual mapping function $F(x)$, where $F(x) = H(x) - x$, with an explicit reference to the layer inputs x. Given that the original mapping function becomes $F(x) + x$, the weights of the stacked layers that correspond to $F(x)$ can be zero, to mimic a "shortcut connection" between layers. As shown in [18], by utilizing residual networks, it is possible to optimize the CNN model even when the number of layers has been increased to 1202, though the use of such a deeper network tends to result in a lower testing accuracy due to overfitting.

In this study, ResNet18 (an 18-layer residual network) was implemented to deeply extract spatial features from our images, with the aims to address the degradation problem of deep CNN model training, while avoiding overfitting due to unnecessarily deeper network. Moreover, the last layer of ResNet18 was modified to provide a single output in the form of a single PM measurement. Figure 5 shows the structure of the modified ResNet18. The block label indicates the block-type, and the kernel size/the stride, whereas the block number indicates the output dimension, including the width, the height, and the channel of the output.

The images were first read by the system and resized to the input dimension ($3 \times 224 \times 224$), where 3 corresponds to three *RGB* channels, whilst the width and height are both 224. After average pooling (after the last convolutional layer), the images were converted to a flattened vector, with a size of 512. It was subsequently passed to a fully connected layer and the $PM_{2.5/10}$ value of the image was estimated. The size of the output of the last fully connected layer was changed to 1 to represent final estimated values, for $PM_{2.5}$ or PM_{10} estimation. Parameters of ResNet were updated through backward propagation.

2) LONG SHORT-TERM MEMORY

Long Short-Term Memory (LSTM) is a time-sequential model, which can extract the temporal domain features from any sequential data [23]. Compared to Recurrent Neural Network (RNN), LSTM addresses the long-term gradient vanishing problem and enhances the model estimation ability. In this study, the output from ResNet was fed into LSTM sequentially. The cell structure of the LSTM model is shown in Figure 6, and its formulation in (1) – (5).

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \quad (1)$$
$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (2)$$
$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \quad (3)$$
$$c_t = f_t * c_{t-1} + i_t * tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \quad (4)$$
$$h_t = o_t * tanh(c_t) \quad (5)$$

With reference to (1) – (5), W represents the weight parameters, b the bias, and σ the sigmoid function. The subscripts of W and b indicate, respectively, the weight and the bias for three different gates. For example, W_{fx} is the weight of input x_t at the gate f_t. c is the cell state, h the hidden state, x the input, and *tanh* the hyperbolic tangent function. The first gate is the forget gate f_t for controlling the information from the previous cell state, as shown in (1). It decides how much information should be kept or passed to the next stage. The second gate is the input gate i_t, which decides how much new information should be used. Meanwhile, the current cell state can be updated by combining the output from the forget gate with the input gate, as shown in (4). Finally, the information in the latest cell state and that from the input are used to update the latest hidden state, while serving also as the output of LSTM (see (5)). The parameters of LSTM were updated through backward propagation.

3) RESNET-LSTM, MET-RESNET-LSTM, RESNET-LSTM-SP, AND MET-RESNET-LSTM-SP

As shown in Figure 1(d), the collected data were pre-processed before model development. First, the sequential images captured by the smartphone camera were turned into an image vector. The vectorized data were then fed into the proposed models for $PM_{2.5}$ and PM_{10} estimation. Meanwhile, the meteorological information was packed with the image vector to estimate $PM_{2.5}$ and PM_{10} values.

Our proposed ResNet-LSTM model is shown in Figure 1(e) (without the dash line), based on the following methodology. First, ten consecutive images, taken once every single minute and the previous nine minutes, were fed into the ResNet model. Then, ten corresponding outputs of the same ResNet model were fed into the LSTM model subsequently. In order to match the output of the ResNet, the input size

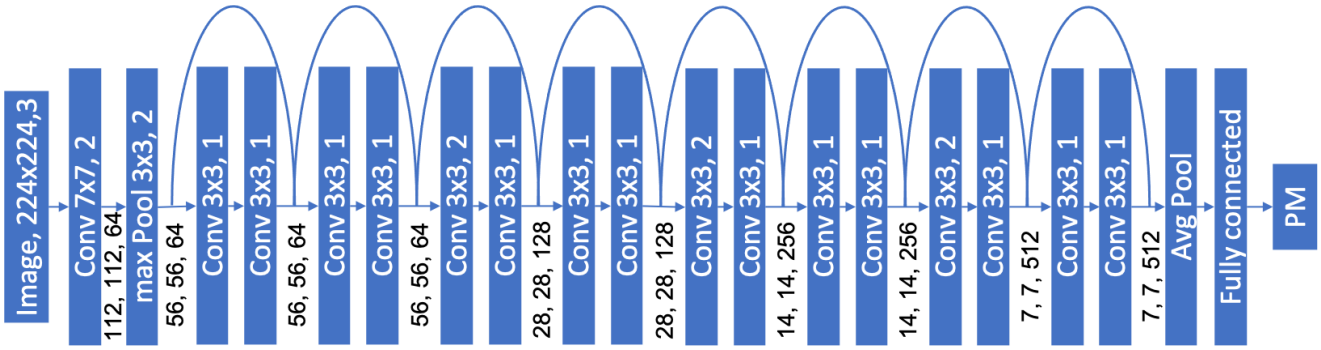


FIGURE 5. Structure of the modified ResNet18.

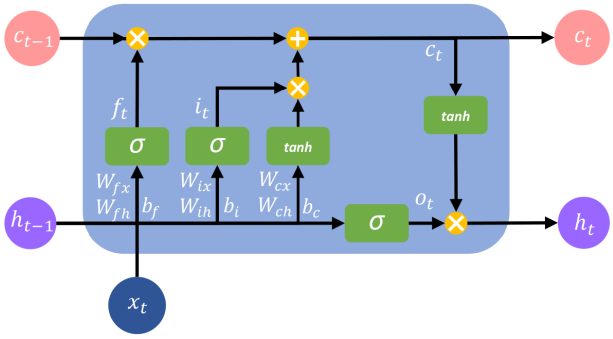


FIGURE 6. Structure of a typical LSTM cell.

and the hidden unit size of the LSTM model were set as 512. Finally, the final hidden state of the LSTM model was used to predict $PM_{2.5}$ or PM_{10} values, via a fully connected layer.

Moreover, in order to better account for the meteorological information, we developed a new Met-ResNet-LSTM model as shown in Figure 1(f) (without the dash line). It consists of the proposed ResNet-LSTM model and a parallel LSTM model. For the parallel LSTM model, the six meteorological features observed at every single minute real-time and the previous nine minutes were used as inputs. The two final hidden states from the ResNet-LSTM model and the parallel LSTM model were concatenated and used to estimate $PM_{2.5}$ or PM_{10} values, via a fully connected layer.

Furthermore, inspired by the deeply-supervised net [40], we developed a new ResNet-LSTM-SP model. As shown in Figure 1(e), compared to the ResNet-LSTM model, one short path (dash line) was constructed between the output of LSTM and the latest output from ResNet. This is based on the intuition that the latest output generated from ResNet should be more representative than those generated from LSTM. More specifically, the features extracted from the latest image (image captured once every single minute real-time) through ResNet, were combined with the final hidden state of LSTM. The concatenated vector, consisting of both the latest features and the historical information, was finally passed to a fully connected layer for estimating $PM_{2.5}$ or PM_{10}. Similarly, as shown in Figure 1(f) (with the dash line), we developed a new Met-ResNet-LSTM-SP model by applying one short path to the proposed Met-ResNet-LSTM model.

The size of the input and output for each part of our model is shown in Table 1. ResNet represents ResNet18 used in our proposed model; LSTM-Img represents the LSTM for the sequential images (as represented by the green "LSTM" boxes and the grey "Img" boxes shown in Figure 1(e) and Figure 1(f)); LSTM-Met represents the LSTM for the meteorological conditions (as represented by the light blue "LSTM" boxes and the grey "Met" boxes shown in Figure 1(f)); FC-RL, FC-MRL, FC-RL-SP, and FC-MRL-SP represent the fully connected layer for the proposed ResNet-LSTM, Met-ResNet-LSTM, ResNet-LSTM-SP, and Met-ResNet-LSTM-SP, respectively.

TABLE 1. Model input and output dimension.

	Input Dimension	Output Dimension
ResNet	3x224x224	512
LSTM-Img	512	512
LSTM-Met	6	512
FC-RL	512	1
FC-MRL	1024	1
FC-RL-SP	512	1
FC-MRL-SP	1024	1

F. BASELINE SELECTION

Our proposed ResNet-LSTM model was compared with the conventional approach and the machine learning and deep learning approach for image-based PM estimation.

1) CONVENTIONAL METHOD

Conventional pollution estimation method adopts a haze-image model, which is integrated with the dark channel prior techniques. The haze-image model developed by [31] shows how the haze image can be overlaid by pure sceneries and scattering effects. This model has been widely applied for

image dehazing and light scattering coefficient estimation, as shown in (6) below:

$$\boldsymbol{I}(x) = \boldsymbol{O}t(x) + \boldsymbol{A}(1 - t(x)) \tag{6}$$

where $\boldsymbol{I}(x)$ represents the observed scenery captured by the camera, $\boldsymbol{O}$ represents the pure scenery without attenuation, $\boldsymbol{A}$ represents the atmospheric light, and $t(x)$ is the scattering ability of the medium over the light pathlength. The transmission $t(x)$ can be expressed as $e^{-\beta x}$, where β is the scattering coefficient of the atmosphere and x the distance between the scenery and the observer (camera).

Based on this conventional method, a conventional baseline is developed as follows. First, the ambient scattering coefficient $t(x)$ is estimated from a single image, based on a well-known method, the dark channel prior. [32] used multiple images and found that the minimum intensity among all R, G, B channels of a single low-haze image is close to $\mathbf{0}$, where the dark channel of the image $\boldsymbol{O}^{dark}$ is defined in (7).

$$\boldsymbol{O}^{dark}(x) = \min_{y\in\Omega}(\min_{c\in R,G,B} \boldsymbol{O}^{c}(y)) \tag{7}$$

(3) makes the assumption that $\boldsymbol{O}^{dark} \rightarrow \mathbf{0}$ if there is no haze and the captured image shows pure scenery. c is used to represent the R, G, B channels of an image, and Ω is a selected portion of the image. Combining the dark channel approach with the haze-image model, $t(x)$ can be estimated in (8).

$$t(x) = 1 - \min_{y\in\Omega}(\min_{c\in R,G,B} \frac{\boldsymbol{I}^{c}(y)}{\boldsymbol{A}}) \tag{8}$$

Next, the ambient scattering coefficient β is calculated for each single pixel, and the average β is calculated for the entire image. Finally, based on the estimated ambient scattering coefficients, linear regression models are used to estimate the $PM_{2.5}$ and PM_{10} values.

2) MACHINE LEARNING AND DEEP LEARNING APPROACHES

Machine learning and deep learning models that estimate the $PM_{2.5}$ and PM_{10} values of single images were selected for model comparison. The feature-based machine learning models, including FEBM [7] and SILM [11], and the CNN-based deep learning models, including VGG19 [17] and ResNet18 [18], were selected as baselines.

IV. RESULTS

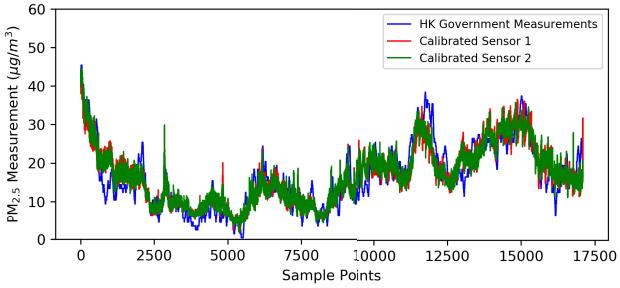
A. SENSOR CALIBRATION

We trained a calibration model to calibrate measurements from highly sophisticated portable sensors. Key calibration models were applied and compared, including SVR with different kernels and PR of different degrees. MAE of different models based on individual sensors are shown in Table 2.

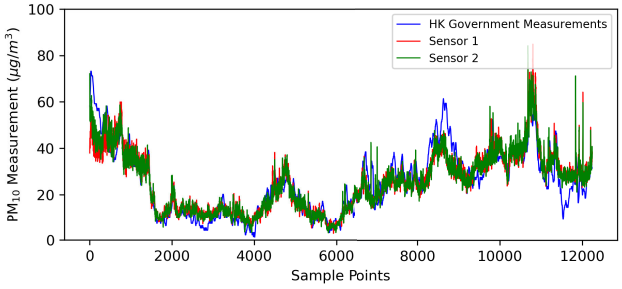
Based on results, a PR with degree 2 model gives the best results and hence is used as our calibration model. Although the degree 3 model achieves a smaller MAE in the training set, the test set gives a larger MAE when compared to the degree 2 model.

TABLE 2. MAE ($\mu g/m^3$) of PR and SVR.

PR	Degree 2				Degree 3			
	$PM_{2.5}$		PM_{10}		$PM_{2.5}$		PM_{10}	
	train	test	train	test	train	test	train	test
Sensor 1	3.03	**4.23**	3.43	**7.19**	2.75	6.80	2.70	32.29
Sensor 2	2.94	**4.29**	3.51	**6.05**	2.69	5.41	2.76	30.63
SVR	Poly Kernel Degree 2				Radial Basis Function			
	$PM_{2.5}$		PM_{10}		$PM_{2.5}$		PM_{10}	
	train	test	train	test	train	test	train	test
Sensor 1	3.29	4.85	3.77	7.33	3.32	5.50	4.03	7.66
Sensor 2	3.29	4.48	3.86	7.04	3.39	5.14	4.03	6.89



(a) $PM_{2.5}$ measurements after calibration



(b) PM_{10} measurements after calibration

FIGURE 7. Sensor calibration.

The raw $PM_{2.5}$ and PM_{10} measurements were obtained from our sensor (see Figure 7), which was calibrated on a government reference instrument. After training the full dataset, MAE of $PM_{2.5}$ values was reduced to 3.0 $\mu g/m^3$ for Sensor 1, and 3.09 $\mu g/m^3$ for Sensor 2, and MAE of PM_{10} values was reduced to 5.60 $\mu g/m^3$ for Sensor 1, and 5.43 $\mu g/m^3$ for Sensor 2. The calibrated measurements for $PM_{2.5}$ and PM_{10} are shown in Figure 7(a) and Figure 7(b), respectively.

B. $PM_{2.5}$ AND PM_{10} MEASUREMENTS OVER DIFFERENT DISTANCES

Two sensors were placed at different distances apart from each other and the readings from the two sensors were compared. The distance between the two sensors was increased from 0 to 500 meters with a 100-meter increment. The measurements obtained from the two sensors were calibrated based on the calibration model described in Section IV.A. Mean absolute difference (MAD) in $PM_{2.5}$ and PM_{10} values obtained across 500 meters are shown in Table 3.

TABLE 3. Mean absolute differences ($\mu g/m^3$) in $PM_{2.5}$ and PM_{10} values obtained from two calibrated sensors (S1 and S2), with S2 placed at 0m, 100m, 200m, 300m, 400m, and 500m away from S1.

MAD	0 m	100 m	200 m	300 m	400 m	500 m
$PM_{2.5}$	1.46	1.93	1.64	1.99	1.89	1.98
PM_{10}	1.79	2.21	1.63	1.66	1.25	1.55

Based on our experimental results, the variation in $PM_{2.5}$ and PM_{10} measurements between two sensor units (S1 and S2), across a distance of 500 meters and under an open setting, is small. The result supports our hypothesis that both $PM_{2.5}$ and PM_{10} values remain constant within a short distance across an open area. Hence, images taken with a camera co-located with a calibrated sensor give $PM_{2.5}$ and PM_{10} values of better quality, as compared to the data that carry only city-wide $PM_{2.5}$ measurements and labelled images, in which both distance and environmental conditions are excluded.

C. MODEL PERFORMANCE

The experimental settings for model evaluation are as follows. First, images were taken consecutively, once every minute, across a period of five hours per day, and over a ten-day period. All images were labeled with measurements of the co-located calibrated sensor. Next, the dataset was divided into training set (70%), validation set (10%) and test set (20%). The test set, consisting of 613 day-time and night-time images, was taken during the last two days of a ten-day experiment. As compared to previous studies, for the first time, a test set consisting of both day-time images and night-time images was constructed to provide a more comprehensive evaluation of estimation performance between the proposed ResNet-LSTM model and the baselines for estimating $PM_{2.5}$ and PM_{10} values.

The model performance was evaluated using MAE, symmetric mean absolute percentage error (SMAPE), and Pearson correlation (r). MAE is calculated by (9), SMAPE by (10), and r by (11). In addition to absolute and percentage errors, Pearson correlation is selected because it can measure how strong an association between ground truths and predicted $PM_{2.5}$ or PM_{10} values is, which can give us a better evaluation on whether or not the image (and meteorology) features are helpful in predicting PM concentrations.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t^{pr} - y_t^{re}| \tag{9}$$

$$SMAPE = \frac{2}{n}\sum_{t=1}^{n}\frac{|y_t^{pr} - y_t^{re}|}{y_t^{pr} + y_t^{re}} \tag{10}$$

$$r = \frac{\sum_{t=1}^{n}(y_t^{pr} - \overline{y_t^{pr}})(y_t^{re} - \overline{y_t^{re}})}{\sqrt{\sum_{t=1}^{n}(y_t^{pr} - \overline{y_t^{pr}})^2}\sqrt{\sum_{t=1}^{n}(y_t^{re} - \overline{y_t^{re}})^2}} \tag{11}$$

where y^{pr} refers to the predicted value, y^{re} refers to the ground truth value, $\overline{y^{pr}}$ refers to the mean predicted value, $\overline{y^{re}}$ refers to the mean ground truth value, and n refers to the sample size.

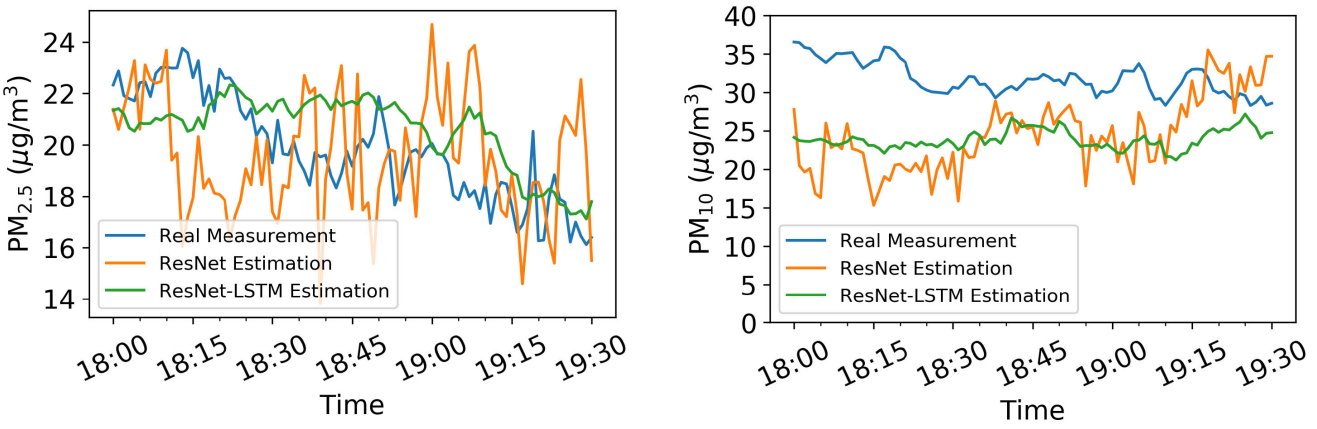
Experimental results have shown that our proposed ResNet-LSTM models outperformed the baselines. As shown in Table 4, deep learning models have outperformed the conventional and feature-based machine learning models, especially when Pearson correlation is used for evaluation. This suggests that feature representations automatically learned by deep learning models are more helpful in predicting $PM_{2.5}$ and PM_{10} concentrations, as compared to manually selected features adopted by conventional machine learning models. Moreover, as compared to ResNet, the best deep learning baseline, our proposed ResNet-LSTM model has achieved a lower MAE and SMAPE and a higher Pearson correlation. This verifies that ResNet-LSTM can better capture the temporal correlation of sequential images, thus outperforming ResNet based on single images only. As compared to the best baselines, for $PM_{2.5}$ estimation, our ResNet-LSTM has achieved a reduction in MAE and SMAPE respectively by 6.56% and 6.74%, whereas for PM_{10} estimation, our ResNet-LSTM has achieved a reduction in MAE and SMAPE respectively by 13.25% and 11.03%.

TABLE 4. MAE, SMAPE, and Pearson correlation of $PM_{2.5}$ and PM_{10} estimation performed by ResNet-LSTM and baselines, trained on both day-time and night-time images.

Estimation Model	Test Set for $PM_{2.5}$ (Day + Night)		
	MAE	SMAPE	r*
Conventional	8.28	0.3586	0.16 (<0.05)
FEBM	8.30	0.3665	0.35 (<0.05)
SILM	8.29	0.3170	0.30 (<0.05)
VGG	7.61	0.3324	0.43 (<0.05)
ResNet	7.01	0.3099	0.55 (<0.05)
ResNet-LSTM	**6.55**	**0.2890**	**0.56 (<0.05)**
Estimation Model	Test Set for PM_{10} (Day + Night)		
	MAE	SMAPE	r*
Conventional	13.57	0.3548	0.27 (<0.05)
FEBM	14.40	0.3870	0.41 (<0.05)
SILM	13.90	0.3644	0.22 (<0.05)
VGG	13.64	0.3678	0.59 (<0.05)
ResNet	12.83	0.3463	0.60 (<0.05)
ResNet-LSTM	**11.13**	**0.3081**	**0.64 (<0.05)**

* p-value in parenthesis.

As shown in Figure 8, for night-time images, the estimation results for both $PM_{2.5}$ and PM_{10} of the proposed ResNet-LSTM are flatter as compared to ResNet. This verifies that sequential images can help the model stabilize any estimation. In addition, our proposed model ResNet-LSTM tends to give under- and overestimations for $PM_{2.5}$ values, while PM_{10} values tend to be underestimated. Similar patterns can be observed for ResNet, the baseline. This is likely linked to the representativeness of our dataset, where certain ranges of particulates concentrations have not been covered during model training.



(a) $PM_{2.5}$ Estimation

(b) PM_{10} Estimation

FIGURE 8. PM estimation by ResNet, ResNet-LSTM versus PM real measurement.

TABLE 5. MAE, SMAPE, and Pearson correlation of $PM_{2.5}$ and PM_{10} estimation performed by ResNet-LSTM after incorporating meteorological features and one short path, trained on both day-time and night-time images.

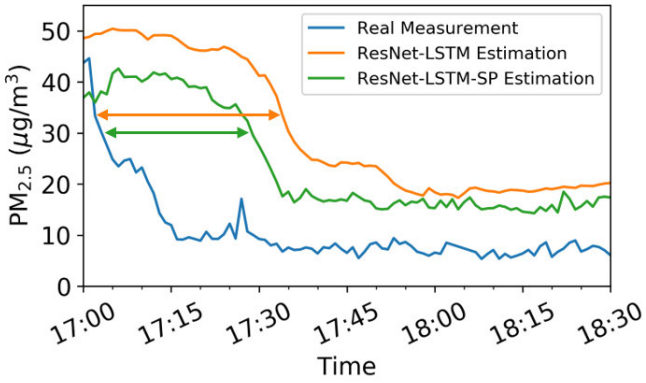
Estimation Model	Test Set for $PM_{2.5}$ (Day + Night)		
	MAE	SMAPE	r*
ResNet-LSTM	6.55	0.2890	0.56 (<0.05)
Met-ResNet-LSTM	6.29	0.2752	0.56 (<0.05)
ResNet-LSTM-SP	5.45	0.2528	0.70 (<0.05)
Met-ResNet-LSTM-SP	**5.31**	**0.2474**	**0.79 (<0.05)**
Estimation Model	Test Set for PM_{10} (Day + Night)		
	MAE	SMAPE	r*
ResNet-LSTM	11.13	0.3081	0.64 (<0.05)
Met-ResNet-LSTM	10.59	0.2892	0.72 (<0.05)
ResNet-LSTM-SP	10.61	0.2924	0.71 (<0.05)
Met-ResNet-LSTM-SP	**9.23**	**0.2612**	**0.73 (<0.05)**
* p-value in parenthesis			

D. MODEL PERFORMANCE AFTER INCORPORATING METEOROLOGICAL FEATURES AND ONE SHORT PATH

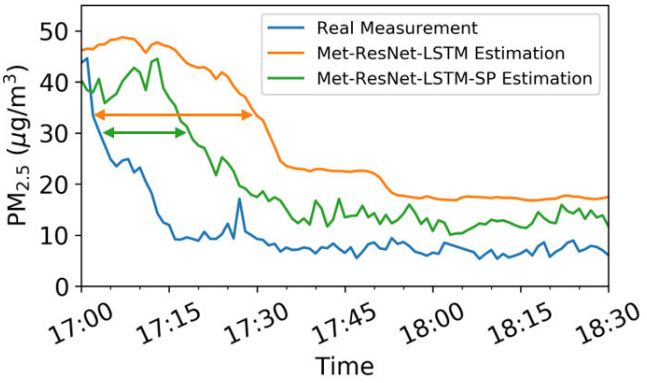
After incorporating the meteorological information into our proposed ResNet-LSTM model, the estimation performance is further improved (see Table 5). This highlights the benefit of using domain-specific meteorological features for image-based PM estimation. As compared to ResNet-LSTM, for $PM_{2.5}$ estimation, our Met-ResNet-LSTM has achieved a reduction in MAE and SMAPE, respectively, by 3.97% and 4.78%, whereas for PM_{10} estimation, our Met-ResNet-LSTM has achieved a reduction in MAE and SMAPE, respectively, by 4.85% and 6.13%. Moreover, Met-ResNet-LSTM has achieved a higher Pearson correlation in PM_{10} estimation, though the corresponding Pearson correlation is not further improved after including meteorological features in $PM_{2.5}$ estimation.

Although the proposed ResNet-LSTM and Met-ResNet-LSTM models have achieved the lowest estimation errors when compared to the baselines, it remains a challenge to accurately predict PM values when they are changing sharply. As shown in Figure 9(a), the real $PM_{2.5}$ measurements dropped sharply starting from 17:00 onwards. However, it was not until 17:36, a 36-min time lag, that the predicted $PM_{2.5}$ measurements from ResNet-LSTM started to drop sharply. More specifically, ResNet-LSTM had not started to drop sharply from 48.57 to 28.29 $\mu g/m^3$ until 17:36. In comparison, Met-ResNet-LSTM had started to drop sharply from 46.18 to 29.89 $\mu g/m^3$ 4 mins earlier, at 17:32. Similar patterns can be observed for PM_{10} (see Figure 9(c)). It takes both ResNet-LSTM and Met-ResNet-LSTM a longer time to respond to the sudden changes in $PM_{2.5}$ and PM_{10} concentrations, though for Met-ResNet-LSTM, such time lag is comparably shorter.

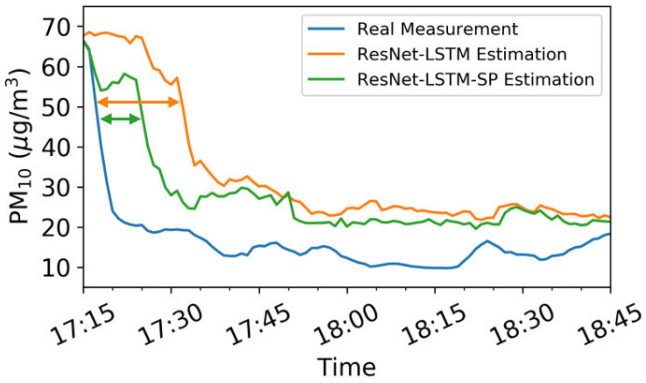
After incorporating one short path into our proposed models, estimation errors for $PM_{2.5}$ and PM_{10} have been further reduced. As shown in Figures 9(b) and 9(d), adding one short path to the model improves the model's sensitivity to sudden changes in PM concentrations, as compared to the previous model without any short path. Given that the short path improves the model's capability in detecting any sudden change in PM concentrations, the overall estimation errors have been further reduced (see Table 5). More specifically, the performance of ResNet-LSTM is further improved after incorporating the short path. For $PM_{2.5}$ estimation, MAE and SMAPE have been further reduced by 16.8% and 12.5%, respectively, whereas for PM_{10} estimation, MAE and SMAPE have been further reduced by 4.67% and 5.10% respectively. Similarly, the performance of Met-ResNet-LSTM is further improved after incorporating the short path. For $PM_{2.5}$ estimation, MAE and SMAPE have been further reduced by 15.6% and 10.1%, respectively, whereas for PM_{10} estimation, MAE and SMAPE has been further reduced by 12.8% and 9.68%, respectively. Furthermore, for both $PM_{2.5}$ and PM_{10} estimation, as compared to the non-SP counterparts, both ResNet-LSTM-SP and Met-ResNet-LSTM-SP have achieved a higher Pearson



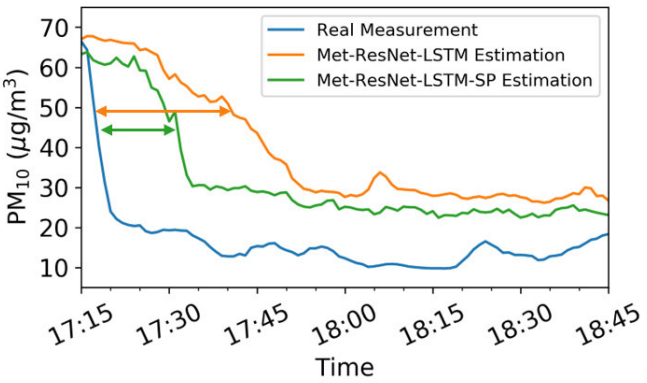
(a) $PM_{2.5}$ estimation of ResNet-LSTM and ResNet-LSTM-SP models versus real measurement



(b) $PM_{2.5}$ estimation of Met-ResNet-LSTM and Met-ResNet-LSTM-SP models versus real measurement



(c) PM_{10} estimation of ResNet-LSTM and ResNet-LSTM-SP models versus real measurement



(d) PM_{10} estimation of Met-ResNet-LSTM and Met-ResNet-LSTM-SP models versus real measurement

FIGURE 9. MAE and SMAPE of PM estimation based on models with and without one short path.

correlation, suggesting that the image and meteorological features can be better learnt by the deep learning models via the short path.

E. LIMITATIONS AND FUTURE WORK

This study aims to estimate $PM_{2.5}$ and PM_{10} concentrations of a specific location in HK using sequential images taken consecutively once every minute, addressing also the night-time image estimation challenge. Our results show that the proposed ResNet-LSTM and the Met-ResNet-LSTM model can achieve a better $PM_{2.5}$ and PM_{10} estimation when compared to the baselines, especially after incorporating the short path. However, a few limitations remain.

First, although deep-learning models can learn the most important features that are proxies to $PM_{2.5}$ and PM_{10} pollutant concentrations, the estimation performance can still be affected by other pollutants, as high O_3 and NO_2 values may also degrade the visibility. For example, [41] noted that image visibility can be affected by other pollutants, such as O_3, NO_2, and humidity. In the future, we can develop a model based on the same methodology to estimate values of other pollutants, such as O_3 and NO_2, based on sequential images taken.

Moreover, though a large image set with labelled $PM_{2.5}$ and PM_{10} measurements has been collected for this study, the range of $PM_{2.5}$ and PM_{10} values covered remains small. The representativeness of the dataset may decrease the estimation performance. In our experiment, images from the full dataset covering the last two days are selected as the test set, however, the test set consists of images labeled with very low $PM_{2.5}$ and PM_{10} values, which are not available in the training set. As a result, significant overestimations of $PM_{2.5}$ and PM_{10} can be observed, though such overestimations are partially addressed by incorporating one short path in the model structure. As shown in Figure 9, ResNet-LSTM-SP and Met-ResNet-LSTM-SP have achieved better estimation results as compared to ResNet. However, if the actual $PM_{2.5}$ and PM_{10} values become too low, i.e., less than 10 $\mu g/m^3$ and 20 $\mu g/m^3$, respectively, the estimation results are comparatively less accurate (see the gaps between the actual and the predicted values in Figure 9). Although the short path further reduces estimation errors, estimation errors that associate with very low $PM_{2.5}$ and PM_{10} values remain large. To overcome this limitation, more sequential images covering a wider range of $PM_{2.5}$ and PM_{10} values across

different time periods (e.g., different hours of day, days of week, and seasons of year) and locations will be collected in the future.

Furthermore, the transferability of our proposed models to other locations should be examined in the future. We plan to collect more images from other official traffic surveillance cameras across different locations in HK. Sequential images will be collected over a longer period to capture the geographical and seasonal variations of air pollution. A larger dataset that covers multiple locations and seasons can be constructed and be used to train and evaluate our proposed model. After the generalization performance of the proposed model is validated on a holdout dataset, a large city-based estimation system will be constructed to provide air pollution estimation in those locations where air quality monitoring stations are not available.

V. CONCLUSION

An end-to-end ResNet-LSTM model has been proposed to estimate $PM_{2.5}$ and PM_{10} values from smartphone-taken images directly. Reliable estimation can be obtained for both day-time and night-time images. Our study consists of four stages. First, we have calibrated two low-cost portable sensors to provide reliable high accuracy pollutant measurements. Second, we have conducted an experiment to show that PM measurements within a distance of up to 500 meters are nearly constant. Third, based on our calibrated sensors and the empirical experiment, a comprehensive dataset containing 3024 images have been constructed. It has covered both day-time and night-time images of the same building (up to 500 meters away), with all images taken by a smartphone camera labeled with the ground truth $PM_{2.5}$ and PM_{10} values obtained from co-located calibrated sensors of the AQMS. Finally, our proposed ResNet-LSTM model have been constructed.

Experimental results have shown that our proposed ResNet-LSTM models and their extended counterparts have outperformed the best deep learning baseline, ResNet, and other conventional baselines. Both ResNet and our proposed ResNet-LSTM have achieved lower estimation errors when compared to conventional baselines. By exploiting the temporal correlation of sequential images, our proposed ResNet-LSTM models have further outperformed ResNet, especially for the night-time image-based estimation. Built upon ResNet-LSTM, our Met-ResNet-LSTM has attempted to account for the meteorological effects on air pollution. We have also incorporated one short path, and developed ResNet-LSTM-SP and Met-ResNet-LSTM-SP. The short path increases the ability of ResNet-LSTM and Met-ResNet-LSTM to capture any sudden changes in pollutant concentrations and improves our image-based estimation performance.

Our proposed deep-learning image-based air pollution models is capable of providing air pollutants estimation based on night-time images and make full-day air pollutants estimation from smartphone- or surveillance camera-taken images possible. In the future, our model can be extended for estimating pollutant values from sequential images taken by government-run traffic surveillance cameras, making full-day estimation for $PM_{2.5}$ and PM_{10}, and other pollutants, such as O_3 and NO_2, across the whole city of HK possible. Our study aims to provide HK citizens a fine-grained pollutants estimation and better health advice based on a fuller coverage of pollutants levels in HK. Our image-based deep learning model can be transferred to other cities to provide timely personal air pollutants report and advice, wherever location-specific and/or timely air pollution reports are not readily available.

ACKNOWLEDGMENT

The authors would like to thank the valuable comments of three anonymous reviewers. They also gratefully acknowledge the Environmental Protection Department in facilitating portable sensor calibration. They acknowledge the assistance of the World Green Organisation in coordinating with the Environmental Protection Department for sensor calibration. They also thank the following PhD students, including Qi Zhang, Along Jin, Andong Wang, and Wenhui Huang (HKU) for their help in sensor calibration and experimental study.

REFERENCES

[1] *Environmental Protection Department*. Accessed: Oct. 29, 2020. [Online]. Available: https://www.aqhi.gov.hk/en/what-is-aqhi/about-aqhi.html
[2] H. Akimoto, *Atmospheric Reaction Chemistry* (Springer Atmospheric Sciences). Tokyo, Japan: Springer, 2016.
[3] M. Kampa and E. Castanas, "Human health effects of air pollution," *Environ. Pollut.*, vol. 151, no. 2, pp. 362–367, Jan. 2008.
[4] X. Meng, Y. Zhang, K.-Q. Yang, Y.-K. Yang, and X.-L. Zhou, "Potential harmful effects of PM2.5 on occurrence and progression of acute coronary syndrome: Epidemiology, mechanisms, and prevention measures," *Int. J. Environ. Res. Public Health*, vol. 13, no. 8, p. 748, Jul. 2016.
[5] J. C. Chow, J. G. Watson, J. L. Mauderly, D. L. Costa, R. E. Wyzga, S. Vedal, G. M. Hidy, S. L. Altshuler, D. Marrack, J. M. Heuss, G. T. Wolff, C. A. Pope, III, and D. W. Dockery, "Health effects of fine particulate air pollution: Lines that connect," *J. Air Waste Manage. Assoc.*, vol. 56, no. 10, pp. 1368–1380, Oct. 2006.
[6] Y. Xing, Y. Xu, M. Shi, and Y. Lian, "The impact of PM2. 5 on the human respiratory system," *J. Thoracic Disease*, vol. 8, no. 1, p. E69, 2016.
[7] C. Liu, F. Tsow, Y. Zou, and N. Tao, "Particle pollution estimation based on image analysis," *PLoS ONE*, vol. 11, no. 2, Feb. 2016, Art. no. e0145955.
[8] Z. Zhang, H. Ma, H. Fu, L. Liu, and C. Zhang, "Outdoor air quality level inference via surveillance cameras," *Mobile Inf. Syst.*, vol. 2016, pp. 1–10, Jan. 2016.
[9] X. Liu, Z. Song, E. Ngai, J. Ma, and W. Wang, "PM2: 5 monitoring using images from smartphones in participatory sensing," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2015, pp. 630–635.
[10] K. Gu, J. Qiao, and X. Li, "Highly efficient picture-based prediction of PM2.5 concentration," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3176–3184, Apr. 2019.
[11] G. Yue, K. Gu, and J. Qiao, "Effective and efficient photo-based PM2.5 concentration estimation," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 3962–3971, Oct. 2019.
[12] C. Zhang, J. Yan, C. Li, H. Wu, and R. Bie, "End-to-end learning for image-based air quality level estimation," *Mach. Vis. Appl.*, vol. 29, no. 4, pp. 601–615, May 2018.
[13] A. Chakma, B. Vizena, T. Cao, J. Lin, and J. Zhang, "Image-based air quality analysis using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3949–3952.
[14] C. Zhang, J. Yan, C. Li, X. Rui, L. Liu, and R. Bie, "On estimating air pollution from photos using convolutional neural network," in *Proc. ACM Multimedia Conf.*, 2016, pp. 297–301.

[15] L. Liu, W. Liu, Y. Zheng, H. Ma, and C. Zhang, "Third-eye: A mobilephone-enabled crowdsensing system for air quality monitoring," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–26, 2018.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, BMVA Press, 2015, p. 41, doi: 10.5244/C.29.41.

[20] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.

[21] T. Zheng, M. H. Bergin, S. Hu, J. Miller, and D. E. Carlson, "Estimating ground-level PM2.5 using micro-satellite images by a convolutional neural network and random forest approach," *Atmos. Environ.*, vol. 230, Jun. 2020, Art. no. 117451.

[22] K. Y. Hong, P. O. Pinheiro, and S. Weichenthal, "Predicting global variations in outdoor PM2.5 concentrations using satellite images and deep convolutional neural networks," 2019, *arXiv:1906.03975*. [Online]. Available: http://arxiv.org/abs/1906.03975

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

[24] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: http://arxiv.org/abs/1609.08144

[25] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[26] R. Chen, X. Wang, W. Zhang, X. Zhu, A. Li, and C. Yang, "A hybrid CNN-LSTM model for typhoon formation forecasting," *GeoInformatica*, vol. 23, no. 3, pp. 375–396, Jul. 2019.

[27] 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, Jul. 2018.

[28] Q. Zhang, J. C. Lam, V. O. Li, and Y. Han, "Deep-AIR: A hybrid CNN-LSTM framework forFine-grained air pollution forecast," 2020, *arXiv:2001.11957*. [Online]. Available: http://arxiv.org/abs/2001.11957

[29] H. Ozkaynak, A. D. Schatz, G. D. Thurston, R. G. Isaacs, and 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.*, vol. 35, no. 11, pp. 1176–1185, Nov. 1985.

[30] N. P. Hyslop, "Impaired visibility: The air pollution people see," *Atmos. Environ.*, vol. 43, no. 1, pp. 182–195, Jan. 2009.

[31] S. K. Nayar and S. G. Narasimhan, "Vision in bad weather," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2. 1999, pp. 820–827.

[32] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[33] H. Wang, X. Yuan, X. Wang, Y. Zhang, and Q. Dai, "Real-time air quality estimation based on color image processing," in *Proc. IEEE Vis. Commun. Image Process. Conf.*, Dec. 2014, pp. 326–329.

[34] B. Yang and Q. Chen, "PM2.5 concentration estimation based on image quality assessment," in *Proc. 4th IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2017, pp. 676–681.

[35] Y. Li, J. Huang, and J. Luo, "Using user generated online photos to estimate and monitor air pollution in major cities," in *Proc. 7th Int. Conf. Internet Multimedia Comput. Service (ICIMCS)*, 2015, pp. 1–5.

[36] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

[37] J. Ma, K. Li, Y. Han, and J. Yang, "Image-based air pollution estimation using hybrid convolutional neural network," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 471–476.

[38] Z. Wang, W. Zheng, C. Song, Z. Zhang, J. Lian, S. Yue, and S. Ji, "Air quality measurement based on double-channel convolutional neural network ensemble learning," *IEEE Access*, vol. 7, pp. 145067–145081, 2019.

[39] Q. Li and B. Xie, "Image-based air quality estimation," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*. Cham, Switzerland: Springer, 2019, pp. 161–171.

[40] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, in Proceedings of Machine Learning Research, vol. 38, G. Lebanon and S. V. N. Vishwanathan, Eds. San Diego, CA, USA: PMLR, May 2015, pp. 562–570. [Online]. Available: http://proceedings.mlr.press/v38/lee15a.pdf

[41] F.-C. Lin, S.-H. Lu, C.-J. Liang, Y.-H. Chen, and J.-J. Liang, "Quantifying source apportionment for ambient haze: An image haze extraction approach with air quality monitoring data," *Environ. Res.*, vol. 184, May 2020, Art. no. 109216.

SHIGUANG SONG received the B.Eng. (Hons.) degree in Electrical Engineering from University of Liverpool and the M.Sc. degree in Future Power Networks from Imperial College London, United Kingdom, in 2015 and 2016, respectively. He is currently working toward the Ph.D. degree in the Department of Electrical & Electronic Engineering, The University of Hong Kong (HKU). His research interests include deep learning, artificial intelligence, environmental technologies and image processing.

JACQUELINE C. K. LAM (Member, IEEE) is Associate Professor at the Department of Electrical and Electronic Engineering, The University of Hong Kong. Since 2018, she has co-established the HKU-AI WiSe with Prof. Victor O. K. Li, Chair of Information Engineering at the University of Hong Kong. She is the Co-Director of the HKU-Cambridge Clean Energy and Environment Research Platform, and of the HKU-Cambridge AI to Advance Well-being and Society Research Platform. She was the Hughes Hall Visiting Fellow, before she takes up the Visiting Senior Research Fellow and Associate Researcher in Energy Policy Research Group, Judge Business School, The University of Cambridge. Her research studies clean energy and environment using interdisciplinary approaches, with a special focus on China and the UK. Her recent research focuses on the use of big data and machine learning techniques to study personalized air pollution monitoring and health management. Her work is published in IEEE, *Environment International*, *Applied Energy*, *Environmental Science and Policy*, and *Energy Policy*. She has received three times the research grants awarded by the Research Grants Council, HKSAR Government, during 2011–2017. The funded amount was USD 7.8M in PI/Co-PI capacity. Her recent research study, in joint collaboration with Yang Han and Victor O. K. Li on $PM_{2.5}$ pollution and environmental inequality in Hong Kong, has been published in *Environmental Science and Policy*, and widely covered by more than 30 local and overseas newspapers and TVs. She has recently co-organized the HKU-Cambridge AI for Social Good Symposium with Prof. Jonathan Crowcroft, FRS (CST Dept., Cambridge) and Prof. Victor O. K. Li (Clare Hall Life Fellow, Cambridge). She also serves as the co-editor of a special issue on AI for environmental decision-making, published by *Environmental Science and Technology*, an Elsevier SCI journal.

YANG HAN received the M.Sc, degree in Computer Science (with distinction) from the University of Hong Kong (HKU) and the M.Phil. degree in Technology Policy from The University of Cambridge. Currently he is working towards the Ph.D. degree with the Department of Electrical & Electronic Engineering, HKU. His recent research specializes on spatio-temporal analysis and applications in environmental science and policy studies in China. He has published widely in *Environmental Science and Policy*, *Data and Policy*, and several IEEE journals.

VICTOR O. K. LI (Life Fellow, IEEE) received the S.B., S.M., E.E., and Sc.D. degrees in Electrical Engineering and Computer Science from MIT. He is Chair of Information Engineering and Cheng Yu-Tung Professor in Sustainable Development at the Department of Electrical & Electronic Engineering (EEE) at the University of Hong Kong. He is the Director of the HKU-Cambridge Clean Energy and Environment Research Platform, and of the HKU-Cambridge AI to Advance Well-being and Society Research Platform, which are interdisciplinary collaborations with Cambridge University. He was Visiting Professor in the Department of Computer Science and Technology, The University of Cambridge, from April to August 2019. He was the Head of EEE, Assoc. Dean (Research) of Engineering and Managing Director of Versitech Ltd. He serves on the board of Sunevision Holdings Ltd., listed on the Hong Kong Stock Exchange and co-founded Fano Labs Ltd., an artificial intelligence (AI) company with his Ph.D. student. Previously, he was a Professor of Electrical Engineering at the University of Southern California (USC), Los Angeles, CA, USA, and Director of the USC Communication Sciences Institute. His research interests include big data, AI, optimization techniques, and interdisciplinary clean energy and environment studies. In Jan 2018, he was awarded a USD 6.3M RGC Theme-based Research Project to develop deep learning techniques for personalized and smart air pollution monitoring and health management. Sought by government, industry, and academic organizations, he has lectured and consulted extensively internationally. He has received numerous awards, including the PRC Ministry of Education Changjiang Chair Professorship at Tsinghua University, the UK Royal Academy of Engineering Senior Visiting Fellowship in Communications, the Croucher Foundation Senior Research Fellowship, and the Order of the Bronze Bauhinia Star, Government of the HKSAR. He is a Fellow of the Hong Kong Academy of Engineering Sciences, the IEEE, the IAE, and the HKIE.