HVAQ: A High-Resolution Vision-Based Air Quality Dataset

Zuohui Chen, Tony Zhang, *Graduate Student Member, IEEE*, Zhuangzhi Chen, Yun Xiang, *Member, IEEE*, Qi Xuan, *Member, IEEE*, and Robert P. Dick, *Senior Member, IEEE*

Abstract— Air pollutants, such as particulate matter, negatively impact human health. Most existing pollution-monitoring techniques use stationary sensors, which are typically sparsely deployed. However, real-world pollution distributions vary rapidly with position, and the visual effects of air pollution can be used to estimate concentration, potentially at high spatial resolution. Accurate pollution-monitoring requires either densely deployed conventional point sensors, at-a-distance vision-based pollution monitoring, or a combination of both. The main contribution of this article is that, to the best of our knowledge, it is the first publicly available, high temporal and spatial resolution air quality dataset containing simultaneous point sensor measurements and corresponding images. The dataset enables, for the first time, high spatial resolution evaluation of image-based air pollution estimation algorithms. It contains PM2.5, PM10, temperature, and humidity data. We evaluate several state-of-the-art vision-based PM concentration estimation algorithms on our dataset and quantify the increase in accuracy resulting from higher point sensor density and the use of images. It is our intent and belief that this dataset can enable advances by other research teams working on air quality estimation. Our dataset is available at https://github.com/implicitDeclaration/HVAQ-dataset/tree/master.

Index Terms— Air pollution, evaluation, high-resolution dataset, image analysis, particulate matter (PM).

I. INTRODUCTION

AIR pollution is a serious threat to human health and is closely related to disease and death rates [1]–[5]. According to the World Health Organization (WHO), air pollution causes seven million premature deaths per year, largely as a result of increased mortality from stroke, heart disease, and lung cancer. The most common harmful air pollutants are particulate matter (PM), sulfur dioxide, and nitrogen dioxide. This work focuses on PM, which can increase the rate of cardiovascular and respiratory diseases [6], [7].

Air-monitoring stations are mainly used to obtain air pollution data [8]. However, air pollution concentration can

Fig. 1. Distribution of air monitoring stations in Hangzhou. On average, there is one monitoring site for every 35 km^2.

vary within a relatively short distance: the low sensor density (see Fig. 1) can lead to inaccurate estimation of the high-resolution pollution field. Low measurement spatial density makes it especially difficult to estimate human exposure to air pollution.

The PM concentrations are correlated with source distributions. For example, PM has heterogeneous sources [9], for example, automobiles, manufacturing, and building construction. In addition, numerous factors, including wind, humidity, and geography [10], [11], are related to PM distributions. Increasing sensor density or adding image sensors supporting high spatial resolution captures can increase the field estimation accuracy and resolution of pollutant concentrations.

In addition to air quality sensors, image sensors can also be used to estimate PM concentrations, potentially at higher spatial resolutions [12], [13]. Image data can be derived from several sources such as social media and digital cameras [14], [15]. Moreover, consumer-grade cameras are inexpensive and can gather pollution data over a wide field of view (see Fig. 2). PM can be estimated by analyzing the visual haze effect caused by scattering and absorption [13]. On days with high pollution, visibility is low because light is scattered away from the camera by PM.

The ground-truth pollution data are typically obtained from the nearest of several sparsely deployed monitoring stations. Lack of high-resolution ground-truth pollution measurements makes it difficult to evaluate vision-based pollution estimation techniques, as pollution can vary rapidly with position [15].

Fig. 2.   (a) High-altitude quadcopter view and (b) low-altitude camera view.

TABLE I
DATASETS COMPARISON

	Spatial resolution	Temporal resolution	Scale	Images
Apte et al.	High	Second	$30\ km^2$	No
ChinaHighPM10	Low	Hour	$9.6 Mkm^2$	No
ImgSensingNet	High	Hour	$7\ km^2$	Yes
Air monitoring stations (Hangzhou)	Low	Hour	$16\,853\ km^2$	No
HVAQ	High	Second	$1\ km^2$	Yes

Moreover, vision-based pollution estimation techniques generally assume homogeneous distributions of particles and gases within images, implying constant light attenuation [15]. In reality, pollution concentration changes rapidly in space. Therefore, accurate evaluation of vision-based estimation algorithms requires high-resolution datasets [16].

This article presents a dataset containing high spatial (one sensor every 2.5 km^2) and temporal (one second interval) resolution particle counter-based pollution measurements with corresponding images, in addition to auxiliary information including GPS locations, humidity, and temperature. These properties are significant: this is the first publicly available dataset capable of being used to train and evaluate vision-based pollution estimation and forecasting techniques at high spatial resolutions. To the best of our knowledge, there have been no publicly available datasets enabling evaluation in this context. Certainly, there are datasets containing high-resolution [16], [17] and wide coverage [18], [19] air pollution data. However, none of them contain corresponding synchronized images. Based on the dataset, which is the primary contribution of this article, we also make several observations, for example, the rate of spatial variation in pollution concentration and the evaluation of several existing vision-based PM concentration predication algorithms. To evaluate the improvement brought by increasing sensor spatial resolution and using images, we use heterogeneous information for concentration estimation, that is, images and particle counter-based PM concentrations, which were not considered in the previous vision-based algorithms.

The rest of this article is organized as follows. Section II summarizes related work. Sections III and IV describe our data collection and analysis process. Section V presents the experimental results. Section VI concludes this article.

II. RELATED WORK

The related work can be generalized into three categories: environment monitoring methods, spectrum RGB cameras, and vision-based techniques.

A. Environment Monitoring Methods

Most existing air quality monitoring systems [20] have low temporal and spatial resolutions. For example, Janssens-Maenhout et al. [21] provide a harmonized gridded air pollution emission dataset. This dataset includes multiple pollutants on a global scale with $0.1° \times 0.1°$ spatial resolution (latitude and longitude). De Vito et al. [22] collect an air quality dataset containing 9358 instances of hourly averaged responses from metal oxide chemical sensors. Devices are

deployed in a highly polluted area in Italy at high spatial resolution. Li et al. [23] provide a dataset of mobile air quality measurements in Zurich. They use sensor boxes installed on mobile trams. Static installations are also deployed close to high-quality reference stations for calibration. Their sensors move around the city, recording every 5 s. Apte et al. [16] use two Google street view vehicles equipped with data acquisition platforms to collect the air pollution data of a 30 km^2 city area. Their data contain nitrogen oxides and black carbon. Unlike our dataset, it does not contain PM concentrations, images, and weather conditions. Wei et al. [18] describe the ChinaHighPM10 dataset, which integrates multiple data sources and contains hourly PM10 data in China with 1 km resolution. Generally, these datasets typically do not include images for the evaluation of vision-based pollution estimation.

Yang et al. [24] describe ImgSensingNet, an air quality monitoring system consisting of unmanned aerial vehicles (UAVs) and ground sensors. Their system uses UAVs for vision-based air quality index (AQI) monitoring and activates the ground sensing network based on vision-based AQI inference. Their dataset differs from ours as follows: 1) it is not public nor can be expected to be public in the near future; 2) their data have time gaps, that is, there is ground sensor data only when the UAV is uncertain about the AQI in some areas, making the intervals without ground sensor data useless for validating vision-based algorithms. In contrast, our data acquisition system operated continuously; 3) their dataset contains an irreversible AQI index inferred from PM data, while our dataset provides raw PM data, meteorology information, and environmental information; and 4) their point sensor data acquisition period is 1 h (at best, depending on whether the point sensors are activated) while ours is 1 s. Table I compares different datasets. Our HVAQ is unique in providing high spatial and temporal resolution pollution measurements with corresponding images. Moreover, it has been publicly released.

B. Spectrum-Based RGB Cameras

Image data are typically gathered using spectral measurements, imaging systems, and multispectral imaging [25]. Spectral measurements acquire accurate color data without structural information from the scene. Imaging systems such as scanners or digital RGB cameras capture detailed structural information [26] with a limited number of color channels. Mismatches between the camera's spectral sensitivities and desired color matching functions can reduce the camera's color quality.

In contrast, multispectral imaging can produce accurate, detailed representations of the scene for both color

and structure [27]. Nonetheless, multispectral imaging systems are expensive, while spectral measurements provide very limited spatial resolution.

C. Image-Based Techniques

Vision-based pollution estimation is being studied due to its spatial range and low cost. Li *et al.* [14] analyze photographs acquired from social media and establish the correlation between visual haze and PM2.5 concentration. Liu *et al.* [15] estimate PM2.5 using support vector regression on six features extracted from images. Zhang *et al.* [12] describe an air pollution estimation algorithm that uses outdoor photos. They use a convolutional neural network to predict the pollution level. Rijal *et al.* [28] integrate the prediction of multiple neural networks to calculate the PM2.5 of images. The ensemble of neural networks provides more accurate estimation than a single network. Gu *et al.* [29] designed a vision-based estimation of PM2.5 concentration. It requires low-pollution reference images for each area of interest. Zhang and Dick [13] estimate multiple pollution concentrations from images using scattering and absorption models.

The main problem is that it is unclear as to how accurate these methods are because, until now, ground-truth data are of very low spatial resolution. Therefore, our dataset makes it possible to evaluate them in a high-resolution scenario.

III. SENSOR DEPLOYMENT

This section describes the deployment of our sensors and cameras in order to collect both high-resolution ground-truth data and corresponding images.

A. Sensor Calibration

Our sensing platform is equipped with humidity and temperature sensors. The precision is 3% relative humidity (RH) and ± 0.3 °C, respectively. The PM sensor can detect particles with 0.3–10 μm diameters according to the scattered light intensity in a specific direction [30].

We calibrated our particle counting sensors based on the measurements from the air quality monitoring site of Hangzhou Meteorological Bureau in Hemu Primary School using co-location method [31]. Our device is co-located with an air monitoring station, which is equipped with a high-precision sensor. The device collects data for 58 h continuously. The station data are considered ground truth. We use the least-squares method to fit a quadratic function to the ground truth data. According to the measured concentration, we fit the data into a two-stage piecewise linear function. The fitted function is as follows:

$$y = \begin{cases} 1.61x + 16.01 & \text{for } 0 \leqslant x \leqslant 30 \text{ and} \\ 0.13x + 29.48 & \text{for } x > 30 \end{cases} \quad (1)$$

where x is the original value and y is the calibrated one. The calibration data are gathered during a 2-day period and the result is shown in Fig. 3. Calibration reduces root mean squared error from 32.74 to 3.88 μg/m^3, while the variance for all the data over all locations is 9.33 μg/m^3.
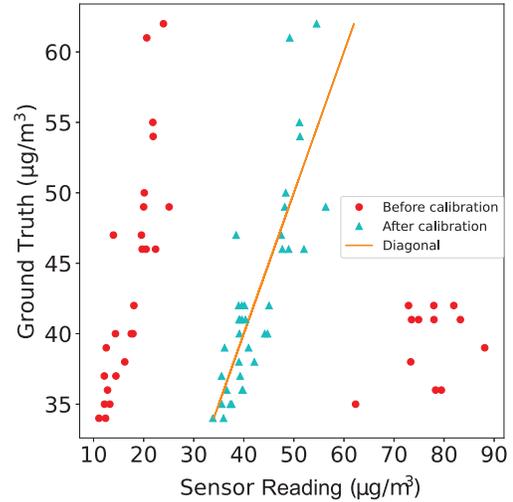
Fig. 3. Pollution concentration calibration.
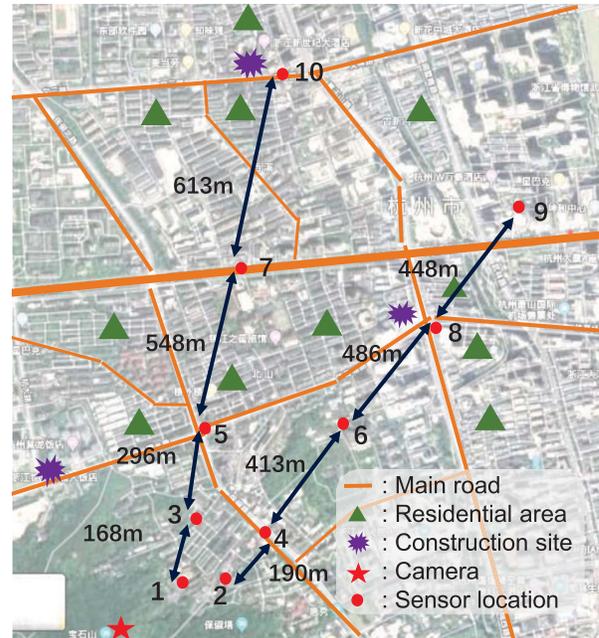
Fig. 4. Distribution of sensors and pollution sources. Sensor locations are numbered in ascending order according to the distance from the camera.

B. Deployment Details

Our dataset[1] contains both high-resolution ground-truth data and wide-view images. The sensors are placed outdoors and close to living areas. The city center is very crowded and the air quality here affects people more significantly. We deploy our sensors in the urban area of Hangzhou, a city of more than eight million residents and frequently affected by high PM2.5 concentrations [32]. Existing research [33] shows that in the main urban area of Hangzhou, the sources of PM2.5 are biomass burning/construction dust (41.6%), vehicle exhaust/metallurgical (metals' production and purification) dust (29.3%), unknown source (11.2%), oil combustion (9.8%), and soil (8.0%). As shown in Fig. 4, the main pollution sources are marked on the map and the sensors are

[1]Available at https://github.com/implicitDeclaration/HVAQ-dataset/tree/master.

TABLE II
GPS LOCATIONS OF OUR SENSORS AND PICTURE

Location	P1	P2	P3
Longitude	120.153173°	120.15488°	120.153894°
Latitude	30.269884°	30.268726°	30.27096°
P4	P5	P6	P7
120.156252°	120.153905°	120.15936°	120.155162°
30.270242°	30.27358°	30.273139°	30.278369°
P8	P9	P10	Photo location
120.161912°	120.164792°	120.156541°	120.153955°
30.276465°	30.279437°	30.283932°	30.267191°

TABLE III
PAIRWISE SENSORS DISTANCE (m)

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	0	113	168	281	464	697	1012	1150	1730	1625
P2		0	211	190	509	603	1040	1089	1537	1660
P3			0	245	296	575	844	1000	1450	1457
P4				0	388	413	871	899	1347	1510
P5					0	464	548	819	1250	1161
P6						0	625	486	934	1170
P7							0	656	918	613
P8								0	448	953
P9									0	877
P10										0

located on two straight lines from the observation point. The GPS locations and distances between each pair of sensors are listed in Tables II and III. The sensors are sampled every second and an image is captured every 20 min. Since the temporal variation of PM concentration is slower than the acquisition rate of the sensors, and the image acquisition rate is also limited by the flight time of quadcopter, images are captured less frequently.

Table IV lists our equipment. To derive a wide-view image covering all sensor locations, we mount a camera on a quadcopter, as shown in Fig. 5. We use two approaches to take photographs. First, we use a quadcopter equipped with a 4864 × 3648 camera at 90 m altitude. Moreover, since the quadcopter has limited carrying and battery capacities, we also take pictures at a fixed location on the mountain top with a resolution of 2592 × 1936 using an iPad and at a resolution of 4000 × 3000 using a phone. The parameters of our cameras are listed in Table V. The quadcopter camera uses the Sony Exmor R CMOS sensor. The other two cameras types are the Apple iSight and Sony IMX586. Hundred and six images and over 300 000 PM samples were gathered in three days.

Apparently, the cost of vision-based approach (about $30) is much less than the total cost of the sensors (about $670). Thus, predicting air pollution concentrations using images is more convenient and less expensive than particle counters, but requires sophisticated image-processing algorithms.

IV. DATASET ANALYSIS

Q1) What is the impact of environmental conditions on measurement accuracy?

Our data collection period spans three days with varying environmental conditions. The impact of the environmental conditions determines the corresponding parameters of the calibration functions. Thus, it is important to investigate the relationship between the impact of environmental conditions on measurement accuracy.

Fig. 5. Quadcopter (Dji Phantom 4 Pro) used in our deployment.

TABLE IV
USED EQUIPMENT

	Camera	Quadcopter	Sensors	Battery	Platform
Price ($)	200-460	1500	3-15	12	37
Model	Quadcopter default camera, iPad, Onplus 7 phone camera	Dji Phantom 4 Pro	PM2.5, PM10, humidity, temperature	4000 mAh	Raspi 3B+

TABLE V
CAMERA PARAMETER

Camera	Pixel	Sensor	Aperture
quadcopter	20,000,000	Sony Exmor R	f/2.8-f/11
pad	8,000,000	iSight	f/2.4
phone	48,000,000	Sony IMX586	f/1.6

Q2) What are the spatial variation characteristics of PM2.5 concentration?

Since the spatial distribution of PM2.5 determines the required sensor density, this information helps us to determine the benefits of increasing the spatial resolution of a pollution sensor network.

Q3) How does the correlation of pollution concentrations at two different locations depend on their separation?

The relationship between PM2.5 concentration and distance provides information relevant to the required number of sensors. For example, if the PM2.5 concentrations at two locations are highly correlated, one sensor might be used to determine the concentrations at both locations.

Q4) How much do additional point sensors and vision-based methods improve estimation accuracy?

This question builds on Q2 and Q3. To evaluate the effectiveness and novelty of our new air quality dataset, it is important to understand how the sensor density and the use of images relates to the PM2.5 estimation accuracy.

A. Temperature and Humidity Correlations With Pollution Concentrations

Environmental factors such as weather conditions can affect sensor readings. For the deployments on October 19 and November 10, we calculate the R^2 correlation coefficients for PM2.5 and several environmental factors.

A1: Environmental Conditions Have Limited Impact: The results in Table VI shows that the correlations between PM2.5 and weather factors are insignificant. Wu *et al.* [34]
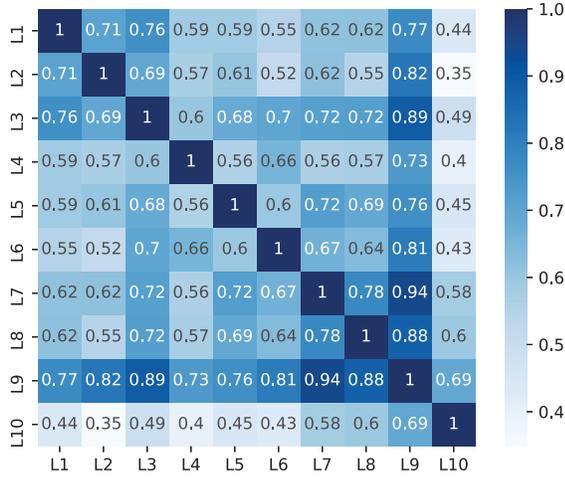
Fig. 6. Confusion matrix for PM2.5 correlation. The number on the axis represents the number of the corresponding location.

TABLE VI

PM2.5 CORRELATION WITH ENVIRONMENTAL FACTORS ON OCTOBER 19 AND NOVEMBER 10

	PM2.5	Temperature	Humidity
PM2.5	1.0	0.298	0.306
Temperature	0.298	1.0	0.808
Humidity	0.306	0.808	1.0

TABLE VII

STATISTICS FOR PM AND ENVIRONMENTAL DATA

Data in Different Days	Standard Deviation	Data Range	Average Value	Date
PM2.5 ($\mu g\,m^{-3}$)	2.03	4.56	13.93	Jul. 24
PM2.5 ($\mu g\,m^{-3}$)	2.22	5.11	25.03	Jul. 06
PM2.5 ($\mu g\,m^{-3}$)	6.68	22.21	56.90	
Temperature (°C)	5.64	16.02	26.40	Oct. 19
RH (%)	12.36	37.63	45.13	
PM2.5 ($\mu g\,m^{-3}$)	4.77	16.51	48.76	
Temperature (°C)	4.53	14.21	24.79	Nov. 10
RH (%)	10.20	34.71	40.16	

report that temperature is not significantly correlated with PM concentrations in Hangzhou and PM2.5 concentration is only significantly elevated when RH is higher than 60%. Moreover, the correlation between environmental factors and PM concentration also depends on the region and season, for example, Zhu *et al.* [35] report that PM concentration and RH show seasonal correlation. However, since temperature and humidity do not directly affect the measurement process of particle counters, we do not include temperature and humidity in our calibration functions.

B. Correlations of PM Readings

Our measurements demonstrate that it typically takes more than 10 min for concentration to change by 10 μg/m³. Our data sampling period is 1 s and is considered adequate given the relatively slow concentration change. We quantify the spatial variation of pollution by calculating the standard deviations over all sensors.

A2: Spatial variation of PM2.5 is high. As shown in Table VII, the two locations have different concentration and variation trends. Moreover, the large differences indicate that multiple pollution levels coexist in a single image.
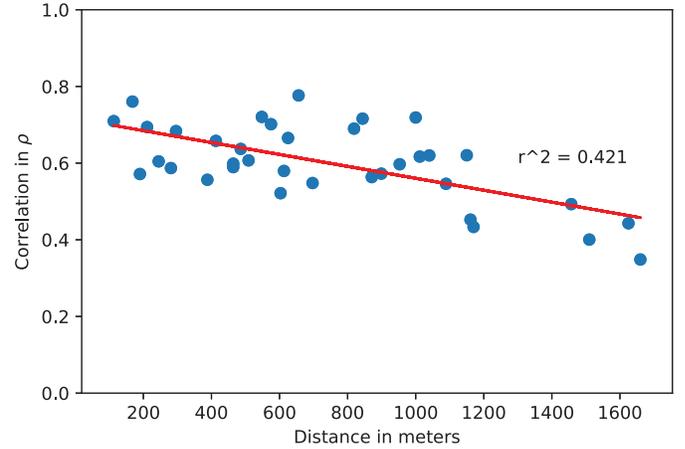
Fig. 7. Pairwise correlation between sensors for October 19 and November 10 as functions of their distances.

Fig. 6 shows the sensor correlations during the November 10 and October 19 deployments. We expect the correlations to decrease with increasing inter-sensor distances. The deployment results confirm our hypothesis.

A3: Pollution concentrations are spatially correlated. We quantify the correlations for further analysis. Fig. 7 shows sensor correlations as a function of pairwise distance. The correlation is measured by the Spearman correlation coefficient

$$\rho = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)} \qquad (2)$$

where d_i represents the position difference of the paired variables after the two variables are sorted separately and N is the total number of samples. The slope of the fitting line is 0.155 /km and the coefficient of determination (R^2) is 0.421, which implies that closer sensors enable more accurate estimates.

V. EXPERIMENTAL RESULTS

In this section, we evaluate the impacts of changing sensor density and using vision-based techniques on estimation accuracy. Specifically, we estimate PM2.5 concentration based on vision-based analysis using transmission information and standard deviation of gray-scale pixel values and compare the performance of several state-of-the-art estimation algorithms.

A. Experimental Setup

We design a portable sensing platform to collect, process, and transmit data, as shown in Fig. 8. The system battery life is 3.5 h. The following algorithms are used to estimate pollutant concentration from the measured data.

1) *Random forest regression (RFR):* This is a widely used algorithm for regression and classification. It combines multiple weak models to form a strong model with much better performance. Random forests contain multiple, unrelated decision trees. The final output depends on the decisions of all trees.

2) *Gradient boosting regression (GBR):* This method combines a group of weak learners with low complexity and low training cost. It reduces the problem of overfitting and modifies the weights at each training round to
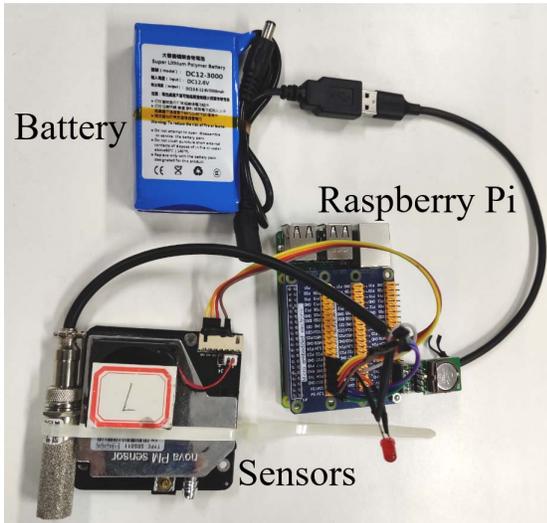
Fig. 8. Sensing platform consisting of battery, Raspberry Pi, and sensors.

Fig. 9. Transmission estimation result.

produce a strong learner. Gradient boosting modifies its models based on the gradient descent direction of the loss functions of the previously established models.

3) *Support vector regression (SVR):* Support vector regression aims to find a regression plane with minimal distance to the dataset. Generally, there is a kernel function mapping data to high-dimensional space for better performance on complex manifolds.

The algorithms are implemented using the Python package *sklearn*. The RFR parameter *n_estimators* is set to 100 and *criterion* is set to mean squared error. We set the GBR parameter *learning_rate* to 0.1 and *n_estimators* to 100, using least-squares regression. We set the SVR parameter *c* to 1.0 and *epsilon* to 0.1, using a radial basis function kernel. The parameters are chosen to maximize the training performance. We use the following equation to combine sensor readings and image properties:

$$S(x, t) = G\big(s_1(t), s_2(t), \ldots, s_{10}(t), t_{\mathrm{dcp}}(t), \beta_{\mathrm{sd}}(t)\big) \quad (3)$$

where x is the location index, $S(x, t)$ is the concentration estimation at time t, $s_1, s_2, \ldots, s_{10}$ are the available ten sensors, $t_{\mathrm{dcp}}(t)$ is the transmission information at time t, $\beta_{\mathrm{sd}}(t)$ is the standard deviation of gray-scale image, and G is the estimation algorithm, which refers to RFR, GBR, or SVR. We use $t_{\mathrm{dcp}}(t)$ for low-altitude data and $\beta_{\mathrm{sd}}(t)$ for high-altitude data. We later describe how $t_{\mathrm{dcp}}(t)$ and $\beta_{\mathrm{sd}}(t)$ are obtained for each image in the dataset in Section V-B.

We use data from July 24, October 19, and November 10 for all time stamps and use the mean absolute error (MAE) as the evaluation criteria

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (4)$$

where y_i is the actual PM2.5 concentration and $\hat{y}_i$ is the predicted PM2.5 concentration.

There are two classes of images in our dataset. Those in the first class were captured from the ground, on top of a mountain (78 m). The second class contains images from
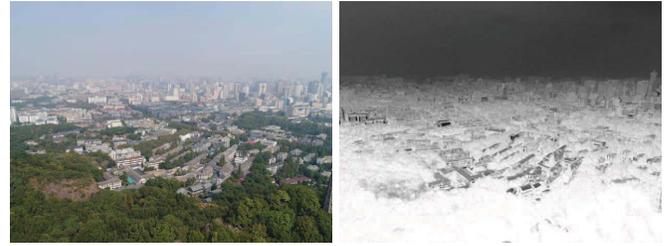
a quadcopter flying above the same mountain (78 m + 90 m). We tried to take the images from the quadcopter and mountain at the same angle and keep the images as similar to each other as possible, but there is some (unavoidable) variation in camera orientation (see Fig. 2).

We divide our data into "low-altitude" and "high-altitude" subsets according to the image class and evaluate our algorithms separately on the two classes. We analyze 19 images from the high-altitude dataset and 26 images from the low-altitude dataset. Furthermore, each dataset is divided as follows: 75% of the data and images are randomly selected as a training set and the rest are the testing set. We run the prediction model 50 times per random split of the training and testing datasets.

B. Image-Enhanced Concentration Estimation

Images can be used to estimate PM concentrations in large areas because images can extract the haziness information. We predict PM2.5 concentrations at locations without sensors. Since PM2.5 attenuates light, we estimate PM2.5 concentration in part through the light attenuation coefficient β from the haze model. We use the following image features to estimate PM2.5: the dark channel $t_{\mathrm{dcp}}(x)$ and the standard deviation β_{sd}. $t_{\mathrm{dcp}}(x)$ is determined using (13) and β_{sd} is determined using (21). Fig. 9 is an example of applying the dark channel prior to an image. The left image is the original hazy image, and the right image is the transmission estimate for each pixel using dark channel prior. Note that certain kinds of weather conditions like rain and snow might be misinterpreted as pollution. Our PM2.5 estimation algorithm mainly works in sunny and cloudy conditions, which are common in Hangzhou.

The atmospheric model describing an image influenced by haze follows [36]:

$$\boldsymbol{I}(x) = \boldsymbol{J}(x)t(x) + \boldsymbol{A}(1 - t(x)) \quad (5)$$

where x is the pixel location, $\boldsymbol{I}$ is the observed image, $\boldsymbol{J}$ is the scene radiance (image without any haze), $\boldsymbol{A}$ is the atmospheric light, t is the transmission function.

Images with higher air pollution tend to look hazier due to lower transmission and contrast. Hence, image features that correlate to haze level enable pollutant concentration estimation.

1) Dark Channel Prior: The dark channel prior, which has been widely used for haze removal, can be used to estimate the transmission of each image pixel. The dark channel prior method is based on the observation that in most haze-free

patches, at least one color channel has some pixels with very low intensities. The dark channel is defined as the minimum of all pixel colors in a local patch and can be calculated using the following [37]:

$$J_{\text{dark}}(x) = \min_{c \in r,g,b} \left(\min_{y \in \Omega_r(x)} J^c(y) \right) \tag{6}$$

where J^c is an RGB channel of $\boldsymbol{J}$, and $\Omega_r(x)$ is a local patch centered at x with the size of 15×15. Assume the atmospheric light $\boldsymbol{A}$ is given and the transmission in a local patch $\Omega_r(x)$ is constant, taking the minimum operation in the local patch on (5), we have

$$\min_{y \in \Omega_r(x)} \left(I^c(y) \right) = \tilde{t}(x) \min_{y \in \Omega_r(x)} \left(J^c(y) \right) + \left(1 - \tilde{t}(x) \right) A^c \tag{7}$$

where $\tilde{t}(x)$ is the patch's transmission. The minimum operation is performed on three color channels independently, which is equivalent to

$$\min_{y \in \Omega_r(x)} \left(\frac{I^c(y)}{A^c} \right) = \tilde{t}(x) \min_{y \in \Omega_r(x)} \left(\frac{J^c(y)}{A^c} \right) + \left(1 - \tilde{t}(x) \right). \tag{8}$$

By taking the minimum of three color channels, we have

$$\min_c \left(\min_{y \in \Omega_r(x)} \left(\frac{I^c(y)}{A^c} \right) \right) = \tilde{t}(x) \min_c \left(\min_{y \in \Omega_r(x)} \left(\frac{J^c(y)}{A^c} \right) \right) + \left(1 - \tilde{t}(x) \right). \tag{9}$$

According to the definition of dark channel prior, the dark channel J_{dark} of the haze-free radiance $\boldsymbol{J}$ tends to be zero

$$J_{\text{dark}}(x) = \min_c \left(\min_{y \in \Omega_r(x)} J^c(y) \right) = 0. \tag{10}$$

Because A^c is always positive, this lead to

$$\min_c \left(\min_{y \in \Omega_r(x)} \frac{J^c(y)}{A^c} \right) = 0. \tag{11}$$

Substituting (11) into (9), we can estimate the transmission as follows:

$$\tilde{t}(x) = 1 - \min_c \left(\min_{y \in \Omega_r(x)} \frac{I^c(y)}{A^c} \right). \tag{12}$$

In practice, the atmosphere always contains some haze, which provides depth information. We can optionally keep a small amount of haze by introducing a constant parameter ω $(0 < \omega < 1)$ into (12)

$$\tilde{t}(x) = 1 - \omega \min_c \left(\min_{y \in \Omega_r(x)} \frac{I^c(y)}{A^c} \right). \tag{13}$$

We fix the value of ω to 0.95 because this approximates the sparse haze present even on relatively clear days. The atmospheric light A is estimated through this procedure: we pick the top 0.1% brightest pixels in the dark channel and the input image $\boldsymbol{I}$ to calculate the atmospheric light. For each hazy image in our dataset, we take the average of the pixel-level transmissions estimated using (13). The resulting average is used for low-altitude data.

2) Standard Deviation: In order to calculate the intensity standard deviation for each image, we convert the RGB image to a gray-scale image and then calculate the standard deviation of all the pixel intensities. The standard deviation is closely related to haze density [38]. The scattering coefficient is a measure of haze density. The higher the scattering coefficient, the higher the haze density. The transmission function is

$$t(x) = e^{\beta d(x)} \tag{14}$$

where β is the scattering coefficient and d is the depth. Substituting (14) into (5), we have

$$\boldsymbol{I}_g(x) = \boldsymbol{J}(x)e^{\beta d(x)} + \boldsymbol{A}\left(1 - e^{\beta d(x)}\right). \tag{15}$$

The variance of a gray-scale image is

$$\sigma_{I_g}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(I_g(i) - \frac{1}{N} \sum_{j=1}^{N} I_g(j) \right)^2$$
$$= e^{2\beta d(x)} \frac{1}{N} \sum_{i=1}^{N} \left(J(i) - \frac{1}{N} \sum_{j=1}^{N} J(j) \right)^2 \tag{16}$$

where I_g is the gray-scale image and N is the number of pixels in the image. When $\beta = 0$, we have

$$\sigma_0^2 = \frac{1}{N} \sum_{i=1}^{N} \left(J(i) - \frac{1}{N} \sum_{j=1}^{N} J(j) \right)^2. \tag{17}$$

Combining (17) and (16) yields

$$\sigma_{I_g} = e^{-2\beta} \sigma_0^2. \tag{18}$$

After taking the logarithm of both sides, the scattering coefficient can be expressed as

$$\beta = \ln \sigma_0 - \ln \sigma_{I_g}. \tag{19}$$

Since σ is changed at each image, (19) can be expressed using the first-order Taylor series approximation

$$\beta = 1 + \ln \sigma_0 - \sigma_{I_g}. \tag{20}$$

When $\beta = 0$, the variance of the scene radiance approximates 1. Thus, we have

$$\beta = 1 - \ln \sigma_{I_g}. \tag{21}$$

Therefore, we can estimate the concentration using the standard deviation of the gray-scale image. The standard deviation is used for high-altitude data.

C. Concentration Estimation Results

For each available sensor deployment location, we speculatively remove one or more sensor's data and use estimation techniques with access to the remaining sensors to infer concentration(s), thereby allowing comparison with ground-truth measurements. We investigate the impact of the number of sensors on the estimation accuracy by using all possible combinations of speculatively removed sensors and averaging the results. We also consider the impact of using image data on estimation accuracy.
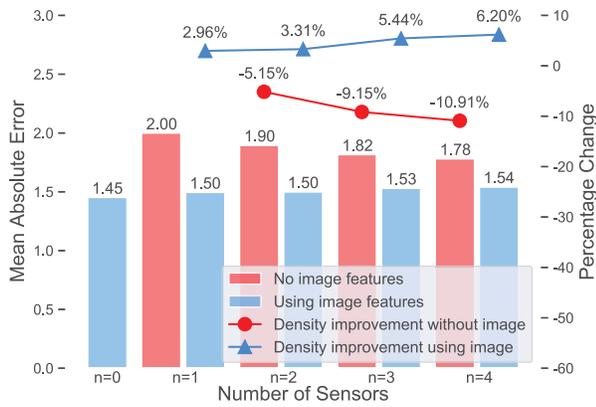
Fig. 10. Relationship between mean average error and sensor density for gradient boosting regression on high-altitude data.
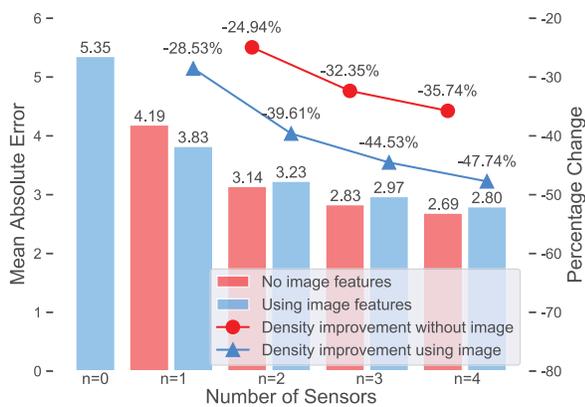
Fig. 11. Relationship between mean average error and sensor density for gradient boosting regression on low-altitude data.

Fig. 12. Relationship between mean average error and using images for gradient boosting regression on high-altitude data.

Fig. 13. Relationship between mean average error and using images for gradient boosting regression on low-altitude data.

We plotted the results from the best-performing algorithm: GBR. The density improvements in Figs. 10 and 11 are compared with the $n = 0$ (no images) case. If image features are used, the improvements are compared with the $n = 1$ case. Note that the lower MAE is, the better the estimation accuracy. Thus, a negative change indicates improvement. The bars in Figs. 10–13 represent the MAE resulting from using different numbers of sensors. The lower bars indicate higher accuracies. The lines in Figs. 10–13 indicate percentage change to MAE in different cases (e.g., using images or more sensors). The lower the percentage change, the larger
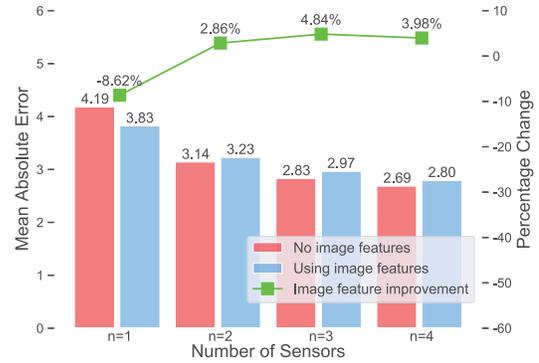
the improvement, and positive percentage changes imply (undesirable) increases in MAE.

As sensor density increases, the estimation MAE decreases. This implies that accuracy improves with sensor density, even at densities much higher than those of modern stationary sensor deployments. Moreover, as shown in Fig. 7, sensor correlations decrease with increasing distance. This is the reason estimation accuracy improves with increasing sensor density. Using the four nearest sensors instead of one sensor improves estimation accuracy by 23.3% on average without using images, and 20.75% when using images. The fact that increasing sensor density improves accuracy less when images are available does not imply that images are unhelpful. In contrast, it implies that using images allows higher accuracy when few sensors are available, leaving less potential for improvement if sensor density is later increased.

A4 vision-based techniques significantly improve estimation accuracy As shown in Figs. 12 and 13, PM2.5 concentration prediction accuracy improves when the images are used. In the case of $n = 0$, we average all the available concentrations. The MAE is 20.821 μg/m^3 for high-altitude data and 6.929 μg/m^3 for low-altitude data. The high-altitude images enable higher accuracy because some sensor locations are occluded in the low-altitude images. When images are used, MAE drops to 1.45 and 5.35 μg/m^3 respectively. For the case where $n = 1$, when we use the PM2.5 concentrations of the nearest available sensor for estimation image data improves prediction accuracy by 16.9% on average. The benefits of using images are greatest when the fewest particle counters are used.

To determine whether the improvement is systemic and statistically significant, we use the Kolmogorov–Smirnov test. We compare the data for the MAE without using images that using images for each number of sensors ($n = 1, \ldots, 4$) to determine whether the two datasets have the same distribution. The test is nonparametric and requires no knowledge about the distribution of data.

We determine that there is a statistically significant difference between the distribution of the MAE when using images and the distribution without using images, that is, the difference between the two distributions is not due only to chance or noise, but due to a genuine improvement in accuracy when using the image data. If the p-value is below 0.10, we can reject the null hypothesis and conclude

TABLE VIII
P-VALUES OF GBR ON HIGH-ALTITUDE DATA

GBR			
n=1	n=2	n=3	n=4
< 0.001	< 0.001	< 0.001	0.064

TABLE IX
P-VALUES OF GBR ON LOW-ALTITUDE DATA

GBR			
n=1	n=2	n=3	n=4
0.068	< 0.001	< 0.001	< 0.001

that using images does improve our results. As shown in Tables VIII and IX, on low-altitude data, the *p*-values of GBR in all cases are less than 0.10, and on high-altitude data, GBR's *p*-values are less than 0.10. As a result, we are confident that using images improves the estimates.

Certain fixed-location images show slightly negative results for the GBR method. Since the fixed-location images are taken from a low altitude, some of the sensor locations are blocked by buildings. For the quadcopter images taken at a higher altitude, all estimation techniques improve accuracy. In general, images decrease MAE by 8.44% on average, when $n \leqslant 1$, adding a camera to collect images helps more than adding more sensors.

Increasing sensor density and using images change the relative accuracy of the estimation algorithms. When we use only one sensor and no images, RFR has lower MAE than GBR on low-altitude data. When the sensor density is increased ($n = 4$) and images are used, GBR outperforms RFR. This result demonstrates that it is important to evaluate estimation algorithms using appropriate sensor densities and access to image data. A sparsely deployed and less accurate sensor network can lead to false conclusions about pollution concentrations, and about which pollution concentration estimation algorithms are most accurate.

To summarize, both higher sensor densities and image data improve estimation accuracy, and adding image data has a similar effect to increasing particle counter density by 0.61 sensors km^{-2}. Of the three estimation techniques evaluated, GBR had the highest accuracy with MAE = 1.45 μg/m^3. In particular, our developed method is unlikely to work as well on night-time images and images with adverse weather conditions such as rain and snow. The main limitation of our dataset is that it does not contain night-time images.

VI. CONCLUSION

This article presents a PM dataset with high spatial and temporal resolutions. In contrast with existing datasets, HVAQ contains images covering the locations of stationary point sensors, making it suitable for evaluating and validating vision-based pollution estimation algorithms. Through our analysis, we find that: 1) the estimation accuracy can be improved significantly using vision-based techniques; 2) the spatial pollutant distribution is spatially correlated; 3) spatial variation of PM2.5 is high; and 4) temperature and humidity had limited impact on PM concentration in our dataset. We also evaluate our data using state-of-the-art prediction methods. Accuracy correlates with density with a coefficient of 0.2875 μg/m^3 MAE per sensor and vision-based estimation improves accuracy by 0.1813 μg/m^3 MAE, on average.

REFERENCES

[1] K. Naddafi *et al.*, "Health impact assessment of air pollution in megacity of Tehran, Iran," *Iranian J. Environ. Health Sci. Eng.*, vol. 9, no. 1, p. 28, Dec. 2012.

[2] H. Nourmoradi *et al.*, "Air quality and health risks associated with exposure to particulate matter: A cross-sectional study in Khorramabad, Iran," *Health Scope*, vol. 5, no. 2, Jan. 2016, Art. no. e31766.

[3] S. Poduri, A. Nimkar, and G. S. Sukhatme, "Visibility monitoring using mobile phones," in *Proc. Annu. Rep., Center Embedded Netw. Sens.*, 2010, pp. 125–127.

[4] B. Chen and H. Kan, "Air pollution and population health: A global challenge," *Environ. Health Preventive Med.*, vol. 13, no. 2, pp. 94–101, Mar. 2008.

[5] F. J. Kelly and J. C. Fussell, "Air pollution and public health: Emerging hazards and improved understanding of risk," *Environ. Geochem. Health*, vol. 37, no. 4, pp. 631–649, 2015.

[6] S. Feng, D. Gao, F. Liao, F. Zhou, and X. Wang, "The health effects of ambient PM$_{2.5}$ and potential mechanisms," *Ecotoxicol. Environ. Saf.*, vol. 128, pp. 67–74, Jun. 2016.

[7] J. O. Anderson, J. G. Thundiyil, and A. Stolbach, "Clearing the air: A review of the effects of particulate matter air pollution on human health," *J. Med. Toxicol.*, vol. 8, no. 2, pp. 166–175, Jun. 2012.

[8] R. A. Rohde and R. A. Müller, "Air pollution in China: Mapping of concentrations and sources," *PLoS ONE*, vol. 10, no. 8, Aug. 2015, Art. no. e0135749.

[9] Y. Zhang, J. Cai, S. Wang, K. He, and M. Zheng, "Review of receptor-based source apportionment research of fine particulate matter and its challenges in China," *Sci. Total Environ.*, vol. 586, pp. 917–929, May 2017.

[10] G. Lin *et al.*, "Spatio-temporal variation of PM$_{2.5}$ concentrations and their relationship with geographic and socioeconomic factors in China," *Int. J. Environ. Res. Public Health*, vol. 11, no. 1, pp. 173–186, Dec. 2013.

[11] X. Zhao, W. Zhou, L. Han, and D. Locke, "Spatiotemporal variation in PM$_{2.5}$ concentrations and their relationship with socioeconomic factors in China's major cities," *Environ. Int.*, vol. 133, Dec. 2019, Art. no. 105145.

[12] 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. Int. Conf. Multimedia*, 2016, pp. 297–301.

[13] T. Zhang and R. P. Dick, "Estimation of multiple atmospheric pollutants through image analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2060–2064.

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

[15] 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.

[16] J. S. Apte *et al.*, "High-resolution air pollution mapping with Google street view cars: Exploiting big data," *Environ. Sci. Technol.*, vol. 51, no. 12, pp. 6999–7008, 2017.

[17] Y. Guan *et al.*, "Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google street view vehicles," *J. Amer. Stat. Assoc.*, vol. 115, no. 531, pp. 1111–1124, Jul. 2020.

[18] J. Wei *et al.*, "The ChinaHighPM$_{10}$ dataset: Generation, validation, and spatiotemporal variations from 2015 to 2019 across China," *Environ. Int.*, vol. 146, Jan. 2021, Art. no. 106290.

[19] J. Khan *et al.*, "Development and performance evaluation of new AirGIS—A GIS based air pollution and human exposure modelling system," *Atmos. Environ.*, vol. 198, pp. 102–121, Feb. 2019.

[20] C. Wen *et al.*, "A novel spatiotemporal convolutional long short-term neural network for air pollution prediction," *Sci. Total Environ.*, vol. 654, pp. 1091–1099, Mar. 2019.

[21] G. Janssens-Maenhout *et al.*, "EDGAR-HTAP: A harmonized gridded air pollution emission dataset based on national inventories," Eur. Commission, Ispra, Italy, Tech. Rep. JRC68434, 2012, p. 299.

[22] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. D. Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sens. Actuators B, Chem.*, vol. 129, no. 2, pp. 750–757, 2008.

[23] J. J. Li, B. Faltings, O. Saukh, D. Hasenfratz, and J. Beutel, "Sensing the air we breathe—The OpenSense Zurich dataset," in *Proc. AAAI Conf. Artif. Intell.*, 2012, pp. 323–325.

[24] Y. Yang, Z. Hu, K. Bian, and L. Song, "ImgSensingNet: UAV vision guided aerial-ground air quality sensing system," in *Proc. IEEE Conf. Comput. Commun.*, Apr. 2019, pp. 1207–1215.

[25] P. G. Herzog and B. Hill, "Multispectral imaging and its applications in the textile industry and related fields," in *Proc. PICS*, 2003, pp. 258–263.

[26] F. Martínez-Verdú, J. Pujol, and P. Capilla, "Characterization of a digital camera as an absolute tristimulus colorimeter," *J. Imag. Sci. Technol.*, vol. 47, no. 4, pp. 279–295, 2003.

[27] A. M. Nahavandi, "Metric for evaluation of filter efficiency in spectral cameras," *Appl. Opt.*, vol. 55, no. 32, pp. 9193–9204, 2016.

[28] N. Rijal, R. T. Gutta, T. Cao, J. Lin, Q. Bo, and J. Zhang, "Ensemble of deep neural networks for estimating particulate matter from images," in *Proc. IEEE 3rd Int. Conf. Image, Vis. Comput.*, Jun. 2018, pp. 733–738.

[29] K. Gu, J. Qiao, and X. Li, "Highly efficient picture-based prediction of $PM_{2.5}$ concentration," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3176–3184, Apr. 2019.

[30] H. Grimm and D. J. Eatough, "Aerosol measurement: The use of optical light scattering for the determination of particulate size distribution, and particulate mass, including the semi-volatile fraction," *J. Air Waste Manage. Assoc.*, vol. 59, no. 1, pp. 101–107, 2009.

[31] M. Kamionka, P. Breuil, and C. Pijolat, "Calibration of a multivariate gas sensing device for atmospheric pollution measurement," *Sens. Actuators, B, Chem.*, vol. 118, nos. 1–2, pp. 323–327, 2006.

[32] G. Liu, J. Li, D. Wu, and H. Xu, "Chemical composition and source apportionment of the ambient $PM_{2.5}$ in Hangzhou, China," *Particuology*, vol. 18, pp. 135–143, Feb. 2015.

[33] Q. Jin, R. Ren, and L. Gong, "Research on the elemental characterization and source apportionment of $PM_{2.5}$ in main urban area of Hangzhou," *Chin. J. Health Lab. Technol.*, vol. 27, no. 22, pp. 3200–3205, 2017.

[34] J. Wu, C. Xu, Q. Wang, and W. Cheng, "Potential sources and formations of the $PM_{2.5}$ pollution in Urban Hangzhou," *Atmosphere*, vol. 7, no. 8, p. 100, Jul. 2016.

[35] Q. Zhu, Y. Liu, W. Xu, and M. Huang, "Analysis on the pollution characteristics and influence factors of $PM_{2.5}$ in Guangzhou," *Environ. Monit. China*, vol. 29, no. 2, p. 15, 2013.

[36] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *Int. J. Comput. Vis.*, vol. 48, no. 3, pp. 233–254, 2002.

[37] 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.

[38] D. Park, D. K. Han, C. Jeon, and H. Ko, "Fast single image de-hazing using characteristics of RGB channel of foggy image," *IEICE Trans. Inf. Syst.*, vol. E96.D, no. 8, pp. 1793–1799, 2013.
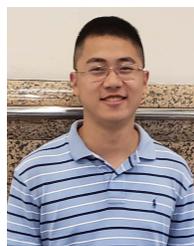
Tony Zhang (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from the University of Maryland, College Park, MD, USA, in 2017, and the M.S. degree in electrical and computer engineering from the University of Michigan, Ann Arbor, MI, USA, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering.

His research interests include machine learning, computer vision, air quality, and data science.

Zhuangzhi Chen received the B.S. degree in electrical engineering and automation from the Zhejiang University of Technology, Hangzhou, China, in 2017, where he is currently pursuing the Ph.D. degree in control theory and engineering.

His current research interests include computer vision, deep learning method including deep convolutional neural networks, and generative adversarial networks.

Yun Xiang (Member, IEEE) received the B.S. degree from Zhejiang University, Hangzhou, China, in 2008, and the Ph.D. degree from the University of Michigan, Ann Arbor, MI, USA, in 2014.

He is currently an Associate Professor with the Institute of Cyberspace Security, College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His current research interests include cyberspace security, machine learning, embedded systems, and computer vision.

Qi Xuan (Member, IEEE) received the B.S. and Ph.D. degrees in control theory and engineering from Zhejiang University, Hangzhou, China, in 2003 and 2008, respectively.

He was a Post-Doctoral Researcher with the Department of Information Science and Electronic Engineering, Zhejiang University, from 2008 to 2010, respectively, and a Research Assistant with the Department of Electronic Engineering, City University of Hong Kong, in 2010 and 2017. From 2012 to 2014, he was a Post-Doctoral Fellow with the Department of Computer Science, University of California at Davis, Davis, CA, USA. He is currently a Professor with the Institute of Cyberspace Security, College of Information Engineering, Zhejiang University of Technology, Hangzhou. His current research interests include network science, graph data mining, cyberspace security, machine learning, and computer vision.

Robert P. Dick (Senior Member, IEEE) received the B.S. degree from Clarkson University, Potsdam, NY, USA, in 1996, and the Ph.D. degree from Princeton University, Princeton, NJ, USA, in 2002.

He worked as a Visiting Professor with the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2002, as a Visiting Researcher with NEC Labs America, Princeton, in 1999, and the Faculty of Northwestern University, Evanston, IL, USA, from 2003 to 2008. He is currently an Associate Professor of electrical engineering and computer science with the University of Michigan, Ann Arbor, MI, USA.

Dr. Dick received an NSF CAREER Award and won his department's Best Teacher of the Year Award in 2004. In 2007, his technology won a Computerworld Horizon Award and his article was selected as one of the 30 in a special collection of DATE Papers appearing during the past ten years. His 2010 work won a Best Paper Award at DATE. He served as the Technical Program Committee Co-Chair for the 2011 International Conference on Hardware/Software Codesign and System Synthesis, as an Associate Editor of the IEEE TRANSACTIONS ON VLSI SYSTEMS, and as a Guest Editor for *ACM Transactions on Embedded Computing Systems*, and the IEEE DESIGN AND TEST OF COMPUTERS. He was also CEO of Stryd, Inc., which produces wearable electronics for athletes.

Zuohui Chen received the B.S. degree in automation from the Zhejiang University of Technology, Hangzhou, China, in 2019, where he is currently pursuing the M.S. degree in control theory and engineering.

His current research interests include computer vision, AI application, and AI security.