Abstract—Given the increasingly serious air pollution problem, air quality index (AQI) monitoring in urban areas has drawn considerable attention. This paper presents ImgSensingNet, a vision guided aerial-ground sensing system, for air quality monitoring and forecasting by the fusion of haze images taken by the unmanned-aerial-vehicle (UAV) and the AQI data collected by an on-ground wireless sensor network. Specifically, ImgSensingNet first leverages the computer vision technique to tell the AQI scale in different regions from the 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 inference model is employed for accurate real-time AQI estimation 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 high estimation accuracy while greatly reduce the battery consumption, compared to other state-of-the-art AQI monitoring approaches.

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 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].

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 applications have been developed to monitor the fine-grained air quality using densely deployed sensors [5], [6]. Although the static sensors may achieve high precision, 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 [8]–[11]. However, the sensor-based approach may induce high energy consumptions for mobile devices to acquire certain amount of data.

The second category of approaches includes the vision-based monitoring approaches. Image-based AQI monitoring stations are set up by researchers at dedicated locations [13], and these static stations can only take photos and infer the AQI values at limited sites over the whole space. Crowd-sourced photos contributed by mobile phones can depict the AQI distribution [12] 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 and reduce 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 [14]. In this work, we seek a way of leveraging both photo-taking and data sensing to monitor and infer the AQI value.

This paper presents ImgSensingNet, a vision guided aerial-ground air quality sensing system, to monitor and predict AQI values. Unlike existing systems, we implement: (1) mobile vision-based sensing over an unmanned-aerial-vehicle (UAV), that realizes 3D AQI monitoring by UAV photo-taking instead of using particle sensors, to infer region-level AQI scale (an interval of possible AQI values) using a convolutional neural network (CNN); (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 WSN in response to AQI scale inference; (4) ground-air image fusion to monitor the AQI distribution 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.

The main contributions are summarized as below.

- We implement ImgSensingNet, a 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;

1207
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.

II. RELATED WORK

A. AQI Monitoring Methods

Existing AQI monitoring methods can be summarized into two categories:

Sensor-based: Stationary stations [2] are set up on dedicated sites in cities, but only provide limited measurement samples. For example, there are only 28 monitoring stations in Beijing. The distance between two nearby stations is typically ten-thousand meters, and the AQI is monitored every 2 hours [3]. AirCloud [5] uses densely distributed sensors in a static way, while [6]–[8] 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 [9]–[11], drones with sensors together with ground sensors are used for AQI profiling. However, they are either restricted in a small area 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 monitoring. In [13], image-based monitoring stations are set up over a city. Again, these methods can only profile AQI at a limited number of locations. In [12], camera-enabled mobile devices are used for generating crowd-sourced photos to monitor AQI. 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 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 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 at unmeasured locations. Existing systems [5], [6], [9] adopt different models for 2D AQI inference based on measured data. A fine-grained AQI distribution model is proposed in [11] for AQI estimation over a 3D space. Deep learning algorithms are used [10], [16] to analyze spatial-temporal correlations and to forecast future distribution. Image-based inference has been used to quantify AQI from haze images [12], [13].

This work investigates two inference models: (1) image-based AQI scale inference in different regions by computer vision, and (2) the fine-grained spatial-temporal AQI value inference at locations inside each region by sensors.

III. SYSTEM OVERVIEW

The ImgSensingNet system includes 200 programmable monitoring devices and a UAV, as shown in Fig. 1. The aerial UAV and ground WSN sensing form a hybrid sensing network.

A. System Components

Ground Devices: As shown in Fig. 1(a), each ground device contains a low-cost A3-IG sensor, a two-layer circuit board, an ATmega128A as the micro-controller, a SIM7000C as the wireless module, a rechargeable battery and a fixed shell structure. To avoid errors, sensors are carefully calibrated through a whole month adjustment by comparing the results with a high-precision calibrating instrument TSI8530. Finally, these devices can provide $\leq \pm 3\%$ monitor error, 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 is presented in Sec. VI.

Aerial Device: We select DJI Phantom 3 as the sensing device (see Fig. 1(g)). The GPS sensor on the UAV can provide real-time 3D positions. In existing systems [11], 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. However, with the built-in HD camera, the UAV vision-based sensing can achieve larger sensing scope and longer flight duration.

B. Two Sensing Mechanisms

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.

Aerial Sensing: The aerial sensing uses 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 analysis and output estimated AQI scale for each region.

Ground Sensing: Ground WSN adopts a spatial-temporal inference model for AQI estimation at unmeasured locations and future air quality prediction. Each time when aerial sensing is finished, each ground device follows a designed wake-up mechanism to decide whether to wake up for data collection based on the inference result and aerial sensing result. In this way, both real-time and future AQI distributions are obtained.
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 computer vision field [18]–[20]. However, to quantify the haze in the image to real AQI value, two challenges 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.

A. Haze-relevant Features Extraction

A haze image can be mathematically described using the haze image formation model [18] as

\[
I(x) = J(x) \cdot t(x) + L_{\infty} (1 - t(x)),
\]

where \(I\) is the hazy image, \(J\) is the haze-free image, \(t\) denotes the medium transmission, \(L_{\infty}\) is the global atmospheric light, and \(x\) represents pixel coordinates. The haze-removal methods have spent large effort estimating \(J\) and \(t\) for haze-free image recovery [18]–[20]. Instead, we propose a new objective to estimate the degree of hazes. The first step is to extract a list of haze-relevant features. Since we 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.

1) Refined Dark Channel: Dark channel [18] is defined as the minimum of all pixel colors in a local patch:

\[
D(x; I) = \min_{y \in \Omega(x)} \left( \min_{c \in \{r,g,b\}} \frac{I_c}{L_{\infty}} \right),
\]

where \(\Omega(x)\) is a local patch centered at \(x\), \(I^c\) is one color channel of \(I\). 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 [18]. 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 [19] \(\hat{G}\) on the estimated medium transmission \(t\), to capture the sharp edge discontinuous and outline the haze profile:

\[
D^R(x; I) = 1 - \hat{G} \left( 1 - \min_{y \in \Omega(x)} \left( \min_{c \in \{r,g,b\}} \frac{I_c}{L_{\infty}} \right) \right),
\]

Fig. 2(b) shows the refined dark channel feature, which 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 haze image is highly reduced. Therefore, the contrast is the perceived feature to detect haze. The local contrast is defined as the variance of pixel intensities in a local \(r \times r\) region compared with the center one. Further, we use the local maximum of local contrast values in a local patch to define the max local contrast feature. Fig. 2(c) shows the visually obvious correlation between haze and the feature.

3) Max Local Saturation: It is observed that the image saturation varies sharply when haze changes in the scene [13]. Therefore, similar to image contrast, we define the max local saturation feature that profiles the maximum saturation value of pixels within a local patch. As shown in Fig. 2(d), the saturation feature is also correlated with the haze.

4) Min Local Color Attenuation: In [19], the scene depth \(d(x; I)\) is found to be positively correlated with the difference between the image brightness and the image saturation. This statistics is regarded as the color attenuation prior. To process the raw depth map for better profiling of haze influence, we define the min local color attenuation feature by considering the minimum pixel-wise depth within a local patch. Fig. 2(e) shows the min local color attenuation feature, where an obvious correlation with haze density can be observed.

5) Hue Disparity: In [20], the hue disparity between the original image and its semi-inverse image is utilized to remove haze. The hue disparity is also reduced by haze, thus can serve as another haze-relevant feature. Fig. 2(f) shows the hue disparity feature for the “forest” haze image.

6) Chroma: In the CIELab color space, the chroma is one of the most representative feature to describe the color degradation by the haze. 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.

B. 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 inference of input haze images. 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 \(128 \times 128\) 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. The rationale for using this hardwired layer is to encode our prior knowledge on different haze-relevant features. This scheme regularizes the 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 $3 \times 3 \times 1$ 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 layer uses $3 \times 3 \times 3$ kernel size, followed by $2 \times 2 \times 1$ 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.

**Training and AQI Scale Inference:** As the output is AQI scale (i.e., $[X_{min}, X_{max}]$), the inference is modelled as a classification problem, where the AQI scale classes are pre-divided based on the 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 INFEERENCE BY GROUND SENSOR MONITORING**

The 3D space is first divided into disjointed cubes, where each cube contains its own geographical coordinates, and each cube is associated with an AQI value. Note that AQI values in a limited number of cubes are sensed from the sensor network, while the AQI in other unobserved cubes need to be estimated.

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 campuses in the experiment), whose AQI 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.

**Feature Selection:** Based on the study for key features in fine-grained scenarios [10], [11], [15], 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.

![Fig. 3. The architecture of the proposed 3D CNN model.](image)

![Fig. 4. An illustration of the proposed multi-layer spatial-temporal correlation graph model.](image)
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 \( \mathcal{E} \) by following steps: (1) Connecting to labeled nodes: Each unlabeled node is connected with all labeled nodes at the same time stamp \( T_k \); (2) Connecting to spatial neighbors: Each unlabeled node is also connected with neighboring nodes within a given spatial radius \( r \); and (3) Connecting to temporal neighbors: 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 \( e = \{ e^{(1)}, e^{(2)}, \ldots, e^{(M)} \} \), the correlation function of each feature between node \( v_1 \) and \( v_2 \) is defined as a linear function

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

In (4), \( \alpha_m \) and \( \beta_m \) are parameters that can be estimated using the maximum likelihood estimation. Based on 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

\[
w_{v_1,v_2} = \exp \left( -\sum_{m=1}^{M} \theta_m^2 \cdot Q_{e(m)}(v_1, v_2) \right), \quad (5)
\]

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

**B. AQI Inference on Unlabeled Nodes**

The objective is to minimize the model’s uncertainty for estimating unlabeled nodes. We show that the distribution \( p_u \) at an unlabeled node is the weighted average of distributions at its neighboring nodes [21]. Then, the objective becomes to minimize the entropy of the whole model, i.e., \( H(p_u) = -\sum_p p_u \log p_u \), since an unlabeled node should have a similar AQI value to its adjacent labeled nodes which are connected to it. Hence, based on the edge weight in (5), we define the loss function of the correlation graph to enable the propagation between highly correlated nodes with higher edge weights:

\[
L(p) = \sum_{(v_1, v_2) \in \mathcal{E}} \frac{1}{2} w_{v_1,v_2} \left\| p_{v_1} - p_{v_2} \right\|^2, \quad (6)
\]

where \( p_{v_1} \) and \( p_{v_2} \) are the AQI distributions at nodes \( v_1 \) and \( v_2 \), \( \left\| p_{v_1} - p_{v_2} \right\| = D_{KL}(p_{v_1} \| p_{v_2}) + D_{KL}(p_{v_2} \| p_{v_1}) \) denotes the similarity of AQI distributions between \( p_{v_1} \) and \( p_{v_2} \), described by the Symmetrical Kullback-Leibler (KL) Divergence [17]. Thus, the objective function is given by:

\[
p^* = \arg \min_p L(p), \quad (7)
\]

By minimizing \( L(p) \), 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 (7) is the average of the distributions at its neighboring nodes.

**Proof:** According to [21], the minimum function in (7) is harmonic. Therefore, we have \( \Delta p_u = 0 \) on unlabeled nodes \( U \), while \( \Delta p_l = P(v_1) \) on labeled nodes \( L \). Here \( \Delta \) 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 (5). The harmonic property provides the form of solution as:

\[
p_u(x) = \frac{1}{d_u} \sum_{(u, j) \in \mathcal{E}} w_{u,j} p_j(x), \quad x \in \{0, 1, 2, \ldots, X\}, \quad (8)
\]

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

\[
p_u(x) = \frac{\sum_{(u, j) \in \mathcal{E}} w_{u,j} p_j(x)}{d_u \sum_x p_j(x)} = \frac{\sum_{(u, j) \in \mathcal{E}} w_{u,j} p_j(x)}{d_u \sum_{x \in \mathcal{X}} p_j(x)}. \quad (9)
\]

Hence, the distribution of unlabeled nodes \( p_u \) is the average of distributions at its neighboring nodes.

**Proposition 2.** \( p_u \) is a probability mass function (PMF).

**Proof:** To be a PMF on \( x \), 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 \( x \).
- \( \forall x \in \mathcal{X}, 0 \leq p_u(x) \leq 1 \).
- \( \sum_{x \in \mathcal{X}} p_u(x) = 1 \).

Considering the expression form in (9), the conclusion is obvious, that \( p_u \) is a PMF on \( x \).

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

**C. 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, \( P(\cdot | \lambda) \) (see Fig. 4). This conditioning allows the learned 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 \( C^{(j)} \) to be inferred. For each \( R_j \), the aerial sensing provides a conditioning \( \lambda_j \) for \( C^{(j)} \):

\[
\lambda_j : \{ \forall x_i \in \left[ X_{\min}^{(j)}, X_{\max}^{(j)} \right], \quad \forall C_i \in C^{(j)} \}. \quad (10)
\]

By applying \( P(\cdot | \lambda) \) to \( p_u \) in (9), we finally induce \( p_u(x | \lambda) \) 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 (5). \( \theta_m \) 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 \( \theta_0 \) is infeasible. Instead, we use model’s 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 \( H(p_u) \) as

\[
H(p_u) = \frac{1}{|U|} \sum_{i=1}^{|U|} H(p_i)
\]

\[
= -\sum_{i=1}^{|U|} \sum_{x=1}^N p_i(x = x|\lambda) \log p_i(x = x|\lambda), \quad \lambda \in \mathcal{C}^{(L)}.
\]

where \(|U|\) is the number of unlabeled nodes. By investigating \( \frac{\partial H}{\partial \theta} \) based on (9) and (5), \( \theta_0 \) is iteratively updated. Thus, the edge weights \( W \) 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 \( P(\lambda) \). When the model converges, we obtain the determined AQI distribution \( \hat{p}_u \) over \( \mathcal{V}_j \), 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 \( x \), we quantify it using the expectation of \( p_u \):

\[
\hat{p}_u = \mathbb{E}_{x \sim p_u} \left[ x \right] = \sum_{x=1}^N x \cdot p_u(x = x|\lambda).
\]

Note that we can obtain \( \hat{p}_u(T_i) \) 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 as real-time inference completed, and further use it as historical data in the future. By doing so, more nodes \( \mathcal{V}_k \) are labeled, thus 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

We only need to wake up a small number of ground sensors in selected regions to sense data at each \( T_j \) for saving power consumption, since the inference model is able to infer AQI at unobserved locations based on the sensed labeled data.

A. Voronoi Diagram based Region Division

Since the total monitoring space can be very large, we first divide it into disjointed regions \( \{R_1, R_2, \ldots, R_k\} \) 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 \( C^{(j)} \) inside each \( R_j \) are provided an AQI scale conditioning \( \lambda_j \) using vision-based inference. Since the distribution of ground devices is heterogeneous and uneven, we implement the division 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 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, 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 [22]. 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 \( \phi_j \) in \( R_j \) using the mean 2D coordinates of \( n_j \) devices inside it. The coordinates of center \( \phi_j \) 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 weighted distance as:

\[
D(y, \phi_j) = \frac{d(y, \phi_j)}{\sqrt{n_j}} = \frac{\|y - \phi_j\|_2}{\sqrt{n_j}},
\]

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

\[
V(\phi_i) = \bigcap_{j \neq i} \{ y \mid D(y, \phi_j) \leq D(y, \phi_j) \}, \quad i, j \in [1, k].
\]
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 stored AQI inference over past \( d \) time stamps. Therefore, for each node \( C \), there are two estimations: (1) AQI scale \( [X_{\text{min}}, X_{\text{max}}] \), and (2) pre-inferred value \( \tilde{X} \) 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, \( \delta \) and \( \Delta \).](image)

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

\[
\delta = \left| \frac{\tilde{X} - X_{\text{min}} + X_{\text{max}}}{2} \right|, \quad (\text{DoB}) \tag{15}
\]

\[
\Delta = X_{\text{max}} - X_{\text{min}}, \quad (\text{DoV}) \tag{16}
\]

where we call \( \delta \) as Degree of Bias (DoB), \( \Delta \) as Degree of Variance (DoV), as shown in Fig. 6. Intuitively, when \( \Delta \) is low, the variance of the AQI scale prior is small, which means a more reliable inference; as for \( \delta \), a low \( \delta \) 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 combining the two metrics, we define the Joint Estimation Error (JE) as:

\[
JE = \frac{1}{2} \left( \frac{\delta}{\delta_m} + \frac{\Delta}{\Delta_m} \right). \tag{17}
\]

where \( \delta_m \) and \( \Delta_m \) 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 \( T_k \), 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.

**Wake-up Mechanism Design:** JE can guide the system waking up selected devices at each time stamp. First, if the two priors \( [X_{\text{min}}, X_{\text{max}}] \) and \( \tilde{X} \) are both less than a pre-given AQI value \( \sigma \), 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 (see Fig. 7(a)). In this case, waking up connected nodes would be redundant as their measurements are similar. Denote \( M \) as the set of selected wake-up nodes using JE, our objective is to find a subset \( S \subseteq M \) such that (1) nodes in \( S \) are not adjacent, and (2) every node not in \( S \) is adjacent to at least one member of \( S \). This problem is well-known as the minimum independent dominating set problem [23], which is NP-hard. Since \( M \) is sparse in our case, the computation overhead is small, we simply apply a greedy method to find \( S \). Note that the algorithm is applied in each region independently. Fig. 7 shows the total process of finding a final wake-up node set \( S \).

![Fig. 7. An example of final wake-up set \( S \) construction: (a) the selected wake-up nodes set \( M \) based on JE; (b) the spatial neighbors within radius \( r \) in graph model; (c) finding the min independent dominating set \( S \).](image)
The vision-based sensing works online campuses since Feb. 2018. Throughout more than a half year’s services and a UA V, and it has been deployed on two university A. Experiment Setup and Data Collection

We collected 17,630 labeled images in different places. We calibrated sensor to label images with ground truth AQI value. Truth data for training, we set up the dataset by carrying video streams between equal time intervals. To get ground continuously and real-timely by sampling images from UA V (for 1000 times to avoid stochastic errors. Validation by randomly choosing the training data, and repeat into training set and testing set, while performing an cross-measured data for most cubes, we divide labeled samples both real-time and near-future estimation. Since there are no measurements 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.

C. Inference Accuracy of ImgSensingNet

We evaluate the inferred values deviate from real values, using root mean square error (RMSE). It is shown that the proposed model performs the best by maintaining a very low deviation. This proves the advances in using 3D model and feature extraction.

D. Energy Efficiency

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. In Fig. 9(b), as JE = 0, our system achieves the maximum consumption, which outperforms AQNet. As JE = 0.5, the consumption scales down.

E. Saving Energy by Wake-up Mechanism

We analyse the impact of r on wake-up mechanism. We vary the number of devices as 30 and 100, and set k = 5.

Average Number of Wake-up Devices: As shown in Fig. 10(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. Moreover, by choosing a proper r, the number of wake-up devices greatly scales down, which is energy efficient.

Average Runtime of Wake-up Mechanism: Further, we study the runtime for obtaining the set of wake-up devices each time. In Fig. 10(b)(d), the runtime also decreases with a greater r. Given more devices, the computation time increases, but it is still completed in real-time (about 1s in Fig. 10(d)).

F. Impact of Degree of Air Pollution

In Fig. 11, 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. 11(a) we compare the inference accuracy when AQI values are different. As a result, our 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: Fig. 11(b) reports the normalized consumption in different AQI degrees. Our system maintains lower consumption when AQI is low, which again validates the energy-efficiency of the wake-up mechanism. By comparing Fig. 11(a) and 11(b), an inherent trade-off between system consumption and inference accuracy can be illustrated.

VIII. Conclusion

This paper presents the design, technologies and implementation of ImgSensingNet, a 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 spatio-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, and 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/.
[4] T. Quang et al., “Vertical particle concentration profiles around urban office buildings,” Atmos. Chem. Phys., vol. 12, pp. 5017-5030, 2012.
[5] Y. Cheng et al., “Aurcloud: a cloud-based air-quality monitoring system for everyone,” in Proc. ACM SenSys, New York, NY, Nov. 2014.
[6] Y. Gao et al., “Mosaic: a low-cost mobile sensing system for urban air quality monitoring,” in Proc. IEEE INFOCOM, Jul. 2016.
[7] D. Hasenfratz et al., “Deriving high-resolution urban air pollution maps using mobile sensor nodes,” Pervasive and Mobile Computing, Jan. 2015.
[8] J. Li et al., “Tethered balloon-based black carbon profiles within the lower troposphere of shanghia in the east china smog,” Atmos. Environ., 2015.
[9] Y. Hu et al., “BlueAer: a fine-grained urban PM2.5 3D monitoring system using mobile sensing,” in Proc. IEEE INFOCOM, Jul. 2016.
[10] Y. Yang et al., “AQNet: fine-grained 3D spatio-temporal air quality monitoring by aerial-ground WSN,” in Proc. IEEE INFOCOM, Apr. 2018.
[11] Y. Yang et al., “Real-time profiling of fine-grained air quality index distribution using UAV sensing,” IEEE Internet of Things Journal, 2018.
[12] Z. Pan, H. Yu, C. Miao, and C. Leung. “Crowdsensing air quality with camera-enabled mobile devices,” in Proc. IIAI, 2017.
[13] S. Li et al., “Inferring fine-grained PM2.5 with bayesian based kernel method for crowdsourcing system,” in Proc. IEEE Globecom, 2017.
[14] R. Gao et al., “Sextant: towards ubiquitous indoor localization service by photo-taking of the environment,” IEEE Trans. Mobile Comput., 2016.
[15] Y. Yang et al., “Arms: a fine-grained 3D AQI realtime monitoring system by UAV,” in Proc. IEEE Globecom, Singapore, Dec. 2017.
[16] V. Li et al., “Deep learning model to estimate air pollution using M-BP to fill in missing proxy urban data,” in Proc. IEEE Globecom, 2017.
[17] 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,” Proc. IEEE ICC, May 2018.
[18] K. He, J. Sun, and X. Tang, “Single image haze removal using dark channel prior,” in Proc. IEEE CVPR, Miami, FL, Jun. 2009.
[19] Q. Zhu, J. Mai, and L. Shao, “A fast single image haze removal algorithm using color attenuation prior,” IEEE Trans. Image Process., 2015.
[20] 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.
[21] X. Zhu et al., “Semi-supervised learning using gaussian fields and harmonic functions,” in Proc. ICML, Washington, DC, Aug. 2003.
[22] F. Aurenhammer, “Voronoi diagrams: a survey of a fundamental geometric data structure,” ACM Comput. Survey, vol. 23, pp. 345-405. 1991.
[23] N. Bourgeois et al., “Fast algorithms for min independent dominating set,” Discrete Applied Mathematics, vol. 161, pp. 558-572. Mar. 2013.