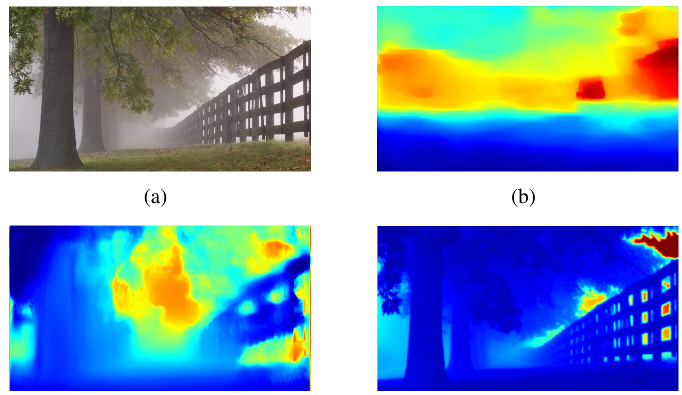
Discern Depth Under Foul Weather: Estimate $\mathrm{PM}_{2.5}$ for Depth Inference

Kun Li, *Member, IEEE*, Jian Ma*, Han Li*, Yahong Han, *Member, IEEE*, Xibin Yue, Zihao Chen, and Jingyu Yang, *Senior Member, IEEE,*

***Abstract*—Nowadays, haze is a common and serious problem and $\mathrm{PM}_{2.5}$ is a main measurement for air quality. Current methods estimate the level of primary pollutant with professional instruments which is expensive and inconvenient. Moreover, with haze, the captured images will be unclear and are difficult to estimate the depth of scene using passive methods. This paper proposes a cheap, fast, and convenient $\mathrm{PM}_{2.5}$ estimation method which only need a captured image using daily-life devices, and further discerns the depth of scene using the estimated $\mathrm{PM}_{2.5}$. We learn haze-relevant classified mapping via hybrid convolutional neural network and combine the high-level features extracted from convolutional layer with ground-truth $\mathrm{PM}_{2.5}$ to train support vector regression (SVR). The transmission map is computed using non-local sparse priors, and the depth map is inferred using the estimated $\mathrm{PM}_{2.5}$ value through the atmospheric scattering model. Experimental results demonstrate that our method achieves accurate $\mathrm{PM}_{2.5}$ estimation and depth inference. This could be very useful in many applications, for both clean and foul weather.**

***Index Terms*—Convolutional neural network, $\mathrm{PM}_{2.5}$ estimation, depth estimation, SVR.**



(a) (b) (c) (d)

Fig. 1. Depth inference results: (a) a color image captured in a hazy environment, (b) estimated depth by conventional method [2], (c) estimated depth by learning-based method [5], and (d) estimated depth by our method.

I. Introduction

AIR pollution is a serious problem nowadays, which is very harmful to people's health. In order to reduce the damage for people, the level of major pollutants (*e.g.* $\mathrm{PM}_{2.5}$) need be quickly and accurately estimated in daily life. Existing methods measure $\mathrm{PM}_{2.5}$ values with special devices, *e.g.* Hanvon M1[1]. However, such special devices cost a lot and they are inconvenient for people to carry everywhere. Although some weather softwares, *e.g.* Moji, can provide $\mathrm{PM}_{2.5}$ values, the detection stations are limited and the provided $\mathrm{PM}_{2.5}$ value may not accurate for the location of user. Therefore, it is very important and urgently needed to accurately measure $\mathrm{PM}_{2.5}$ values using daily-life devices.

On the other hand, it is difficult to estimate the depth of scene in the case of haze using passive image-based methods [1], [2], especially from a single image. Due to the presence of a large amount of solid particles in the air, the atmospheric light is weakened to some extent, resulting in the blurring of the images. This poses a serious challenge for depth estimation, especially for depth estimation from a single image. Traditional vision-based methods, even those based on deep learning, cannot extract the depth from the image under foul weather. Transmission map [3], [4] is proved to be relevant to depth map, and can be used for depth estimation.

This paper proposes a $\mathrm{PM}_{2.5}$ estimation method only using a captured image via a new hybrid convolutional neural network (H-CNN). The image is segmented into two parts: sky and non-sky, and is fed into the the proposed PMnet, together with dark channel maps. A depth estimation method is also proposed by learning the relationship between $\mathrm{PM}_{2.5}$ and atmospheric attenuation coefficient β. We also collect a large dataset containing $\mathrm{PM}_{2.5}$ values and the corresponding images, which will be made publicly available. Our method can work in both clean and foul weather. Experimental results show that our method has low computational complexity and generates best results compared with state-of-the-art methods. Fig. 1 shows an example, in which our estimated $\mathrm{PM}_{2.5}$ value is 180.198 $\mu g/m^3$).

The main contributions are summarized as follows:

- An image-based $\mathrm{PM}_{2.5}$ estimation method based on deep learning. Instead of carrying complex equipment, the user only need to capture a photograph of the current scene by any camera device, such as a smartphone. Our algorithm achieves fast estimation of $\mathrm{PM}_{2.5}$ value.
- A hybrid CNN for learning distance-aware haze-relevant

* Contributed equally.

This work was supported in part by the National Natural Science Foundation of China (Grant 61571322 and Grant 61771339), and Tianjin Research Program of Application Foundation and Advanced Technology under Grant 18JCYBJC19200.

Corresponding author: Jingyu Yang (yjy@tju.edu.cn)

Kun Li, Jian Ma, Han Li, and Yahong Han are with College of Intelligence and Computing, Tianjin University, Tianjin 300350, China.

Xibin Yue and Zihao Chen are with Meteorology Research Department, Moji Co. Ltd, Beijing 100016, China.

Jingyu Yang is with School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China.

[1]http://en.hw99.com/show-7-8-1.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2943631, IEEE Transactions on Industrial Informatics

features with segmentation and classification priors. The accuracy is up to 92.19%.

- A depth estimation method using non-local sparse priors. We learn the relationship between $PM_{2.5}$ and atmospheric attenuation coefficient β using a synthesized dataset.
- A $PM_{2.5}$ image dataset. The dataset contains over 10000 images together with the $PM_{2.5}$ values of the corresponding scenes. The dataset and the code will be available online.

The remainder of this paper is structured as follows. Related work is summarized in Sec. II. We introduce our new $PM_{2.5}$ image dataset in Sec. III, and propose a $PM_{2.5}$ estimation model in Sec. IV. A monocular depth estimation method is proposed in Sec. V, and the proposed method is evaluated with experiments on both synthetic and real datasets in Sec. VI. The paper is concluded in Sec. VII.

II. Related Work

A. $PM_{2.5}$ Estimation

At present, $PM_{2.5}$ is a very important indicator for evaluating air quality. The main methods [6] used special devices to measure the $PM_{2.5}$ values, which is accurate but expensive and inconvenient. Tao *et al.* [7] achieved real-time $PM_{2.5}$ measurement, and Gu *et al.* [8] proposed a heuristic recurrent air quality predictor to infer air quality based on meteorology- and pollution-related factors. Despite accuracy, these methods are expensive and inconvenient for daily-life use. To address this problem, Zhang *et al.* [9] proposed a convolutional neural network to estimate air pollution levels from a single image, and had good classification performance on himself dataset. Chakma *et al.* [10] used VGGNet-19 [11] features and a random forest classifier to classify natural images into different pollution levels. However, these methods have a limited accuracy and cannot estimate the specific $PM_{2.5}$ values. Ma *et al.* [12] proposes an image-based $PM_{2.5}$ estimation method based on VGG features which are not haze-relevant features and hence have limited estimation accuracy.

In this paper, we propose a PMnet to learn haze-relevant features and estimate the specific $PM_{2.5}$ value based on support vector regression (SVR) from a captured image. Our method can work in both clean and foul weather. We also collect a larger scale $PM_{2.5}$ image dataset containing over 10000 images together with the $PM_{2.5}$ values, which will be available online.

B. Depth Estimation from a Single Image

Depth estimation from a single image is a challenging topic in computer vision, which can be divided into four kinds of methods: conditional random fields (CRF)-based methods, non-parametric methods, deep CNN methods and air medium transmission methods.

1) CRF-based Methods : CRF-based methods are the earliest and most classical algorithms, which have profound guiding significance in the filed of computer vision. They usually strictly assume that images are composed of horizontal planes, vertical walls and superpixels. Saxena *et al.* [13] introduced a discriminatively-trained MRF (Markov random fields) into the model and incorporated both the local and global features so that it could model the depth at every point very well. Then, they utilized similar approach to propose a 3-D depth estimation algorithm [14] and obtained satisfactory performance. In order to estimate more accurate depth, Saxena *et al.* [15] presented a more general method by combining monocular and stereo cues together. In addition, simple geometric assumption is made for indoor scenes which are proven to be useful [16].

2) Non-parametric Methods: Non-parametric methods use the similarities between regions and assume that similar regions generally indicate similar depth. Those methods usually migrate depth information from existing RGB-D datasets into the input RGB image via firstly finding the candidate RGB-D that best matches the input image based on the high-level image features, and then aligning the image pairs or other operations to obtain the final depth map. One of the most prominent methods, proposed by Karsch *et al.* [2], transferred depth from the RGB-D dataset to the input RGB image based on SIFT flow [17], and incorporated temporal information into the depth estimation procedure to better optimize the consequent depth map. Konrad *et al.* [18] selected k candidate pairs by kNN (k nearest neighbourhood) searching method, and fuse k depth fields by a median filter followed by smoothing using a cross-bilateral depth filter. Mebtouche *et al.* [19] took local dissimilarities into account and proposed to extract sub-regions which matched the input RGB image best and then used these sub-regions to estimate the desired depth map.

3) Deep CNN Methods: In recent years, deep learning based methods have made remarkable breakthroughs in the field of computer vision, which have also tremendously improved the accuracy of the recovered depth map. Eigen *et al.* [20] proposed a multiscale convolutional neural network including two deep network stacks: one is to estimate a coarse depth map globally and the other is to refine this coarse version locally. Wang *et al.* [21] jointly inferred depth map and semantic segmentation through a hierarchical CRF combining region-wise and pixel-wise potentials generated by a regional CNN and a global CNN, respectively. Liu *et al.* [22] and Xu *et al.* [23] combined deep convolutional neural network and continuous CRF into a unified framework for monocular depth estimation. Godard *et al.* [5] proposes a unsupervised method for monocular depth estimation.

4) Air Medium Transmission Methods: Under foul weather, depth estimation from images becomes more difficult, because the suspended particles affect the clarity of the captured images. None of the above methods can generate satisfactory results. Even deep learning methods cannot accurately estimate the depth of a single image as well due to the lack of ground truth. Fortunately, it is proved that the atmospheric transmission map is relevant to the depth map, thus we can estimate depth map from transmission map. He *et al.* [4] proposed dark channel prior to help compute the transmission map and removed haze from a single image. Berman *et al.* [24] used non-local color-lines to estimate atmospheric transmission. Chen *et al.* [25] refined the transmission map based on total generalized variation (TGV) for reliable dehazing.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2943631, IEEE Transactions on Industrial Informatics

In these methods, the atmospheric attenuation coefficient is randomly selected at $[0.5, 1.5]$ which might limit the accuracy of estimated depth. This paper proposes an estimation method for atmospheric attenuation coefficient based on $\text{PM}_{2.5}$ and achieves promising results for depth estimation.

III. Dataset

Our dataset contains two parts: subdataset-A captured by us in Tianjin, China, and subdataset-B provided by Beijing Moji Wind Technology Co., Ltd, which is a well-known weather information provider.

A. Subdataset-A

Because there is no public dataset with color images and the associated $\text{PM}_{2.5}$ values, we capture 1575 images of different scenes using an Apple 5s mobile phone, and simultaneously measure the corresponding $\text{PM}_{2.5}$ values of the current scenes with Hanvon M1 which measures $\text{PM}_{2.5}$ values with high precision. To better avoid the influence of fog, we collected the images after 10:00 am everyday because fogs naturally evaporate with the increasing temperature. The collected $\text{PM}_{2.5}$ values are between 0-300 $\mu g/m^3$, which are categorized into three classes: Good (PM2.5 < 75), Moderate (PM2.5 $\in [75, 150]$), and Severe (PM2.5 ≥ 150).

B. Subdataset-B

Moji dataset is an air pollution image dataset that contains 9630 captured images by users with associated air pollution parameters and weather conditions. The air pollution parameters include carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), $\text{PM}_{2.5}$ and PM_{10}. The weather conditions include weather, temperature, humidity, wind speed, wind direction and air pressure. All the data is collected by the National Air Quality Monitoring Station and provided by Beijing Moji Wind Technology Co., Ltd [2]. Note that we only use the captured images and the associated $\text{PM}_{2.5}$ values for learning.

IV. $\text{PM}_{2.5}$ Estimation

In order to obtain accurate $\text{PM}_{2.5}$ estimation, haze-relevant features need to be learnt. Hence, we propose a hybrid CNN to extract distance-aware haze-relevant features and then learn the mapping between the features and $\text{PM}_{2.5}$ values by support vector regression (SVR).

A. Haze-relevant Features

Given a single RGB image I, we first segment the image into sky and building parts, and then propose a PMnet to extract the haze-relevant features which need a dark channel map as implicit representation fed into a secondary subnetwork. The dark channel map is computed as the minimum value of light intensity of the region [4]. Fig. 2 shows the architecture of our network. The network extracts haze-relevant features by combining the features of the sky and building parts and the pollution level (0, 1, 2) of the whole image. Then, $\text{PM}_{2.5}$ is estimated by regression learning. The feature of each part is extracted by PMnet. The PMnet contains two subnetworks: a residual network [26] to extract photometric features from the color image, and a VGG network [11] to extract the implicit features from the dark channel map.

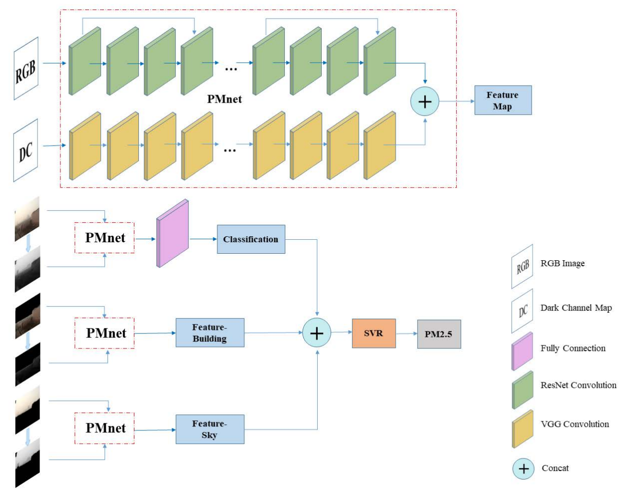


Fig. 2. The architecture of our model.

We extract the haze-relevant features by:

$$F_{Haze} = \omega_1 F_{\text{Sky}} \oplus \omega_2 F_{\text{Building}} \oplus F_{\text{Class}}, \tag{1}$$

where F_{Sky} and F_{Building} represent the features extracted by the sky and building parts of images, respectively. F_{Class} represents the classification result of the whole input image, $\oplus$ is the operation of dimension splicing, and F_{Haze} is the final haze-relevant features.

Distance-aware Segmentation. The selection and fusion of features is important for accurate $\text{PM}_{2.5}$ estimation. It is well-known that the $\text{PM}_{2.5}$ value is proportional to the haze concentration which influences the degree of image blurring. The objects in the distant scene will be most blurred. We assume that the sky is in infinity and the building is in an observable location. Therefore, we segment an image into sky part and building part for feature extraction.

We first use the K-means algorithm to segment the image into two parts, using color information of each part with the prior that the blue channel value in the sky part is higher. However, if the buildings contain glass or other reflective objects, the segmentation by K-means may be wrong. Therefore, we adopt a neighborhood averaging optimization method, which fully considers the neighborhood information of each pixel and assumes that a single pixel has a high degree of similarity to surrounding pixels belonging to the same object. We use a 3×3 kernel to traverse the entire image, iterating and updating. Fig. 3 shows the segmentation results with and without optimization.

Training. We train the PMnet by minimizing the *Softmax* loss function between the estimated level and the ground truth:

$$\delta(y, z) = -\log\left(\frac{e^{z_y}}{\sum_j e^{z_j}}\right), \tag{2}$$

[2]http://www.moji.com

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2943631, IEEE Transactions on Industrial Informatics

(a) (b) (c) (d) (e)

Fig. 3. Segmentation results for a hazy image: (a) raw image, (b) original sky segmentation result, (c) optimized sky segmentation result, (d) original building segmentation result, and (e) optimized building segmentation result.

where z_j is the feature of the j^{th} image and y is the air pollution level of the image.

We use random crop and rotation (90, 180, 270 degrees) to extend our subdataset-A dataset for data augmentation. 1375 images are used for training and the final extended subdataset-A has 18574 images. The other 200 images are taken as the test set. For subdataset-B dataset, we use 7693 images for training and 1937 images for testing. We resize the training images to the size of 224×224 and use mini-batches with size of 8 to best compromise between speed and convergence. We use pre-trained models [11], [26] on ImageNet to initialize weights and train the network using the ADAM solver [27] with a learning rate of 0.0001 and "step" as learning rate of decline strategy. The momentum is set to be 0.9.

B. Regression Machine

We learn the mapping f between the $\mathrm{PM}_{2.5}$ values and the features extracted from our PMnets by SVR:

$$\begin{cases} \min\limits_{\omega,b} \frac{1}{2}||\omega||^2 + c\sum\limits_{i=1}^{m} \ell_{\in}(f(x_i) - y_i) \\ f(x_i) = \omega^T x_i + b \\ \ell(z)_{\in} = \begin{cases} 0, \text{if}|\text{z}| \leq \varepsilon \\ |z| - \varepsilon, \text{otherwise}, \end{cases} \end{cases} \tag{3}$$

where ω is the normal vector representing the direction of the hyperplane, x_i is a feature of the i^{th} image, y_i is the ground-truth $\mathrm{PM}_{2.5}$ value of the i^{th} image, c is a regularization constant, and b is the displacement between the hyperplane and the origin.

V. Monocular Depth Inference Model

A hazy image is usually formulated as [28]:

$$I(\mathbf{x}) = J(\mathbf{x})t(\mathbf{x}) + A[1 - t(\mathbf{x})], \tag{4}$$

where J is the true radiance of scene, t is the medium transmission, A is the global atmospheric light composition, and I is the captured hazy image. The medium transmission t depends on the depth of scene $d(\mathbf{x})$:

$$t(\mathbf{x}) = \mathrm{e}^{-\beta d(\mathbf{x})}, \tag{5}$$

where $\beta (\beta > 0)$ is atmospheric scattering coefficient. Through mathematical transformation from Eq. (5), we can obtain

$$d(\mathbf{x}) = -\frac{1}{\beta} \ln t(\mathbf{x}), \tag{6}$$

Therefore, we need estimate transmission map $t(\mathbf{x})$ and β to compute the depth map of scene.

A. Transmission Estimation

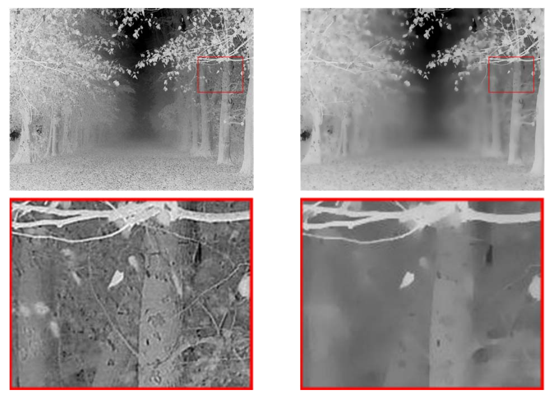
Some methods [4], [24], [25] recovered the clean image $J(\mathbf{x})$ from $I(\mathbf{x})$ by estimating the transmission $t(\mathbf{x})$. In this paper, we propose a new method to estimate the transmission map with non-local sparse priors.

Initial estimation: Using dark channel prior, we can calculate the initial transmission $\widetilde{t}(\mathbf{x})$ [4]:

$$\widetilde{t}(\mathbf{x}) = 1 - w \min_c \{ \min_{\mathbf{y} \in \Omega(\mathbf{x})} \frac{I^c(\mathbf{y})}{A^c} \}, \tag{7}$$

where w is an environmental factor set to be 0.95, $\Omega(\mathbf{x})$ is a local patch centered at $\mathbf{x}$, A^c is the atmospheric light component of each channel c calculated by the method in [4], and I^c is a color channel of the observed hazy image I.

Refinement: The initial transmission map using the dark channel prior shows a good performance on haze removal, but much texture information is kept, as shown in Fig. 4(a). Hence, we optimize the transmission map by minimizing the following cost function:



(a) (b)

Fig. 4. Transmission estimation result: (a) initial transmission $\widetilde{t}(\mathbf{x})$ and (b) optimized transmission $t(\mathbf{x})$.

$$\sum_{\mathbf{x}} \frac{\left(t(\mathbf{x}) - \widetilde{t}(\mathbf{x})\right)^2}{\sigma^2(\mathbf{x})} + \lambda \sum_{\mathbf{x}} \sum_{\mathbf{y} \in N(\mathbf{x})} \sqrt{\alpha_{x,y}} \|t(\mathbf{x}) - t(\mathbf{y})\|_1, \tag{8}$$

where $\sigma(\mathbf{x})$ is the standard deviation of $\widetilde{t}(\mathbf{x})$, $N(\mathbf{x})$ represents the neighborhood set of pixel $\mathbf{x}$, $\|\cdot\|_1$ represents the ℓ_1 norm, λ is a penalization parameter, and $\alpha_{x,y}$ is a pairwise weight calculated by

$$\alpha_{x,y} = \exp\left(\frac{||\mathbf{B}_x \circ (\mathcal{P}_x - \mathcal{P}_y)||_2^2}{\vartheta_1^2} \right), \tag{9}$$

where $\mathcal{P}_x$ ($\mathcal{P}_y$) is an operator extracting a $w \times w$ patch centered at $\mathbf{x}$ ($\mathbf{y}$) on the hazy image I, $\circ$ represents the element-wise multiplication, ϑ_1 determines the decay rate of exponential function, and B_x is a bilateral filter kernel defined as

$$\mathbf{B}_x(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{||\mathbf{x} - \mathbf{y}||_2^2}{\vartheta_2^2} \right) \exp\left(-\frac{\sum\limits_{i \in c} (I_x^i - I_y^i)^2}{\vartheta_3^2} \right), \tag{10}$$

where ϑ_2 and ϑ_3 are constant parameters adjusting the spatial range and the intensity range, respectively. As shown in Fig.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2943631, IEEE Transactions on Industrial Informatics

4(b), our optimized transmission map is more accurate and less messy.

In our formulation, both the distance of local patches and the similarity between the pixel $\mathbf{x}$ and every pixel $\mathbf{y}$ in the neighborhood of $\mathbf{x}$ are evaluated. We use non-local prior weighted by a bilateral kernel on a larger neighborhood to fully exploit structural correlation, and adopt ℓ_1 norm to model the piecewise smoothness of the transmission map. Our optimized transmission map is accurate without loss of smoothness, as shown in Fig. 4 (b).

Minimization: We define the following matrices and variables to reformulate the cost function in Eq. (8):

$$T = [t(1,1), t(1,2), \cdots, t(w,h)], \tag{11}$$

$$\tilde{T} = [\tilde{t}(1,1), \tilde{t}(1,2), \cdots, \tilde{t}(w,h)], \tag{12}$$

$$W = diag(\frac{1}{\sigma(x_1)} \frac{1}{\sigma(x_2)} \cdots \frac{1}{\sigma(x_n)}), \tag{13}$$

$$L = \{e_{x,y} | (x,y) \in M\}, \tag{14}$$

where (w,h) is the size of the image, T is the matrix representation of the transmission $t(\mathbf{x})$, $\tilde{T}$ is the matrix representation of the initial transmission $\tilde{t}(\mathbf{x})$, n denotes the number of pixels, $diag(\cdot)$ represents a diagonal array, and thus W is an n-order diagonal matrix. $\sigma(x)$ is the standard deviation of $\tilde{T}$, and x is a pixel. $e_{x,y}$ represents the edge between pixel $\mathbf{x}$ and pixel $\mathbf{y}$, and M is the collection of pairs of four-neighborhood pixels.

Define a matrix Q, each row of which corresponds to an edge in L and each column of which corresponds to a pixel in the image. Each row in Q has only two nonzero entries. Supposing the r^{th} row of Q associates with edge $e_{x,y}$ of L, the value of (r,x) is $\sqrt{\alpha_{x,y}}$ and the value of (r,y) is $-\sqrt{\alpha_{x,y}}$.

Let $A = QT$, then Eq. (8) can be rewritten as

$$\left\| W(T - \tilde{T}) \right\|_2^2 + \lambda \left\| QT \right\|_1 . \tag{15}$$

We solve the constrained minimization Eq. (15) using the augmented Lagrangian method (ALM), which converts the original problem to an iterative minimization of its augmented Lagrangian function:

$$\left\| W(T - \tilde{T}) \right\|_2^2 + \lambda \left\| A \right\|_1 + \langle Y, A - QT \rangle + \frac{\mu}{2} \left\| A - QT \right\|_2^2, \tag{16}$$

where λ is penalty coefficient, $\|\cdot\|_2$ denotes the l_2 norm, μ is a constant of a positive number, Y is a Lagrangian multiplier, and $\langle \cdot, \cdot \rangle$ denotes the inner product of two matrices considered as long vectors. Y and μ can be updated efficiently using ALM, however, each iteration has to solve T and A simultaneously. Hence, we use alternate direction method [29] to optimize T and A separately at each iteration:

$$\begin{cases} A^{(k+1)} = \arg_A \min \lambda \|A\|_1 + \langle Y^{(k)}, A - QT^{(k)} \rangle \\ \quad + \frac{\mu^{(k)}}{2} \left\| A - QT^{(k)} \right\|_2^2, \\ T^{(k+1)} = \arg_T \min \left\| W(T - \tilde{T}) \right\|_2^2 + \langle Y^{(k)}, A^{(\mathrm{k+1})} - QT \rangle \\ \quad + \frac{\mu^{(k)}}{2} \left\| A^{(k+1)} - QT \right\|_2^2, \\ Y^{(k+1)} = Y^{(k)} + A^{(k+1)} - QT^{(k+1)}, \\ \mu^{(k+1)} = \rho \mu^{(k)}, \rho > 0. \end{cases} \tag{17}$$

B. *Atmospheric Attenuation Coefficient β Estimation*

When electromagnetic waves with various wavelengths propagate in the atmosphere, the absorption and scattering of the atmosphere of gas molecules (water vapor, carbon dioxide, ozone, *etc.*), water vapor condensate (ice crystals, snow, fog, *etc.*) and suspended particles (dust, smoke, salt, microorganisms) will form the absorption band which can weaken the energy of electromagnetic wave. Therefore, different weather with different $\mathrm{PM}_{2.5}$ values will have different atmospheric attenuation coefficient β.

We learn the relationship between $\mathrm{PM}_{2.5}$ and β by synthetic experiments. Specifically, we generate synthetic hazy images by adding artificial haze to haze-free color images in SYNTHIA SAN FRANCISCO dataset [30]. We choose a set of atmospheric attenuation coefficient β which is in the range of $[0.5, 4.5]$, and compute the medium transmission $t(\mathbf{x})$ by Eq. 6. The global atmospheric light composition A is set as a constant of $[0.755, 0.77, 0.77]$ for RGB channels, and the haze images are synthesized via Eq. (4).

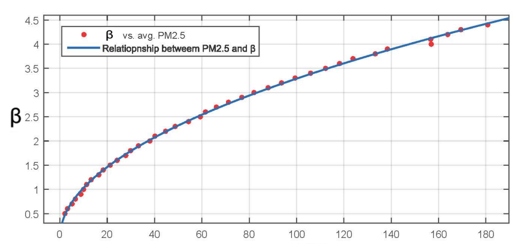


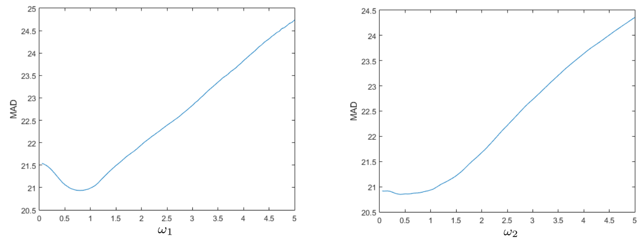
Fig. 5. Relationship between $\mathrm{PM}_{2.5}$ and β. X-axis represents the average of $\mathrm{PM}_{2.5}$ values and Y-axis represents the β.

We estimate the $\mathrm{PM}_{2.5}$ value for each synthesized hazy image using our proposed method, and then find the relationship between $\mathrm{PM}_{2.5}$ and β through a lot of statistical experiments as shown in Fig. 5. Considering the existence of errors, we take the average values of all the $\mathrm{PM}_{2.5}$ values which corresponds to the same β and fit the relationship between β and $\mathrm{PM}_{2.5}$. Fitted by least squares method, the relationship between $\mathrm{PM}_{2.5}$ and β is

$$\beta = ax^b, \tag{18}$$

where a and b are the parameters, and are learnt as 0.324 and 0.5032, respectively.

Hence, given a captured image, we first estimate the $\mathrm{PM}_{2.5}$ value by the proposed deep learning method in Sec. IV, and then compute the β according to Eq. (18). Finally, we estimate the transmission map using the proposed optimization method and obtain the depth map of the scene using Eq. (6).



(a) (b)

Fig. 7. $\mathrm{PM}_{2.5}$ estimation result using different values of ω_1 (a) and ω_2 (b).

VI. Experimental Results

In this section, we first evaluate the proposed $\mathrm{PM}_{2.5}$ estimation method with ablation study and comparison (Section VI-B), and then compare our proposed transmission estimation method with state-of-the-art methods (Section VI-C). Finally, we test the proposed depth estimation method on synthetic datasets and real datasets (Section VI-D). The running times of the proposed method are reported in Section VI-E.

A. Experimental Setup

We use libsvm [31], an open source library, to learn the relationship between features and $\mathrm{PM}_{2.5}$ values. The penalty coefficient c is learnt by cross entropy. In addition, The parameters for transmission optimization are set as follows: λ=0.1, ϑ_1= 3.05, ϑ_2= 1000, and ϑ_3= 0.2. The window size of bilateral filter is set at 5 and the number of ALM iterations is set at 7 (normally, 5 to 8 iterations can achieve the desired results). All comparison experiments use default parameters and KITTI pre-trained model is used to carry out the comparison experiments for deep learning method [5].

B. $PM_{2.5}$ Estimation

To evaluate regression performance, we randomly choose 3002 images for training and 647 images for test. Fig. 6 shows the comparison results between VGG-based method [12] and our proposed method. It can be seen that our method is more accurate than the VGG-based method [12]. Our predicted values are more concentrated and more accurate in each class (good, moderate, severe), and the result is best by using weights for segments for our method. The MADs (Mean Absolute Deviations) of VGG-based method [12], our method without weights and our method with weights are 59.42, 24.35, and 20.8478, respectively. Although VGG network generates distinguishable features, the features of RGB images are far from enough for more complex scenes, such as illumination changes and different weather conditions. We obtain more detailed and reasonable features by combining the features of dark channel maps and segmentation.

In order to evaluate the influence of the haze-relevant feature maps, we also compare five variants with different feature combinations in Table I. Sky and Building represent the features of sky part and building part, respectively. Sky *concat* Building with classification represents our method without weights for segments. It can be seen that our method with weights achieves the most accurate $\mathrm{PM}_{2.5}$ estimation.

We also evaluate the influence of weights ω_1 and ω_2 in Eq. (1) on the estimation accuracy by tuning each parameter over the interesting part of the parameter space while setting other parameters at the fixed reasonable values. Fig. 7 shows the MADs of estimation results using different parameters, which suggests that more accurate results can be achieved by setting $\omega_1 = 0.8$ and $\omega_2 = 0.4$. This demonstrates that the features of sky part are more critical to our task because sky is distant and is more visually distinguishable for different $\mathrm{PM}_{2.5}$ values.

TABLE I
Quantitative comparison with different variants.

Feature Combinations	MAD
Sky only	45.93
Building only	50.89
Sky + Building	46.10
Sky *concat* Building	45.26
Sky *concat* Building with classification	24.35
our method	**21.95**

C. Transmission Estimation

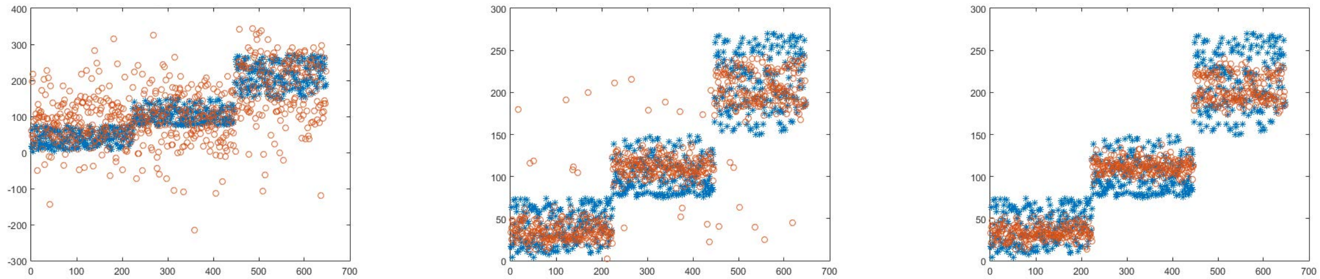
We evaluate our transmission estimation method quantitatively on a synthetic dataset, compared with three state-of-the-arts methods. The synthetic dataset is generated by artificially adding haze using the depth images according to Eq. (4) and Eq. (5) on the color images in the NYU-Depth v2 dataset [32]. Table II gives quantitative evaluation result. We use five commonly-used measurements for quantitative evaluation:

- Relative error (Rel): $\frac{1}{T}\sum_p \frac{|t_p^{gt}-t_p^{est}|}{t_p^{gt}}$;
- Root mean squared error (RMSE): $\sqrt{\frac{1}{T}\sum_p \left(t_p^{gt}-t_p^{est}\right)^2}$;
- $\log_{10}$ error (log10): $\frac{1}{T}\sum_p |\log_{10} t_p^{gt} - \log_{10} t_p^{est}|$;
- PSNR: $10 \times log_{10}\frac{255^2}{RMSE^2}$;
- SSIM: $\frac{(2\mu_{gt}\mu_{est}+c_1)(2\sigma_{gt_est}+c_2)}{(\mu_{gt}^2+\mu_{est}^2+c_1)(\sigma_{gt}^2+\sigma_{est}^2+c_2)}$;

where t_p^{gt} is the ground-truth transmission at pixel p, t_p^{est} is the corresponding estimated transmission, and T is the number of image pixels. For SSIM, μ_{gt} and μ_{est} represent the mean values of the ground-truth transmission and the estimated transmission, σ_{gt} and σ_{est} represent the standard deviations of the two images, and σ_{gt_est} represents the covariance of the the ground-truth transmission and the estimated transmission. c_1 and c_1 are constants, which are set as $c_1 = (K_1 * L)^2$ and $c_2 = (K_2 * L)^2$ with $K_1 = 0.01$, $K_2 = 0.03$ and $L = 255$.

As shown in Table II, our method achieves the best results for all the measurements except Rel. Fig. 8 and Fig. 9 show some transmission estimation results on natural scenery images collected from Internet and a challenge dataset [33],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2943631, IEEE Transactions on Industrial Informatics



(a) (b) (c)

Fig. 6. PM$_{2.5}$ estimation results (PM$_{2.5}$ values *w.r.t.* indices of images): (a) VGG based method [12], (b) our method without weights, and (c) our method with weights. The blue $*$ is ground-truth and the orange o is predicted value.

respectively. Method [25] and method [24] generate coarse over-smoothed transmission maps without details, *e.g.*, the blurred leaves. Although method [4] can reflect the detailed depth of near scene, it retains a lot of texture information. On the contrary, the transmission maps generated by our method are accurate without loss of smoothness for both near scene and distant scene.

TABLE II
QUANTITATIVE EVALUATION FOR TRANSMISSION ESTIMATION.

Method	Lower is better			Higher is better	
	Rel	log10	RMSE	PSNR	SSIM
[24]	0.6633	0.6335	0.2896	59.2652	0.9866
[25]	0.5697	0.4596	0.1756	63.5177	0.9953
[4]	**0.4470**	0.3709	0.0980	68.8108	0.9984
Our Method	0.4886	**0.3280**	**0.0830**	**70.4581**	**0.9988**

D. Monocular Depth Estimation

In order to compare with other depth estimation methods, we use three commonly-used measurements for quantitative evaluation:

- Relative error (Rel): $\frac{1}{T}\sum_p \frac{|d_p^{gt}-d_p^{est}|}{d_p^{gt}}$;
- Root mean squared error (RMSE): $\sqrt{\frac{1}{T}\sum_p \left(d_p^{gt}-d_p^{est}\right)^2}$;
- $\log_{10}$ error (log10): $\frac{1}{T}\sum_p |\log_{10} d_p^{gt} - \log_{10} d_p^{est}|$;

where d_p^{gt} is the ground-truth depth at pixel p, d_p^{est} is the corresponding estimated depth, and T is the number of image pixels.

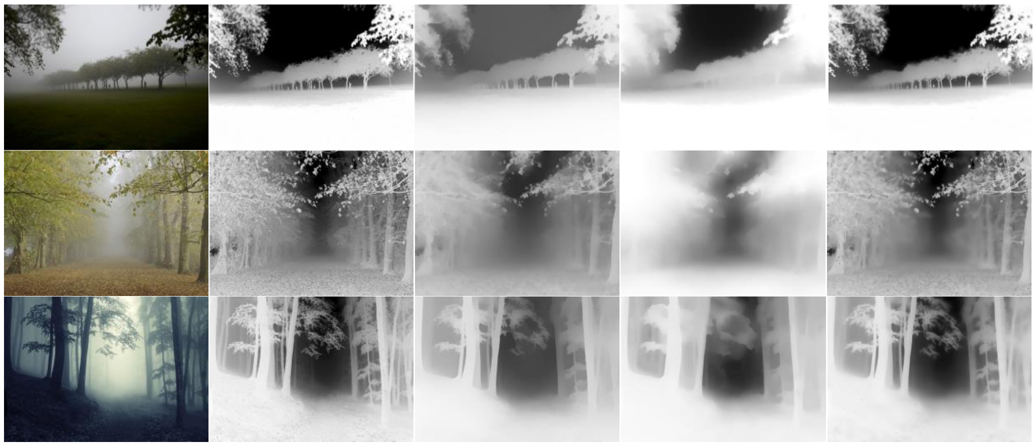
1) Synthetic Data: Synthetic dataset is generated by artificially adding haze using the depth images according to Eq. (4) and Eq. (5) on the color images in the NYU-Depth v2 dataset [32]. Table III gives quantitative evaluation result, compared with seven state-of-the-art methods. As shown in the table, our method has the smallest errors for all the measurements except Rel. Fig. 10 shows some visual effect of depth estimation of all the methods. In this case, the PM$_{2.5}$ value estimated by our method is 189.128 $\mu g/m^3$. Our method achieved the best results, which suggests that combining PM$_{2.5}$ to estimate depth is very effective.

TABLE III
QUANTITATIVE EVALUATION FOR DEPTH ESTIMATION. LOWER IS BETTER.

Method	Rel	log10	RMSE
[5]	13.870	0.4615	0.507
[35]	2.0153	0.5097	0.5276
[2]	1.5458	0.7160	0.5567
[4]	0.921	0.459	0.479
[24]	0.896	0.349	0.427
[25]	0.614	0.3556	0.3748
[34]	**0.5865**	0.2838	0.3740
Our Method	0.613	**0.233**	**0.265**

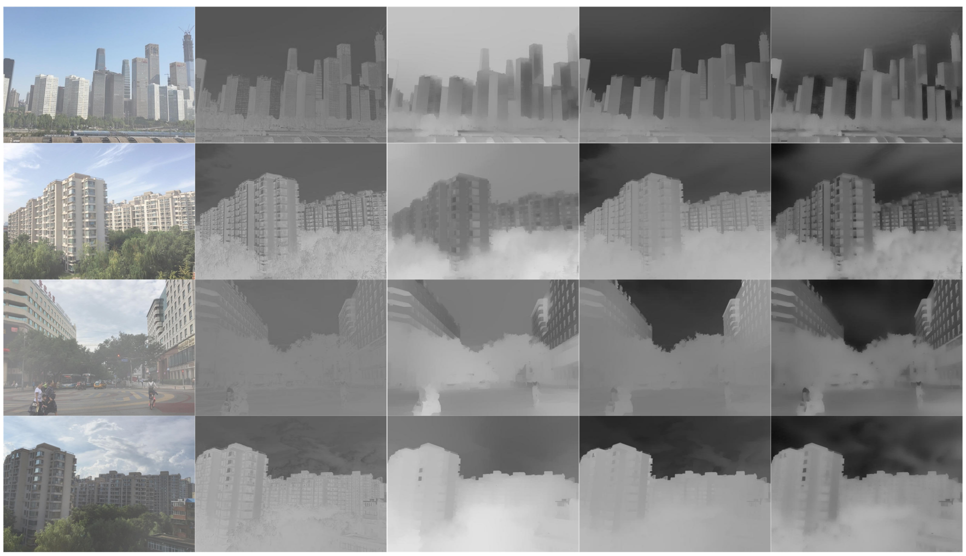
2) Real Data: We first evaluate our method on Make3D dataset [13], [14], which contains 534 outdoor images with the corresponding depth maps scanned by a laser. Fig.11 provides a qualitative comparison of our method with seven state-of-the-art methods. In this case, the PM$_{2.5}$ value estimated by our method is 43.573 $\mu g/m^3$. As shown in Fig.11, the vision-based depth estimation methods [2], [5], [34], [35] have the worst results for the hazy weather, even for learning based methods [2], [34], [35]. The ground truth is captured by the laser, but for distant objects, such as the far house marked by blue rectangle, it does not show the accurate depth. On the contrary, our method achieves the most accurate depth estimation which is even better than the ground truth captured using laser scan. In addition, our method is also very robust in dealing with objects that are close in distance. More specifically, as the yellow box shown, the two trees can be recognized in our depth map. Moreover, the depth of the holes on the tree, actually the sky, is accurately estimated by our method. In a word, our method gives the accurate depth, no matter how far objects are. Note that there is no haze in this dataset, and our method also achieves the most accurate depth estimation, which demonstrates that our method also works in good air quality conditions.

We also compare our method with seven state-of-the-art methods on real images downloaded from Internet. As shown in Fig. 12, when the haze is heavy, the haze will make the whole image appear blurred, especially for distant objects. Our method shows an excellent performance in wild images, especially for distant objects. The tranditional depth estimation methods [2] cannot preserve the outline information of objects



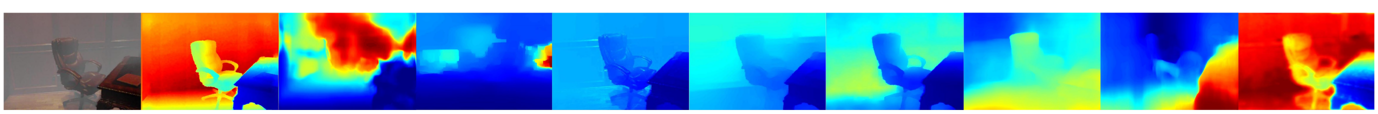
(a) (b) (c) (d) (e)

Fig. 8. Transmission maps estimated by (b) method [4], (c) method [24], (d) method [25], and (e) our method for (a) the input image collected from Internet.



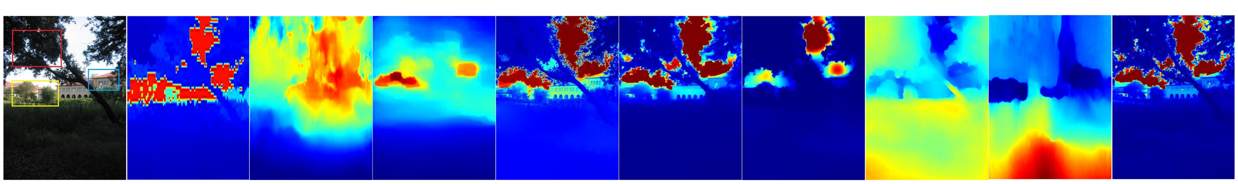
(a) (b) (c) (d) (e)

Fig. 9. Transmission maps estimated by (b) method [4], (c) method [24], (d) method [25], and (e) our method for (a) the input image in a challenge dataset [33].



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 10. Depth inference results of (c) method [5], (d) method [2], (e) method [4], (f) method [24], (g) method [25] , (h) method [34], (i) method [35] and (j) our method on (a) NYU synthetic dataset, compared with (b) ground truth.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 11. Depth inference results of (c) method [5], (d) method [2], (e) method [4], (f) method [24], (g) method [25] , (h) method [34], (i) method [35] and (j) our method on (a) Make3D dataset, compared with (b) ground truth.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2943631, IEEE Transactions on Industrial Informatics

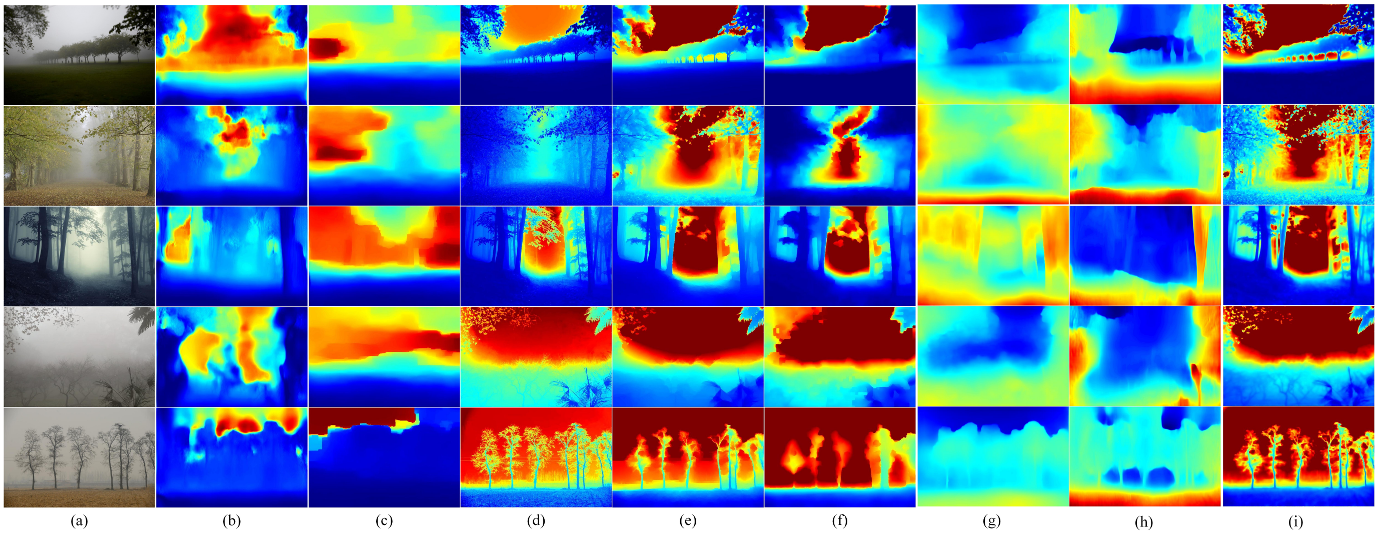


Fig. 12. Depth inference results of (b) method [5], (c) method [2], (d) method [4], (e) method [24], (f) method [25] , (g) method [34], (h) method [35] and (i) our method on (a) Internet images. $PM_{2.5}$ values from top to bottom are 248.127 $\mu g/m^3$, 200.619 $\mu g/m^3$, 184.285 $\mu g/m^3$, 264.328 $\mu g/m^3$ and 34.754 $\mu g/m^3$, which are estimated by our method.

in the images and the estimated depth maps do not reflect the distance information of distant objects as well. The results of learning-based methods [5], [34], [35] are also over-smooth, because they rely on the training data and are difficult to estimate depths for complex scenarios even with haze. The results of transmission-based methods [4], [24], [25] are more accurate than the vision-based depth estimation methods [2], [5], [34], [35], but losing some details. On the contrary, our method outperforms these methods with accurate and smooth depth maps. For example, for dense branches, our method can fully show the contours of the branches, not affected by other trunks. Besides, our method can also accurately estimate the depth of distant scene, not only the close scene.

E. Running Times

We test the running times of our method on a desktop with an Intel Core i7-4800K CPU, a NVIDIA GeForce GTX TITAN X GPU, and a 32GB RAM. The results on different datasets are shown in Table IV. The $PM_{2.5}$ estimation module takes less than 1 second for all test images, which is appealing for practical application. Smartphones installed with our method can be used as palm air quality monitors. The transmission estimation module is implemented with Matlab code, which can be significantly accelerated with optimized complied code. Our method provides a promising alternative depth estimation approach for handheld devices with a single color camera for both clean and foul weather.

TABLE IV
THE RUNNING TIMES OF THE PROPOSED METHOD FOR DIFFERENT DATASETS.

Procedure \ Dataset	Make3D (343 × 458)	NYU (561 × 427)	Internet (1019 × 624)
$PM_{2.5}$ Est - proposed net	0.129s	0.105s	0.225s
$PM_{2.5}$ Est - SVR	0.705s	0.708s	0.702s
Transmission Estimation	7.404s	10.678s	28.361s
Depth Estimation	0.004s	0.004s	0.009s
Total	8.242s	11.495s	29.297s

VII. CONCLUSION

In this paper, we propose a new $PM_{2.5}$ estimation method only using a captured image based on high-level features of hybrid convolutional neural network, the performance of which is comparable to the professional measuring instrument. A new transmission estimation method is proposed to estimate the depth of scene through the atmospheric scattering model with our estimated $PM_{2.5}$ value using non-local sparse priors. We fit the relationship between $PM_{2.5}$ and atmospheric attenuation coefficient β by simulation. Experimental results show that our method achieves accurate $PM_{2.5}$ estimation and depth inference. Our method can work for both clean and foul weather.

REFERENCES

[1] J. Lei, J. Sun, Z. Pan, S. Kwong, J. Duan, and C. Hou, "Fast mode decision using inter-view and inter-component correlations for multiview depth video coding," *IEEE Trans. II*, vol. 11, pp. 978–986, 2015.
[2] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non-parametric sampling," in *Proc. ECCV*, 2012, pp. 775–788.
[3] X. Jiang, J. Sun, C. Li, and H. Ding, "Video image defogging recognition based on recurrent neural network," *IEEE Trans. II*, vol. 14, pp. 3281–3288, 2018.
[4] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. PAMI*, vol. 33, no. 12, pp. 2341–2353, 2010.
[5] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. CVPR*, 2017.
[6] W. Zhang, S. Gao, X. Song, J. Liu, W. Liu, and Z. Chen, "Concentration measurement and metrical technology of fine particulate matter pm_(2.5)," *China Powder Science & Technology*, 2013.
[7] Z. Tao, D. Liu, Z. Wang, X. Ma, Q. Zhang, C. Xie, G. Bo, S. Hu, and Y. Wang, "Measurements of aerosol phase function and vertical backscattering coefficient using a charge-coupled device side-scatter lidar." *Optics Express*, vol. 22, no. 1, pp. 1127–34, 2014.
[8] K. Gu, J. Qiao, and W. Lin, "Recurrent air quality predictor based on meteorology- and pollution-related factors," *IEEE Trans. II*, vol. 14, p. 3946, 2018.
[9] C. Zhang, J. Yan, C. Li, X. Rui, L. Liu, and R. Bie, "On estimating air pollution from photos using convolutional neural network," in *Proc. ACM MM*, 2016, pp. 297–301.

[10] A. Chakma, B. Vizena, T. Cao, J. Lin, and J. Zhang, "Image-based air quality analysis deep convolutional neural network," in *Proc. IEEE ICIP*, 2017, pp. 3949–3952.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[12] J. Ma, K. Li, Y. Han, P. Du, and J. Yang, "Image-based PM2.5 estimation and its application on depth estimation," in *Proc. ICASSP*, 2018.

[13] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. NIPS*, 2005, pp. 1161–1168.

[14] ——, "3-D depth reconstruction from a single still image," *IJCV*, vol. 76, no. 1, pp. 53–69, 2008.

[15] A. Saxena, J. Schulte, and A. Y. Ng, "Depth estimation using monocular and stereo cues," in *Proc. IJCAI*, 2007, pp. 2197–2203.

[16] A. G. Schwing and R. Urtasun, "Efficient exact inference for 3D indoor scene understanding," in *Proc. ECCV*. Springer, 2012, pp. 299–313.

[17] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. PAMI*, vol. 33, no. 5, pp. 978–994, 2011.

[18] J. Konrad, M. Wang, and P. Ishwar, "2D-to-3D image conversion by learning depth from examples," in *Proc. CVPR Workshops*, 2012, pp. 16–22.

[19] N. E.-D. Mebtouche, A. Boumahdi, and N. Baha, "Depth estimation from a single 2D image," in *Proc. ISPS*, 2018.

[20] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. NIPS*, 2014, pp. 2366–2374.

[21] P. Wang, X. Shen, Z. Lin, and S. Cohen, "Towards unified depth and semantic prediction from a single image," in *Proc. CVPR*, 2015, pp. 2800–2809.

[22] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. CVPR*, 2015, pp. 5162–5170.

[23] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Monocular depth estimation using multi-scale continuous CRFs as sequential deep networks," *IEEE Trans. PAMI*, vol. PP, no. 99, pp. 1–1, 2018.

[24] D. Berman, T. treibitz, and S. Avidan, "Non-local image dehazing," in *Proc. CVPR*, June 2016.

[25] C. Chen, M. N. Do, and J. Wang, "Robust image and video dehazing with visual artifact suppression via gradient residual minimization," in *Proc. ECCV*, 2016.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[28] W. E. K. Middleton, "Vision through the atmosphere," *Handbuch Der Physik*, 1952.

[29] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[30] D. Hernandez-Juarez, L. Schneider, A. Espinosa, D. Vazquez, A. M. Lopez, U. Franke, M. Pollefeys, and J. C. Moure, "Slanted stixels: Representing san franciscos steepest streets," in *Proc. BMVC*, 2017.

[31] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[32] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. ECCV*, 2012.

[33] IEEE ICIP 2019 Grand Challenge, https://pkustruct.github.io/icip2019.

[34] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proc. CVPR*, 2018.

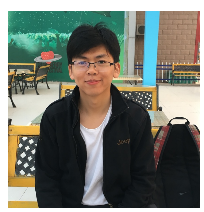
[35] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth prediction," *arXiv:1806.01260*, 2018.

Kun Li (M'14) received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2006, and the master and Ph.D. degrees from Tsinghua University, Beijing, in 2011. She is currently an Associate Professor with the School of Computer Science and Technology, Tianjin University, Tianjin, China. Her research interests include dynamic scene 3D reconstruction and image/video processing.

Jian Ma received the B.S. degree in computer science and technology from the School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, China, in 2016 and the M.S. degree in mould identification and intelligent devices from the School of Computer Science and Technology, Tianjin University, China, in 2019. He is currently working toward the PH.D. degree in University of Bristol, UK.

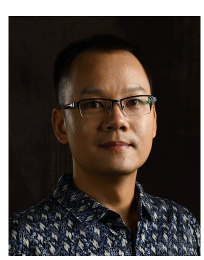
Han Li received the B.S. degree in software engineering, Changchun University of science and technology, Changchun, China, in 2018. She is currently working toward the M.S. degree in Tianjin University, Tianjin. Her research interests include computer vision and image dehazing.

Yahong Han (M'15) received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2012. From 2014 to 2015, he visited Prof. B. Yus Group, UC Berkeley, as a Visiting Scholar. He is currently a Full Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. His current research interests include multimedia analysis, computer vision, and machine learning.



Xibin Yue received the M.S Degree from Beijing Forest University China in 2016. From 2016,he is an Algorithm research staff in Meteorological Institute of MoJi Weather China. His research interest include computer vision, sequence prediction, deep learning towards meteorology.

Zihao Chen received the B.S. degree in software development from Zhengzhou University, Zhengzhou, China, in 2017. He is currently working on algorithmic research in Beijing Moji Fengyun Technology Co., Ltd. The main research direction is nowcasting and artificial intelligence.



Jingyu Yang (M'10-SM'17) received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2003, and Ph.D. (Hons.) degree from Tsinghua University, Beijing, in 2009. He has been a Faculty Member with Tianjin University, China, since 2009, where he is currently a Professor with the School of Electrical and Information Engineering. His research interests include image/video processing, 3D imaging, and computer vision.