ImgSensingNet: UAV Vision Guided Aerial-Ground Air Quality Sensing System

Yuzhe Yang, *Student Member, IEEE,* Zhiwen Hu, *Student Member, IEEE,*
Kaigui Bian, *Member, IEEE,* and Lingyang Song, *Fellow, IEEE*

***Abstract*—Given the increasingly serious air pollution problem, the monitoring of air quality index (AQI) in urban areas has drawn considerable attention. This paper presents *ImgSensingNet*, a vision guided aerial-ground sensing system, for fine-grained air quality monitoring and forecasting using the fusion of haze images taken by the unmanned-aerial-vehicle (UAV) and the AQI data collected by an on-ground three-dimensional (3D) wireless sensor network (WSN). Specifically, *ImgSensingNet* first leverages the computer vision technique to tell the AQI scale in different regions from the taken haze images, where haze-relevant features and a deep convolutional neural network (CNN) are designed for direct learning between haze images and corresponding AQI scale. Based on the learnt AQI scale, *ImgSensingNet* determines whether to wake up on-ground wireless sensors for small-scale AQI monitoring and inference, which can greatly reduce the energy consumption of the system. An entropy-based model is employed for accurate real-time AQI inference at unmeasured locations and future air quality distribution forecasting. We implement and evaluate *ImgSensingNet* on two university campuses since Feb. 2018, and has collected 17,630 photos and 2.6 millions of AQI data samples. Experimental results confirm that *ImgSensingNet* can achieve higher inference accuracy while greatly reduce the energy consumption, compared to state-of-the-art AQI monitoring approaches.**

***Index Terms*—Mobile sensing, air quality index, computer vision, sensor networks, unmanned aerial vehicle (UAV).**

I. Introduction

Air pollution has been proved to have significantly negative effects on human health and sustainable development [1]. Air pollution is caused by gaseous pollutants that are harmful to humans and ecosystem. To quantify the degree of air pollution, government agencies have defined the air quality index (AQI). AQI is calculated based on the concentration of a number of air pollutants, such as $\mathrm{PM}_{2.5}$ and PM_{10} particles. A higher AQI indicates that air pollution is more severe and people are more likely to experience harmful health effects [2]. Thus, AQI monitoring is a critical issue. The more accurate AQI distribution that can be obtained in a region, the more effective methods we can find to deal with the air pollution.

Existing AQI monitoring approaches can be classified into two categories. The first category includes the *sensor-based monitoring* approaches, wherein government agencies have set up monitoring stations on dedicated sites in a city [3]. However, these fixed stations only provide coarse-grained 2D monitoring, with several kilometers between two monitoring stations. Existing study has shown that AQI distribution has intrinsic variation within meters [4]. Large scale Internet-of-Things (IoT) applications have been developed to monitor the fine-grained air quality using densely deployed sensors [6], [7]. Although the static sensors may achieve the high precision of monitoring, they suffer from the high cost as well as lack of mobility. Mobile devices or vehicles, such as phones, cars, balloons are utilized to carry sensors for AQI monitoring [9]–[12]. However, the sensor-based approach may induce high energy consumptions for mobile devices or vehicles to acquire certain amount of data.

The second category of approaches includes the *vision-based monitoring*. Image-based AQI monitoring stations are set up by researchers at dedicated locations [16], and these static stations can only take photos and infer the AQI at limited sites over the whole region. Crowd-sourced photos contributed by mobile phones can depict the AQI distribution [15] at more locations. However, the performance of the crowd sourcing approach is usually restricted by the low quality photos contributed by many non-savvy users.

Previous works have separated the two categories of methods in AQI monitoring; however, *sensor-based* and *vision-based* methods can be combined to promote the performance of the mobile sensing system, while reducing the power consumption. For example, the combination of computer vision and inertial sensing has been proved to be successful in the task of localization and navigation by phones [17], [18]. In this work, we seek a way of leveraging both photo-taking and data sensing to monitor and infer the AQI value.

In this paper, we present *ImgSensingNet*, a UAV vision guided aerial-ground air quality sensing system, to monitor and forecast AQI distributions in spatial-temporal perspectives. Unlike existing systems, we implement: (1) mobile vision-based sensing over an unmanned-aerial-vehicle (UAV), which realizes three-dimensional (3D) AQI monitoring by UAV photo-taking instead of using particle sensors, to infer region-level AQI scale (an interval of possible AQI values) by applying a deep convolutional neural network (CNN) over the taken hazy photos; (2) ground sensing over a wireless sensor network (WSN) for small-scale accurate spatial-temporal AQI inference, using an entropy-based inference model; (3) an energy-efficient wake-up mechanism that powers on the sensors in a region when small-scale monitoring is needed in that region, based on the result of vision-based AQI inference, which greatly reduces energy consumption while maintaining high inference accuracy. We implement and evaluate ImgSens-

Y. Yang is with the Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139 (email: yuzhe@mit.edu).

Z. Hu, K. Bian and L. Song are with School of Electrical Engineering and Computer Science, Peking University, Beijing, China (email: {zhiwen.hu, bkg, lingyang.song}@pku.edu.cn).

TABLE I
COMPARISON OF STATE-OF-THE-ART AIR QUALITY MONITORING APPROACHES

Systems	Scale	Dimension	Monitoring Methods	Resolution	Mobility	Costs	Real-Time Capability	Accuracy
Official stations [3]	∼ 100 km	2-D	Sensor	Low	Static	High	No	Low
AirCloud [6]	∼ 5 km	2-D	Sensor	Medium	Static	Low	No	Medium
Mosaic [7]	∼ 5 km	2-D	Sensor	Medium	Mobile	Low	No	Medium
Mobile nodes [8]	1 km	2-D	Sensor	Medium	Mobile	Medium	Yes	Medium
Balloons [9]	1 km	3-D	Sensor	High	Mobile	Medium	No	Low
BlueAer [10]	∼ 10 km	3-D	Sensor	Medium	Static+Mobile	High	No	High
ARMS [14]	100 m	3-D	Sensor	High	Mobile	High	Yes	High
AQNet [12]	∼ 2 km	3-D	Sensor	High	Static+Mobile	Low	Yes	High
Cell phones [15]	4 km	2-D	Vision	Low	Mobile	Low	Yes	Medium
IBAQMS [16]	∼ 1 km	2-D	Vision	Low	Static	Medium	No	Medium
ImgSensingNet	**∼ 10 km**	**3-D**	**Sensor+Vision**	**High**	**Static+Mobile**	**Low**	**Yes**	**High**

ingNet on two university campuses (i.e., Peking University and Xidian University) since Feb. 2018. We have collected 17,630 photos and 2.6 millions of data samples. Compared to state-of-the-art methods, evaluation results confirm that ImgSensingNet can save the energy consumptions by 50.5% while achieving an accuracy of 95.2% for inference.

The main contributions are summarized as below.

- We implement ImgSensingNet, a UAV vision guided aerial-ground AQI sensing system, and we deploy and evaluate it in the real-world testbed;
- The proposed vision-based sensing method can learn the direct correlation between raw haze images and corresponding AQI scale distribution;
- The proposed entropy-based inference model for ground WSN can achieve a high accuracy in both real-time AQI distribution estimation and future AQI prediction;
- The wake-up mechanism connects the aerial vision technique with the on-ground WSN, which can greatly save the energy consumptions of the on-ground sensor network while ensuring high inference and prediction accuracy

The rest of this paper goes as follows. Related works are introduced in Section II. In Section III, we present the system overview of ImgSensingNet. Section IV introduces the UAV vision-based aerial sensing. In Section V, we propose the AQI inference model for ground WSN. Section VI introduces the energy-efficient wake-up mechanism. In Section VII, we detail the system implementation. Experimental results and conclusions are provided in Section VIII and Section IX.

II. TAXONOMY

A. AQI Monitoring Methods

In Table I, we show state-of-the-art works on air quality monitoring systems. Existing AQI monitoring methods can be summarized into two categories.

Sensor-based: Stationary stations [2] are set up on dedicated sites in a city, but only provide a limited number of measurement samples. For example, there are only 28 monitoring stations in Beijing. The distance between two nearby stations is typically several ten-thousand meters, and the AQI is monitored every 2 hours [3]. AirCloud [6] uses densely distributed sensors in a static way, while Mosaic [7] and [8], [9] adopt mobile devices such as buses or balloons to carry low-cost sensors. However, they all fail to consider the heterogeneous 3D AQI distribution. In [10]–[13], drones with sensors together with ground sensors are used for AQI profiling. However, they are either restricted in a small scale region or may induce high costs, without designing energy-efficient schemes for integrating aerial sensing with ground sensing.

Vision-based: Instead of various particle sensors, image-based approaches are also used for AQI estimation. In [16], image-based air quality monitoring stations are set up at dedicated sites over a city. Again, these methods can only profile AQI at a limited number of locations. In [15], camera-enabled mobile devices are used for generating crowd-sourced photos for AQI monitoring. However, the incentive to stimulate users for volunteer high-quality photo-taking is the pain point for such a crowd-sourced system. Without precise correlations between haze images and AQI values, they cannot generalize well and may introduce low accuracy.

ImgSensingNet overcomes the above shortcomings by using vision guided aerial sensing to extend sensing scope, while also combining it with ground WSN for accurate AQI distribution inference. An energy-efficient wake-up mechanism is designed to switch on or off the on-ground WSN by examining the aerial sensing results, which greatly lowers the systems energy consumption.

B. AQI Inference at Unmeasured Locations

In real-world sensing applications, it is not feasible to acquire AQI data samples at all locations within a region. Hence, AQI modeling and inference are used to estimate AQI distributions at unmeasured locations. Again, the inference models can be summarized into two categories.

Inference by sensor data: Zheng et al. [5] propose to infer air quality based on data from official air quality stations and other features such as the meteorological data. In [6], [7], crowd-sourcing and a Gaussian process model are used for 2D AQI inference. [10] extends the inference to 3D space by using a random walk model. A fine-grained AQI distribution model is proposed in [14] for real-time AQI estimation over a 3D space. Long-short term memory (LSTM) networks are used in [19] to utilize historical data for more accurate inference. To do the temporal inference, neural networks (NN) are used [12] to analyze spatial-temporal correlations and to forecast future distribution.

Inference by image data: Image-based inference has been used to estimate AQI from haze images by designing appropriate inference models. Classical image processing methods as well as machine learning techniques are used in [15], [16] to model the correlation between haze images and the degree of air pollution.

In this work, we investigate two novel inference models: (1) image-based AQI scale inference in different monitoring regions by computer vision, and (2) the fine-grained spatial-temporal AQI value inference at locations inside each region by ground WSN.

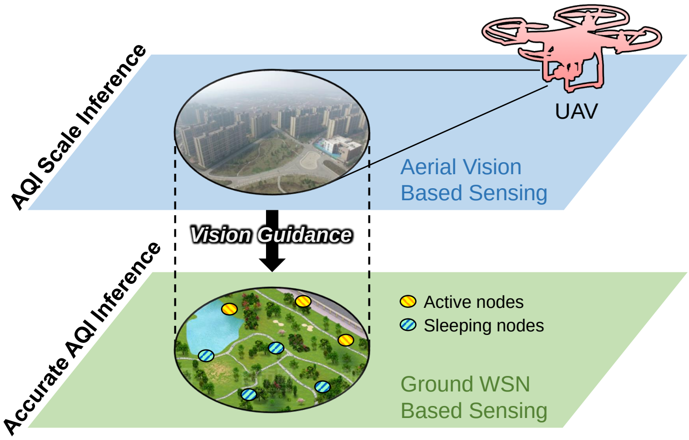


Fig. 1. The overall framework of *ImgSensingNet*.

III. System Overview

The ImgSensingNet system includes on-ground programmable monitoring devices and a UAV. The aerial UAV sensing and the ground WSN sensing form a hybrid sensing network, as illustrated in Fig. 1.

The central idea of ImgSensingNet is to trigger aerial sensing and ground sensing sequentially during one measurement, which can provide coarse-to-fine grained AQI value inference. This operation can not only achieve high accuracy, but also scale down the monitoring overhead, which can guarantee a long battery duration without external power supply.

A. Aerial Sensing

Fig. 1 shows the overall framework of ImgSensingNet. The aerial sensing utilizes the UAV camera to capture a series of haze images in different monitoring regions. The raw image data is streamed back to the central server, where a well-trained deep learning model performs real-time image data analysis and output the inferred AQI scale for each region.

B. Ground Sensing

Ground WSN adopts a spatial-temporal inference model for AQI estimation at unmeasured locations and future air quality prediction. Every time when aerial sensing is finished, each ground device follows a designed wake-up mechanism to decide whether to wake up for data sensing based on both the inference result at last time and the aerial sensing result. In this way, the real-time fine-grained AQI distribution is obtained and the future distribution can also be forecasted.

IV. Aerial Sensing: Learning AQI Scale from Images Captured by UAV

ImgSensingNet performs vision-based sensing using UAV, because: (1) the UAV has intrinsic advantages in flexible 3D space sensing over different heights and angles, which avoids possible obstacles, and also guarantees certain scene depths; (2) with built-in camera, the UAV does not need to carry extra sensors, which enables longer monitoring time; and (3) instead of hovering at different locations to collect data by sensors, the UAV can keep flying and video recording by cameras through monitoring regions, which greatly extends the sensing scope.

Recent works have well studied how to remove haze from images in the computer vision field [21], [23]–[25]. However, there has not been works on quantifying the haze in the image to real AQI value. To do the direct learning from raw haze images to quantified AQI values, two main problems should be answered: *(1) how to extract the haze components from origin images to eliminate the influence of image content*, and *(2) how to quantify the AQI based on the haze components*.

This section details the method to solve these problems. Specifically, we investigate content-nonspecific haze-relevant features for raw haze images. With the haze features extracted, a novel 3D CNN model is designed to better process feature maps and output the inferred AQI scale for each single image.

A. Overview of Haze Image Processing

In image processing, a haze image can be mathematically described using the haze image formation model [21] as

$$\boldsymbol{\mathcal{I}}(\boldsymbol{x}) = \boldsymbol{\mathcal{J}}(\boldsymbol{x})\, t(\boldsymbol{x}) + \boldsymbol{L}_\infty (1 - t(\boldsymbol{x})), \tag{1}$$

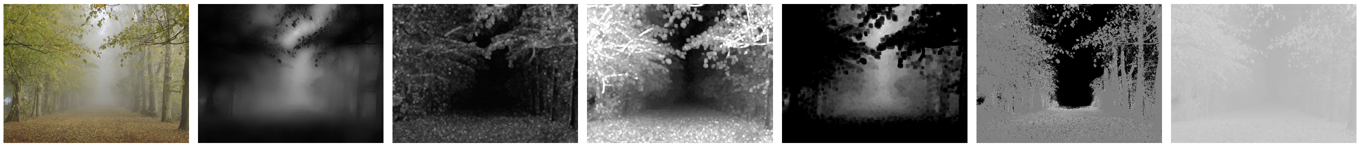
where $\boldsymbol{\mathcal{I}}$ is the observed hazy image, $\boldsymbol{\mathcal{J}}$ is the haze-free image, t denotes the medium transmission, $\boldsymbol{L}_\infty$ is the global atmospheric light, and $\boldsymbol{x}$ represents pixel coordinates. The haze-removal methods have spent large effort estimating $\boldsymbol{\mathcal{J}}$ and t for haze-free image recovery [21], [23]–[25].

Instead, in this work we propose a new objective to estimate the degree of haze in a single image.

B. Haze-relevant Features Extraction

The first step is to extract a list of haze-relevant statistical features. Since we want to investigate general approach for all image inputs regardless of their contents, the features that correlate well with haze density in images but do not correlate well with image contents should be selected.

In the following, we investigate six content-nonspecific haze-relevant features, and an example is illustrated in Fig. 2.



(a) Origin image (b) F_1 (c) F_2 (d) F_3 (e) F_4 (f) F_5 (g) F_6

Fig. 2. An example of the extracted feature maps for raw haze image: (a) origin input image; (b) refined dark channel; (c) max local contrast; (d) max local saturation; (e) min local color attenuation; (f) hue disparity; (g) chroma.

1) Refined Dark Channel: Dark channel [21] is an informative feature for haze detection, defined as the minimum of all pixel colors in a local patch:

$$D\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) = \min_{\boldsymbol{y}\in\Omega(\boldsymbol{x})}\left(\min_{c\in\{r,g,b\}}\frac{\mathcal{I}^c\left(\boldsymbol{y}\right)}{L_\infty^c}\right), \tag{2}$$

where $\Omega\left(\boldsymbol{x}\right)$ is a local patch centered at $\boldsymbol{x}$, $\mathcal{I}^c$ is one color channel of $\boldsymbol{\mathcal{I}}$. It is found that most local patches in outdoor haze-free images contain some pixels whose intensity is very low in at least one color channel [21]. Therefore, the dark channel is a rough approximation of the thickness of the haze.

To obtain a better estimation of haze density, we propose the *refined dark channel* by applying the guided filter [22] $\mathcal{G}$ on the estimated medium transmission $\tilde{t}$, to capture the sharp edge discontinuous and outline the haze profile. Note that by applying the min operation on (1), the dark channel of $\boldsymbol{\mathcal{J}}$ tends to be zero, and we have $\tilde{t}(\boldsymbol{x}) = 1 - D\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right)$. Hence, the refined dark channel can be expressed as

$$D^R\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) = 1 - \mathcal{G}\left(1 - \min_{\boldsymbol{y}\in\Omega(\boldsymbol{x})}\left(\min_{c}\frac{\mathcal{I}^c\left(\boldsymbol{y}\right)}{L_\infty^c}\right)\right). \tag{3}$$

Fig. 2(b) shows the refined dark channel feature. As we can see, the feature has a high correlation to the amount of haze in the image.

2) Max Local Contrast: Since haze can scatter the light reaching cameras, the contrast of the haze image can be highly reduced. Therefore, the contrast is one of the most perceived features to detect haze in the scene. The local contrast is defined as the variance of pixel intensities in a local $r \times r$ region compared with the center pixel. Inspired by [23], we further use the local maximum of local contrast values in a local patch $\Omega\left(\boldsymbol{x}\right)$ to form the *max local contrast* feature as

$$C^T\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) = \max_{\boldsymbol{y}\in\Omega(\boldsymbol{x})}\sqrt{\frac{\sum_{\boldsymbol{z}\in\Omega_r(\boldsymbol{y})}\left\|\boldsymbol{\mathcal{I}}\left(\boldsymbol{z}\right)-\boldsymbol{\mathcal{I}}\left(\boldsymbol{y}\right)\right\|^2}{\pi\left|\Omega_r\left(\boldsymbol{y}\right)\right|}}, \tag{4}$$

where $\left|\Omega_r\left(\boldsymbol{y}\right)\right|$ denotes the size of the local region $\Omega_r\left(\boldsymbol{y}\right)$, and π is a constant that equals to the number of channels. Fig. 2(c) shows the contrast feature, in which the correlation between haze and the contrast feature are visually obvious.

3) Max Local Saturation: It is observed that the image saturation varies sharply with the change of haze in the scene [16]. Therefore, similar to image contrast, we define the *max local saturation* feature that represents the maximum saturation value of pixels within a local patch, written as

$$S\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) = \max_{\boldsymbol{y}\in\Omega(\boldsymbol{x})}\left(1 - \frac{\min_c \mathcal{I}^c\left(\boldsymbol{y}\right)}{\max_c \mathcal{I}^c\left(\boldsymbol{y}\right)}\right). \tag{5}$$

The max local saturation feature for the "forest" image is shown in Fig. 2(d), which is also correlated with the haze.

4) Min Local Color Attenuation: In [24], the scene depth is found to be positively correlated with the difference between the image brightness and the image saturation by numerous experiments on haze images. This statistics is regarded as the color attenuation prior, expressed as

$$d\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) = \theta_0 + \theta_1 \cdot \boldsymbol{\mathcal{I}}^v\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) + \theta_2 \cdot \boldsymbol{\mathcal{I}}^s\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) + \epsilon\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right), \tag{6}$$

where $\boldsymbol{\mathcal{I}}^v$ and $\boldsymbol{\mathcal{I}}^s$ denote the brightness and the saturation, respectively. Let $\epsilon(\boldsymbol{x}) \sim \mathcal{N}(0, \sigma^2)$, and θ_0, θ_1, θ_2, σ can be estimated through maximum likelihood. To process the raw depth map for better representation of the haze influence, we define the *min local color attenuation* feature by considering the minimum pixel-wise depth within a local patch $\Omega\left(\boldsymbol{x}\right)$:

$$A\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) = \min_{\boldsymbol{y}\in\Omega(\boldsymbol{x})} d\left(\boldsymbol{y};\boldsymbol{\mathcal{I}}\right). \tag{7}$$

Fig. 2(e) shows the min local color attenuation feature, where an obvious correlation with haze density can be observed.

5) Hue Disparity: In [25], the hue disparity between the original image and its semi-inverse image is utilized to remove haze. The semi-inverse image is defined as the max value between original image and its inverse, expressed as

$$\mathcal{I}_{si}^c\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) = \max_{\boldsymbol{x}\in\boldsymbol{\mathcal{I}}}\left\{\mathcal{I}^c\left(\boldsymbol{x}\right), 1-\mathcal{I}^c\left(\boldsymbol{x}\right)\right\}, \quad c\in\{r,g,b\}. \tag{8}$$

The *hue disparity* is also reduced by haze, thus can serve as another haze-relevant feature, written as

$$H\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) = \left|\boldsymbol{\mathcal{I}}_{si}^h\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) - \boldsymbol{\mathcal{I}}^h\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right)\right|, \tag{9}$$

where $\boldsymbol{\mathcal{I}}^h$ denotes the hue channel of the image. Fig. 2(f) shows the hue disparity feature for haze image.

6) Chroma: In the CIELab color space, the chroma is one of the most representative image feature to describe the color degradation by the haze in the atmosphere. Let $\left[L\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right)\ a\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right)\ b\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right)\right]^{\mathrm{T}}$ denotes the haze image $\boldsymbol{\mathcal{I}}$ in the CIELab space, the $chroma$ feature is defined as

$$C^H\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) = \sqrt{a^2\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right) + b^2\left(\boldsymbol{x};\boldsymbol{\mathcal{I}}\right)}. \tag{10}$$

As shown in Fig. 2(g), chroma is an excellent haze-relevant feature since it strongly correlates with the haze density but is not affected by the image contents.

C. 3D CNN-based Learning for AQI Scale Inference

With the above haze-relevant features extracted, we design a 3D CNN model to perform direct learning for precise AQI scale estimation of input haze images. CNN is a type of deep learning model in which trainable filters and local neighborhood pooling operations are applied alternatively on the raw input images, resulting in a hierarchy of increasingly

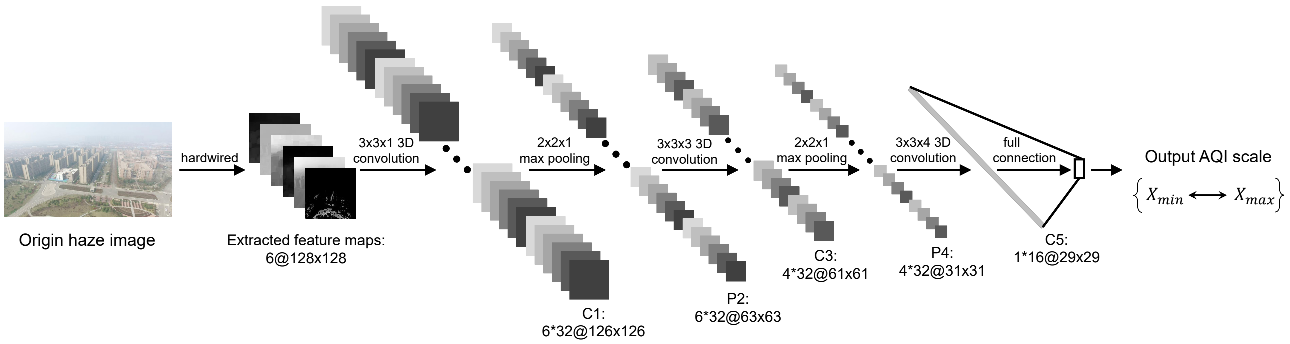


Fig. 3. The architecture of the proposed 3D CNN model.

complex features. CNN has been widely used for image processing and vision applications, and has been proved to achieve superior performance compared to classical methods.

In this work, to better fit the extracted features for high accuracy, we introduce a 3D CNN model by adding a "prior feature map" dimension. The advantage behind 3D convolution is the utilization of haze prior information, which is encoded in the six feature maps.

Preprocessing: For each input haze image, we first resize it spatially to 128×128 pixels. The resized image is then performed with feature maps extraction and rescaled into $[0,\ 1]$ in grayscale. We normalize each dimension except the prior feature map dimension of all training haze images to be of zero mean, which can help our model converge faster.

Model Architecture: Fig. 3 presents the architecture of the 3D CNN model. The first layer is called the "hardwired" layer that extracts feature maps from original haze image, consisting of six feature frames stacked together to be a $128 \times 128 \times 6$ sized tensor. The rationale for using this hardwired layer is to encode our prior knowledge on different haze-relevant features. This scheme regularizes the model training constrained in the prior haze feature space, which leads to better performance compared to random initialization.

We then apply 3D convolutions with a kernel size of $3\times3\times1$ and 32 kernels, to extract complex features in different feature map domains separately. In the subsequent pooling layer, $2 \times 2 \times 1$ max pooling is applied. The next convolution layer uses $3\times3\times3$ kernel size, followed by another $2\times2\times1$ max pooling. 3D convolution with $3 \times 3 \times 4$ kernel size is then applied and it contains 13,456 trainable parameters. Finally, the vector is densely connected to the output layer that consists pre-divided AQI scale classes. This architecture has been verified to give the best performance compared to other 3D CNN architectures.

Training and AQI Scale Inference: As the output is AQI scale (i.e., $[X_{min},\ X_{max}]$), the inference is modeled as a classification problem, where the AQI scale classes are pre-divided based on the number of different AQI values in training data. Given new image input, the model finds images in training set with most similar haze degrees, and uses the corresponding AQI ground truth values to generate an AQI scale. With more data of different AQI values collected, the number of class will increase, resulting in more fine-grained scale labels.

V. Ground Sensing: AQI Inference by Ground Sensor Monitoring

Given the 3D target monitoring space, we utilize ground WSN for accurate AQI inference that enables both the real-time inference spatially, and future distribution forecasting temporally. This section illustrates how to do accurate inference based on (1) sparse historical ground WSN data, and (2) the prior AQI scale knowledge by aerial sensing.

The target 3D space is first divided into disjointed cubes, which form the basic unit in our inference. Each cube contains its own geographical coordinates in 3D space, and each cube is associated with an AQI value. Note that AQI values in a limited number of cubes are observed/sensed from the WSN, while the AQI values in other unobserved cubes need to be estimated using the proposed model. Here we define a set of cubes $\{C_1, C_2, \ldots, C_s\}$ over a series of time stamps $\{T_1, T_2, \ldots, T_d\}$ with equal intervals (e.g., one hour). Most cubes do not have observed/sensed data (e.g., $\geq 99\%$ in both Peking University and Xidian University), whose AQI values can be estimated using a probability function, p_u. The objective is to infer p_u of any unobserved location C at any given time stamp T_i (including both the current and future time stamps).

Why a semi-supervised learning model: Since the data observed using the sensor network can be extremely sparse, prevailing deep learning methods for time series processing (e.g., RNN and LSTM) are not feasible in our task. Hence, a semi-supervised learning method is designed to achieve the goal. We first establish a multi-layer spatial-temporal graph to model the correlation between cubes. The weights of edges are represented by the correlations of features between cubes, based on the fact that cubes whose features are similar tend to share similar AQI values. The model iteratively learns and adjusts the edge weights to achieve the inference.

A. Feature Selection

Based on the study for key features in fine-grained scenarios [11]–[14], we select nine highly correlated features as: *3D coordinates*, *current time stamp*, *weather condition*, *wind speed*, *wind direction*, *humidity* and *temperature*. These features can be obtained either by our monitoring devices or crawling data from online websites.

B. Multi-Layer Spatial-Temporal Inference Model

The AQI values at different locations are correlated with each other in a spatial-temporal manner. For example, the AQI value at one location is highly similar to that at its neighboring location; the AQI value at a location depend on its values in past few hours.

Based on this observation, we propose a multi-layer graph model to characterize the correlations between cubes. Each cube is represented by a node in the graph, as shown in Fig. 4. These nodes are connected in both spatial and temporal dimensions to form a multi-layer weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each layer represents one spatial graph at a specific time stamp T_k. We name the nodes with observed data from the sensors as $labeled$ nodes, while nodes without observed data as $unlabeled$ nodes. Each labeled node l has the ground truth AQI value, while the AQI value of each unlabeled node u is estimated through a probability distribution p_u.

We construct the edges $\mathcal{E}$ in the graph by following steps: **(1) Connecting to labeled nodes,** where each unlabeled node is connected with all labeled nodes at the same time stamp T_k; **(2) Connecting to spatial neighbors,** where each unlabeled node is also connected with neighboring nodes within a given spatial radius r; and **(3) Connecting to temporal neighbors,** where each unlabeled node is connected to nodes in the same location but at neighboring time stamps. Fig. 4 shows an example of edge construction.

For every edge $(v_1, v_2) \in \mathcal{E}$, it has a corresponding weight. The weight of edge denotes how much the features between v_1 and v_2 are correlated. The correlation is defined by:

Definition 1. Correlation Function. *Given a set of features* $\boldsymbol{e} = \{e^{(1)}, e^{(2)}, \ldots, e^{(M)}\}$, *the correlation function of each feature between node* v_1 *and* v_2 *is defined as a linear function*

$$Q_{e^{(m)}}(v_1, v_2) = \alpha_m + \beta_m \left\| e^{(m)}(v_1) - e^{(m)}(v_2) \right\|_1, \quad m = 1, 2, \ldots, M. \tag{11}$$

In (11), α_m and β_m are parameters that can be estimated using the maximum likelihood estimation. Based on the correlation modeling between feature difference and AQI similarity, we define the weight matrix $\mathcal{W} = \{w_{i,j}\}$, where the weight on edge $\{(v_1, v_2) \in \mathcal{E}\}$ is expressed as

$$w_{v_1,v_2} = \exp\left(- \sum_{m=1}^{M} \theta_m^2 \cdot Q_{e^{(m)}}(v_1, v_2) \right), \tag{12}$$

where θ_m is the weight of feature $e^{(m)}$, and needs to be further learned to determine the AQI distribution of unlabeled nodes.

C. AQI Inference on Unlabeled Nodes

The objective for the model's convergence is to minimize the model's uncertainty for inferring unlabeled nodes. We show that the distribution p_u at an unlabeled node is the weighted average of distributions at its neighboring nodes [27]. Then, the objective becomes to minimize the entropy of the whole model, i.e., $H(p_u) = -\sum_u p_u \log p_u$, to achieve accurate estimation. This idea comes from the fact that an unlabeled node should possess a similar AQI value of its adjacent labeled nodes which are connected to it. Therefore, based on the edge weight function in (12), we define the loss function of the correlation graph to enable the propagation between highly correlated nodes with higher edge weights:

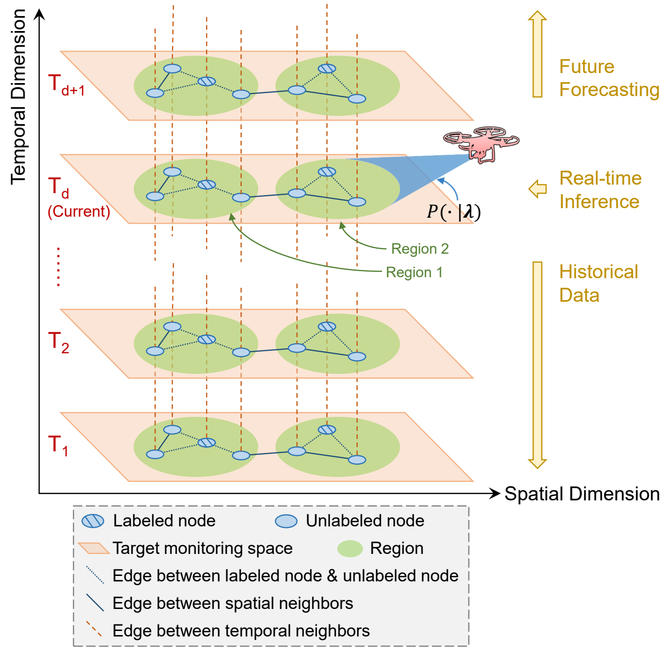


Fig. 4. An illustration of the proposed multi-layer spatial-temporal correlation graph model.

$$\mathcal{L}(\boldsymbol{p}) = \sum_{(v_1,v_2)\in\mathcal{E}} \frac{1}{2} w_{v_1,v_2} \left\| p_{v_1} - p_{v_2} \right\|^2, \tag{13}$$

where p_{v_1} and p_{v_2} are the AQI distribution at node v_1 and v_2, $\|p_{v_1} - p_{v_2}\| = \mathcal{D}_{KL}(p_{v_1} \| p_{v_2}) + \mathcal{D}_{KL}(p_{v_2} \| p_{v_1})$ denotes the similarity of AQI distributions between p_{v_1} and p_{v_2}, described by the Symmetrical Kullback-Leibler (KL) Divergence [26]. Thus, the objective function is given by:

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \ \mathcal{L}(\boldsymbol{p}). \tag{14}$$

By minimizing $\mathcal{L}(\boldsymbol{p})$, the nodes with higher edge weights would possess more similar AQI value while the nodes with lower edge weights would be more independent. Thus, the objective function can enable the AQI propagation between highly correlated nodes, thus improving inference accuracy.

Proposition 1. *The solution of* p_u *for (14) is the average of the distributions at its neighboring nodes.*

Proof: According to [27], the minimum function in (14) is $harmonic$. Therefore, we have $\Delta p_u = 0$ on unlabeled nodes U, while $\Delta p_l = P(v_l)$ on labeled nodes L. Here Δ is the $combinatorial\ Laplacian$, which is defined by $\Delta = D - W$. $D = \text{diag}(d_i)$ is the diagonal matrix with d_i denotes the degree of i; $W = \{w_{i,j}\}$ is the weight matrix defined in (12). The harmonic property provides the form of solution as:

$$p_u(x) = \frac{1}{d_u} \sum_{(u,l)\in\mathcal{E}} w_{u,l}\, p_l(x), \quad x \in \{0, 1, 2, \ldots, X\}, \tag{15}$$

where X is the maximum possible AQI value. To normalize the solution, we redefine it as

$$\begin{aligned} p_u(x) &= \frac{1}{d_u \sum_x p_u(x)} \sum_{(u,l)\in\mathcal{E}} w_{u,l} p_l(x) \\ &= \frac{\sum_{(u,l)\in\mathcal{E}} w_{u,l} p_l(x)}{\sum_x \sum_{(u,l)\in\mathcal{E}} w_{u,l} p_l(x)}. \end{aligned} \tag{16}$$

Hence, the distribution of unlabeled nodes p_u is the average of distributions at its neighboring nodes. ■

Proposition 2. *p_u in (16) is a probability mass function (PMF) on $\mathbf{x}$.*

Proof: To be a PMF on $\mathbf{x}$, we test the satisfaction of p_u on the following three properties:

- The domain of p_u is the set of all possible states of $\mathbf{x}$.
- $\forall x \in \mathbf{x}$, $0 \leq p_u(x) \leq 1$.
- $\sum_{x\in\mathbf{x}} p_u(x) = 1$.

Considering the expression form in (16), the conclusion is obvious, that p_u is a PMF on $\mathbf{x}$. ■

The solution again shows the influence of the highly correlated nodes that are connected by high-weight edges.

D. Entropy-based Learning with AQI Scale Prior

AQI Scale Prior: A key characteristic of our model is the conditioning of prior AQI scale knowledge on unlabeled nodes at current time stamp, $P(\cdot|\boldsymbol{\lambda})$ (see Fig. 4). This conditioning allows the learnt AQI scale from vision-based sensing to guide ground WSN sensing, providing faster convergence and more accurate inference. Specifically, target space is divided into disjointed regions $\{R_1, R_2, \ldots, R_k\}$ for aerial sensing. Each R_j contains a number of cubes $\boldsymbol{C}^{(j)}$ to be inferred. For each R_j, the aerial sensing provides a conditioning λ_j for $\boldsymbol{C}^{(j)}$:

$$\lambda_j : \left\{ x_i \in \left[X^{(j)}_{min},\ X^{(j)}_{max} \right], \quad \forall C_i \in \boldsymbol{C}^{(j)} \right\}. \tag{17}$$

By applying $P(\cdot|\boldsymbol{\lambda})$ to p_u in (16), we finally induce $p_u(x|\boldsymbol{\lambda})$ as the inferred distribution. The conditioning brings faster convergence during training, and also enables more accurate inference. Sec. VI will detail the region division method, which helps lead out the low-cost wake-up mechanism design.

So far, the expression of p_u is determined, the next step is to investigate the *learning* weight functions given by (12). θ_m is learned from both labeled and unlabeled data, which forms a semi-supervised mechanism.

Learning Criterion: Since the labeled nodes are sparse, maximizing the likelihood of labeled nodes data to learn θ_m is infeasible. Instead, we use models entropy as the criterion, since high entropies can be regarded as unpredicted values, resulting in poor capability of inference and low accuracy. Thus, the objective is to minimize the entropy $H(p_u)$ of unlabeled nodes:

$$\min_{\boldsymbol{\theta}} \ H(p_u) \ = \ \min_{\boldsymbol{\theta}} \ \frac{1}{|U|} \sum_{i=1}^{|U|} H_i(p_i), \tag{18}$$

where $|U|$ is the number of unlabeled nodes. By unfolding the objective function, we have

$$H(p_u) = -\sum_{i=1}^{|U|} \sum_{x=1}^{X} \frac{p_i\left(\mathbf{x} = x|\lambda_j\right)}{|U|} \log p_i\left(\mathbf{x} = x|\lambda_j\right). \tag{19}$$

For simplicity, we denote $\sum_x p_j(\mathbf{x} = x|\lambda_j) \log p_j(\mathbf{x} = x|\lambda_j)$ as $p_j \log p_j$, the gradient can be derived as

$$\frac{\partial H}{\partial \theta_m} = \frac{1}{|U|} \sum_{j=1}^{|U|} \left(\log \frac{1}{p_j} - \frac{1}{\ln 2} \right) \frac{\partial p_j}{\partial \theta_m}. \tag{20}$$

For every unlabeled p_j, we investigate $\frac{\partial p_j}{\partial \theta_m}$ based on (16) and (12). By applying the chain rule of differentiation, the final gradient can be derived as

$$\frac{\partial w_{u,l}}{\partial \theta_m} = 2 w_{u,l} Q_{e^{(m)}} \theta_m. \tag{21}$$

Thus, by iteratively learning and updating θ_m using (21), the edge weights $\mathcal{W}$ can be studied and further generate the final AQI distribution when the iteration converges.

Real-time Inference: As illustrated in Fig. 4, the real-time inference is based on (1) historical ground WSN data over last d time stamps, and (2) the conditioning of prior AQI scale knowledge $P(\cdot|\boldsymbol{\lambda})$. When the model converges, we obtain the determined AQI distribution $\hat{p}_u$ over $\mathcal{V}_U$, which is called as *soft* labeling. To provide an exact or *hard* labeling value of inference, as is proofed in Proposition 2 that p_u is a PMF on $\mathbf{x}$, we quantize it using the expectation of p_u:

$$\widehat{P}_u = \mathbb{E}_{\mathbf{x}\sim p_u}\left[\mathbf{x}\right] = \sum_{x=1}^{X} x \cdot P_u\left(\mathbf{x} = x|\boldsymbol{\lambda}\right). \tag{22}$$

Note that we can obtain $\widehat{P}_u(T_i)$ on each unlabeled node u over d time stamps. However, only data at current time T_d is needed for real-time inference. Inspired by this idea, we store the whole inferred distribution map each time when real-time inference is completed, and further use it as historical data in the future sensing. By doing so, more labeled nodes $\mathcal{V}_L$ are known to get better inference results, which can accelerate the convergence speed and improve the accuracy.

Future Forecasting: Our model is also capable of future inference. In Fig. 4, the edge can be extended to following time stamps and more. With the entropy-based learning procedure, it can maintain sufficient accuracy for near-future distribution forecasting even without the prior by aerial sensing.

VI. Energy-efficient Wake-up Mechanism

Since our ground inference model is able to operate with very sparse labeled data, we only need to wake up a small number of ground sensors in selected regions to sense data at each T_i. This scheme can greatly save the battery of devices and extend our system's working duration, while also ensure high inference accuracy.

Recent methods [6], [7], [12] which utilize ground WSN for inference, have employed all of their sensors to wake up simultaneously for data collection. This can lead to short working duration even with low-cost sensors. For example, devices

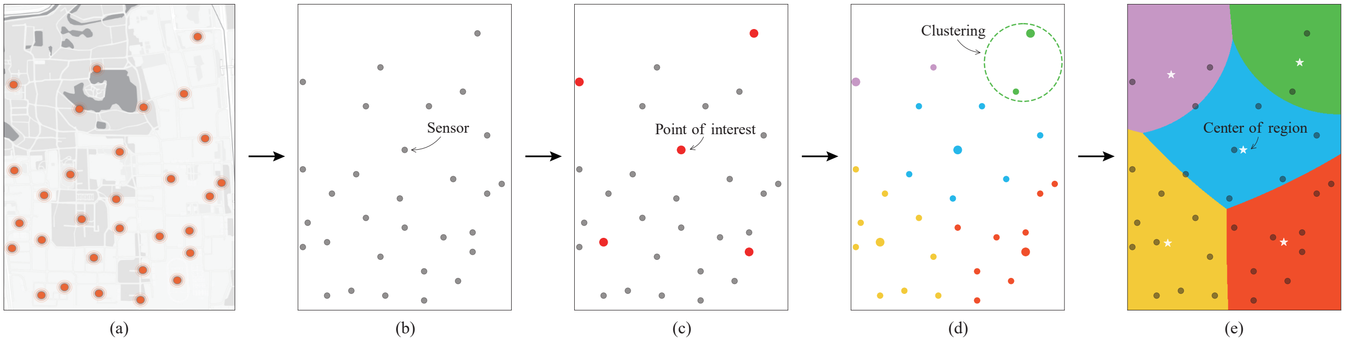


Fig. 5. An example of ground region division process: (a) ground sensors deployed in Peking University; (b) simplified graph illustrations; (c) dynamically select k points of interests each time; (d) result of k-means clustering to different classes; (e) result of multi-site weighted voronoi division.

in [7] can only last for less than 5 days before recharging, causing high consumption of battery power and human labor. Yet there have scarcely been works in asynchronous wake-up for AQI monitoring. In fact, due to the spatial-temporal correlations of AQI distribution, waking up a specific number of sensors is enough to realize high inference accuracy, while greatly reducing the power consumption. Thus, an energy-efficient wake-up mechanism is designed for ImgSensingNet to connect the aerial sensing and ground sensing, and to guide system selecting specific devices to wake up at current time stamp for energy saving.

A. Voronoi Diagram based Region Division

Since the total monitoring space can be very large, we first divide it into disjointed regions $\{R_1, R_2, \dots, R_k\}$ for aerial sensing. Note that even if devices are deployed in 3D (e.g., different floors of buildings), we only consider 2D coordinates for region division. The height and the camera angle of UAV are fixed in advance, in order to make sure the region is covered in images. Cubes $\boldsymbol{C}^{(j)}$ inside each R_j are provided an AQI scale conditioning λ_j using vision-based inference. Since the distribution of ground devices is heterogeneous and uneven, we implement the division (as shown in Fig. 5) by following steps:

Initialization: Fig. 5(a)(b)(c) present an example of the initialization process. Given a target space with ground devices deployed, k points of interests (POIs), e.g., a hospital or an office building, are selected dynamically at different time stamps.

Clustering: With k POIs selected, we cluster each device to its nearby POI in spatial dimension based on the spatial correlation of AQI distribution, where k-means clustering is used. We obtain k classes after the clustering, each containing n_j devices ($j \in [1, k]$), as shown in Fig. 5(d).

Multi-site Weighted Voronoi Diagram: Voronoi diagram is a partitioning of a plane into regions based on distance to sites in a specific subset [28]. The original voronoi diagram only considers one site in a region, and using the Euclidean distance for division. As we have multiple devices in one region, we propose a *multi-site weighted voronoi diagram* that enables division with (1) multiple sites inside one region, and (2) different weights assigned to each region for calculating the division boundary.

As shown in Fig. 5(e), we first calculate the center ϕ_j in R_j using the mean 2D coordinates of n_j devices inside it. The coordinates of center ϕ_j is used for division on behalf of R_j. Since the number of devices n_j varies over different regions, they should possess different weights when calculating the division boundary. Hence, we define the *weighted distance* as:

$$D(y, \phi_j) = \frac{d(y, \phi_j)}{\sqrt{n_j}} = \frac{\left\| \vec{y} - \vec{\phi_j} \right\|_2}{\sqrt{n_j}}, \tag{23}$$

where $d(y, \phi_j)$ is the Euclidean distance between location $\vec{y}$ and region center $\vec{\phi_j}$, n_j is the number of devices inside region R_j. Thus, the weighted voronoi division can be written as

$$V(\phi_i) = \bigcap_{j \neq i} \{y \mid D(y, \phi_i) \leq D(y, \phi_j)\}, \quad i, j \in [1, k]. \tag{24}$$

Proposition 3. *The complexity of the region division algorithm is* $\Theta(k \cdot n)$.

Proof: Denote the total number of 2D grids as $a \times b$, where a and b are constants for a specific monitoring area. In the first stage, all devices need to be clustered to a nearby POI, which computes for $k \cdot n$ times. In the second stage, we calculate the center of each class, which will compute k times in the worst case. In the last stage, the assignments for each grid will take $a \cdot b \cdot k$ times for computing. Note that we always have $a \cdot b \geq n$, while there always exists a constant c such that $c \cdot n \geq a \cdot b$ ($c = a \cdot b$ in the worst case). Hence, the complexity for the last stage is $\Theta(a \cdot b \cdot k) = \Theta(k \cdot n)$. By combining three stages, we derive the final complexity as $\Theta(k \cdot n) + \mathcal{O}(k) + \Theta(k \cdot n) = \Theta(k \cdot n)$. ■

Algorithm 1: Voronoi Diagram based Region Division Algorithm

```
Input: current POIs $\{O_1, O_2, \dots, O_k\}$; device set $\boldsymbol{C}$
Output: regions $\{R_1, R_2, \dots, R_k\}$
(a) Initialize device set $\boldsymbol{C}^{(1)}, \boldsymbol{C}^{(2)}, \dots, \boldsymbol{C}^{(k)}$ in each region;
forall $C_i \in \boldsymbol{C}$ do
    (b) Cluster $C_i$ to a nearest $O_j$, add $C_i$ into $\boldsymbol{C}^{(j)}$ ;
end
forall $\boldsymbol{C}^{(j)}$ do
    (c) Calculate the center $\phi_j$;
end
(d) Generate weighted voronoi diagram $\{R_1, R_2, \dots, R_k\}$.
```

The best achieved complexity for classical voronoi diagram is $\mathcal{O}(n^2)$ [28], which only fits for one-site division. In contrast, our algorithm can generate multi-site division as well as can reduce the computation overhead. Algorithm 1 shows the procedure of the region division algorithm.

B. When to Wake up

At each time stamp, we first perform vision-based aerial sensing over k regions to obtain the AQI scale inference for each region. Before triggering ground devices, we first utilize the semi-supervised learning model to give hard labeling on all nodes at current time stamp, based on the stored AQI inference maps over past d time stamps. Therefore, for each node C, there are two estimations: (1) AQI scale $[X_{min},\ X_{max}]$, and (2) pre-inferred value $\widetilde{X}$ using historical data. Based on the two priors, we propose an indicator to analyse the inference reliability and further decide which devices to wake up at current time stamp.

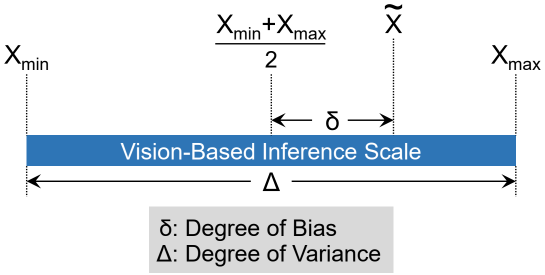


Fig. 6. An illustration of the two defined metrics, δ and Δ.

Joint Estimation Error: We first define two metrics of correlations between the two priors:

$$\delta = \left| \widetilde{X} - \frac{X_{min} + X_{max}}{2} \right|, \qquad (DoB) \qquad (25)$$

$$\Delta = X_{max} - X_{min}, \qquad (DoV) \qquad (26)$$

where we call δ as *Degree of Bios* (DoB), Δ as *Degree of Variance* (DoV), as shown in Fig. 6. Intuitively, when Δ is low, the variance of the AQI scale prior is small, which means a more reliable inference; as for δ, a low δ induces small deviation between the two priors, which in turn guarantees the inference reliability. Hence, DoB and DoV can both reflect the degree of estimation errors. By merging the two metrics, we define the *Joint Estimation Error* (JE) as:

$$JE = \frac{1}{2}\left[\frac{\delta}{\delta_m} + \frac{\Delta}{\Delta_m}\right], \qquad (27)$$

where δ_m and Δ_m denote the maximum value of DoB and DoV for all nodes with devices. As a result, *JE* is normalized into $[0, 1]$, and each node has a corresponding *JE*. In general, *JE* reflects the degree of average inference error for labeled nodes before waking up for ground sensing. For i^{th} cube, a greater $JE^{(i)}$ indicates higher uncertainty for inference at C_i, which signifies C_i should be measured currently if $JE^{(i)}$ exceeds a threshold. Hence, given a specific *JE* as threshold, sensors/nodes with $JE^{(i)} \geq JE$ should wake up for data collection at current time stamp. These nodes are then labeled with measured data at layer T_d, which can best reduce the model's entropy and are sufficient for real-time and future inference. In this way, by only measuring a small number of cubes, ImgSensingNet can greatly reduce the measurement overhead while maintaining high inference accuracy.

In general, *JE* is adjusted manually for different scenarios, which forms a tradeoff. When *JE* is low, the threshold for inference error declines, indicating the measuring cubes will increase and brings promotion in inference accuracy. However, it can cause great battery consumption. On the other hand, as *JE* is high, the measuring cubes will decrease. This may cause a decline in accuracy, but can highly reduce consumption. In summary, the tradeoff between accuracy and consumption should be studied to acquire a better performance.

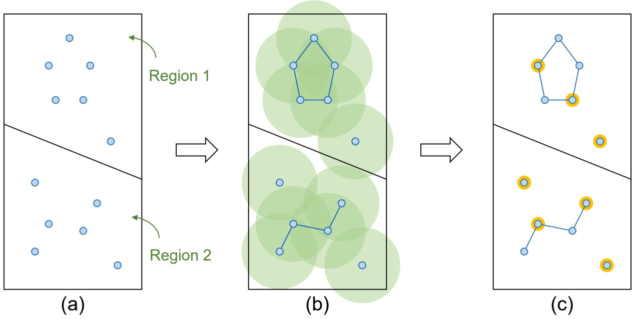


Fig. 7. An example of the final wake-up set $\mathcal{S}$ construction: (a) the selected wake-up nodes set $\mathcal{M}$ based on *JE*; (b) the spatial neighbors within radius r in graph model; (c) finding the min independent dominating set $\mathcal{S}$.

Wake-up Mechanism Design: *JE* can guide system waking up selected devices at each time stamp. First, if the two priors $[X_{min},\ X_{max}]$ and $\widetilde{X}$ are both less than a pre-given value σ, then the current air quality is too good to wake up the device for measurement, and switching off the device in such a case can help save the battery.

Second, since we construct the graph model by connecting nodes within a spatial radius r, it is possible that two nodes that are selected to wake-up are adjacent and connected in the model (Fig. 7(a) provides an example). In this case, waking up connected nodes would be redundant as their measurements are similar. Denote $\mathcal{M}$ as the set of selected wake-up nodes using *JE*, our objective is to find a subset $\mathcal{S} \subseteq \mathcal{M}$ such that (1) nodes in $\mathcal{S}$ are not adjacent, and (2) every node not in $\mathcal{S}$ is adjacent to at least one member of $\mathcal{S}$. This problem is well-known as the *minimum independent dominating set* problem [29], which is NP-hard. Since $\mathcal{M}$ is sparse in our case, the computing overhead is small, we simply apply a greedy-based method to find $\mathcal{S}$. Note that the algorithm is applied in each region independently. The total process of finding a final wake-up set $\mathcal{S}$ based on $\mathcal{M}$ is shown in Fig. 7.

Lemma 1. *The number of final wake-up devices $|\mathcal{S}|$ decreases monotonically when r increases.*

Proof: When r increases to $r^{'}$, the number of edge $\mathcal{E}^{'}$ can increase, which induces larger local connectivity graphs. For the extreme case, we choose $\mathcal{S}^{'}$ as the same as $\mathcal{S}$, which at least forms an independent dominating set. Since $\mathcal{E}^{'}$ can increase, nodes in $\mathcal{S}$ can be directly connected, which cannot be a minimum set. Hence, we have $|\mathcal{S}^{'}| \leq |\mathcal{S}|$ with $r^{'}$, that $|\mathcal{S}|$ decreases monotonically when r gets larger. ■

Algorithm 2: Wake-up Mechanism Design

```
Input: device set $\boldsymbol{C}^{(1)}, \boldsymbol{C}^{(2)}, \ldots, \boldsymbol{C}^{(k)}$; JE; $\sigma$
Output: target node set $\{\mathcal{S}\}$
forall $\boldsymbol{C}^{(j)}$ do
    forall $C_i^{(j)} \in \boldsymbol{C}^{(j)}$ do
        (a) Compute $JE^{(i)}$, $\widetilde{X}^{(j)}$, $(X_{min}^{(j)},\ X_{max}^{(j)})$;
        (b) Initialize $\mathcal{M}^{(j)} \leftarrow \varnothing$;
        if $\underline{JE^{(i)} \geq JE}$ and $\underline{\max\{\widetilde{X}^{(j)}, X_{max}^{(j)}\} \leq \sigma}$ then
            (c) Add $C_i^{(j)}$ into $\mathcal{M}^{(j)}$ ;
        end
    end
    while $\mathcal{M}^{(j)} \neq \varnothing$ do
        (d) Choose node $S \in \mathcal{M}^{(j)}$ with max $d_s$, add $S$ into $\mathcal{S}^{(j)}$;
        (e) Remove $S$ and all it's neighbors from $\mathcal{M}^{(j)}$;
    end
end
$\{\mathcal{S}\} = \bigcup\{\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \ldots, \mathcal{S}^{(k)}\}$.
```

Proposition 4. *The complexity of the wake-up mechanism algorithm decreases monotonically when r increases, which is $\Theta(|\mathcal{S}| \cdot n)$.*

Proof: For the inner loop, we compute n_j times to generate $\mathcal{M}^{(j)}$ for each region R_j. As we need to traverse $\mathcal{M}^{(j)}$ to find a node S with largest degree d_s each time, and the total times are represented by $|\mathcal{S}^{(j)}|$, thus the complexity to find $\mathcal{S}^{(j)} \subseteq \mathcal{M}^{(j)}$ is $\mathcal{O}(|\mathcal{S}^{(j)}| \cdot n_j)$. As for the outer loop, k times are needed for each region. Hence, the total complexity is $\mathcal{O}(k \cdot (|\mathcal{S}^{(j)}| \cdot n_j + |\mathcal{S}^{(j)}|)) = \mathcal{O}(k \cdot |\mathcal{S}^{(j)}| \cdot n_j)$. It's obvious to find both the upper bound and lower bound can be denoted as $|\mathcal{S}| \cdot n$ multiply by a constant c, which induces the final complexity as $\Theta(|\mathcal{S}| \cdot n)$.

Based on Lemma 1, the complexity also decreases monotonically when r gets larger. ∎

Algorithm 2 shows the procedure of the wake-up mechanism. The target node set is determined based on *JE* as well as two conditions we studied, which contains the nodes with the highest uncertainty currently. Thus, by only monitoring $\mathcal{S}$, ImgSensingNet achieves high inference accuracy while greatly reducing the measurement overhead.

Proposition 5. *The overall complexity for wake-up mechanism each time is $\Theta\left((|\mathcal{S}| + k) \cdot n\right)$.*

Proof: The proposition is obvious based on the results of Proposition 3 and 4. ∎

The overall wake-up mechanism contains dynamic region division and wake-up devices selection each time. In practice, since k and $|\mathcal{S}|$ are small due to the device sparsity in deployment, by choosing proper r, the computation of the whole mechanism can be completed in real-time, which will be shown in the evaluation.

VII. Implementation

In this section, we detail the implementation of ImgSensingNet system. Specifically, we first introduce the components of ImgSensingNet, including both the ground devices and the aerial device. Based on the devices, we then illustrate the data collection of both aerial image data and ground sensing data.

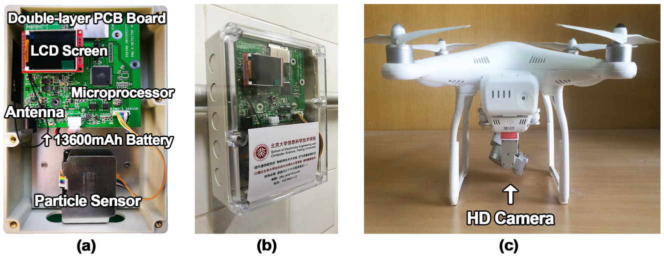


Fig. 8. The components of ImgSensingNet: (a) the sensor-based monitoring device; (b) the device is installed on the wall; (c) the UAV for vision-based sensing.

A. System Components

Ground Devices: Fig. 8(a) shows the components of the ground AQI monitoring device. Each device contains a low-cost `A3-IG` sensor, a two-layer circuit board, a `ATmega128A` working as the micro-controller unit (MCU), a `SIM7000C` as the wireless communication module, a 13600mAh rechargeable battery and a fixed shell structure. Considering the intrinsic lack of precision for the small-scale laser-based sensors, these devices are carefully calibrated through a whole month adjustment by comparing the results with a high-precision calibrating instrument `TSI8530`. Finally, these devices can provide $\leq$ 3% monitor error for common pollutants in AQI calculation, such as $\text{PM}_{2.5}$ and PM_{10}, and send the real-time data back to the central server for further data analysis. To realize high energy-efficiency, the devices are programmed to sleep during most of the time and wake up for data collection based on adjustable time intervals that are controlled by a designed wake-up mechanism, which will be discussed in Sec. VI. Thus, an online tradeoff between data quantity and battery endurance can be implemented.

Aerial Device: For the UAV, we select `DJI Phantom 3` Quadcopter as the sensing device, as shown in Fig. 8(c). The GPS sensor on the UAV can provide the real-time 3D position. In existing systems [14], the UAV can keep flying for at most 10∼20 minutes due to both the *load consumption* (carrying sensors can significantly reduce the UAV's battery life), and the *loitering consumption* (to acquire sensing data, the UAV needs to stay still at every measuring location), which restricts the monitoring scope within one measurement [30]. However, as the UAV contains a built-in HD camera, when we focus on vision-based sensing, the extra loading and the hovering time can be eliminated. Hence, the sensing scope as well as the flight duration can be greatly increased.

B. Experiment Setup and Data Collection

ImgSensingNet system prototype includes 200 ground devices and a UAV, and it has been deployed on two university campuses (i.e., Peking University and Xidian University), since Feb. 2018. Throughout more than a half years measurement, 2.6 millions of ground data samples and and a number of 17,630 haze images are collected, covering from good air quality cases to hazardous air quality cases, which are used for evaluation in this paper.

Aerial Image Data: The vision-based sensing works online continuously and real-timely by sampling images from the

TABLE II
AVERAGE ESTIMATION ERRORS BY DIFFERENT METHODS

Methods	2D				3D			
	Real-time	*After 1 hour*	*After 3 hours*	*After 10 hours*	*Real-time*	*After 1 hour*	*After 3 hours*	*After 10 hours*
ImgSensingNet	**3.540**	**6.178**	**9.330**	**20.269**	**5.529**	**9.928**	**13.341**	**31.409**
ARMS [14]	5.412	–	–	–	7.384	–	–	–
LSTM Nets [19]	4.217	7.804	10.672	**19.873**	–	–	–	–
AQNet [12]	4.493	7.562	13.695	25.192	6.481	14.735	19.634	39.790
S-T kNN [20]	7.039	9.667	11.882	26.055	9.147	12.954	18.065	44.256

Fig. 9. An overview of the aerial image dataset: more than 10,000 labeled haze images taken in Beijing, Xi'an and Wuhan.

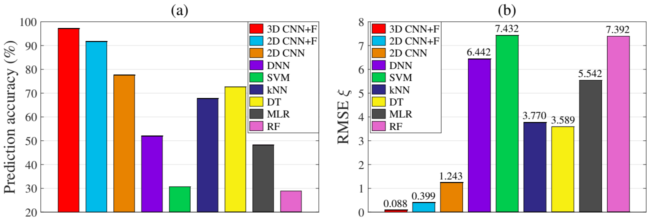


Fig. 10. The image-based inference comparison between different methods: (a) the inference accuracy; (b) the robustness of inference.

UAV video streams between equal time intervals. To get ground truth data for training the CNN model, we set up the dataset by carrying calibrated sensor to label the image with ground truth AQI value. We collected 17,630 labeled images in different places to make the data generalize well. Fig. 9 shows an overview of the image dataset.

Ground Sensing Data: The testing areas are on campus of Peking University ($\sim$ 2km×2km) and Xidian University ($\sim$ 2km×1.5km). The ground devices are deployed in 3D space with a 50m maximum height. We divide the areas into 20m×20m×10m cubes, where a small number of cubes are deployed with our devices. We manually set the minimum time intervals as 30 minutes.

VIII. EVALUATION

In this section, we present the performance analysis of ImgSensingNet in various aspects.

A. Vision-based Aerial Sensing

We evaluate the proposed AQI scale inference model in two aspects: *accuracy* and *robustness* in predictions. We randomly divide the image dataset with 7:3 training set to testing set ratio. We compare the proposed inference model with the following models from two categories: (1) three *deep learning* methods: 2D CNN with our extracted features, 2D CNN without features and a 50-layer deep neural network (DNN); the 2D CNN architecture is the same with our 3D CNN, but with only 2D kernels. (2) five *classical training* methods: support vector machine (SVM), k-nearest neighbors (kNN), decision tree (DT), multi-variable linear regression (MLR) and random forest (RF).

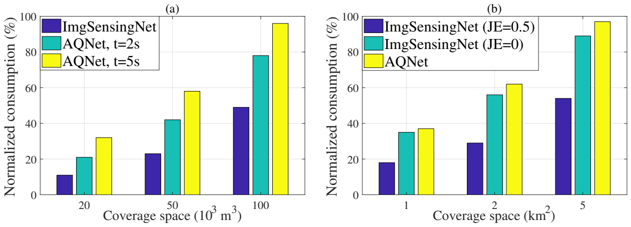


Fig. 11. The energy-efficiency comparison between different methods: (a) the consumption of aerial UAV sensing; (b) the consumption of ground WSN sensing.

Accuracy of Inference: As shown in Fig. 10(a), in general our method outperforms all other models. We can achieve a 96% accuracy for image-based AQI scale inference by the proposed model. Moreover, when the features are considered, the 2D CNN model also outperforms the one without features, which confirms the effectiveness of haze-relevant feature extraction.

Robustness of Inference: In Fig. 10(b), we test how much the inferred values deviate from the real values, using root mean square error (RMSE). The results show that the proposed model outperforms other models by maintaining a very low deviation, i.e., 0.088 classification deviation in average. This again proves the advances in using 3D model and feature extraction.

B. Inference Accuracy

We evaluate the inference accuracy of ImgSensingNet in both real-time estimation and near-future forecasting. Since there are no measured data for most cubes, we divide labeled samples into training set and testing set, while performing an cross-validation by randomly choosing the training data, and repeat for 1000 times to avoid stochastic errors.

We use inference models in state-of-the-art AQI monitoring systems as ARMS [14], LSTM Nets [19], AQNet [12] and spatio-temporal kNN [20] for comparison. These models are all evaluated using the same data each time.

In Table II, we report the average estimation errors of real-time inference and near-future forecasting (i.e., after 1, 3, and 10 hours respectively), in both 2D and 3D scenarios. As a result, ImgSensingNet can achieve the best inference accuracy (referred as the lowest RMSE in the table) in both

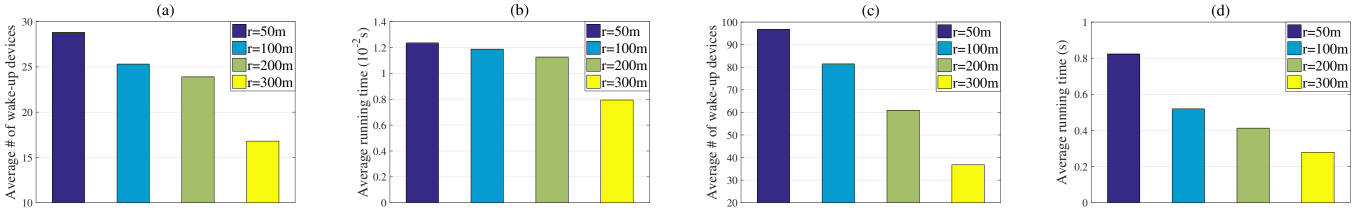


Fig. 12. The wake-up mechanism performance versus different r: (a) the average number of wake-up devices, with 30 total devices; (b) the average runtime, with 30 total devices; (c) the average number of wake-up devices, with 100 total devices; (d) the average runtime, with 100 total devices.

real-time inference and future AQI forecasting. Even with high accuracy, the competitors may either lack the ability of future prediction (e.g., ARMS) or 3D inference (e.g., LSTM Nets).

C. Energy Efficiency

The energy-efficiency is analysed in two aspects: (1) the consumption of aerial UAV sensing, and (2) the consumption of ground WSN sensing. We choose AQNet [12] that has similar components (using UAV and ground WSN) for comparison in the two aspects, respectively.

Consumption of Aerial Sensing: We set up experiments by comparing the normalized system consumption in monitoring tasks with different coverage spaces. As shown in Fig. 11(a), ImgSensingNet uses UAV that does not suffer from both the *load* and *loitering* consumptions, hence can greatly save the battery. Compared to AQNet system with different loitering time t for data sensing, ImgSensingNet consumes about 50% less energy than that of AQNet, with different coverage space. Thus, the energy-efficiency of the proposed system is demonstrated.

Consumption of Ground Sensing: We further study the normalized consumption of ground sensing using the same method. We compare one day's consumption of all ground devices within different coverage spaces, using the same detection time and uploading time for each method. Fig. 11(b) presents the experimental results. When $JE = 0$, our ground sensing achieves the maximum consumption, which still slightly outperforms AQNet system. As $JE = 0.5$, the normalized consumption of the WSN significantly reduces to only 53%, which again validates the energy-efficiency of ImgSensingNet.

D. Wake-up Mechanism

We analyse the impact of r on wake-up mechanism in two aspects: (1) the average number of devices that wake up each time, and (2) the average computing time for devices selection. We vary the number of devices as 30 and 100, and set $k = 5$. For each instance, we perform 1000 independent runs to get the average values.

Average Number of Wake-up Devices: As shown in Fig. 12(a)(c), we plot the average number of wake-up devices with different values of r, by setting $JE = 0$ as an invariant. The number of selected devices decreases monotonically when r increases. Specifically, when we choose $r = 300$ m, the average number of wake-up devices can greatly reduce to less than 50% of total devices (e.g., 38.5 on average when there are 100 devices in total). Thus, by choosing a proper r, the number of wake-up devices greatly scales down, which is energy efficient.

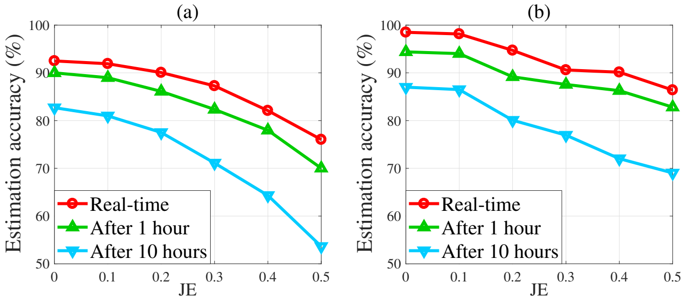


Fig. 13. The system estimation accuracy in real-time, after 1 hour, and after 10 hours, versus different *JE*. (a) with 30 total devices; (b) with 100 total devices.

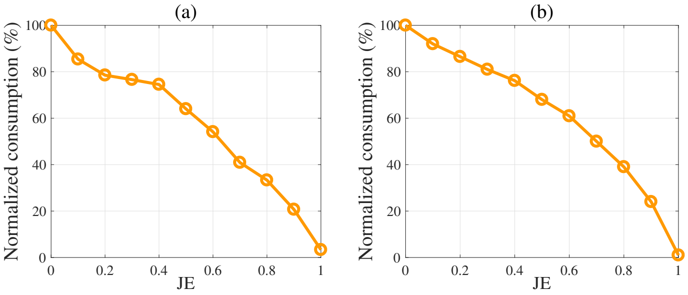


Fig. 14. The system normalized energy consumption versus different *JE*. (a) with 30 total devices; (b) with 100 total devices.

Average Runtime of Wake-up Mechanism: Further, we study the runtime for obtaining the set of wake-up devices each time. As shown in Fig. 12(b)(d), the runtime also decreases with a greater r. Specifically, the average running time is about 0.01 s when there are 30 devices in total. When there are more devices, the computation time will increase, but it is still completed in real-time (about 1s in Fig. 12(d)).

E. Impacts of Different Joint Estimation Errors

In this section, we investigate the impacts of different *JE* values on ImgSensingNet, in three aspects as (1) estimation accuracy, (2) energy consumption, and (3) working durations, respectively.

Estimation Accuracy: As shown in Fig. 13, the estimation accuracy gradually decreases when *JE* increases. From the figure, we can see that ImgSensingNet achieves high accuracy in both real-time inference and future forecasting. Moreover, the system can achieve higher inference accuracy when there are more devices deployed.

Normalized Energy Consumption: In Fig. 14 we report the relationship between energy consumption and different *JE*

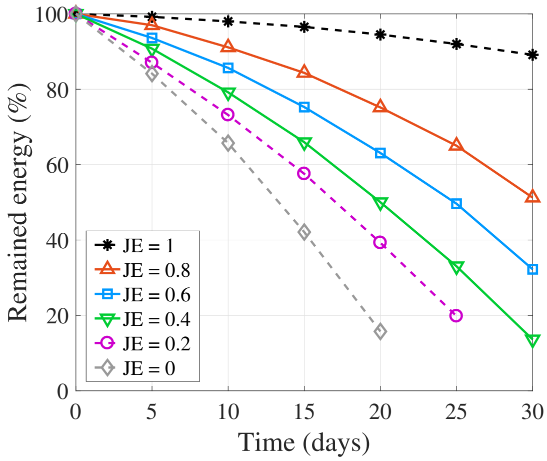


Fig. 15. The impact of different *JE* on system working durations.

values. By comparing Fig. 14(a) and Fig. 14(b), the energy consumption scales down as *JE* increases, while a more stable procedure is obtained when there are more devices.

System Working Durations: In Fig. 15 we study the impacts of different *JE* on system working durations, over a fixed area inside Peking University. It is shown that ImgSensingNet can guarantee a long battery duration for more than one month when $JE \geq 0.4$, which greatly outperforms state-of-the-art systems. As *JE* decreases, the monitoring overhead would increase, while it can also bring high inference accuracy. Hence, there exists a tradeoff between consumption and accuracy caused by different *JE*, which needs to be further studied.

F. Impact of Degree of Air Pollution

In Fig. 16, we study the impact of the degree of air pollution on ImgSensingNet. We first manually divide our dataset into three degrees as slightly, moderately and highly polluted (i.e., AQI ≤ 50, $50 <$ AQI < 200 and AQI ≥ 200), and evaluate the performance of our model separately.

Estimation Accuracy: In Fig. 16(a) we compare the inference accuracy when AQI value varies. As a result, out system performs the best when AQI ≥ 200. Moreover, the performance tends to be better when AQI value is higher, as most devices are scheduled to sleep when air quality is good.

Normalized Energy Consumption: We further study the normalized energy consumption in different AQI degrees with various values of *JE*. From Fig. 16(b), we can see that our system maintains the lowest consumption when AQI value is low, which again validates the energy-efficiency of the wake-up mechanism. By comparing Fig. 16(a) and (b), the tradeoff can also be illustrated.

G. Tradeoff between Accuracy and Consumption

In Fig. 17, an inherent tradeoff between system consumption and inference accuracy is illustrated, versus *JE*. As *JE* becomes higher, the average inference error grows rapidly while consumption can drop fairly. Given the average error, for example, when RMSE is 25, the corresponding $JE = 0.6$, which indicates the power consumption can be reduced to as little as 60%. Hence, by choosing proper *JE* value, the measuring cost can greatly scale down.

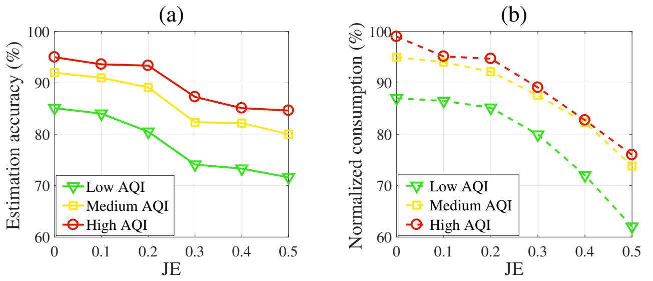


Fig. 16. The impact of degree of air pollution: (a) the estimation accuracy; (b) the normalized consumption.

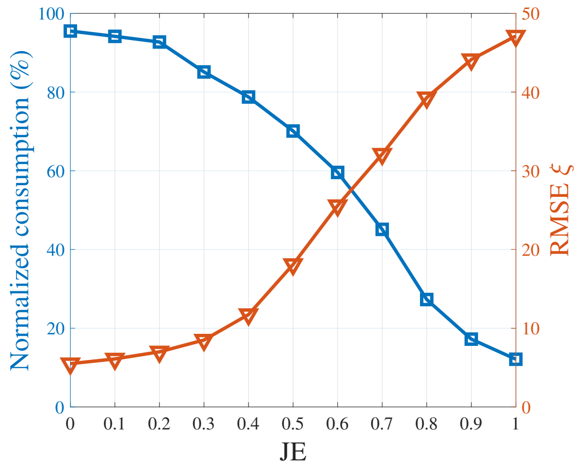


Fig. 17. The tradeoff between system power consumption and inference accuracy.

IX. Conclusion

This paper presents the design, technologies and implementation of ImgSensingNet, a UAV vision guided aerial-ground AQI sensing system, to monitor and forecast the air quality in a fine-grained manner. We first utilize vision-based aerial UAV sensing for AQI scale inference, based on the proposed haze-relevant features and 3D CNN model. Ground WSN sensing are then used for accurate AQI inference in spatial-temporal perspectives using an entropy-based model. Further, an energy-efficient wake-up mechanism is designed to greatly reduce the energy consumption while achieving high inference accuracy. ImgSensingNet has been deployed on two university campuses for daily monitoring and forecasting. Experimental results show that ImgSensingNet outperforms state-of-the-art methods, by achieving higher inference accuracy while best reducing the energy consumption.

References

[1] W. H. O., "7 million premature deaths annually linked to air pollution," *Air Quality & Climate Change*, vol. 22, no. 1, pp. 53-59, Mar. 2014.

[2] B. Zou *et al.*, "Air pollution exposure assessment methods utilized in epidemiological studies," *J. Environ. Monit.*, vol. 11, pp. 475-490, 2009.

[3] Beijing MEMC. [Online]. Available: *http://www.bjmemc.com.cn/*. 2018.

[4] T. Quang *et al.*, "Vertical particle concentration profiles around urban office buildings," *Atmos. Chem. Phys.*, vol. 12, pp. 5017-5030. 2012.

[5] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-Air: When urban air quality inference meets big data," in *Proc. ACM KDD'13*, Chicago, IL, Aug. 2013.

[6] Y. Cheng *et al.*, "Aircloud: a cloud-based air-quality monitoring system for everyone," in *Proc. ACM SenSys'14*, New York, NY, Nov. 2014.

[7] Y. Gao, W. Dong, K. Guo *et al.*, “Mosaic: a low-cost mobile sensing system for urban air quality monitoring,” in *Proc. IEEE INFOCOM’16*, San Francisco, CA, Jul. 2016.

[8] D. Hasenfratz, O. Saukh, C. Walser *et al.*, “Deriving high-resolution urban air pollution maps using mobile sensor nodes,” *Pervasive and Mobile Compting*, vol. 16, no. 2, pp. 268-285, Jan. 2015.

[9] J. Li *et al.*, “Tethered balloon-based black carbon profiles within the lower troposphere of shanghai in the east china smog,” *Atmos. Environ.*, vol. 123, pp. 327-338. Sept. 2015.

[10] Y. Hu, G. Dai, J. Fan, Y. Wu and H. Zhang, “BlueAer: a fine-grained urban PM2.5 3D monitoring system using mobile sensing,” in *Proc. IEEE INFOCOM’16*, San Francisco, CA, Jul. 2016.

[11] Y. Yang *et al.*, “Arms: a fine-grained 3D AQI realtime monitoring system by UAV,” in *Proc. IEEE Globecom’17*, Singapore, Dec. 2017.

[12] Y. Yang, Z. Bai, Z. Hu, Z. Zheng, K. Bian, and L. Song, “AQNet: fine-grained 3D spatio-temporal air quality monitoring by aerial-ground WSN,” in *Proc. IEEE INFOCOM’18*, Honolulu, HI, Apr. 2018.

[13] Z. Hu *et al.*, “UAV Aided Aerial-Ground IoT for Air Quality Sensing in Smart City: Architecture, Technologies and Implementation,” *IEEE Network Magazine*, accepted, available on https://arxiv.org/abs/1809.03746.

[14] Y. Yang, Z. Zheng, K. Bian, L. Song, and Z. Han, “Real-time profiling of fine-grained air quality index distribution using UAV sensing,” *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 186-198, Feb. 2018.

[15] Z. Pan, H. Yu, C. Miao, and C. Leung. “Crowdsensing air quality with camera-enabled mobile devices,” in *Proc. Thirty-First AAAI Conf. on Artificial Intell.*, San Francisco, CA, Feb. 2017.

[16] S. Li, T. Xi, Y. Tian, and W. Wang. “Inferring fine-grained PM2.5 with bayesian based kernel method for crowdsourcing system,” in *Proc. IEEE Globecom’17*, Singapore, Dec. 2017.

[17] R. Gao *et al.*, “Sextant: towards ubiquitous indoor localization service by photo-taking of the environment,” *IEEE Trans. Mobile Comput.*, vol. 15, no. 2, pp. 460-474, Feb. 2016.

[18] H. Kim and K. G. Shin. “In-band spectrum sensing in cognitive radio networks: energy detection or feature dtection?” in *Proc. ACM MobiCom’08*, 2008.

[19] V. O.K. Li, J. Lam, Y. Chen, and J. Gu. “Deep learning model to estimate air pollution using M-BP to fill in missing proxy urban data,” in *Proc. IEEE Globecom’17*, Singapore, Dec. 2017.

[20] Y. Yang, Z. Zheng, K. Bian, L. Song, and Z. Han, “Sensor deployment recommendation for 3D fine-grained air quality monitoring using semi-supervised learning,” in *Proc. IEEE ICC’18*, Kansas City, MO, May 2018.

[21] K. He, J. Sun, and X. Tang, “Single image haze removal using dark channel prior,” in *Proc. IEEE CVPR’09*, Miami, FL, Jun. 2009.

[22] K. He, J. Sun, and X. Tang, “Guided image filtering,” in *Proc. ECCV’10*, Crete, Greece, Sept. 2010.

[23] R. Tan, “Visibility in bad weather from a single image,” in *Proc. IEEE CVPR’08*, Anchorage, AK, Jun. 2008.

[24] Q. Zhu, J. Mai, and L. Shao, “A fast single image haze removal algorithm using color attenuation prior,” *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3522-3533, Jun. 2015.

[25] C. Ancuti *et al.*, “A fast semi-inverse approach to detect and remove the haze from a single image,” in *Proc. ACCV*, Pondicherr, India, Aug. 2011.

[26] I. Goodfellow, Y. Bengio, and A. Courville, “Applied Math and Machine Learning,” in *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[27] X. Zhu *et al.*, “Semi-supervised learning using gaussian fields and harmonic functions,” in *Proc. ICML’03*, Washington, DC, Aug. 2003.

[28] F. Aurenhammer, “Voronoi diagrams: a survey of a fundamental geometric data structure,” *ACM Comput. Survey*, vol. 23, pp. 345-405. 1991.

[29] N. Bourgeois *et al.*, “Fast algorithms for min independent dominating set,” *Discrete Applied Mathematics*, vol. 161, pp. 558-572. Mar. 2013.

[30] P. Zhao *et al.*, “Optimal trajectory planning of drones for 3d mobile sensing,” in *Proc. IEEE Globecom’18*, Abu Dhabi, UAE, Dec. 2018.